arXiv:2304.01849v2 [stat.ME] 2 Jun 2023

基于半参数效率的遗传相关性估计：机器学习方法

Xu Guo[1], Yiyuan Qian[1], Hongwei Shi[1],Weichao Yang[1],和 Niwen Zhou[2*]

[1] 中国北京北京师范大学统计学院

[2] 中国珠海北京师范大学统计与数据科学中心

摘要

本文在无模型框架下，提出了用于估计两个性状之间遗传相关性的半参数有效估计量。现有的大多数方法都需要指定涉及性状和遗传变异的某些参数模型。然而，模型设定错误导致的偏差可能会产生误导性的统计结果。此外，目前尚缺乏遗传相关性估计量的半参数有效界。本文采用机器学习方法开发半参数有效估计量，并为遗传相关性的两个重要度量——遗传协方差和遗传相关性——构建有效置信区间，适用于连续和离散响应变量。基于推导出的遗传相关性有效影响函数，只要其中一个遗传值能够被一致估计，我们就提出了遗传协方差的一致估计量。两个性状的数据可以来自同一组或不同组的个体。我们开展了多项数值研究，以展示所提出的方法。我们

*通讯作者：Niwen Zhou。所有作者对本研究的贡献相同，作者姓名按字母顺序排列。

还将所提出的方法应用于卡沃斯农场白鼠全基因组关联研究数据的分析。

关键词： 模型设定错误；遗传协方差；半参数有效界；置信区间。

1 引言

理解复杂性状之间的遗传相关性是人类遗传学研究中的重要课题。在实际遗传学研究中，研究人员已在许多复杂疾病中发现了共享的常见遗传变异，例如各种自身免疫性疾病（Zhernakova 等人，2009年）和精神障碍（Craddock 和 Owen，2005）。遗传相关性分析有多种下游应用，有助于发现与疾病相关的遗传变异、改进多基因风险预测，并可能有助于改进疾病分类学与诊断、风险分层及生活方式干预（Van Rheenen 等人，2019）。

遗传协方差和遗传相关性是衡量遗传相关性的两种常用指标。考虑两个响应变量 $Y \in \mathbb{R}$ 和 $Z \in \mathbb{R}$，例如复杂性状、疾病结局或基因表达，以及作为p维预测变量的 p个遗传变异，记为 $X \in \mathbb{R}^p$。正如 Van Rheenen 等人（2019）和 Wang 等人（2021）所述，Y 和Z 的遗传协方差可定义为它们的条件均值函数之间的协方差

$$I = \mathrm{cov}\{m(X), h(X)\},$$

其中，$m(X) = E(Y \mid X)$ 和 $h(X) = E(Z \mid X)$ 分别是Y 和Z的遗传值。随后，遗传相关性可定义如下：

$$\rho = \frac{\mathrm{cov}\{m(X), h(X)\}}{\sqrt{\mathrm{var}\{m(X)\}\mathrm{var}\{h(X)\}}}.$$

因此，ρ 被标准化为 $-1 \le \rho \le 1$，从而可以用于比较多对个体之间的遗传相关性。

在遗传学文献中，遗传相关性估计方法学进展主要基于家系研究或全基因组关联研究（GWAS）。

与传统的家系研究方法相比，基于全基因组关联研究的方法不要求在同一个体中测量所研究的表型（Zhang 等人，2021）。因此，基于全基因组关联研究数据量化性状对之间遗传效应的重叠颇具潜力。大多数基于全基因组关联研究的方法均基于某些特定的回归模型推导而来，例如线性混合效应模型（Lee 等人，2012年；Vattikuti 等人，2012年；Yang 等人，2013年）、线性固定效应模型（Guo 等人，2019）以及广义线性模型（Ma 等人，2022；Wang 等人，2021）。

然而，实际中可能并不满足这些严格的模型结构假设，从而导致估计量有偏、推断结果不准确。因此，开发不依赖任何参数模型假设的估计与推断方法十分重要。为避免模型设定错误，可以采用灵活的机器学习方法估计回归函数 $m(X)$ 和 $h(X)$。若直接采用这些方法，由于机器学习方法复杂且甚至具有黑箱性质，相应的估计量可能不一致。此外，在许多研究中，响应变量 Y 和 Z 是离散型的。现有的大多数方法侧重于连续性状，无法直接用于估计二元性状之间的遗传相关性。Weissbrod 等人 (2018年) 提出了一种改进的混合效应模型来处理二元性状。最近，Ma 等人(2022) 和 Wang 等人 (2021年) 也致力于利用二元性状和高维全基因组关联研究数据估计遗传相关性。然而，这些方法同样会受到模型设定错误的影响。此外，尽管遗传相关性分析领域的文献中已提出多种估计量，但遗传协方差和遗传相关性估计量的半参数有效界限仍未确立。所有这些关键问题都亟需新的方法和理论结果。

为应对上述挑战与问题，本文针对连续性状和离散性状，提出遗传协方差与遗传相关性的高效、无模型估计量。本文的主要贡献总结如下。首先，我们推导出遗传协方差和遗传相关性的有效影响函数，从而得到遗传协方差和遗传相关性估计量的半参数有效界。其次，基于有效影响函数，只要 $m(X)$ 或 $h(X)$ 中任意一个得到一致估计，我们就提出了遗传协方差的一致估计量。第三，为确保在全基因组关联研究数据可能具有高维的情况下，对遗传协方差和遗传相关性进行有效的统计推断，我们通过结合有效影响函数的结果与样本拆分策略来估计

遗传协方差和遗传相关性。所提出的估计量无模型且具有半参数有效性。第四，我们的方法具有很强的通用性。事实上，我们的程序适用于两个连续性状或离散性状的数据，无论样本是否重叠。

本文的结构安排如下。第2节介绍遗传协方差的估计和推断程序。在第3节中，还针对遗传相关性建立了这些程序。此外，第4节还考虑了离散响应的情形。第5节展示了所提出方法的数值表现。在第6节中，我们给出了一个真实数据示例。第7节给出了结论与讨论。理论结果的证明见附录。

2 遗传协方差的估计

本节提出遗传协方差 I 的估计量。现在假设$\mathcal{D}_y = \{(X_i, Y_i),\ i = 1, \ldots, N_y\}$和$\mathcal{D}_z = \{(Xj, Z\,j),\ j = 1, \ldots, N_z\}$分别是来自总体 (X, Y) 和 (X, Z) 的两个独立同分布随机样本。令 $\mathcal{D}_0 = \mathcal{D}_y \cap \mathcal{D}_z$, $\mathcal{D} = \mathcal{D}_y \cup \mathcal{D}_z$ 和 $N_0 = |\mathcal{D}_0|$, $N = |\mathcal{D}| = N_y + N_z - N_0$。本文允许两个性状的数据来自同一组或不同组的个体。具体而言，当数据来自两个独立样本且 $N_0 = 0$ 时，我们称之为非重叠；而当数据来自两个样本且 $N_0 > 0$ 时，我们称之为重叠。特别地，如果数据来自相同的样本且 $N_0 = N_y = N_z$，ˆ 则称之为完全重叠。假设 ˆm(X) 和 h(X) 分别是 m(X) 和 h(X) 的两个合适估计量。回顾 I = E{m(X)h(X)} − E(Y)E(Z)，可以考虑以下自然的 I 的代入估计量：

$$I_N = N^{-1} \sum_{i \in \mathcal{D}} \hat{m}(X_i)\hat{h}(X_i) - \bar{Y}_N \bar{Z}_N.$$

这里 $\bar{Y}_N = N_y^{-1} \sum_{=1i}^{N_y} Y_i$, $\bar{Z}_N = N_z^{-1} \sum_{=1i}^{N_z} Z_i$。然而，当我们采用灵活的机器学习方法来获得 $\hat{m}(X)$ 和 $h(X)$ 时，由于第一项 $N^{-1} \sum_{i \in \mathcal{D}} \hat{m}(X_i)\hat{h}(X_i)$ 存在过拟合和偏差，这个自然估计量甚至无法保持一致性。

为获得一致且渐近正态的估计量，研究表明有效影响函数（EIF）发挥着关键作用。在此之前，我们先介绍一些记号。回顾一下，观测数据为 $\mathcal{D} = \mathcal{D}_y \cup \mathcal{D}_z$，且某些个体在两个样本中可能都未被观测到。若观测到 Y，则令 $\mathrm{T}_{y,i} = 1$；否则令 $\mathrm{T}_{y,i} = 0$。$\mathrm{T}_{z,i}$ 的定义类似。请注意，我们能否观测到 Y 或 Z 与数据无关。因此，Y 或 Z 的缺失完全随机，即 $\mathrm{T}_y \perp (\mathrm{Y}, \mathrm{X})$ 和 $\mathrm{T}_z \perp (\mathrm{Z}, \mathrm{X})$。采用缺失数据文献中的记号，令 $\mathrm{Y}_i^* = \mathrm{T}_{y,i}\mathrm{Y}_i$，$\mathrm{Z}_i^* = \mathrm{T}_{z,i}\mathrm{Z}_i$。更多内容请参阅 Tsiatis（2007）。因此，观测数据为 $\{\mathrm{O}_i\}^N \mathrm{i}_{=1} = \{\mathrm{X}_i, \mathrm{Y}_i^*, \mathrm{Z}_i^*, \mathrm{T}_{y,i}, \mathrm{T}_{z,i}\}^N \mathrm{i}_{=1}$。为推导 I 的有效影响函数，需要满足以下正则条件。

条件 2.1。 $E\{\sigma^4(X)\}$、$E\{\delta^4(X)\}$、$E[\{m(X) - E(Y)\}^4]$，以及$E[\{h(X) - E(Z)\}^4]$ 均为有限值。

这里，$\sigma^2(X) = E(\epsilon^2 \mid X)$，且 $\delta^2(X) = E(\eta^2 \mid X)$，其中 $\epsilon = Y - m(X)$，$\eta = Z - h(X)$。该条件非常宽松，并保证I 的有效影响函数的方差有限。

定理 2.1。在条件 2.1下，效率影响函数I 为

$$
\begin{aligned}
S_{yz} = {} & \frac{\mathbb{I}(T_y = 1)}{\Pr(T_y = 1)}\{Y^* - E(Y^* \mid X, T_y = 1)\}\{E(Z^* \mid X, T_z = 1) - E(Z^* \mid T_z = 1)\} \\
& + \frac{\mathbb{I}(T_z = 1)}{\Pr(T_z = 1)}\{Z^* - E(Z^* \mid X, T_z = 1)\}\{E(Y^* \mid X, T_y = 1) - E(Y^* \mid T_y = 1)\} \\
& + \{E(Y^* \mid X, T_y = 1) - E(Y^* \mid T_y = 1)\}\{E(Z^* \mid X, T_z = 1) - E(Z^* \mid T_z = 1)\} - I,
\end{aligned}
\tag{2.1}
$$

其中 $\mathbb{I}(\cdot)$是指示函数。因此，I 的半参数效率界等于

$$
\begin{aligned}
\mathrm{var}(S_{yz}) = {} & \frac{1}{\Pr(T_y = 1)}E[\sigma^2(X)\{h(X) - E(Z)\}^2] + \frac{1}{\Pr(T_z = 1)}E[\delta^2(X)\{m(X) - E(Y)\}^2] \\
& + E([\{m(X) - E(Y)\}\{h(X) - E(Z)\} - I]^2) \\
& + \frac{\Pr(T_y = 1, T_z = 1)}{\Pr(T_y = 1)\Pr(T_z = 1)}E[\epsilon\eta\{m(X) - E(Y)\}\{h(X) - E(Z)\}].
\end{aligned}
$$

注 2.1。在非重叠情形下，I 的半参数效率界为

$$
\begin{aligned}
\mathrm{var}(S_{yz}) = {} & \frac{1}{\Pr(T_y = 1)}E[\sigma^2(X)\{h(X) - E(Z)\}^2] + \frac{1}{\Pr(T_z = 1)}E[\delta^2(X)\{m(X) - E(Y)\}^2] \\
& + E([\{m(X) - E(Y)\}\{h(X) - E(Z)\} - I]^2).
\end{aligned}
$$

W而在完全重叠情形下，I 的半参数效率界为

$$\begin{aligned}\mathrm{var}(S_{yz}) = {} & E[\sigma^2(X)\{h(X) - E(Z)\}^2] + E[\delta^2(X)\{m(X) - E(Y)\}^2] \\ & + E([\{m(X) - E(Y)\}\{h(X) - E(Z)\} - I]^2) \\ & + E[\epsilon\eta\{m(X) - E(Y)\}\{h(X) - E(Z)\}].\end{aligned}$$

˜根据上述针对I的有效影响函数，通过令估计影响函数的I_N 样本均值为零，即可得到I 的一致估计量。

$$\begin{aligned}\tilde{I}_N = {} & N_y^{-1}\sum_{i=1}^{N_y}\{Y_i - \hat{m}(X_i)\}\{\hat{h}(X_i) - \bar{Z}_N\} + N_z^{-1}\sum_{i=1}^{N_z}\{\hat{m}(X_i) - \bar{Y}_N\}\{Z_i - \hat{h}(X_i)\} \\ & + N^{-1}\sum_{i=1}^{N}\{\hat{m}(X_i) - \bar{Y}_N\}\{\hat{h}(X_i) - \bar{Z}_N\}.\end{aligned}$$

ˆ ˜假设ˆm(X$_i$)和h(X$_i$)是 ˜m(X$_i$)和h(X$_i$)的一致估计量，满足E[{ˆm(X$_i$) − ˜m(X$_i$)}^{2}] = o(1)且E[{ˆh(X$_i$) − ˜h(X$_i$)}^{2}] = o(1)，但 ˜m(X$_i$)和 ˜h(X$_i$)可能分别不等于m(X$_i$)和h(X$_i$)。在下述定理中，我们证明在温和条件下I$_N$ 的一致性。

定理 2.2. 在条件 2.1 下，E{˜m^2(X)} 和E{˜h^2(X)} 均有限，且进一步有˜ ˜E[{˜m(X) − m(X)}{h(X) − h(X)}] = 0，I$_N$ 是I的一致估计量。

˜定理 2.2 给出了非常有意思的结果。它表明，在一般情况下，上述 I$_N$ 是一致的。尽管 ˆm(X$_i$) 和 h(X$_i$) 都是对 m(X$_i$) 和 h(X$_i$) 完全不一致的估计量˜，但只要 ˜m(X) − m(X)˜ ˆ 和 h(X) − h(X) 不相关，上述 I$_N$ 仍然可以是 I 的一致估计量。如果 ˆm(X$_i$) 或 h(X$_i$) 中任意一个（不必两者都满足）是 m(X$_i$) 或 h(X$_i$) 的一致估计量˜，那么 I$_N$ 就是遗传协方差 I 的一致估计量。

˜然而，我们注意到，要进行统计推断，上述 I$_N$ 还不够。事实上，诸如 N$_y^{-1}$ Σ^{N_y} i=1 ϵ_i{ˆh(X$_i$) − ˜h(X$_i$)} 这样的项并非 o$_P$(N$^{-1/2}$)。为此，需要进行样本拆分。具体而言，我们将整个数据集 $\mathcal{D}_y$ 随机拆分为两个相互独立的数据集 $\mathcal{D}_{1y}$ 和 $\mathcal{D}_{2y}$。令 $|\mathcal{D}_{2y}| = |\mathcal{D}_{1y}| = n_y$。类似地， $\mathcal{D}_z$ 被划分为 $\mathcal{D}_{1z}$ 和$\mathcal{D}_{2z}$ ，其中 $|\mathcal{D}_{2z}| = |\mathcal{D}_{1z}| = n_z$。记 $\mathcal{D}_{0k} = \mathcal{D}_{ky} \cap \mathcal{D}_{kz}$、 $\mathcal{D}_k = \mathcal{D}_{ky} \cup \mathcal{D}_{kz}$、 $n_0 = |\mathcal{D}_{0k}|$、 $n = |\mathcal{D}_k| = n_y + n_z - n_0$, k = 1, 2。记基于 $\mathcal{D}_1$ 得到的 m(X) 和 h(X) 的估计量分别为

$\hat{m}_{\mathcal{D}_1}(X)$ 和 $\hat{h}_{\mathcal{D}_1}(X)$。考虑

$$I_n = n_y^{-1} \sum_{i\in\mathcal{D}_{2y}} \{Y_i - \hat{m}_{\mathcal{D}_1}(X_i)\}\{\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n\} + n_z^{-1} \sum_{i\in\mathcal{D}_{2z}} \{\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n\}\{Z_i - \hat{h}_{\mathcal{D}_1}(X_i)\}$$
$$+ n^{-1} \sum_{i\in\mathcal{D}_2} \{\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n\}\{\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n\}. \tag{2.2}$$

这里 $\bar{Y}_n = n_y^{-1}\sum_{i\in\mathcal{D}_{2y}} Y_i$，$\bar{Z}_n = n_z^{-1}\sum_{i\in\mathcal{D}_{2z}} Z_i$。

理论分析需要满足以下条件。

条件 2.2。 $E[\{\hat{m}_{\mathcal{D}_1}(X) - m(X)\}^2] = o(n_y^{-1/2})$，且 $E[\{\hat{h}_{\mathcal{D}_1}(X) - h(X)\}^2] = o(n_z^{-1/2})$。

条件 2.3。 $E[\sigma^2(X)\{\hat{h}_{\mathcal{D}_1}(X) - h(X)\}^2] = o(1)$，且 $E[\delta^2(X)\{\hat{m}_{\mathcal{D}_1}(X) - m(X)\}^2] = o(1)$。

条件 2.4。 $E[\{m(X) - E(Y)\}^2\{\hat{h}_{\mathcal{D}_1}(X) - h(X)\}^2] = o(1)$，$E[\{h(X) - E(Z)\}^2\{\hat{m}_{\mathcal{D}_1}(X) - m(X)\}^2] = o(1)$，以及 $E[\{\hat{m}_{\mathcal{D}_1}(X) - m(X)\}^2\{\hat{h}_{\mathcal{D}_1}(X) - h(X)\}^2] = o(1)$。

上述条件对矩和估计误差的要求较为宽松。条件2.2 相当一般，在文献中也被广泛采用。例如参见Chernozhukov 等 (2018年) 以及 Vansteelandt 和 Dukes(2020年)。当 $\sigma^2(X)$ 和$\delta^2(X)$ 均有限时，条件 2.3 在条件 2.2下直接成立。条件 2.3 和2.4 对于在渐近分析中适当控制二次项及更高阶项都是必需的。

进一步定义$I_n^* = (I_n + I_n')/2$，其中I_n' 的定义类似，只需交换 $\mathcal{D}_1$ 和 $\mathcal{D}_2$的角色。由于I_n 和I_n' 渐近独立，以下定理为构造遗传协方差的置信区间提供了一种解法。

定理 2.3。假设条件 2.1–2.4 均满足，则有

$$I_n^* - I = \sum_{i\in\mathcal{D}} S_{yz,i}^* + o_p(N_z^{-1/2} + N_y^{-1/2}).$$

Here

$$S_{yz,i}^* = N_y^{-1}\epsilon_i^*\mathbb{I}(T_{y,i} = 1) + N_z^{-1}\eta_i^*\mathbb{I}(T_{z,i} = 1) + N^{-1}\xi^*.$$

其中 $\epsilon_i^* = \epsilon_i\{h(X_i) - E(Z)\}$，$\eta_i^* = \eta_i\{m(X_i) - E(Y)\}$ 以及 $\xi_i^* = \{m(X_i) - E(Y)\}\{h(X_i) - E(Z)\} - I$。此外，$I_n^*$ 的渐近正态性为

$$\sigma^{-1}(I_n^* - I) \xrightarrow{d} N(0,1).$$

这里 $\sigma^2 = N_y^{-1}E(\epsilon^{*2}) + N_z^{-1}E(\eta^{*2}) + N^{-1}E(\xi^{*2}) + (N_yN_z)^{-1}N_0E(\epsilon^*\eta^*)$。

注 2.2。假设

$$\Pr(T_y = 1) = \lim_{(N,N_y)\to\infty} N_y/N,\ \Pr(T_z = 1) = \lim_{(N,N_z)\to\infty} N_z/N,\ \text{and}\ \Pr(T_y = 1, T_z = 1) = \lim_{(N,N_0)\to\infty} N_0/N.$$

我们可以得到 $N\sigma^2 \to \mathrm{var}(S_{yz})$，这意味着所提出的估计量 I_n^* 是半参数有效的，也就是说，其渐近方差达到了半参数有效界。

在上述定理中，σ^2 通常未知，需要进行估计。为此，记

$$\hat{\delta}_i = n_y^{-1}\{Y_i - \hat{m}_{\mathcal{D}_1}(X_i)\}\{\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n\}\mathbb{I}(T_{y,i} = 1) + n_z^{-1}\{\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n\}\{Z_i - \hat{h}_{\mathcal{D}_1}(X_i)\}\mathbb{I}(T_{z,i} = 1)$$
$$+ n^{-1}\{\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n\}\{\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n\}.$$

$\hat{}$ 那么$I_n = \sum_{i\in\mathcal{D}_2}\delta_i$。$I_n$ 的方差估计量为

$$K_1 = \sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - n^{-1}I_n)^2.$$

K_2 的定义类似，只需交换 $\mathcal{D}_1$ 和 $\mathcal{D}_2$的角色。令$\hat{}\sigma^2 = (K_1 + K_2)/4$。可以证明，$\hat{}\sigma/\sigma \xrightarrow{p} 1$。因此，$I$ 在置信水平为$1-\alpha$ 时的有效置信区间为

$$\mathrm{CI}(\alpha) = \left(I_n^* - z_{\alpha/2}\cdot\hat{\sigma},\ I_n^* + z_{\alpha/2}\cdot\hat{\sigma}\right). \tag{2.3}$$

下述定理确立了上述I置信区间的有效性。

定理 2.4。假设 $E(\epsilon_i^{*4})$、$E(\eta_i^{*4})$ 和 $E(\xi_i^{*4})$ 均有限，且满足条件 2.2–2.4 。我们有

$$\hat{\sigma}^{-1}(I_n^* - I) \xrightarrow{d} N(0,1),$$

and

$$\lim_{(N_y,N_z)\to\infty} \Pr\{I \in \mathrm{CI}(\alpha)\} = 1 - \alpha.$$

3 遗传相关性的估计

本节中，我们考虑遗传相关性的推断。遗传相关性是遗传协方差的标准化形式，定义为

$$\rho = \frac{\mathrm{cov}\{m(X), h(X)\}}{\sqrt{\mathrm{var}\{m(X)\}\mathrm{var}\{h(X)\}}},$$

其中，$m(X) = E(Y \mid X)$，$h(X) = E(Z \mid X)$。为简化表述，我们在此给出一些记号。记 $B_0^y = \mathrm{var}\{m(X)\}$，$B_0^z = \mathrm{var}\{h(X)\}$。进一步令

$$S_{yy} = 2\frac{\mathbb{I}(T_y = 1)}{\Pr(T_y = 1)}\{Y^* - m(X)\}\{m(X) - E(Y)\} + \{m(X) - E(Y)\}^2 - B_0^y,$$
$$S_{zz} = 2\frac{\mathbb{I}(T_z = 1)}{\Pr(T_z = 1)}\{Z^* - h(X)\}\{h(X) - E(Z)\} + \{h(X) - E(Z)\}^2 - B_0^z,$$

它们分别是根据定理 2.1得到的 B_0^y 和 B_0^z 的有效影响函数。于是，我们在以下命题中得到其有效影响函数。 ρ

命题 3.1。在条件 2.1下，的有效影响函数为 ρ

$$\phi_1(O, \mathcal{P}) = \frac{S_{yz}}{\sqrt{B_0^y B_0^z}} - \rho\frac{S_{yy}}{2B_0^y} - \rho\frac{S_{zz}}{2B_0^z}. \tag{3.4}$$

受B_0^y 和B_0^z 的有效影响函数启发，我们考虑以下估计量：

$$\hat{B}_0^y = n_y^{-1}\sum_{i\in\mathcal{D}_{2y}} 2\{Y_i - \hat{m}_{\mathcal{D}_1}(X_i)\}\{\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n\} + n^{-1}\sum_{i\in\mathcal{D}_2}\{\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n\}^2,$$
$$\hat{B}_0^z = n_z^{-1}\sum_{i\in\mathcal{D}_{2z}} 2\{Z_i - \hat{h}_{\mathcal{D}_1}(X_i)\}\{\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n\} + n^{-1}\sum_{i\in\mathcal{D}_2}\{\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n\}^2.$$

此外，通过令估计得到的影响函数的样本均值为零，我们可以得到 ρ_n 关于 ρ 的一致估计量，即

$$\rho_n = \frac{I_n}{\sqrt{\hat{B}_0^y \hat{B}_0^z}}. \tag{3.5}$$

进一步，我们定义 $\rho_n^* = (\rho_n + \rho_n')/2$，其中 ρ_n' 的定义类似，只需交换 $\mathcal{D}_1$ 和 $\mathcal{D}_2$ 的角色。 ρ_n^* 的渐近性质见下述定理。

定理 3.1。假设条件 2.1–2.4 均成立。则有

$$\rho_n^* - \rho = \sum_{i\in\mathcal{D}} \phi_1^*(O_i) + o_p(N_z^{-1/2} + N_y^{-1/2}). \tag{3.6}$$

Here

$$\phi_1^*(O_i) = \frac{S_{yz,i}^*}{\sqrt{B_0^y B_0^z}} - \rho\frac{S_{yy,i}^*}{2B_0^y} - \rho\frac{S_{zz,i}^*}{2B_0^z}$$

with

$$S_{yy,i}^* = 2\epsilon_{y,i}^*\mathbb{I}(T_{y,i}=1) + \xi_{y,i}^*, \text{ and } S_{zz,i}^* = 2\eta_{y,i}^*\mathbb{I}(T_{z,i}=1) + \xi_{z,i}^*,$$

其中 $\epsilon_{y,\ i}^* = \epsilon_i\{m(X_i) - E(Y)\}$, $\xi_{y,\ i}^* = \{m(X_i) - E(Y)\}^2 - B_0^y$。$\eta_{y,\ i}^*$和 $\xi_{z,\ i}^*$的定义类似。并且ρ_n^* 的渐近正态性为

$$\sigma_r^{-1}(\rho_n^* - \rho) \xrightarrow{d} N(0,1). \tag{3.7}$$

此处 $\sigma_r^2 = \text{var}\{\sum_{i\in\mathcal{D}}\phi_1^*(O_i)\}$。

根据定理3.1 和命题3.1，我们可以看出，所提出的估计量具有半参数效率。此外，我们也可以类似地构造的估计量。定义

$$J_1 = \sum_{i\in\mathcal{D}_2}\left(\frac{\hat{S}_{yz,i}}{\sqrt{\hat{B}_0^y\hat{B}_0^z}} - \rho_n\frac{\hat{S}_{yy,i}}{2\hat{B}_0^y} - \rho_n\frac{\hat{S}_{zz,i}}{2\hat{B}_0^z}\right)^2,$$

其中

$$\begin{aligned}\hat{S}_{yz,i} = {} & n_y^{-1}\{Y_i - \hat{m}_{\mathcal{D}_1}(X_i)\}\{\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n\}\mathbb{I}(T_{y,i}=1)\\ & + n_z^{-1}\{\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n\}\{Z_i - \hat{h}_{\mathcal{D}_1}(X_i)\}\mathbb{I}(T_{z,i}=1)\\ & + n^{-1}[\{\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n\}\{\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n\} - I_n],\end{aligned}$$

$\hat{}\ \hat{}$ 和 $S_{yy,\ i}$ 以及 $S_{zz,\ i}$的定义类似。交换 $\mathcal{D}_1$和 $\mathcal{D}_2$的作用后，可类似地定义J_2。令 $\hat{}\sigma_r^2 = (J_1 + J_2)/4$。因此，对于 ρ，在置信水平为 $1-\alpha$ 时，有效置信区间为

$$\text{CI}_r(\alpha) = (\rho_n^* - z_{\alpha/2}\cdot\hat{\sigma}_r,\ \rho_n^* + z_{\alpha/2}\cdot\hat{\sigma}_r). \tag{3.8}$$

同样，正如定理2.4所示，所提出置信区间的有效性可以得到验证。为节省篇幅，证明细节从略；如有需要，可向作者索取。

4 广义遗传协方差与广义遗传相关

在前面的章节中，我们关注连续响应Y和Z。在某些应用中，感兴趣的响应变量Y或Z可能是离散型。因此，本节考虑广义遗传协方差和广义遗传相关，它们允许结果变量具有不同类型。回顾一下，$m(X) = E(Y \mid X)$和$h(X) = E(Z \mid X)$分别是Y和Z的遗传值。在这种情况下，稍微滥用一下符号，我们定义

$$I = \mathrm{cov}[g_1\{m(X)\}, g_2\{h(X)\}], \quad \text{and} \quad \rho = \frac{\mathrm{cov}[g_1\{m(X)\}, g_2\{h(X)\}]}{\sqrt{\mathrm{var}[g_1\{m(X)\}]\mathrm{var}[g_2\{h(X)\}]}}.$$

这里，$g_1(\cdot)$和$g_2(\cdot)$是两个已知的链接函数。例如，对于二元响应，链接函数可以取为$g_1(x) = \log x/(1-x)$。

为简化说明，令$E(g_1) = E[g_1\{m(X)\}]$，并令$E(g_2) = E[g_2\{h(X)\}]$。为推导I的有效影响函数，需要满足以下正则条件。

条件 4.1. 函数$g_1(\cdot)$和$g_2(\cdot)$连续可微，且$|g_i'(\cdot)| \leq C < \infty, i = 1, 2$。此外，$g_i'(\cdot)$、$i = 1$、2 满足利普希茨条件，其中正数$L > 0$,

$$|g_i'(x_1) - g_i'(x_2)| \leq L|x_1 - x_2| \quad \text{for all } x_1, x_2 \in \mathbb{R}.$$

条件 4.2. $E\{\sigma^4(X)\}$、$E\{\delta^4(X)\}$、$E([g_1\{m(X)\} - E(g_1)]^4)$和$E([g_2\{h(X)\} - E(g_2)]^4)$均为有限值。

回想一下，$\sigma^2(\mathrm{X}) = \mathrm{E}(\epsilon^2 \mid \mathrm{X})$且$\delta^2(\mathrm{X}) = \mathrm{E}(\eta^2 \mid \mathrm{X})$，其中$\epsilon = \mathrm{Y} - \mathrm{m(X)}$，$\eta = \mathrm{Z} - \mathrm{h(X)}$。条件 4.1 是温和且正则的。常见模型的链接函数均满足这一条件，包括标准线性模型、逻辑回归模型和多项逻辑回归模型。Wang et al. (2021) 也要求链接函数满足这一正则条件。条件 4.2 与条件 2.1 起相同作用，以确保 I 的有效影响函数的方差有限。

命题 4.1。在条件 4.1 和 4.2下，我们有

(a) I 的有效影响函数为

$$\begin{aligned} S_{yz} = & \frac{\mathbb{I}(T_y = 1)}{\Pr(T_y = 1)} g_1'\{m(X)\}\{Y^* - m(X)\}[g_2\{h(X)\} - E(g_2)] \\ & + \frac{\mathbb{I}(T_z = 1)}{\Pr(T_z = 1)} g_2'\{h(X)\}\{Z^* - h(X)\}[g_1\{m(X)\} - E(g_1)] \\ & + [g_1\{m(X)\} - E(g_1)][g_2\{h(X)\} - E(g_2)] - I. \end{aligned}$$

(b) ρ 的有效影响函数为

$$\phi_1(O, \mathcal{P}) = \frac{S_{yz}}{\sqrt{B_0^y B_0^z}} - \rho \frac{S_{yy}}{2B_0^y} - \rho \frac{S_{zz}}{2B_0^z},$$

其中

$$S_{yy} = 2\frac{\mathbb{I}(T_y = 1)}{\Pr(T_y = 1)} g_1'\{m(X)\}\{Y^* - m(X)\}[g_1\{m(X)\} - E(g_1)] + [g_1\{m(X)\} - E(g_1)]^2 - B_0^y,$$

$$S_{zz} = 2\frac{\mathbb{I}(T_z = 1)}{\Pr(T_z = 1)} g_2'\{h(X)\}\{Z^* - h(X)\}[g_2\{h(X)\} - E(g_2)] + [g_2\{h(X)\} - E(g_2)]^2 - B_0^z.$$

因此，I 和 ρ 的半参数效率界分别等于 $\mathrm{var}(S_{yz})$ 和 $\mathrm{var}\{\phi_1(O, \mathcal{P})\}$。

有效影响函数的表达式因条件不同而异。以下是两个示例 s.

示例 4.1. 当 $g_1(x) = g_2(x) = x$ 时，上述 I 和 ρ 的有效影响函数分别与（2.1）和（3.4）一致。

示例 4.2. 当 Y 和 Z 服从逻辑回归模型且为二元变量时，即

$$g_1\{m(X)\} = \log\left\{\frac{m(X)}{1 - m(X)}\right\} = X^\top \beta, \text{ and } g_2\{h(X)\} = \log\left\{\frac{h(X)}{1 - h(X)}\right\} = X^\top \gamma,$$

这意味着 $m(X) = \exp(X^\top\beta)/\{1 + \exp(X^\top\beta)\}$，$h(X) = \exp(X^\top\gamma)/\{1 + \exp(X^\top\gamma)\}$，且 $I = \beta^\top \Sigma \gamma$ with $\Sigma = \mathrm{var}(X)$。由此可得，I 的有效影响函数为

$$\begin{aligned} S_{yz} = & \frac{\mathbb{I}(T_y = 1)}{\Pr(T_y = 1)} \frac{1}{m(X)\{1 - m(X)\}} \{Y^* - m(X)\}\{X - E(X)\}^\top \gamma \\ & + \frac{\mathbb{I}(T_z = 1)}{\Pr(T_z = 1)} \frac{1}{h(X)\{1 - h(X)\}} \{Z^* - h(X)\}\{X - E(X)\}^\top \beta \\ & + \beta^\top \{X - E(X)\}\{X - E(X)\}^\top \gamma - \beta^\top \Sigma \gamma. \end{aligned}$$

基于上述有效影响函数，我们考虑I的以下估计量，

$$I_n = n_y^{-1} \sum_{i \in \mathcal{D}_{2y}} \hat{g}'_{1i}(Y_i - \hat{m}_i)(\hat{g}_{2i} - \bar{g}_2) + n_z^{-1} \sum_{i \in \mathcal{D}_{2z}} \{\hat{g}'_{2i}(Z_i - \hat{h}_i)\}(\hat{g}_{1i} - \bar{g}_1)$$
$$+ n^{-1} \sum_{i \in \mathcal{D}_2} (\hat{g}_{2i} - \bar{g}_2)(\hat{g}_{1i} - \bar{g}_1).$$

这里，$\hat{m}_i = \hat{m}_{\mathcal{D}_1}(X_i)$、$\hat{g}_{1i} = g_1(\hat{m}_i)$、$\hat{g}'_{1i} = g'_1(\hat{m}_i)$、$\bar{g}_1 = n^{-1} \sum_{i \in \mathcal{D}_2} \hat{g}_{1i}$、$\hat{h}_i$、$\hat{g}_{2i}$、$\hat{g}'_{2i}$ 和 $\bar{g}_2$ 的记号定义类似。此外，我们定义 $I_n^* = (I_n + I'_n)/2$，其中 I'_n 的定义类似，只需交换 $\mathcal{D}_1$ 和 $\mathcal{D}_2$ 的角色。

类似地，我们可以通过令估计得到的影响函数的样本均值为零，得到 ρ 的一致估计量 ρ_n，即

$$\rho_n = \frac{I_n}{\sqrt{\hat{B}_0^y \hat{B}_0^z}}. \tag{4.9}$$

其中

$$\hat{B}_0^y = 2n_y^{-1} \sum_{i \in \mathcal{D}_{2y}} \hat{g}'_{1i}(Y_i - \hat{m}_i)(\hat{g}_{1i} - \bar{g}_1) + n^{-1} \sum_{i \in \mathcal{D}_2} (\hat{g}_{1i} - \bar{g}_1)^2,$$
$$\hat{B}_0^z = 2n_z^{-1} \sum_{i \in \mathcal{D}_{2z}} \hat{g}'_{2i}(Z_i - \hat{h}_i)(\hat{g}_{2i} - \bar{g}_2) + n^{-1} \sum_{i \in \mathcal{D}_2} (\hat{g}_{2i} - \bar{g}_2)^2.$$

此外，$\rho_n^* = (\rho_n + \rho'_n)/2$，其中 ρ'_n 的定义类似，只需交换 $\mathcal{D}_1$ 和$\mathcal{D}_2$ 的角色。

在陈述I_n 和 ρ_n的渐近行为之前，需要满足以下正则条件：

条件4.3。 $E[\{\hat{m}_{\mathcal{D}_1}(X) - m(X)\}^4] = o(n_y^{-1})$，且 $E[\{\hat{h}_{\mathcal{D}_1}(X) - h(X)\}^4] = o(n_z^{-1})$。

条件4.4。 $E[\sigma^2(X)\{\hat{m}_{\mathcal{D}_1}(X) - m(X)\}^2\{\hat{h}_{\mathcal{D}_1}(X) - h(X)\}^2] = o(1)$，且 $E[\delta^2(X)\{\hat{m}_{\mathcal{D}_1}(X) - m(X)\}^2\{\hat{h}_{\mathcal{D}_1}(X) - h(X)\}^2] = o(1)$。

条件4.5。 $E(\sigma^2(X)\{\hat{m}_{\mathcal{D}_1}(X) - m(X)\}^2[g_2\{h(X)\} - E(g_2)]^2) = o(1)$，且 $E(\delta^2(X)\{\hat{h}_{\mathcal{D}_1}(X) - h(X)\}^2[g_1\{m(X)\} - E(g_1)]^2) = o(1)$。

条件4.3比条件2.2 更强，这是由链接函数导致的。条件4.4 和4.5 也是保证渐近正态性的技术条件。当 $\sigma^2(X)$ 和 $\delta^2(X)$ 均为有限值时，在条件4.4 下，该条件也可由条件4.3直接推出。

定理 4.1。假设条件4.1–4.5 均成立。我们有：(a) I_n^* 和 ρ_n^* 分别的渐近线性表示为

$$I_n^* - I = \sum_{i\in\mathcal{D}} S_{yz,i}^* + o_p(N_z^{-1/2} + N_y^{-1/2})$$
$$\rho_n^* - \rho = \sum_{i\in\mathcal{D}} \phi_1^*(O_i) + o_p(N_z^{-1/2} + N_y^{-1/2}).$$

Here

$$S_{yz,i}^* = N_y^{-1}\epsilon_i^*\mathbb{I}(T_{y,i}=1) + N_z^{-1}\eta_i^*\mathbb{I}(T_{z,i}=1) + N^{-1}\xi^*,$$
$$\phi_1^*(O_i) = \frac{S_{yz,i}^*}{\sqrt{B_0^y B_0^z}} - \rho\frac{S_{yy,i}^*}{2B_0^y} - \rho\frac{S_{zz,i}^*}{2B_0^z},$$

with

$$S_{yy,i}^* = 2N_y^{-1}g_{1i}'\{Y_i - m(X_i)\}\{g_{1i} - E(g_1)\}\mathbb{I}(T_{y,i}=1) + N^{-1}g_{1i}'[\{g_{1i} - E(g_1)\}^2 - B_0^y],$$

$$S_{zz,i}^* = 2N_z^{-1}g_{2i}'\{Z_i - h(X_i)\}\{g_{2i} - E(g_2)\}\mathbb{I}(T_{z,i}=1) + N^{-1}g_{2i}'[\{g_{2i} - E(g_2)\}^2 - B_0^z].$$

其中 $\epsilon_i^* = g_1'\{Y_i - m(X_i)\}\{g_{2i} - E(g_2)\}$，$\eta_i^* = g_2'\{Z_i - h(X_i)\}\{g_{1i} - E(g_1)\}$，且 $\xi_i^* = \{g_{1i} - E(g_1)\}\{g_{2i} - E(g_2)\} - I$。并且 $g_{1i} = g_1\{m(X_i)\}$、$g_{2i} = g_2\{h(X_i)\}$、$g_1' = g_1'\{m(X_i)\}$ 和 $g_2' = g_2'\{h(X_i)\}$。回顾一下，$E(g_1) = E[g_1\{m(X)\}]$，且 $E(g_2) = E[g_2\{h(X)\}]$。

（b）I_n^* 和 ρ_n^* 分别渐近服从正态分布：

$$\sigma^{-1}(I_n^* - I) \xrightarrow{d} N(0,1) \quad \text{and} \quad \sigma_r^{-1}(\rho_n^* - \rho) \xrightarrow{d} N(0,1).$$

这里，$\sigma^2 = \text{var}(\sum_{i\in\mathcal{D}} S_{yz,i}^*)$ 和 $\sigma_r^2 = \text{var}\{\sum_{i\in\mathcal{D}} \phi_1^*(O_i)\}$。

基于定理4.1，我们同样可以推导出，所提出的广义遗传协方差和遗传相关性估计量是半参数有效的。显然，定理4.1 也是定理2.3 和3.1的推广版本。因此，我们同样可以进一步开展推断程序，例如分别构造I 和 ρ的置信区间。

5 模拟研究

在本节中，我们开展了一些模拟研究，以说明所提出的方法。为估计$m(X)$和$h(X)$，我们采用了一些灵活的机器学习算法，例如

最小绝对收缩与选择算子（LASSO）和神经网络。对于 LASSO 方法，使用 R 软件包 glmnet 实现估计，并采用 10 折交叉验证选择相应的调优参数。对于神经网络，我们使用来自 sklearn 的多层感知机神经网络（MLP）。更具体地说，我们采用包含两个隐藏层的 MLP，其中每个隐藏层的神经元数量设为 100。此外，最大迭代次数为 5000，学习率选择为自适应。在每次实验中，我们重复模拟 500 次。

我们将响应变量Y_i、Z_i 视为连续型或离散型。连续响应变量分别由线性模型和非线性模型生成，而离散响应则由逻辑回归模型生成。预测变量X_i由均值为 $\mathcal{N}0$、协方差为 Σ的多元正态分布生成，其中$\Sigma_{ij} = 0.6^{|i-j|}$。

5.1 连续响应的遗传相关性

本小节中，我们关注连续响应情况下遗传协方差和遗传相关系数估计量的有限样本表现。

首先，我们考虑高维线性回归模型，其中$p = 400$为600或800，且$n_y = n_z = 400$。

例5.1。连续型结果变量由以下线性模型生成，

$$Y_i = X_i^\top \beta + \epsilon_i, \text{ and } Z_i = X_i^\top \gamma + \eta_i.$$

β的信号满足$\beta_j = 0.4(1 + j/2s_1)$，而 γ的信号满足 $\gamma_j = 0.3(1 - j/2s_2)$。这里，$s_1$ 和s_2 分别是 β 和 γ的稀疏度。此外，误差项ϵ_i 和 η_i 由标准正态分布$N(0, 1)$生成。在此例中，遗传协方差为$I = \beta^\top \Sigma \gamma$。

˜我们首先考察I_N 的一致性。结果列于表1 中，采用重叠设置，即两个性状的数据来自同一组个体；表2采用非重叠设置，即两个性状的数据来自不同个体组，且$N_0 = 0$。每一行报告了在不同稀疏度和维度大小设置下的平均偏差绝对值和相对偏差。这里，我们将相对偏差（记为 rBIAS）定义为平均偏差绝对值除以真实值I。我们

˜表1：估计遗传协方差的性能 I_N 当 $m(X)$ 被一致估计且$h(X)$ 被估计为 0 时：**示例 5.1 重叠设置**

				MLP		LASSO	
n_y	n_z	p	(秒$_1$, 秒$_2$)	BIAS	相对偏差	BIAS	相对偏差
400	400	400	(20,20)	0.04201	0.0053	0.1305	0.0164
			(30,30)	0.0570	0.0046	0.0460	0.0037
			(40,40)	0.0643	0.0038	0.1456	0.0087
400	400	600	(20,20)	0.0218	0.0027	0.0279	0.0035
			(30,30)	0.0334	0.0027	0.1698	0.0137
			(40,40)	0.0757	0.0045	0.1236	0.0074
400	400	800	(20,20)	0.0091	0.0012	0.0941	0.0118
			(30,30)	0.0628	0.0051	0.1072	0.0087
			(40,40)	0.0237	0.0014	0.2273	0.0136

MLP：采用 MLP 方法估计；LASSO：采用 LASSO 方法估计。

比较了两种用于估计$E(Y \mid X)$和$E(Z \mid X)$的方法：MLP 方法和 LASSO 方法。从表1中可以看出，无论p、s_1 和s_2 如何变化，我们的方法表现都很好，偏差也相对较小。从表2可以看出，无论p、s_1 和s_2 如何变化，相对偏差都在可接受范围内。此外，在重叠和非重叠˜情形下，MLP 方法的估计结果都可以与 LASSO 方法相媲美。从结果可以看出，I_N 具有一致性，且相对偏差较小。

接下来，我们研究 I_n^* 的表现。此外，我们还将本方法与 Wang 等人 (2021年提出的方法进行比较。我们提出的方法的结果汇总于表 3 和表 4。每一行报告了不同稀疏度和维度大小设置下的经验覆盖概率（记为 CP）、平均偏差的绝对值（记为 BIAS）以及估计的置信区间长度（记为 LEN）。本部分中，我们设定 $s_1 = s_2 = s$。根据表 3-4，我们得到以下发现。首先，可以看到，对于不同的 p 和 s，我们的方法表现良好。无论 p和 s 如何变化，覆盖概率始终约为 0.95，且偏差相对较小。其次，对于高维线性回归模型，LASSO 和 MLP 方法均表现良好。第三，与 Wang 等人 (2021年的方法相比，当两种方法的覆盖概率均约为 0.95 时，我们的方法偏差更小，置信区间更短。

˜表2：估计遗传协方差的表现 I_N 当 $m(X)$ 得到一致估计，且$h(X)$ 估计为 0：**非重叠设置下的例 5.1**

				MLP		LASSO	
n_y	n_z	p	(秒$_1$, 秒$_2$)	BIAS	相对偏差	BIAS	相对偏差
400	400	400	(20,20)	0.1317	0.0166	0.1465	0.0184
			(30,30)	0.1114	0.0090	0.2661	0.0215
			(40,40)	0.1891	0.0113	0.3790	0.0226
400	400	600	(20,20)	0.1003	0.0126	0.1416	0.0178
			(30,30)	0.1296	0.0105	0.2456	0.0199
			(40,40)	0.0179	0.0011	0.3697	0.0221
400	400	800	(20,20)	0.1123	0.0141	0.1838	0.0231
			(30,30)	0.2033	0.0165	0.2993	0.0242
			(40,40)	0.1160	0.0069	0.4233	0.0253

MLP：采用多层感知机方法估计；LASSO：采用 LASSO 方法估计。

表3：估计遗传协方差的性能 I_n^* 生成自 **示例 5.1重叠设置**

			MLP			LASSO			Wang		
n	p	s	覆盖概率	偏差	长度	覆盖概率	偏差	长度	覆盖概率	偏差	长度
400	400	10	0.930	0.027	0.870	0.924	0.017	0.851	0.944	0.044	1.232
		20	0.946	0.003	1.733	0.944	0.004	1.729	0.946	0.097	2.513
		30	0.950	0.081	2.587	0.958	0.015	2.595	0.950	0.026	3.721
400	600	10	0.944	0.011	0.913	0.934	0.006	0.851	0.928	0.065	1.241
		20	0.956	0.073	1.721	0.950	0.030	1.717	0.950	0.009	2.480
		30	0.932	0.102	2.066	0.950	0.003	2.615	0.958	0.093	3.718
400	800	10	0.936	0.046	0.866	0.940	0.023	0.855	0.930	0.047	1.228
		20	0.948	0.061	1.725	0.952	0.032	1.717	0.942	0.113	2.499
		30	0.944	0.096	2.587	0.940	0.091	2.575	0.938	0.144	3.728

MLP：采用 MLP 方法进行估计；LASSO：采用 LASSO 方法进行估计；Wang：所提出的方法Wang 等人（2021年）。

表4：估计遗传协方差的性能 I_n^* 生成自 **示例 5.1非重叠设置**

			MLP			LASSO			Wang		
n	p	s	覆盖概率	偏差	长度	覆盖概率	偏差	长度	覆盖概率	偏差	长度
400	400	10	0.922	0.025	0.808	0.922	0.011	0.741	0.926	0.049	0.904
		20	0.930	0.038	1.443	0.940	0.003	1.376	0.888	0.127	1.772
		30	0.924	0.044	2.070	0.954	0.028	1.999	0.920	0.201	2.638
400	600	10	0.924	0.016	0.815	0.938	0.013	0.753	0.930	0.059	0.895
		20	0.944	0.035	1.439	0.942	0.022	1.380	0.930	0.119	1.777
		30	0.932	0.030	2.066	0.940	0.017	2.011	0.900	0.276	2.628
400	800	10	0.938	0.023	0.815	0.922	0.016	0.745	0.908	0.077	0.893
		20	0.926	0.035	1.439	0.946	0.024	1.380	0.888	0.206	1.759
		30	0.918	0.033	2.066	0.944	0.060	2.003	0.892	0.287	2.618

MLP：通过 MLP 方法估计；LASSO：通过 LASSO 方法估计；Wang：由Wang 等人提出的方法（2021年）。

接下来，我们探讨使用线性模型估计连续响应变量遗传相关性的模拟研究。

Exam例 5.2。连续响应变量由以下线性模型生成。

$$Y_i = X_i^\top \beta + \epsilon_i \text{ and } Z_i = X_i^\top \gamma + \eta_i.$$

这里，β 的系数由 $\beta_i = 1 + \mathrm{i}/2\mathrm{s}_1$ 生成，其中 $1 \le \mathrm{i} \le \mathrm{s}_1,$；$\gamma$ 的信号满足 $\gamma_j = 2(1 - \mathrm{j}/2\mathrm{s}_2)$，其中 $\mathrm{s}_1 + 1 \le \mathrm{j} \le \mathrm{s}_1 + \mathrm{s}_2,$，且 $\beta\mathrm{j} = 0$，否则。s_1 和 s_2 分别表示 β 和 γ 的稀疏度。误差项 ϵ_i 和 η_i 与之前一样，服从正态分布 N(0, 1)。在本例中，遗传相关性为 $\rho = \beta^\top \Sigma\gamma / \sqrt{\beta^\top \Sigma\beta \cdot \gamma^\top \Sigma\gamma}$。

我们设定样本量 $ny = n_z = 400$，维度 $p = 400$，600 或 800。我们考虑不同的稀疏参数 $s_1 = 20$，25，30 和 $s_2 = 5$。表 ρ 分别展示了重叠设置和非重叠设置下遗传相关性的估计结果，见表 5 和表 6。首先可以看到，所提出的方法表现良好；当 p 和 s_1 在不同设置下取值时，覆盖概率约为 0.95，绝对偏差也相对较小。其次，我们发现 MLP 方法的覆盖概率接近 0.95，且

绝对偏差得到了控制，这与 LASSO 方法相似。因此，我们得出结论：尽管 MLP 方法的绝对偏差和置信区间长度均大于 LASSO 方法，但它仍能提供准确的估计结果。第三，总体而言，非重叠设置下的偏差和置信区间长度均小于重叠设置。原因是，尽管两种设置的样本设置$n^y = n_z = 400$相同，但非重叠设置的总样本量大于重叠设置。

表 5：估计遗传相关性的表现 ρ_n^* 生成自**示例5.2重叠设置**

			MLP			LASSO		
n	p	(秒$_1$, 秒$_2$)	CP	BIAS	SE	CP	BIAS	SE
400	400	(20,5)	0.926	0.0161	0.0408	0.928	0.0003	0.0363
		(25,5)	0.938	0.0160	0.0405	0.944	0.0006	0.0363
		(30,5)	0.944	0.0121	0.0406	0.950	0.0005	0.0365
400	600	(20,5)	0.926	0.0155	0.0409	0.942	0.0002	0.0363
		(25,5)	0.952	0.0105	0.0406	0.940	0.0014	0.0364
		(30,5)	0.926	0.0127	0.0406	0.950	0.0004	0.0364
400	800	(20,5)	0.938	0.0155	0.0409	0.950	0.0029	0.0363
		(25,5)	0.924	0.0157	0.0404	0.936	0.0002	0.0364
		(30,5)	0.932	0.0133	0.0412	0.926	0.0004	0.0365

MLP：由 MLP 方法估计；LASSO：由 LASSO 方法估计。

接下来，我们展示非线性模型设定下遗传协方差的模拟结果。误差项 ϵ_i 和 η_i 服从标准正态分布。

例5.3。 连续型结局Y_i 由一个非线性模型生成

$$Y_i = -5 + 2\sin(\pi x_{i1} x_{i2}) + 4(x_{i3} - 0.5)^2 + 2x_{i5} + x_{i6} + \epsilon_i,$$

且Z_i 由另一个非线性模型生成

$$Z_i = 3\sin(x_{i1}) + x_{i3}^3 + \exp(x_{i4}) + \eta_i.$$

此处x_j、$j = 1$、$\cdots$、6 是X的元素。回顾一下，预测变量X_i是从多元正态分布 $N(0, \Sigma)$生成的，其中$\Sigma_{ij} = 0.6^{|i-j|}$。

表6：估计遗传相关性 ρ_n^* 生成于**示例5.2非重叠设置**

			MLP			LASSO		
n	p	$(秒_1, 秒_2)$	CP	BIAS	SE	CP	BIAS	SE
400	400	(20,5)	0.922	0.0088	0.0329	0.934	0.0010	0.0272
		(25,5)	0.926	0.0079	0.0330	0.932	0.0010	0.0271
		(30,5)	0.922	0.0050	0.0330	0.940	0.0015	0.0270
400	600	(20,5)	0.924	0.0073	0.0331	0.928	0.0012	0.0272
		(25,5)	0.942	0.0067	0.0329	0.966	0.0009	0.0271
		(30,5)	0.954	0.0047	0.0326	0.950	0.0020	0.0270
400	800	(20,5)	0.924	0.0094	0.0331	0.930	0.0016	0.0272
		(25,5)	0.940	0.0069	0.0331	0.938	0.0017	0.0271
		(30,5)	0.932	0.0044	0.0327	0.942	0.0023	0.0271

MLP：通过多层感知机方法估计；LASSO：通过 LASSO 方法估计。

表7 报告了基于非线性模型的经验结果。标记为重叠和非重叠的列分别表示两个数据集来自相同样本和来自不同样本。每行报告不同维度大小设定下的经验覆盖概率（覆盖概率）、平均偏差的绝对值（偏差）和估计标准差（标准误）。我们使用多层感知机方法和LASSO 方法估计非线性模型。此外，我们还将本方法与 Wang 等人（2021年）提出的方法进行了比较。我们观察到，采用多层感知机方法的本方法，其覆盖概率仍可近似达到理想的0.95水平，并且对不同的p值具有稳健性。然而，采用LASSO 方法的结果存在有偏估计，估计覆盖概率低于0.90，且绝对偏差远大于多层感知机方法。作为比较，我们发现，Wang 等人（2021年）提出的方法采用LASSO 方法时，表现不如本方法采用多层感知机算法，在某些情形下偏差更大、覆盖概率更低。因此，在非线性模型设定下，无论是重叠情形还是非重叠情形，多层感知机方法都优于LASSO 方法。

表7：估计遗传协方差的表现I_n^*生成自**示例5.3**

			MLP			LASSO			Wang		
	n	p	CP	BIAS	SE	CP	BIAS	SE	CP	BIAS	SE
重叠	400	400	0.926	0.555	3.206	0.846	2.727	2.435	0.938	1.995	3.522
		600	0.928	0.513	3.258	0.866	2.892	2.460	0.930	2.338	3.667
		800	0.928	0.558	3.206	0.858	2.837	2.434	0.918	2.062	3.506
		1000	0.932	0.676	3.220	0.862	2.833	2.438	0.930	2.212	3.645
非重叠	400	400	0.918	0.230	2.713	0.884	2.441	2.467	0.872	2.552	3.158
		600	0.910	0.605	2.781	0.854	2.475	2.450	0.898	2.274	3.119
		800	0.912	0.519	2.766	0.862	2.360	2.445	0.924	2.336	3.172
		1000	0.912	0.478	2.710	0.882	2.123	2.046	0.870	2.401	3.080

重叠：两个数据集来自相同样本；非重叠：两个数据集来自不同样本。多层感知机：由 MLP 方法估计；LASSO：由 LASSO 方法估计；Wang：由 Wang 等人提出的方法估计（Wang 等人2021年）。

5.2 离散响应的遗传协方差

在本小节中，我们考察所提出的离散型响应估计量在有限样本下的表现。

例5.4。二元结果 Y_i 和 Z_i 由以下逻辑回归模型生成，

$$\Pr(Y_i = 1|X_i) = \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)}, \ \Pr(Z_i = 1|X_i) = \frac{\exp(X_i^\top \gamma)}{1 + \exp(X_i^\top \gamma)}.$$

我们考虑高维逻辑回归模型，其中 $p = 400$、600 或 800，且 $n_y = n_z = 400$。系数向量 β 和 γ 与线性模型类似。对于真实回归系数，给定支撑集 $\mathcal{S} = \{j : 1 \le j \le s\}$，$\beta$ 的信号满足 $\beta_j = 0.2(1 + j/2s)$，而 γ的信号满足 $\gamma_j = 0.3(1 - j/2s)$，对所有 $j \in \mathcal{S}$均成立。

数值结果汇总于重叠情形的表8 和非重叠情形的表 9。每一行报告不同维度大小设置及不同稀疏度下的经验覆盖概率（CP）、平均偏差的绝对值（BIAS）和估计的置信区间长度（LEN），其中维度大小分别为不同不同稀疏度水平下的 β 和 γ，取值为（s_1, s_2） $\in \{(5,5),(10,10)\}$。我们比较了两种用于估计函数的方法：MLP 算法和 LASSO 方法。此外，我们还将我们的方法与该方法进行了比较

由Ma 等人(2022)提出。

我们得到以下观察结果。首先，当 p 和 s 变化时，我们的方法所得的所有估计结果都处于可控范围内。覆盖概率始终接近正态水平，偏差也相对较小。其次，MLP 方法与 LASSO 方法的表现没有显著差异。具体而言，在重叠情形下，LASSO 方法的覆盖概率更高、偏差更小，置信区间长度也比 MLP 方法更短。对于非重叠设置，当覆盖概率处于相近水平时，MLP 方法比 LASSO 方法的偏差更小，置信区间长度也更短。第三，在所有设置下，我们的方法都比Ma 等人(2022) 提出的估计方法表现更令人满意。无论是在重叠情形还是非重叠情形下，我们的方法的覆盖概率都约为 0.95，而他们的结果约为 0.90。此外，与他们的方法相比，我们的方法偏差小得多，置信区间宽度也短得多。

表 8：重叠设置下，由 I_n^* 生成的**示例 5.4估计遗传协方差的性能**

			MLP			LASSO			Ma		
n	p	s	覆盖概率	偏差	长度	覆盖概率	偏差	长度	覆盖概率	偏差	长度
400	400	5	0.930	0.023	0.463	0.934	0.018	0.349	0.782	0.598	1.685
		10	0.934	0.065	1.125	0.944	0.076	0.886	0.800	1.171	2.991
400	600	5	0.942	0.022	0.447	0.952	0.015	0.345	0.800	0.769	2.058
		10	0.944	0.071	1.121	0.956	0.080	0.878	0.744	1.403	3.397
400	800	5	0.938	0.020	0.482	0.954	0.009	0.349	0.778	0.858	2.262
		10	0.940	0.075	0.933	0.962	0.064	0.882	0.728	1.514	3.554

MLP：采用 MLP 方法估计；LASSO：采用 LASSO 方法估计；Ma：由Ma 等人（2022）提出的方法。

总体而言，数值结果显示我们提出的方法具有优越性。我们注意到，在使用线性模型和逻辑回归模型时，MLP 方法的表现与 LASSO 方法一样好；但在使用非线性模型时，MLP 方法的估计结果比 LASSO 方法更准确。

表9：估计遗传协方差的表现 I_n^* 生成自**示例 5.4非重叠设置**

			MLP			LASSO			Ma		
n	p	s	覆盖概率	偏差	长度	覆盖概率	偏差	长度	覆盖概率	偏差	长度
400	400	5	0.938	0.017	0.474	0.942	0.028	0.517	0.860	0.535	1.671
		10	0.932	0.089	1.094	0.924	0.139	1.384	0.854	1.091	3.067
400	600	5	0.934	0.025	0.478	0.942	0.029	0.521	0.832	0.660	2.040
		10	0.946	0.076	1.129	0.928	0.149	1.392	0.874	1.216	3.402
400	800	5	0.926	0.025	0.470	0.944	0.035	0.521	0.888	0.707	2.264
		10	0.912	0.053	1.070	0.928	0.162	1.407	0.890	1.238	3.592

多层感知机：由 MLP 方法估计；LASSO：由 LASSO 方法估计；Ma：Ma 等人提出的方法 Ma 等人(2022)。

6 真实数据分析

为展示我们方法的实用性，我们分析了 Parker 等人 (2016 报道的一组 Carworth Farms (CFW) 白色 Parker 等人 (2016)小鼠数据集。根据 Parker 等人（2016）的研究，研究人员使用 1,200 只雄性 CFW 小鼠，对行为、生理和基因表达表型进行了全基因组关联研究（GWAS）。这些 CFW 小鼠的表型包括条件性恐惧、焦虑行为、甲基苯丙胺敏感性、前脉冲抑制、空腹血糖水平、体重、尾长、睾丸重量等。这些表型可分为三类：行为性、生理性和表达数量性状。此外，由于 CFW 小鼠不存在隐匿亲缘关系，我们可以将这些小鼠视为相互独立的个体。

Parker 等人 (2016)开展了一系列实验来测量行为性状。对于甲基苯丙胺敏感性性状，通过记录每只小鼠每天注射后在场地中央停留的时间（秒）来测量其运动活动。第一天在场地中央停留的时间可用于衡量对新环境的基线响应。第三天记录的时间可用于衡量甲基苯丙胺敏感性。条件性恐惧性状通过测量小鼠对刺激的冻结行为来衡量其对恐惧记忆的回忆。生理性状包括甲基苯丙胺敏感性测试期间测得的体重，以及

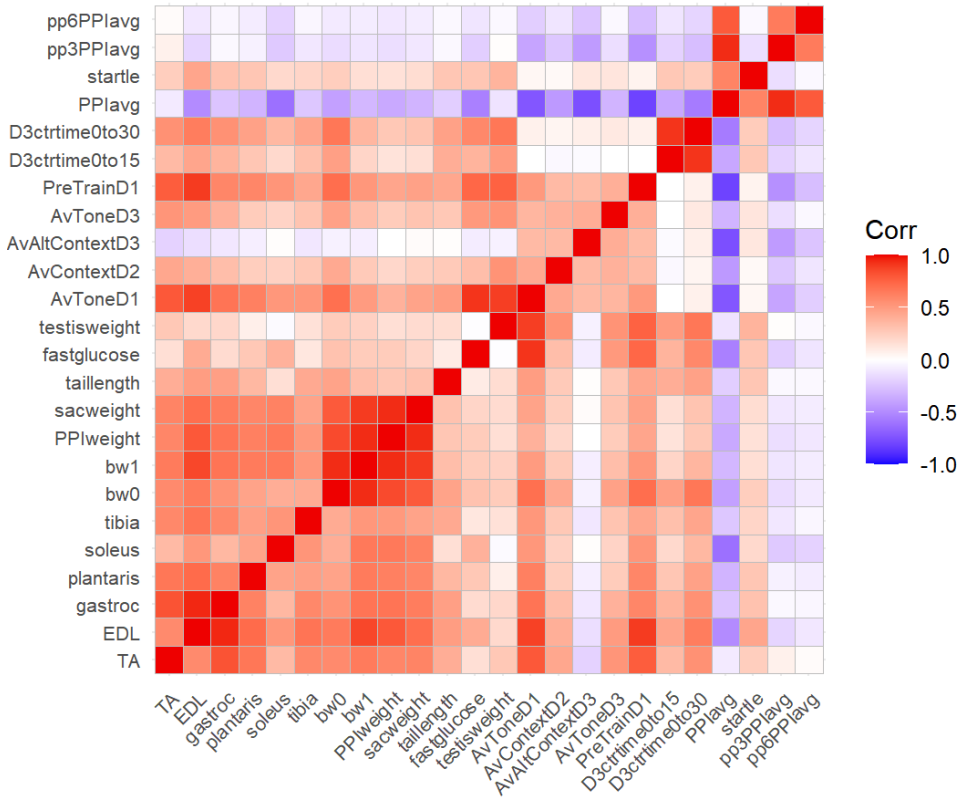

图1：24种表型的表型相关性热图。

各种肌肉的重量。

部分小鼠存在健康问题或其他问题，因此我们将这些观测值排除在后续分析之外。由于流动池处理不当或标记错误而被标记为可能存在样本混淆的观测值也被剔除。该数据集中的表型缺失情况不一，因此我们选择了缺失最少的24种表型。预处理后，数据集包含898只小鼠、92734个遗传变异（单核苷酸多态性）和24种不同的表型。我们计算了276对性状的两两遗传相关性，并在图1中报告了结果。

根据真实数据的估计结果，我们可以得出几项重要结论。首先，热图显示这些表型相关系数通常为正，因为热图中的大多数色块都是红色。其次，生理性状

具有共同的遗传结构，而行为性状也与其所属的家族存在遗传关联。然而，生理性状与行为性状之间的遗传相关性并不显著。例如，肌肉性状 *TA* 与其他肌肉性状密切相关，但与行为性状共享的常见遗传变异较少，这一点也在Wang 等人 (2021年) 的研究中有类似报道。同样，我们注意到，前脉冲抑制性状与其他行为性状存在显著遗传相关性，但与生理性状的关联并不显著。特别是，我们发现，前脉冲抑制性状 *PPlavg* 是对所有振幅下前脉冲反应与单独脉冲惊跳振幅之比所表示的抑制强度取平均值；除若干同类性状外，它与其他性状均呈强负相关。此外，*pp3PPlavg* 和 *pp6PPlavg* 分别是在 3 dB 和 6 dB 前脉冲试验中，抑制强度相对于单独脉冲惊跳振幅的平均值；二者表现出相似的特征。

7 结论与讨论

本文针对连续和离散响应变量，提出了遗传协方差和遗传相关性的半参数有效估计量。我们首先推导了遗传相关性的有效影响函数。只要遗传值中的一个能够得到一致估计，我们就提出了遗传协方差的一致估计量。基于有效影响函数，我们提出的估计量具有半参数有效性，且不存在模型设定错误的风险。我们的方法允许两个性状的数据来自两个不同的个体群体。

未来还有一些值得探讨的课题。例如，如何对非线性相关效应进行建模。在本文中，遗传协方差可以为零，而$EY|X$ 和$EZ|X$之间仍存在非线性依赖。此外，当 Y 和Z 是多变量时，如何定义合适的度量并进行推断也很值得研究。我们将在不久的将来探讨这些问题。

附录

记号：对于两个正数列a_n 和b_n，$a_n \lesssim b_n$ 表示对所有n，$a_n \leq Cb_n$；$a_n \gtrsim b_n$

若$b_n \lesssim a_n$且$a_n \asymp b_n$；若$a_n \lesssim b_n$且$a_n \gtrsim b_n$。C表示通用正常数，其取值可能随处变化。此外，我们用$a_n \lesssim_P b_n$表示$a_n \lesssim b_n$依概率收敛。

定理的证明 2.1：记 $\mathcal{P}$为 $O = (X^\top、Y^*、Z^*、T_y、T_z)^\top$。为便于说明，我们可以将 I 改写为

$$I = \Psi(\mathcal{P}) = E_{\mathcal{P}}\{[E_{\mathcal{P}}(Y|X) - E_{\mathcal{P}}(Y)][E_{\mathcal{P}}(Z|X) - E_{\mathcal{P}}(Z)]\}.$$

注意，指示变量 T_y和 T_z与 (X,Y,Z) 相互独立。因此我们有$E_{\mathcal{P}}(Y|X) = E_{\mathcal{P}}(Y^*|X,T_y = 1)$，以及$E_{\mathcal{P}}(Y) = E_{\mathcal{P}}(Y^*|T_y = 1)$。$Z$ 的情形类似。因此可得

$$I = \Psi(\mathcal{P}) = E_{\mathcal{P}}\{[E_{\mathcal{P}}(Y^*|X, T_y = 1) - E_{\mathcal{P}}(Y^*|T_y = 1)][E_{\mathcal{P}}(Z^*|X, T_z = 1) - E_{\mathcal{P}}(Z^*|T_z = 1)]\}$$

考虑以下由t索引的参数子模型，即

$$\mathcal{P}_t = t\tilde{\mathcal{P}} + (1-t)\mathcal{P},$$

其中 $t \in [0,1]$，且 $\tilde{\mathcal{P}}$是在单个观测值 $\tilde{o} = (\tilde{x}^\top, \tilde{y}^*, \tilde{z}^*, \tilde{t}_y, \tilde{t}_z)^\top$处的点质量。如 Hines 等人（2022）所述，I 在观测值$\tilde{o}$ 处的有效影响函数（EIF）直接表示为

$$\phi(\tilde{o}, \mathcal{P}) = \left.\frac{d\Psi(\mathcal{P}_t)}{dt}\right|_{t=0},$$

其中 $\Psi(\mathcal{P}_t) = E_{\mathcal{P}_t}\{[E^{\mathcal{P}_t}(Y^*|X, T_y = 1) - E^{\mathcal{P}_t}(Y^*|T_y = 1)][E^{\mathcal{P}_t}(Z^*|X, T_z = 1) - E^{\mathcal{P}_t}(Z^*|T_z = 1)]\}$。

记 $m_t^*(X) = E_{\mathcal{P}_t}\{[E^{\mathcal{P}_t}(Y^*|X, T_y = 1) - E^{\mathcal{P}_t}(Y^*|T_y = 1)]$，以及 $h_t^*(X) = [E^{\mathcal{P}_t}(Z^*|X, T_z = 1) - E^{\mathcal{P}_t}(Z^*|T_z = 1)]$。此外，算子 ∂_t 作用于任意函数 g(t) 时，定义如下：

$$\partial_t g(t) = \left.\frac{dg(t)}{dt}\right|_{t=0}.$$

简单计算可得

$$\partial_t \Psi(\mathcal{P}_t) = E_{\mathcal{P}}\{\partial_t[m_t^*(X)h_t^*(X)]\} + m_0^*(\tilde{x})h_0^*(\tilde{x}) - \Psi(\mathcal{P}).$$

根据以下事实

$$\partial_t E_{\mathcal{P}_t}(Y^*|X, T_y = 1) = \frac{\mathbb{I}_{(\tilde{x},\tilde{t}_y)}(X, T_y = 1)}{f(X, T_y = 1)}[\tilde{y}^* - E(Y^*|X, T_y = 1)],$$

and

$$\partial_t E_{\mathcal{P}_t}(Y|T_y = 1) = \frac{\mathbb{I}_{\tilde{t}_y}(T_y = 1)}{\Pr(T_y = 1)}[\tilde{y}^* - E(Y^*|T_y = 1)],$$

我们有

$$\begin{aligned} E_{\mathcal{P}}\{\partial_t[m_t^*(X)]h_0^*(X)\} &= \int \frac{\mathbb{I}_{(\tilde{x},\tilde{t}_y)}(X, T_y = 1)}{f(X, T_y = 1)}[\tilde{y}^* - E(Y^*|X, T_y = 1)]h_0^*(X)f(X)dX \\ &= \frac{\mathbb{I}_{\tilde{t}_y}(T_y = 1)}{\Pr(T_y = 1|\tilde{x})}[\tilde{y}^* - E(Y^*|\tilde{x}, T_y = 1)]h_0^*(\tilde{x}), \end{aligned}$$

其中第一个等式成立是因为$E[h_0^*(X)] = 0$。注意到$T_y \perp X$，我们有 $\Pr(T_y = 1|X) = \Pr(T_y = 1)$，这意味着

$$E_{\mathcal{P}}\{\partial_t[m_t^*(X)]h_0^*(X)\} = \frac{\mathbb{I}_{\tilde{t}_y}(T_y = 1)}{\Pr(T_y = 1)}[\tilde{y}^* - E(Y^*|\tilde{x}, T_y = 1)]h_0^*(\tilde{x}).$$

因此，根据导数的链式法则和商法则，我们得到

$$\begin{aligned} E_{\mathcal{P}}\{\partial_t[m_t^*(X)h_t^*(X)]\} = &\frac{\mathbb{I}_{\tilde{t}_y}(T_y = 1)}{\Pr(T_y = 1)}[\tilde{y}^* - E(Y^*|\tilde{x}, T_y = 1)][E(Z^*|\tilde{x}, T_z = 1) - E(Z^*|T_z = 1)] \\ &+ \frac{\mathbb{I}_{\tilde{t}_z}(T_z = 1)}{\Pr(T_z = 1)}[\tilde{z} - E(Z^*|\tilde{x}, T_z = 1)][E(Y^*|\tilde{x}, T_y = 1) - E(Y^*|T_y = 1)]. \end{aligned}$$

注意，$m_0^*(\tilde{x})h_0^*(\tilde{x}) = [E(Y^*|\tilde{x}, T_y = 1) - E(Y^*|T_y = 1)][E(Z^*|\tilde{x}, T_z = 1) - E(Z^*|T_z = 1)]$。这意味着

$$\begin{aligned} \phi(\tilde{o}, \mathcal{P}) = &\frac{\mathbb{I}_{\tilde{t}_y}(T_y = 1)}{\Pr(T_y = 1)}[\tilde{y}^* - E(Y^*|\tilde{x}, T_y = 1)][E(Z^*|\tilde{x}, T_z = 1) - E(Z^*|T_z = 1)] \\ &+ \frac{\mathbb{I}_{\tilde{t}_z}(T_z = 1)}{\Pr(T_z = 1)}[\tilde{z}^* - E(Z^*|\tilde{x}, T_z = 1)][E(Y^*|\tilde{x}, T_y = 1) - E(Y^*|T_y = 1)] \\ &+ [E(Y^*|\tilde{x}, T_y = 1) - E(Y^*|T_y = 1)][E(Z^*|\tilde{x}, T_z = 1) - E(Z^*|T_z = 1)] - \Psi(\mathcal{P}). \end{aligned} \tag{A1}$$

由于 $\phi(O,\mathcal{P})$ 具有有限方差，我们得出 I 的有效影响函数为

$$\begin{aligned}\phi(O,\mathcal{P}) = &\frac{\mathbb{I}_{T_y}(1)}{\Pr(T_y=1)}[Y^* - E(Y^*|X,T_y=1)][E(Z^*|X,T_z=1) - E(Z^*|T_z=1)]\\ &+ \frac{\mathbb{I}_{T_z}(1)}{\Pr(T_z=1)}[Z^* - E(Z^*|X,T_z=1)][E(Y^*|X,T_y=1) - E(Y^*|T_y=1)]\\ &+ [E(Y^*|X,T_y=1) - E(Y^*|T_y=1)][E(Z^*|X,T_z=1) - E(Z^*|T_z=1)] - \Psi(\mathcal{P}).\end{aligned}$$

定理的证明 2.2：设 $\epsilon = Y - m(X)$ 和 $\eta = Z - h(X)$。注意到

$$\begin{aligned}G_1 &= \frac{1}{N_y}\sum_{i=1}^{N_y}[Y_i - \hat{m}(X_i)][\hat{h}(X_i) - \bar{Z}_N]\\ &= \frac{1}{N_y}\sum_{i=1}^{N_y}[\epsilon_i + m(X_i) - \tilde{m}(X_i) + \tilde{m}(X_i) - \hat{m}(X_i)][\hat{h}(X_i) - \tilde{h}(X_i) + \tilde{h}(X_i) - EZ + EZ - \bar{Z}_N]\\ &= \frac{1}{N_y}\sum_{i=1}^{N_y}[\epsilon_i + m(X_i) - \tilde{m}(X_i)][\tilde{h}(X_i) - EZ] + \frac{1}{N_y}\sum_{i=1}^{N_y}[\tilde{m}(X_i) - \hat{m}(X_i)][\tilde{h}(X_i) - EZ]\\ &\quad + \frac{1}{N_y}\sum_{i=1}^{N_y}\epsilon_i[\hat{h}(X_i) - \tilde{h}(X_i)] + \frac{1}{N_y}\sum_{i=1}^{N_y}[m(X_i) - \tilde{m}(X_i)][\hat{h}(X_i) - \tilde{h}(X_i)]\\ &\quad + \frac{1}{N_y}\sum_{i=1}^{N_y}[\tilde{m}(X_i) - \hat{m}(X_i)][\hat{h}(X_i) - \tilde{h}(X_i)] + [EZ - \bar{Z}_N]\frac{1}{N_y}\sum_{i=1}^{N_y}[\epsilon_i + m(X_i) - \tilde{m}(X_i)]\\ &\quad + [EZ - \bar{Z}_N]\frac{1}{N_y}\sum_{i=1}^{N_y}[\tilde{m}(X_i) - \hat{m}(X_i)] := \sum_{i=1}^{8} G_{1i}.\end{aligned}$$

对于项G_{12},，由柯西-施瓦茨不等式可得

$$\begin{aligned}&\frac{1}{N_y}\sum_{i=1}^{N_y}[\tilde{m}(X_i) - \hat{m}(X_i)][\tilde{h}(X_i) - EZ]\\ &\le \left(\frac{1}{N_y}\sum_{i=1}^{N_y}[\tilde{m}(X_i) - \hat{m}(X_i)]^2\right)^{1/2}\left(\frac{1}{N_y}\sum_{i=1}^{N_y}[\tilde{h}(X_i) - EZ]^2\right)^{1/2} = o_p(1),\end{aligned}$$

其中，最后一个等式在以下条件下成立：$E[\tilde{h}(X_i) - EZ]^2 < \infty$，且 $E[\hat{h}(X_i) - h(X_i)]^2 = o(1)$。类似地，由于 $\hat{m}(X_i)$、$\hat{h}(X_i)$ 分别一致收敛于 $\tilde{m}(X_i)$、$\tilde{h}(X_i)$，因此我们得出，

分别地，我们得出

$$G_1 = \frac{1}{N_y}\sum_{i=1}^{N_y}[\epsilon_i + m(X_i) - \tilde{m}(X_i)][\tilde{h}(X_i) - EZ] + o_p(1)$$
$$= E\{[m(X) - \tilde{m}(X)][\tilde{h}(X) - EZ]\} + o_p(1),$$

˜最后一个等式由$E\{\epsilon[h(X_i) - EZ]\} = 0$这一事实得出。类似地，对于$I_N$的第二项，我们有$G_2 = E\{[\tilde{m}(X) - EY][h(X) - h(X)]\} + o_p(1)$。

˜对于I_N的第三项，注意到

$$G_3 = \frac{1}{N}\sum_{i=1}^{N}[\hat{m}(X_i) - \bar{Y}_N][\hat{h}(X_i) - \bar{Z}_N]$$
$$= \frac{1}{N}\sum_{i=1}^{N}[\hat{m}(X_i) - \tilde{m}(X_i) + \tilde{m}(X_i) - EY + EY - \bar{Y}_N][\hat{h}(X_i) - \tilde{h}(X_i) + \tilde{h}(X_i) - EZ + EZ - \bar{Z}_N]$$
$$= \frac{1}{N}\sum_{i=1}^{N}[\tilde{m}(X_i) - EY][\tilde{h}(X_i) - EZ] + o_p(1)$$
$$= E[\tilde{m}(X) - EY][\tilde{h}(X) - EZ] + o_p(1).$$

第三个等式成立是因为这些估计量具有一致性。

综上，我们有

$$\tilde{I}_N = G_1 + G_2 + G_3 = E[\tilde{m}(X_i) - m(X_i)][h(X_i) - \tilde{h}(X_i)] + I + o_p(1).$$

于是定理得证。

定理2.3的证明；注意，本定理的证明是定理4.1 在$g_1(x) = g_2(x) = x$证明的一个特例。因此省略详细证明。

定理的证明 2.4：若 $(\hat{\sigma}^2 - \sigma^2)/\sigma^2 = o_p(1)$ 成立，请注意

$$\frac{I_n^* - I}{\sigma} \xrightarrow{d} N\Big(0, 1\Big),$$

这意味着

$$\frac{I_n^* - I}{\hat{\sigma}} \xrightarrow{d} N\Big(0, 1\Big).$$

因此，我们可以完成证明。

接下来，我们的目标是证明 $(\hat{\sigma}^2 - \sigma^2)/\sigma^2 = o_P(1)$。记

$$\delta_i - n^{-1}I = \frac{1}{n_y}\epsilon_i^* I(i \in \mathcal{D}_{2y}) + \frac{1}{n_z}\eta_i^* I(i \in \mathcal{D}_{2z}) + \frac{1}{n}\xi_i^*,$$

以及 $\tilde{K}_1 = \sum_{i \in \mathcal{D}_2}(\delta_i - n^{-1}I)^2$，，其中 $\epsilon_i^* = \epsilon_i[h(X_i) - EZ]$，$\eta_i^* = \eta_i[m(X_i) - EY]$，，且 $\xi_i^* = [m(X_i) - EY][h(X_i) - EZ] - I$。类似地，我们可以定义 K_2。记

$$\tilde{\sigma}^2 = \frac{\tilde{K}_1 + \tilde{K}_2}{4}.$$

我们有

$$\frac{\hat{\sigma}^2 - \sigma^2}{\sigma^2} = \frac{\tilde{\sigma}^2 - \sigma^2}{\sigma^2} + \frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\sigma^2}. \tag{A2}$$

我们首先考虑第一项。注意，$E(\tilde{\sigma}^2 - \sigma^2) = 0$，并且

$$\mathrm{Var}(\tilde{\sigma}^2 - \sigma^2) \lesssim \mathrm{Var}(\tilde{K}_1) + \mathrm{Var}(\tilde{K}_2).$$

我们有

$$\mathrm{Var}(\tilde{K}_1) \lesssim E[\sum_{i \in \mathcal{D}_2}(\delta_i - n^{-1}I)^4] \lesssim \frac{1}{n_y^3}E(\epsilon_i^{*4}) + \frac{1}{n_z^3}E(\eta_i^{*4}) + \frac{1}{n^3}E(\xi_i^{*4}),$$

并且 $\mathrm{Var}(\tilde{K}_2) \lesssim \frac{1}{n_y^3}E(\epsilon_i^{*4}) + \frac{1}{n_z^3}E(\eta_i^{*4}) + \frac{1}{n^3}E(\xi_i^{*4})$，这意味着

$$\tilde{\sigma}^2 - \sigma^2 = O_p\Big(\frac{1}{n_y^{3/2}} + \frac{1}{n_z^{3/2}}\Big).$$

由于

$$\sigma^2 = \frac{1}{2n_y}E(\epsilon_i^{*2}) + \frac{1}{2n_z}E(\eta_i^{*2}) + \frac{1}{2n}E(\xi_i^{*2}) + \frac{N_0}{N_y N_z}E[\epsilon^*\eta^*] = O_p\left(\frac{1}{n_y} + \frac{1}{n_z}\right),$$

由此可得

$$\frac{\tilde{\sigma}^2 - \sigma^2}{\sigma^2} = o_p(1).$$

接下来，我们考虑 ($\hat{}\sigma^2 - \tilde{\sigma}^2$)/$\sigma^2$ 的界。注意，

$$\frac{\hat{\sigma}^2 - \tilde{\sigma}^2}{\sigma^2} = \frac{K_1 - \tilde{K}_1}{4\sigma^2} + \frac{K_2 - \tilde{K}_2}{4\sigma^2}.$$

˜我们可以将K_1 –K_1 改写为

$$\begin{aligned}
K_1 - \tilde{K}_1 &= \sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - n^{-1}I_n)^2 - \sum_{i\in\mathcal{D}_2}(\delta_i - n^{-1}I)^2 \\
&= \sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - n^{-1}I_n - \delta_i + n^{-1}I)^2 + 2\sum_{i\in\mathcal{D}_2}(\delta_i - n^{-1}I)\sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - n^{-1}I_n - \delta_i + n^{-1}I) \\
&\le \sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - n^{-1}I_n - \delta_i + n^{-1}I)^2 + 2\sqrt{\tilde{K}_1}\sqrt{\sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - n^{-1}I_n - \delta_i + n^{-1}I)^2} \\
&=: K_{11} + 2K_{12}.
\end{aligned}$$

对于K_{11},这一项，我们有

$$\sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - n^{-1}I_n - \delta_i + n^{-1}I)^2 \le 2n^{-1}(I_n - I)^2 + 2\sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - \delta_i)^2.$$

根据I_n的渐近正态性，我们有（I_n –I)^2 = O_p(2σ^2)，这意味着n^{-1}(I_n –I)^2/4σ^2 =o_p(1)。

因此，只需证明 $\sum_{i\in\mathcal{D}_2}$($\hat{}\delta_i - \delta_i$)^2/4$\sigma^2$ = o_p(1)。注意

$$\begin{aligned}
\sum_{i\in\mathcal{D}_2}(\hat{\delta}_i - \delta_i)^2 \lesssim\ & \frac{1}{n_y^2}\sum_{i\in\mathcal{D}_{2y}}\{[Y_i - \hat{m}_{\mathcal{D}_1}(X_i)][\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n] - [Y_i - m(X_i)][h(X_i) - EZ]\}^2 \\
&+ \frac{1}{n_z^2}\sum_{i\in\mathcal{D}_{2z}}\{[\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n][Z_i - \hat{h}_{\mathcal{D}_1}(X_i)] - [m(X_i) - EY][Z_i - h(X_i)]\}^2 \\
&+ \frac{1}{n^2}\sum_{i\in\mathcal{D}_2}\{[\hat{m}_{\mathcal{D}_1}(X_i) - \bar{Y}_n][\hat{h}_{\mathcal{D}_1}(X_i) - \bar{Z}_n] - [m(X_i) - EY][h(X_i) - EZ]\}^2 \\
=:\ & K_{111} + K_{112} + K_{113}.
\end{aligned}$$

我们首先考虑项 K_{111} 的界。注意

$$
\begin{aligned}
K_{111} &= \frac{1}{n_y^2}\sum_{i\in\mathcal{D}_{2y}}\{[Y_i-\hat{m}_{\mathcal{D}_1}(X_i)][\hat{h}_{\mathcal{D}_1}(X_i)-\bar{Z}_n]-[Y_i-m(X_i)][h(X_i)-EZ]\}^2 \\
&\lesssim_P \frac{1}{n_y^2}\sum_{i\in\mathcal{D}_{2y}}\epsilon_i^2[\hat{h}_{\mathcal{D}_1}(X_i)-h(X_i)]^2+[EZ-\bar{Z}_n]^2\frac{1}{n_y^2}\sum_{i\in\mathcal{D}_{2y}}\epsilon_i^2 \\
&\quad+\frac{1}{n_y^2}\sum_{i\in\mathcal{D}_{2y}}[\hat{m}_{\mathcal{D}_1}(X_i)-m(X_i)]^2[h(X_i)-EZ]^2+\frac{1}{n_y^2}\sum_{i\in\mathcal{D}_{2y}}[\hat{m}_{\mathcal{D}_1}(X_i)-m(X_i)]^2[\hat{h}_{\mathcal{D}_1}(X_i)-h(X_i)]^2 \\
&\quad+[EZ-\bar{Z}_n]^2\frac{1}{n_y^2}\sum_{i\in\mathcal{D}_{2y}}[\hat{m}_{\mathcal{D}_1}(X_i)-m(X_i)]^2.
\end{aligned}
$$

根据 $[EZ-\bar{Z}_n]^2=O_P(n_z^{-1})$这一事实，并在条件2.1–2.4下，容易证明

$$
E(|K_{111}|)=o_p\left(\frac{1}{n_y}\right),
$$

由此可得$K_{111}=o_p\left(n_y^{-1}\right)$。

类似地，我们可以得到

$$
K_{112}=o_p\left(\frac{1}{n_z}\right),\ \text{and}\ K_{113}=o_p\left(\frac{1}{n}\right).
$$

因此

$$
\frac{\sum_{i\in\mathcal{D}_2}(\hat{\delta}_i-\delta_i)^2}{4\sigma^2}=o_p(1),
$$

这意味着$K_{11}/4\sigma^2=o_P(1)$。

与对 $\tilde{\sigma}^2$的讨论类似，我们可以证明

$$
\frac{\tilde{K}_1}{4\sigma^2}=o_p(1).
$$

因此，对于项 K_{12}，我们有

$$
\frac{K_{12}}{4\sigma^2}=\sqrt{\frac{\tilde{K}_1}{4\sigma^2}}\sqrt{\frac{K_{11}}{4\sigma^2}}=o_p(1).
$$

根据上述结果，可得

$$(K_1 - \tilde{K}_1)/4\sigma^2 = o_p(1).$$

同样，我们可以证明

$$(K_2 - \tilde{K}_2)/4\sigma^2 = o_p(1).$$

由此完成证明。

命题 *3.1*的证明：同样，我们可以将 ρ 重写为

$$\rho = \Psi_1(\mathcal{P}) = \frac{Cov[m(X), h(X)]}{\sqrt{Var[m(X)]Var[h(X)]}}.$$

考虑相同的参数子模型，

$$\mathcal{P}_t = t\tilde{\mathcal{P}} + (1-t)\mathcal{P},$$

其中 $t \in [0,1]$。因此， ρ 在观测值 $\tilde{o}$ 处的有效影响函数（EIF）可直接写为

$$\phi_1(\tilde{o}, \mathcal{P}) = \frac{d\Psi_1(\mathcal{P}_t)}{dt}|_{t=0} = \frac{\phi(\tilde{o}, \mathcal{P})}{\sqrt{B_0^y B_0^z}} - \frac{\Psi_1(\mathcal{P})}{2B_0^y B_0^z}\partial_t(B_t^y B_t^u),$$

其中 $\phi(\tilde{o}, \mathcal{P})$ 是 I 的有效影响函数，$B_t^y = E_{\mathcal{P}_t}\{[E^{\mathcal{P}_t}(Y^*|X, T_y = 1) - E^{\mathcal{P}_t}(Y^*|T_y = 1)]^2\}$，且 $B_0^y = Var(E[Y^*|X, T_y = 1])$，$B_t^z$ 的定义与此类似。同样，按照定理 2.2 中的论证，我们可以证明 $\partial_t B_t^u$ 等于

$$\begin{aligned}\phi_u(\tilde{o}, \mathcal{P}) = 2\frac{\mathbb{I}_{\tilde{t}_u}(T_u = 1)}{P(T_u = 1)}[\tilde{u}^* - E(U^*|\tilde{x}, T_u = 1)][E(U^*|\tilde{x}, T_u = 1) - E(U^*|T_u = 1)] \\ + [E(U^*|\tilde{x}, T_u = 1) - E(U^*|T_u = 1)]^2 - B_0^u.\end{aligned}$$

根据导数的链式法则和商法则，我们得到

$$\begin{aligned}\phi_1(\tilde{o}, \mathcal{P}) &= \frac{\phi(\tilde{o}, \mathcal{P})}{\sqrt{B_0^y B_0^z}} - \frac{\Psi_1(\mathcal{P})}{2B_0^y B_0^z}[\phi_y(\tilde{o}, \mathcal{P})B_0^z + B_0^y \phi_z(\tilde{o}, \mathcal{P})] \\ &= \frac{S_{yz}}{\sqrt{B_0^y B_0^z}} - \rho\frac{S_{yy}}{2B_0^y} - \rho\frac{S_{zz}}{2B_0^z}.\end{aligned}$$

注意 $\phi_1(O, \mathcal{P})$的方差是有限的，这意味着 ρ 的有效影响函数是 $\phi_1(O, \mathcal{P})$。

定理证明 *3.1*：在给出证明之前，先提出一个必要的引理 .

引理 .1.假设条件 2.1–2.2 均成立。$\hat{}B_0^y$ 和 $\hat{}B_0^z$ 分别的渐近线性表示为

$$\hat{B}_0^y - B_0^y = \sum_{i\in\mathcal{D}_2} S_{yy,i}^2 + o_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n_y}}\right),$$
$$\hat{B}_0^z - B_0^z = \sum_{i\in\mathcal{D}_2} S_{zz,i}^2 + o_p\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{n_z}}\right).$$

Here

$$S_{yy,i}^2 = \frac{2}{n_y}\epsilon_i[h(X_i) - EZ]\mathbb{I}(T_{y,i} = 1) + \frac{1}{n}\{[m(X_i) - EY]^2 - B_0^y\},$$

and

$$S_{zz,i}^2 = \frac{2}{n_z}\eta_i[m(X_i) - EY]\mathbb{I}(T_{z,i} = 1) + \frac{1}{n}\{[h(X_i) - EZ]^2 - B_0^z\}.$$

证明。 该引理的证明与定理 2.3 的证明类似，因此此处省略。 □

注意

$$\begin{aligned}\rho_n - \rho &= \Big(\frac{I_n}{\sqrt{\hat{B}_0^y\hat{B}_0^z}} - \frac{I}{\sqrt{B_0^yB_0^z}}\Big)\\ &= \frac{(I_n - I)}{\sqrt{B_0^yB_0^z}} + I\Big(\frac{1}{\sqrt{\hat{B}_0^y\hat{B}_0^z}} - \frac{1}{\sqrt{B_0^yB_0^z}}\Big) + (I_n - I)\Big(\frac{1}{\sqrt{\hat{B}_0^y\hat{B}_0^z}} - \frac{1}{\sqrt{B_0^yB_0^z}}\Big)\\ &=: C_1 + C_2 + C_3.\end{aligned}$$

根据定理 2.3的证明，我们有

$$C_1 = \sum_{i\in\mathcal{D}_2} \frac{S_{yz,i}}{\sqrt{B_0^yB_0^z}} + o_p\left(\frac{1}{\sqrt{n_z}} + \frac{1}{\sqrt{n_y}}\right),$$

其中

$$S_{yz,i} = \frac{1}{n_y}\epsilon_i[h(X_i) - EZ]\mathbb{I}(T_{y,i} = 1) + \frac{1}{n_z}\eta_i[m(X_i) - EY]I(D_{z,i} = 1) + \frac{1}{n}[m(X_i) - EY][h(X_i) - EZ].$$

通过泰勒级数展开，我们得到

$$\begin{aligned}\frac{1}{\sqrt{\hat{B}_0^y\hat{B}_0^z}} - \frac{1}{\sqrt{B_0^yB_0^z}} = &-\frac{1}{2}(B_0^yB_0^z)^{-3/2}B_0^z(\hat{B}_0^y - B_0^y) - \frac{1}{2}(B_0^yB_0^z)^{-3/2}B_0^y(\hat{B}_0^z - B_0^z)\\ &+ o_p(|\hat{B}_0^y - B_0^y|^2 + |\hat{B}_0^z - B_0^z|^2).\end{aligned}$$

由引理.1，可得

$$\frac{1}{\sqrt{\hat{B}_0^y \hat{B}_0^z}} - \frac{1}{\sqrt{B_0^y B_0^z}} = -\frac{1}{2}(B_0^y B_0^z)^{-3/2} B_0^z (\hat{B}_0^y - B_0^y) - \frac{1}{2}(B_0^y B_0^z)^{-3/2} B_0^y (\hat{B}_0^z - B_0^z) + o_p\left(\frac{1}{\sqrt{n_z}} + \frac{1}{\sqrt{n_y}}\right)$$
$$= -\frac{(B_0^y B_0^z)^{-3/2}}{2} \sum_{i \in \mathcal{D}_2} \{B_0^z S_{yy,i} + B_0^y S_{zz,i}\} + o_p\left(\frac{1}{\sqrt{n_z}} + \frac{1}{\sqrt{n_y}}\right).$$

因此，

$$C_2 = -\frac{\rho}{2B_0^y B_0^z} \sum_{i \in \mathcal{D}_2} \{B_0^z S_{yy,i} + B_0^y S_{zz,i}\} + o_p\left(\frac{1}{\sqrt{n_z}} + \frac{1}{\sqrt{n_y}}\right).$$

另请注意，$I_n - I = o_P(1)$。因此很容易得到 $C_3 = o_p\left(\frac{1}{\sqrt{n_z}} + \frac{1}{\sqrt{n_y}}\right)$。

结合上述结果，我们可以得出

$$\rho_n - \rho = \sum_{i \in \mathcal{D}_2} \frac{S_{yz,i}}{\sqrt{B_0^y B_0^z}} - \rho \frac{S_{yy,i}}{2B_0^y} - \rho \frac{S_{zz,i}}{2B_0^z} + o_p\left(\frac{1}{\sqrt{n_z}} + \frac{1}{\sqrt{n_y}}\right).$$

类似地，对于ρ_n'，我们有

$$\rho_n' - \rho = \sum_{i \in \mathcal{D}_1} \frac{S_{yz,i}}{\sqrt{B_0^y B_0^z}} - \rho \frac{S_{yy,i}}{2B_0^y} - \rho \frac{S_{zz,i}}{2B_0^z} + o_p\left(\frac{1}{\sqrt{n_z}} + \frac{1}{\sqrt{n_y}}\right).$$

显然， ρ_n 和 ρ_n' 渐近独立。因此我们可以得出

$$\rho_n^* - \rho = \sum_{i \in \mathcal{D}} \frac{S_{yz,i}^*}{\sqrt{B_0^y B_0^z}} - \rho \frac{S_{yy,i}^*}{2B_0^y} - \rho \frac{S_{zz,i}^*}{2B_0^z} + o_p\left(\frac{1}{\sqrt{N_z}} + \frac{1}{\sqrt{N_y}}\right).$$

Here

$$S_{yz,i}^* = \frac{1}{N_y} \epsilon_i [h(X_i) - EZ] \mathbb{I}(T_{y,i} = 1) + \frac{1}{N_z} \eta_i [m(X_i) - EY] I(D_{z,i} = 1)$$
$$+ \frac{1}{N} \{[m(X_i) - EY][h(X_i) - EZ] - I\},$$

并且 $S_{yy,i}^*$ 以及 $S^*{}_{zz,i}$ 的定义类似。

命题 *4.1* 的证明：我们首先将 I 改写为

$$I = \Psi(\mathcal{P}) = E_{\mathcal{P}}\{g_1^*(X)g_2^*(X)\},$$

其中 $g_1^*(X) = g_1[E_{\mathcal{P}}(Y^*|X, T_y = 1)] - E_{\mathcal{P}}\{g_1[E_{\mathcal{P}}(Y^*|X, T_y = 1)]\}$，且 $g_2^*(X) = g_2[E_{\mathcal{P}}(Z^*|X, T_z = 1)] - E_{\mathcal{P}}\{g_2[E_{\mathcal{P}}(Z^*|X, T_z = 1)]\}$。考虑由 t 索引的同一参数子模型，即

$$\mathcal{P}_t = t\tilde{\mathcal{P}} + (1-t)\mathcal{P},$$

其中 $t \in [0,1]$，且 $\mathcal{P}$ 是位于单个观测值 $\tilde{o}$ 处的点质量。因此，I 在观测值 $\tilde{o}$ 处的有效影响函数（EIF）可直接写为

$$\phi(\tilde{o}, \mathcal{P}) = \frac{d\Psi(\mathcal{P}_t)}{dt}|_{t=0} = E_{\mathcal{P}}\{\partial_t[g_{1t}^*(X)g_{2t}^*(X)]\} + \partial_t E_{\mathcal{P}_t}\{g_1^*(X)g_2^*(X)\}.$$

其中 $g_{1t}^*(X) = g_1[E_{\mathcal{P}_t}(Y^*|X, T_y = 1)] - E_{\mathcal{P}_t}\{g_1[E_{\mathcal{P}_t}(Y^*|X, T_y = 1)]\}$，且 $g_{2t}^*(X) = g_2[E_{\mathcal{P}_t}(Z^*|X, T_z = 1)] - E_{\mathcal{P}_t}\{g_2[E_{\mathcal{P}_t}(Z^*|X, T_z = 1)]\}$。

首先，我们关注项 $\partial_t g_{1t}^*$。注意，

$$\begin{aligned}
\partial_t g_{1t}^*(X) &= \partial_t\Big[g_1[E_{\mathcal{P}_t}(Y^*|X, T_y = 1)]\Big] - \partial_t\Big(\int g_1[E_{\mathcal{P}_t}(Y^*|X, T_y = 1)]f_t(X)dX\Big) \\
&= g_1'[E_{\mathcal{P}}(Y^*|X, T_y = 1)]\partial_t E_{\mathcal{P}_t}(Y^*|X, T_y = 1)] - g_1[E_{\mathcal{P}}(Y^*|\tilde{x}, T_y = 1)] \\
&\quad + E_{\mathcal{P}}(g_1[E_{\mathcal{P}}(Y^*|X, T_y = 1)]) - E_{\mathcal{P}}\left(\partial_t\Big[g_1[E_{\mathcal{P}_t}(Y^*|X, T_y = 1)]\Big]\right) =: \psi_1(\tilde{o}, \mathcal{P}).
\end{aligned}$$

类似地，我们可以得到

$$\begin{aligned}
\partial_t g_{2t}^*(X) &= \partial_t\Big[g_2[E_{\mathcal{P}_t}(Z^*|X, T_z = 1)]\Big] - \partial_t\Big(\int g_2[E_{\mathcal{P}_t}(Z^*|X, T_z = 1)]f_t(X)dX\Big) \\
&= g_2'[E_{\mathcal{P}}(Z^*|X, T_z = 1)]\partial_t E_{\mathcal{P}_t}(Z^*|X, T_z = 1)] - g_2[E_{\mathcal{P}}(Z^*|\tilde{x}, T_z = 1)] \\
&\quad + E_{\mathcal{P}}(g_2[E_{\mathcal{P}}(Z^*|X, T_z = 1)]) - E_{\mathcal{P}}\left(\partial_t\Big[g_1[E_{\mathcal{P}_t}(Z^*|X, T_z = 1)]\Big]\right) =: \psi_2(\tilde{o}, \mathcal{P}).
\end{aligned}$$

根据导数的链式法则和商法则，可以得出

$$\begin{aligned}
E_{\mathcal{P}}\{\partial_t[g_{1t}^*(X)g_{2t}^*(X)]\} &= E_{\mathcal{P}}\{\psi_1(\tilde{o}, \mathcal{P})g_2^*(X)\} + E_{\mathcal{P}}\{g_1^*(X)\psi_2(\tilde{o}, \mathcal{P})\} \\
&= E_{\mathcal{P}}\{g_1'[E_{\mathcal{P}}(Y^*|X, T_y = 1)]\partial_t E_{\mathcal{P}_t}(Y^*|X, T_y = 1)]g_2^*(X)\} \\
&\quad + E_{\mathcal{P}}\{g_1^*(X)g_2'[E_{\mathcal{P}}(Z^*|X, T_z = 1)]\partial_t E_{\mathcal{P}_t}(Z^*|X, D_z)]\}.
\end{aligned}$$

最后一个等式成立是因为 $E_{\mathcal{P}}[g_i^*(X)] = 0,\ i = 1,\ 2$. 根据定理 2.1，我们有

$$\partial_t E_{\mathcal{P}_t}(Y^*|X, T_y = 1) = \frac{\mathbb{I}_{(\tilde{x},\tilde{t}_y)}(X, T_y = 1)}{f(X, T_y = 1)}[\tilde{y}^* - E(Y^*|X, T_y = 1)],$$

这意味着

$$\begin{aligned}
&E_{\mathcal{P}}\{g_1'[E_{\mathcal{P}_t}(Y^*|X, T_y = 1)]\partial_t E_{\mathcal{P}_t}(Y^*|X, T_y = 1)]g_2^*(X)\} \\
=&\frac{\mathbb{I}_{\tilde{t}_y}(T_y = 1)}{P(T_y = 1)}g_1'[E_{\mathcal{P}}(Y^*|\tilde{x}, T_y = 1)][\tilde{y}^* - E(Y^*|\tilde{x}, T_y = 1)]g_2^*(\tilde{x}),
\end{aligned}$$

其中，最后一个等式成立是因为$T_y \perp X$。

类似地，我们有

$$\begin{aligned}
&E_{\mathcal{P}}\{g_2'[E_{\mathcal{P}_t}(Z^*|X, T_z = 1)]\partial_t E_{\mathcal{P}_t}(Z^*|X, T_z = 1)]g_1^*(X)\} \\
=&\frac{\mathbb{I}_{\tilde{t}_z}(T_z = 1)}{P(T_z = 1)}g_2'[E_{\mathcal{P}}(Z^*|\tilde{x}, T_z = 1)][\tilde{z}^* - E(Z^*|\tilde{x}, T_z = 1)]g_1^*(\tilde{x}).
\end{aligned}$$

此外，我们有

$$\partial_t E_{\mathcal{P}_t}\{g_1^*(X)g_2^*(X)\} = g_1^*(\tilde{x})g_2^*(\tilde{x}) - \Psi(\mathcal{P}).$$

基于上述结果，我们得到

$$\begin{aligned}
\phi(\tilde{o}, \mathcal{P}) = &\frac{d\Psi(\mathcal{P}_t)}{dt}|_{t=0} \\
= &\frac{\mathbb{I}_{\tilde{t}_y}(T_y = 1)}{P(T_y = 1)}g_1'[E_{\mathcal{P}}(Y^*|\tilde{x}, T_y = 1)][\tilde{y}^* - E(Y^*|\tilde{x}, T_y = 1)]g_2^*(\tilde{x}) \\
&+ \frac{\mathbb{I}_{\tilde{t}_z}(T_z = 1)}{P(T_z = 1)}g_2'[E_{\mathcal{P}}(Z^*|\tilde{x}, T_z = 1)][\tilde{z}^* - E(Z^*|\tilde{x}, T_z = 1)]g_1^*(\tilde{x}) \\
&+ g_1^*(\tilde{x})g_2^*(\tilde{x}) - \Psi(\mathcal{P}).
\end{aligned}$$

注意，$\phi(O, \mathcal{P})$ 的方差有限，这意味着 I的有效影响函数为 $\phi(O, \mathcal{P})$。

给定I的有效影响函数，我们可以 ρ 通过采用定理3.1 证明中的类似论证得到结果，故此处省略。

定理4.1的证明：令

$$I_n = \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}'_{1i}[Y_i - \hat{m}_i][\hat{g}_{2i} - \bar{g}_2] + \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} \hat{g}'_{2i}[Z_i - \hat{h}_i][\hat{g}_{1i} - \bar{g}_1] + \frac{1}{n}\sum_{i\in\mathcal{D}_2}[\hat{g}_{2i} - \bar{g}_2][\hat{g}_{1i} - \bar{g}_1]$$

$$=: I_{n1} + I_{n2} + I_{n3}.$$

对于项 I_{n1}，我们有

$$\begin{aligned} I_{n1} &= \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}'_{1i}[Y_i - \hat{m}_i][\hat{g}_{2i} - \bar{g}_2] \\ &= \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}'_{1i}[\epsilon_i + m_i - \hat{m}_i][\hat{g}_{2i} - g_{2i} + g_{2i} - Eg_2 + Eg_2 - \bar{g}_2] \\ &= \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}'_{1i}\epsilon_i(\hat{g}_{2i} - g_{2i}) + \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}'_{1i}\epsilon_i(g_{2i} - Eg_2) + \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}'_{1i}\epsilon_i(Eg_2 - \bar{g}_2) \\ &\quad + \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}'_{1i}(m_i - \hat{m}_i)(\hat{g}_{2i} - \bar{g}_2) =: \sum_{i=1}^{4} D_i. \end{aligned}$$

在接下来的部分中，我们将证明D_1 和D_3 均可忽略，而D_2 和D_4 均包含主导项。

对于D_1, 这一项，可以将其改写为

$$D_1 = \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{g}'_{1i} - g'_{1i})\epsilon_i(\hat{g}_{2i} - g_{2i}) + \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}\epsilon_i(\hat{g}_{2i} - g_{2i}) =: D_{11} + D_{12}.$$

对于第一项，我们有$E(D_{11}) = 0$，并且

$$E(D_{11}^2|\mathcal{D}_1) = \frac{1}{n_y^2}\sum_{i\in\mathcal{D}_{2y}} E\left[\sigma^2(X)(\hat{g}'_{1i} - g'_{1i})^2(\hat{g}_{2i} - g_{2i})^2|\mathcal{D}_1\right].$$

在条件 $|g'_1(m_i) - g'_1(\hat{m}_i)| \le L|m_i - \hat{m}_i|$ 以及$g(\cdot) < C < \infty$下，我们有

$$E(D_{11}^2) \lesssim \frac{1}{n_y}E\left[\sigma^2(X)(\hat{m}_i - m_i)^2(\hat{h}_i - h_i)^2\right] = o\left(\frac{1}{n_y}\right),$$

其中，最后一个等式在条件$E\left[\sigma^2(X)(\hat{m}_i - m_i)^2(\hat{h}_i - h_i)^2\right] = o(1)$下成立。此外，类似地可以证明$E(D_{12}) = 0$，并且

$$E(D_{12}^2) \lesssim \frac{1}{n_y}E\left[\sigma^2(X)(\hat{h}_i - h_i)^2\right] = o\left(\frac{1}{n_y}\right),$$

当条件$E[(\hat{h}_i - h_i)^4] = o(n_y^{-1})$以及$g_2'(\cdot) < C < \infty$满足时。由此可得

$$D_1 = D_{11} + D_{12} = o_p\left(\frac{1}{\sqrt{n_y}}\right).$$

对于项D_{2y}，可将其改写为

$$D_2 = \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}}(\hat{g}_{1i}' - g_{1i}')\epsilon_i(g_{2i} - Eg_2) + \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g_{1i}'\epsilon_i(g_{2i} - Eg_2)$$

在条件 $|g_1'(m_i) - g_1'(\hat{m}_i)| \le L|m_i - \hat{m}_i|$ 以及 $E[\sigma^2(X)(\hat{m}_i - m_i)^2(g_2 - Eg_2)^2] = o(1)$ 下，同理可得

$$D_2 = \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g_{1i}'\epsilon_i(g_{2i} - Eg_2) + o_p\left(\frac{1}{\sqrt{n_y}}\right). \tag{A3}$$

现在我们转而考虑$Eg_2 - \bar{g}_2$。在条件 $g_2'(\cdot) \le C < \infty$ 且 $E[(h_i - \hat{h}_i)^4] = o_p(n_z^{-1})$ 下，我们有

$$\begin{aligned} Eg_2 - \bar{g}_2 &= Eg_2 - \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_2(h_i) + \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_2(h_i) - g_2(\hat{h}_i) \\ &= Eg_2 - \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_2(h_i) + \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_2'(\tilde{h}_i)(h_i - \hat{h}_i) = o_p(n_z^{-1/4}), \end{aligned}$$

其中 $\tilde{h}_i$ 介于h_i 和 $\hat{h}_i$之间。同理可得$Eg_1 - \bar{g}_1 = o_p(n_y^{-1/4})$。

对于项D_{3y}，可将其改写为

$$D_3 = (Eg_2 - \bar{g}_2)\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}_{1i}'\epsilon_i.$$

已知 $\mathcal{D}_{1y}$，容易证明

$$\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}_{1i}'\epsilon_i = O_p\left(\frac{1}{\sqrt{n_y}}\right).$$

同样根据$Eg_2 - \bar{g}_2 = o_p(n_z^{-1/4})$这一事实，我们很容易得到$D_3 = o_p(n_y^{-1/2})$。
到目前为止，我们有

$$I_{n1} = \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g_{1i}'\epsilon_i(g_{2i} - Eg_2) + D_4 + o_p\left(\frac{1}{\sqrt{n_y}}\right).$$

接下来，我们考虑D_4的界。注意到

$$\begin{aligned} D_4 &= \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} \hat{g}'_{1i}(m_i-\hat{m}_i)(\hat{g}_{2i}-\bar{g}_2) \\ &= \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{g}'_{1i}-g'_{1i})(m_i-\hat{m}_i)(\hat{g}_{2i}-\bar{g}_2) + \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}(m_i-\hat{m}_i)(\hat{g}_{2i}-\bar{g}_2) \\ &=: D_{41}+D_{42}. \end{aligned}$$

对于项 D_{41},，我们有

$$\begin{aligned} D_{41} = \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{g}'_{1i}-g'_{1i})(m_i-\hat{m}_i)(\hat{g}_{2i}-g_{2i}) + \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{g}'_{1i}-g'_{1i})(m_i-\hat{m}_i)(g_{2i}-Eg_2) \\ + (Eg_2-\bar{g}_2)\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{g}'_{1i}-g'_{1i})(m_i-\hat{m}_i). \end{aligned}$$

对于第一项,

$$\begin{aligned} &\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{g}'_{1i}-g'_{1i})(m_i-\hat{m}_i)(\hat{g}_{2i}-g_{2i}) \\ &\le \left[\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{g}'_{1i}-g'_{1i})^2(m_i-\hat{m}_i)^2\right]^{1/2}\left[\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{g}_{2i}-g_{2i})^2\right]^{1/2} \\ &\lesssim \left[\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (m_i-\hat{m}_i)^4\right]^{1/2}\left[\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} (\hat{h}_i-h_i)^2\right]^{1/2} = o_p\left(\frac{1}{\sqrt{n_y}}\right), \end{aligned}$$

其中，最后一个等式在 $E[(\hat{m}_i-m_i)^4] = o(n_y^{-1})$ 和 $E[(\hat{h}_i-h_i)^4] = (n_z^{-1})$ 时成立。类似地，我们可以将相同的论证应用于另外两项。因此可得

$$D_{41} = o_p\left(\frac{1}{\sqrt{n_y}}\right).$$

接下来，我们来看这一项 D_{42}。注意,

$$\begin{aligned} D_{42} = \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}(m_i-\hat{m}_i)(\hat{g}_{2i}-\bar{g}_2) = \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}(m_i-\hat{m}_i)(\hat{g}_{2i}-g_{2i}) \\ + \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}(m_i-\hat{m}_i)(g_{2i}-Eg_2) + (Eg_2-\bar{g}_2)\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}(m_i-\hat{m}_i) =: \sum_{i=1}^{3} D_{42i}. \end{aligned}$$

对于第一项 D_{421}，

$$\begin{aligned}
|D_{421}| &\leq \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} |g'_{1i}(m_i-\hat{m}_i)(\hat{g}_{2i}-g_{2i})| \\
&\leq \left[\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}}(g'_{1i})^2(m_i-\hat{m}_i)^2\right]^{1/2}\left[\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}}(\hat{g}_{2i}-g_{2i})^2\right]^{1/2} \\
&\lesssim \left[\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}}(m_i-\hat{m}_i)^2\right]^{1/2}\left[\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}}(\hat{h}_i-h_i)^2\right]^{1/2}.
\end{aligned}$$

注意，在条件 4.3 下，我们有 $E[(\hat{m}_i-m_i)^2]=o(n_y^{-1/2})$ 且 $E[(\hat{h}_i-h_i)^2]=o(n_z^{-1/2})$。由此可得

$$D_{421}=o_p\left(\frac{1}{n_y^{1/4}n_z^{1/4}}\right)=o_p\left(\frac{1}{\sqrt{n_y}}+\frac{1}{\sqrt{n_z}}\right).$$

类似地，根据 $Eg_2-\bar{g}_2=o_p(n_z^{-1/4})$ 和 $E[(\hat{m}_i-m_i)^2]=o(n_y^{-1/2})$，可以很容易地得到

$$D_{423}=o_p\left(\frac{1}{\sqrt{n_y}}+\frac{1}{\sqrt{n_z}}\right).$$

由此可得

$$D_4=\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}(m_i-\hat{m}_i)(g_{2i}-Eg_2)+o_p\left(\frac{1}{\sqrt{n_y}}+\frac{1}{\sqrt{n_z}}\right).$$

因此，我们可以得出结论

$$I_{n1}=\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}\epsilon_i(g_{2i}-Eg_2)+\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g'_{1i}(m_i-\hat{m}_i)(g_{2i}-Eg_2)+o_p\left(\frac{1}{\sqrt{n_y}}+\frac{1}{\sqrt{n_z}}\right).$$

同样，我们可以得到

$$I_{n2}=\frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g'_{2i}\eta(g_{1i}-Eg_1)+\frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g'_{2i}(h_i-\hat{h}_i)(g_{1i}-Eg_1)+o_p\left(\frac{1}{\sqrt{n_y}}+\frac{1}{\sqrt{n_z}}\right).$$

对于项 I_{n3}，我们有

$$
\begin{aligned}
I_{n3} &= \frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{2i}-g_{2i}+g_{2i}-Eg_2+Eg_2-\bar{g}_2)(\hat{g}_{1i}-g_{1i}+g_{1i}-Eg_1+Eg_1-\bar{g}_1)\\
&= \frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{2i}-Eg_2)(g_{1i}-Eg_1)+\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{2i}-g_{2i})(g_{1i}-Eg_1)+\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{1i}-g_{1i})(g_{2i}-Eg_2)\\
&\quad+\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{2i}-g_{2i})(\hat{g}_{1i}-g_{1i})+(Eg_1-\bar{g}_1)\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{2i}-g_{2i})+(Eg_1-\bar{g}_1)\frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{2i}-Eg_2)\\
&\quad+(Eg_1-\bar{g}_1)(Eg_2-\bar{g}_2)+(Eg_2-\bar{g}_2)\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{1i}-g_{1i})+(Eg_2-\bar{g}_2)\frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{1i}-Eg_1)=:\sum_{i=1}^{9}I_{n3i}.
\end{aligned}
$$

对于项 I_{n34}，

$$
\begin{aligned}
|I_{n34}| &\le \frac{1}{n}\sum_{i\in\mathcal{D}_2}|(\hat{g}_{2i}-g_{2i})(\hat{g}_{1i}-g_{1i})| \le \left[\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{2i}-g_{2i})^2\right]^{1/2}\left[\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{1i}-g_{1i})^2\right]^{1/2}\\
&\lesssim \left[\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{h}_i-h_i)^2\right]^{1/2}\left[\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{m}_i-m_i)^2\right]^{1/2}=o_p\left(\frac{1}{n_y^{1/4}n_z^{1/4}}\right)=o_p\left(\frac{1}{\sqrt{n_y}}+\frac{1}{\sqrt{n_z}}\right).
\end{aligned}
$$

同样，很容易证明

$$
\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{1i}-g_{1i})=o_p(n_y^{-1/4}),\ \text{and}\ \frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{2i}-g_{2i})=o_p(n_z^{-1/4}).
$$

此外，根据 $Eg_1-\bar{g}_1=o_p(n_y^{-1/4})$ 和 $Eg_2-\bar{g}_2=o_p(n_z^{-1/4})$ 这些事实，很容易证明

$$
\begin{aligned}
I_{n35}+I_{n37}+I_{n38} &= (Eg_1-\bar{g}_1)\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{2i}-g_{2i})+(Eg_1-\bar{g}_1)(Eg_2-\bar{g}_2)\\
&\quad+(Eg_2-\bar{g}_2)\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{1i}-g_{1i})=o_p\left(\frac{1}{\sqrt{n_y}}+\frac{1}{\sqrt{n_z}}\right).
\end{aligned}
$$

此外，请注意，

$$
\frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{2i}-Eg_2)=O_p(n^{-1/2}),\ \text{and}\ \frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{1i}-Eg_1)=O_p(n^{-1/2}),
$$

这意味着

$$I_{n36} + I_{n39} = (Eg_1 - \bar{g}_1)\frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{2i} - Eg_2) + (Eg_2 - \bar{g}_2)\frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{1i} - Eg_1) = o_p\left(\frac{1}{\sqrt{n_y}} + \frac{1}{\sqrt{n_z}}\right).$$

注意，根据泰勒级数展开

$$\begin{aligned} I_{n32} &= \frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{2i} - g_{2i})(g_{1i} - Eg_1) \\ &= \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_2'(h_i)(\hat{h}_i - h_i)(g_{1i} - Eg_1) + O_p\left(\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{h}_i - h_i)^2(g_{1i} - Eg_1)\right) \\ &= \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_{2i}'(\hat{h}_i - h_i)(g_{1i} - Eg_1) + o_p\left(\frac{1}{\sqrt{n_z}}\right). \end{aligned}$$

当满足以下条件时，最后一个等式成立：$E[(\hat{h}_i - h_i)^4] = o(n_z^{-1})$ 以及条件 4.2 。类似地，我们可以得到

$$\begin{aligned} I_{n33} &= \frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{g}_{1i} - g_{1i})(g_{2i} - Eg_2) \\ &= \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_{1i}'(\hat{m}_i - m_i)(g_{2i} - Eg_2) + O_p\left(\frac{1}{n}\sum_{i\in\mathcal{D}_2}(\hat{m}_i - m_i)^2(g_{2i} - Eg_2)\right) \\ &= \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_{1i}'(\hat{m}_i - m_i)(g_{2i} - Eg_2) + o_p\left(\frac{1}{\sqrt{n_y}}\right). \end{aligned}$$

根据上述结果，我们可以得出结论：

$$\begin{aligned} I_n = &\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g_{1i}'\epsilon_i(g_{2i} - Eg_2) + \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_{2i}'\eta(g_{1i} - Eg_1) + \frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{2i} - Eg_2)(g_{1i} - Eg_1) \\ &+ \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_{2i}'(h_i - \hat{h}_i)(g_{1i} - Eg_1) + \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_{2i}'(\hat{h}_i - h_i)(g_{1i} - Eg_1) \\ &+ \frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g_{1i}'(m_i - \hat{m}_i)(g_{2i} - Eg_2) + \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_{1i}'(\hat{m}_i - m_i)(g_{2i} - Eg_2) + o_p\left(\frac{1}{\sqrt{n_y}} + \frac{1}{\sqrt{n_z}}\right) \end{aligned}$$

记

$$R_1 = \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_{2i}'(h_i - \hat{h}_i)(g_{1i} - Eg_1) + \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_{2i}'(\hat{h}_i - h_i)(g_{1i} - Eg_1).$$

ˆ 进一步令$l_i = g_{2i}^{'}(h_i - h_i)(g_{1i} - Eg_1)$。注意

$$R_1 = \frac{1}{n}\sum_{i\in\mathcal{D}_2} l_i - \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} l_i = \frac{n-n_z}{n}\left(\frac{1}{n-n_z}\sum_{i\in\mathcal{D}_2/\mathcal{D}_{2z}} l_i - \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} l_i\right).$$

与定理2.3证明中的讨论类似，可以得到$E(R_1) = 0$，并且

$$\mathrm{Var}(R_1) \le \frac{2(n-n_z)^2}{n^2}\left(\frac{1}{n-n_z} + \frac{1}{n_z}\right)E(l_i^2).$$

注意到 $El_i^2 = o(1)$，由此可得

$$R_1 = \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_{2i}'(h_i - \hat{h}_i)(g_{1i} - Eg_1) + \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_{2i}'(\hat{h}_i - h_i)(g_{1i} - Eg_1) = o_p\left(\frac{1}{\sqrt{n_z}}\right).$$

类似地，我们可以证明

$$\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g_{1i}'(m_i - \hat{m}_i)(g_{2i} - Eg_2) + \frac{1}{n}\sum_{i\in\mathcal{D}_2} g_{1i}'(\hat{m}_i - m_i)(g_{2i} - Eg_2) = o_p\left(\frac{1}{\sqrt{n_y}}\right).$$

因此，我们有

$$\begin{aligned} I_n = &\frac{1}{n_y}\sum_{i\in\mathcal{D}_{2y}} g_{1i}'\epsilon_i(g_{2i} - Eg_2) + \frac{1}{n_z}\sum_{i\in\mathcal{D}_{2z}} g_{2i}'\eta(g_{1i} - Eg_1) \\ &+ \frac{1}{n}\sum_{i\in\mathcal{D}_2}(g_{2i} - Eg_2)(g_{1i} - Eg_1) + o_p\left(\frac{1}{\sqrt{n_y}} + \frac{1}{\sqrt{n_z}}\right). \end{aligned}$$

对于 I_n'，类似地，我们有：

$$\begin{aligned} I_n' = &\frac{1}{n_y}\sum_{i\in\mathcal{D}_{1y}} g_{1i}'\epsilon_i(g_{2i} - Eg_2) + \frac{1}{n_z}\sum_{i\in\mathcal{D}_{1z}} g_{2i}'\eta(g_{1i} - Eg_1) \\ &+ \frac{1}{n}\sum_{i\in\mathcal{D}_1}(g_{2i} - Eg_2)(g_{1i} - Eg_1) + o_p\left(\frac{1}{\sqrt{n_y}} + \frac{1}{\sqrt{n_z}}\right). \end{aligned}$$

基于上述结果，我们可以得出结论

$$\begin{aligned} I_n^* = &\frac{1}{N_y}\sum_{i\in\mathcal{D}_y} g_{1i}'\epsilon_i(g_{2i} - Eg_2) + \frac{1}{N_z}\sum_{i\in\mathcal{D}_z} g_{2i}'\eta(g_{1i} - Eg_1) + \frac{1}{N}\sum_{i\in\mathcal{D}}(g_{2i} - Eg_2)(g_{1i} - Eg_1) \\ &+ o_p\left(\frac{1}{\sqrt{N_y}} + \frac{1}{\sqrt{N_z}}\right). \end{aligned}$$

根据I_n^*的渐近结果，我们可以通过类似于定理3.1证明中 ρ_n^* 的论证得到相应结果，因此此处省略。

参考文献

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018年). 双重/去偏机器学习在治疗和结构参数估计中的应用. 《计量经济学杂志》, 21(1):C1–C68. Craddock, N. and Owen, M. J. (2005). 克雷佩林二分法终结之始. 《英国精神病学杂志》, 186(5):364–366. Guo, Z., Wang, W., Cai, T. T., and Li, H. (2019). 高维线性模型中遗传相关性的最优估计. 《美国统计学会杂志》, 114(525):358–369. Hines, O., Dukes, O., Diaz-Ordaz, K., and Vansteelandt, S. (2022). 揭开基于有效影响函数的统计学习之谜. 《美国统计学家》, pages 1–13. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M., and Wray, N. R. (2012年). 利用单核苷酸多态性衍生的基因组关系和限制性最大似然估计复杂疾病之间的多效性. 《生物信息学》, 28(19):2540–2542. Ma, R., Guo, Z., Cai, T. T., and Li, H. (2022). 基于高维逻辑回归的遗传相关性统计推断. *arXiv* 预印本 *arXiv:2202.10007*. Parker, C. C., Gopalakrishnan, S., Carbonetto, P., Gonzales, N. M., Leung, E., Park, Y. J., Aryee, E., Davis, J., Blizard, D. A., Ackert-Bicknell, C. L., 等. (2016). 非近交 CFW 小鼠行为、生理及基因表达性状的全基因组关联研究. 《自然·遗传学》, 48(8):919–926. Tsiatis, A. (2007). 半参数理论与缺失数据. Springer Science & Business Media. Van Rheenen, W., Peyrot, W. J., Schork, A. J., Lee, S. H., and Wray, N. R. (2019). 多基因疾病性状的遗传相关性：从理论到实践. 《自然综述：遗传学》, 20(10):567–581.

Vansteelandt, S. 和 Dukes, O.（2020）。广义线性模型参数的弱假设推断。《皇家统计学会杂志：*B*辑（统计方法学）》。Vattikuti, S.、Guo, J. 和 Chow, C. C.（2012）。常见单核苷酸多态性对代谢综合征性状遗传力和遗传相关性的解释。《公共科学图书馆·遗传学》，8(3):e1002637。Wang, J.、Li, S. 和 Li, H.（2021）。遗传协方差稳健推断的统一方法。*arXiv*预印本 *arXiv:2106.04106*。Weissbrod, O.、Flint, J. 和 Rosset, S.（2018）。在病例对照研究中直接估计基于单核苷酸多态性的遗传力和遗传相关性，以及使用汇总统计量进行估计。《美国人类遗传学杂志》，103(1):89–99。Yang, L.、Neale, B. M.、Liu, L.、Lee, S. H.、Wray, N. R.、Ji, N.、Li, H.、Qian, Q.、Wang, D.、Li, J. 等（2013）。注意缺陷多动障碍的多基因传递与复杂神经发育网络：针对常见和罕见变异的全基因组关联研究。《美国医学遗传学杂志*B*辑：神经精神遗传学》，162(5):419–430。Zhang, Y.、Cheng, Y.、Jiang, W.、Ye, Y.、Lu, Q. 和 Zhao, H.（2021）。利用全基因组关联研究汇总统计量估计复杂性状间遗传相关性的方法比较。《生物信息学简报》，22(5):bbaa442。Zhernakova, A.、Van Diemen, C. C. 和 Wijmenga, C.（2009）。从免疫相关疾病的共同遗传基础中发现共同的发病机制。《自然综述：遗传学》，10(1):43–55。