

多重检验中的错误发现率控制：理论和方法

统计与数据科学系 何兼良

1. 问题陈述：多重假设检验

基础设定: 考虑 m 个零假设 $H_{0,1}, \dots, H_{0,m}$ 和摘要统计量 X_1, \dots, X_m 其具有 **已知的零分布**，例如在零假设 $H_{0,i}$ 下 p -值 $P_i \sim \text{Unif}[0,1]$ 和 z -值 $Z_i \sim \mathcal{N}(0,1)$ 。多重检验涉及对 m 个假设同时推断：

数学记号: 令 \mathcal{H}_0 为零假设的集合， \mathcal{H}_1 为非零假设的集合，以及 \mathcal{R} 为拒绝集合。令 $\theta_i = \mathbb{I}(i \in \mathcal{H}_0)$ 为指示函数，给出第 i 个测试问题的真实状态。

I类错误率: $\text{FDR}(\mathcal{R}) = \mathbb{E}[\text{FDP}(\mathcal{R})]$, $\text{FDP}(\mathcal{R}) = \frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}| \vee 1}$

假设检验目标:

功效: $\text{TDR}(\mathcal{R}) = \mathbb{E}[\text{TDP}(\mathcal{R})]$, $\text{TDP}(\mathcal{R}) = \frac{|\mathcal{R} \cap \mathcal{H}_1|}{|\mathcal{H}_1| \vee 1}$.

$$\begin{aligned} & \max \text{TDR}(\mathcal{R}) \\ & \text{s.t. } \text{FDR}(\mathcal{R}) \leq \alpha, \quad \alpha \in (0,1). \end{aligned}$$

2. 多重假设检验面临的问题

图1: 稀疏性

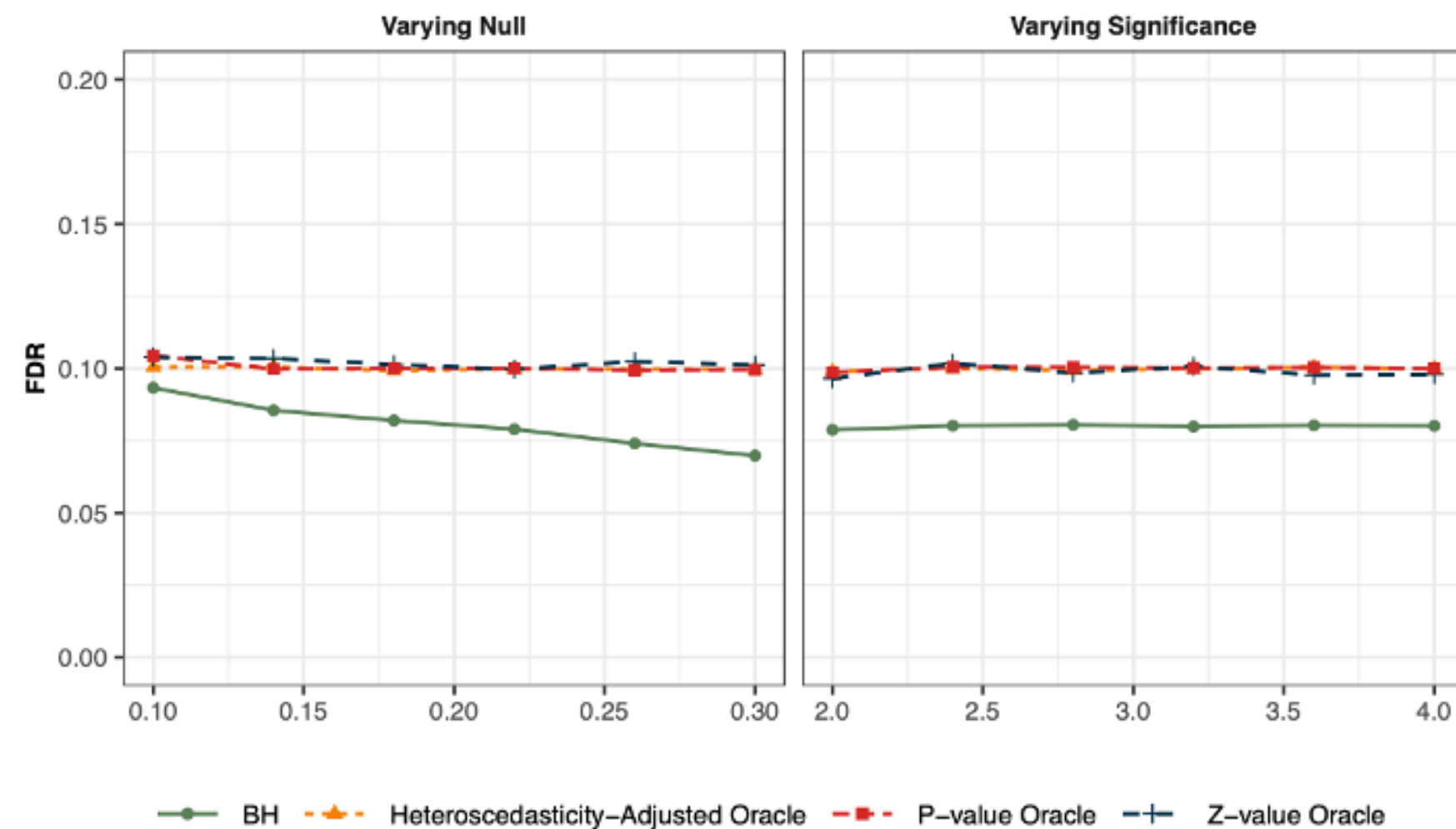
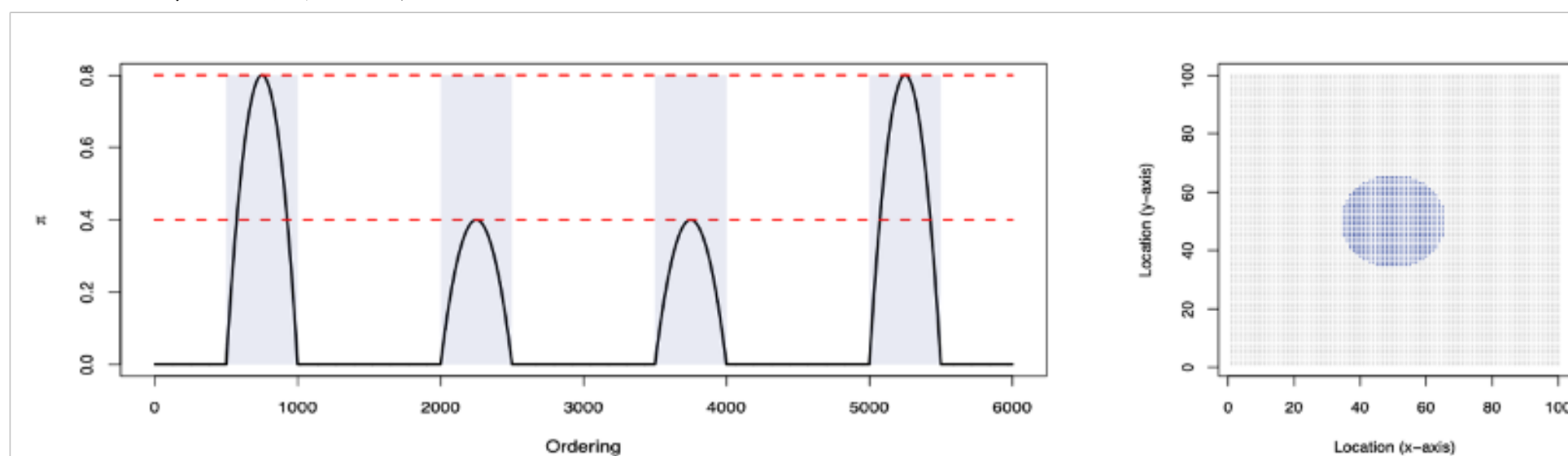


图2: 辅助信息



在现代应用统计学的许多领域，从遗传学和神经影像 (Pe'er et al., 2008) 到在线广告和金融 (Harvey & Liu, 2015)，在数据收集范式转变后，出现了几种现象: **稀疏性**、**辅助序列** 和 **非独立性**。这些结构可能具有信息，但也给测试方法的设计带来了挑战：

- (1) **摘要统计量选择**：使用p值是普遍的选择，但是否有 **更强大或更稳健的统计量** 可以更好地保留结构信息或适应复杂情况？
- (2) **弱假设下的检验**：传统方法，例如BH过程，**删除强假设** 后，确保统计性质保证的最小条件是什么？无额外假设下，我们能得到什么样的统计性质保证？
- (3) **结构特征**：如何通过从 **内部**（稀疏性和依赖性）或 **外部**（辅助序列）结构中提取信息来设计更强大的检验过程？
- (4) **对估计误差稳定的检验方法**：强大的检验过程依赖于一系列复杂的待估计的排序统计量。我们如何改进方法，甚至进一步适应 **估计的不准确性**？

2.1 研究问题：统一框架

Q: 目前有大量的研究从不同的切入点给出了不同场景下更好的假设检验的过程:

稀疏性: e.g., Benjamini et al., (2006); Jin & Cai, (2007); Sun and Cai (2007);

非独立性: e.g., Benjamini and Yekutieli, (2001); Finner et al., (2009); Ramdas et al., (2017a);

辅助信息: e.g., Scott et al., (2015); Ignatiadis et al., (2016); Zhang and Chen, (2022);

我们能否通过一个统一的框架去研究假设检验问题?

A: (排序-估计-阈值设定)

步骤1(排序): 为每个假设生成排序统计量 $T_i = c_i(X_i)$, 其中 $c_i: \mathbb{R} \mapsto \mathbb{R}$ 表示转换函数。

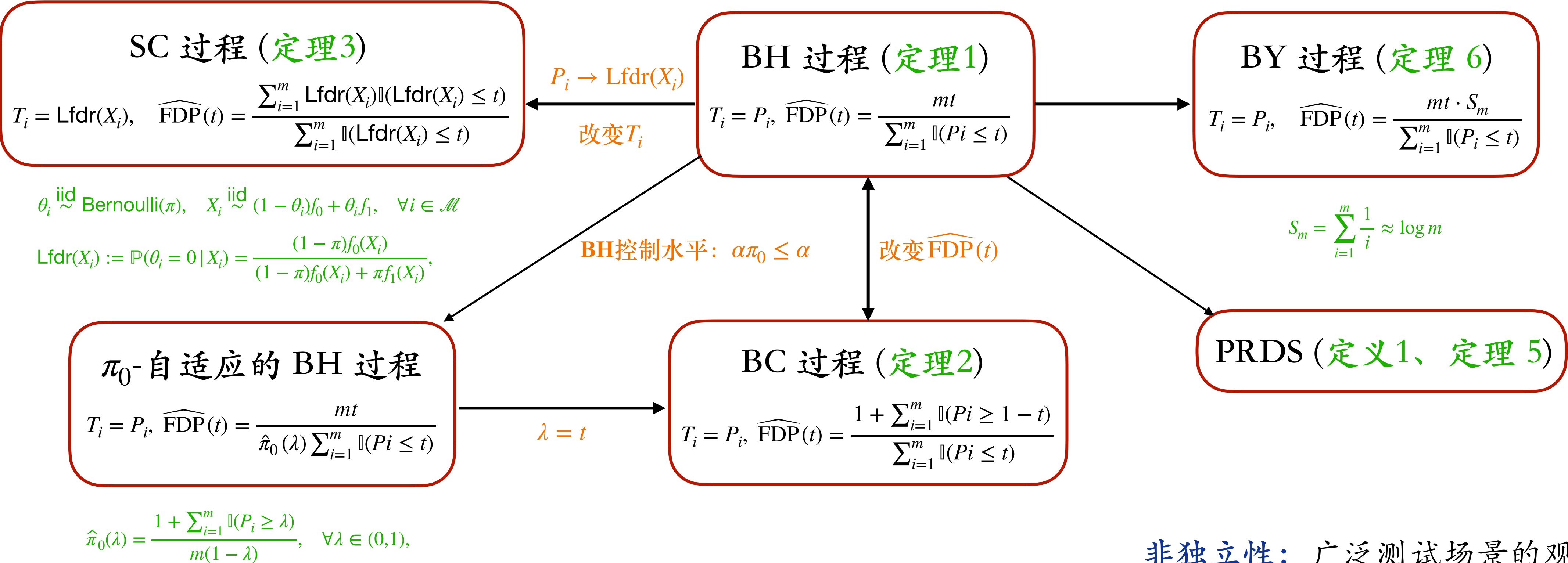
步骤2(估计): 利用预先确定的估计函数 $\widehat{\text{FDP}}: \mathbb{R} \times \mathbb{R}^m \mapsto \mathbb{R}$ 来估计 $(T_i)_{i \in [m]}$ 的 $\text{FDP}(\mathcal{R}_t)$, 记为 $\widehat{\text{FDP}}(t)$, 其中 $\mathcal{R}_t = \{T_i \leq t: i \in \mathcal{M}\}$ 。

步骤3(阈值设定): 对于给定指定的FDR控制水平 $\alpha \in (0,1)$, 选择在估计的FDP控制水平为 α 的排名统计量中的最大阈值:

$$t_\alpha = \sup \{t \in \mathcal{T} : \widehat{\text{FDP}}(t) \leq \alpha\},$$

其中 \mathcal{T} 是备选的阈值集。最终, 将输出拒绝集合。

2.2 研究问题：稀疏性和非独立性

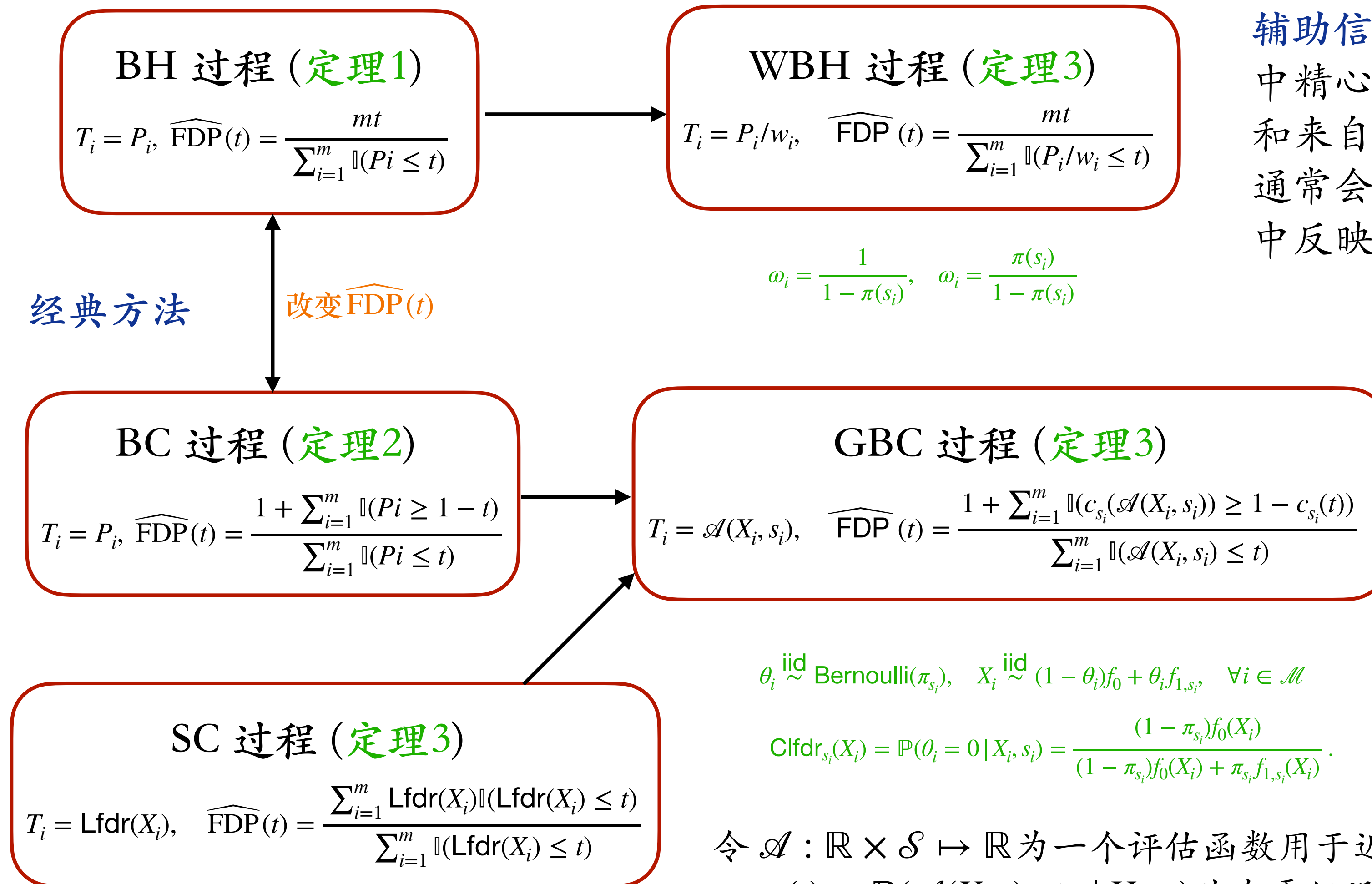


稀疏性问题： $\pi_0 = \frac{|\mathcal{H}_0|}{m}$ ，因而BH过程的控制水平 $\alpha\pi_0$ 是过度保守的

经典方法

非独立性： 广泛测试场景的观察数据往往表现出非独立性。然而，传统的FDR方法往往严重依赖于独立假设，常常忽视假设之间的相关性。

2.3 研究问题：辅助信息



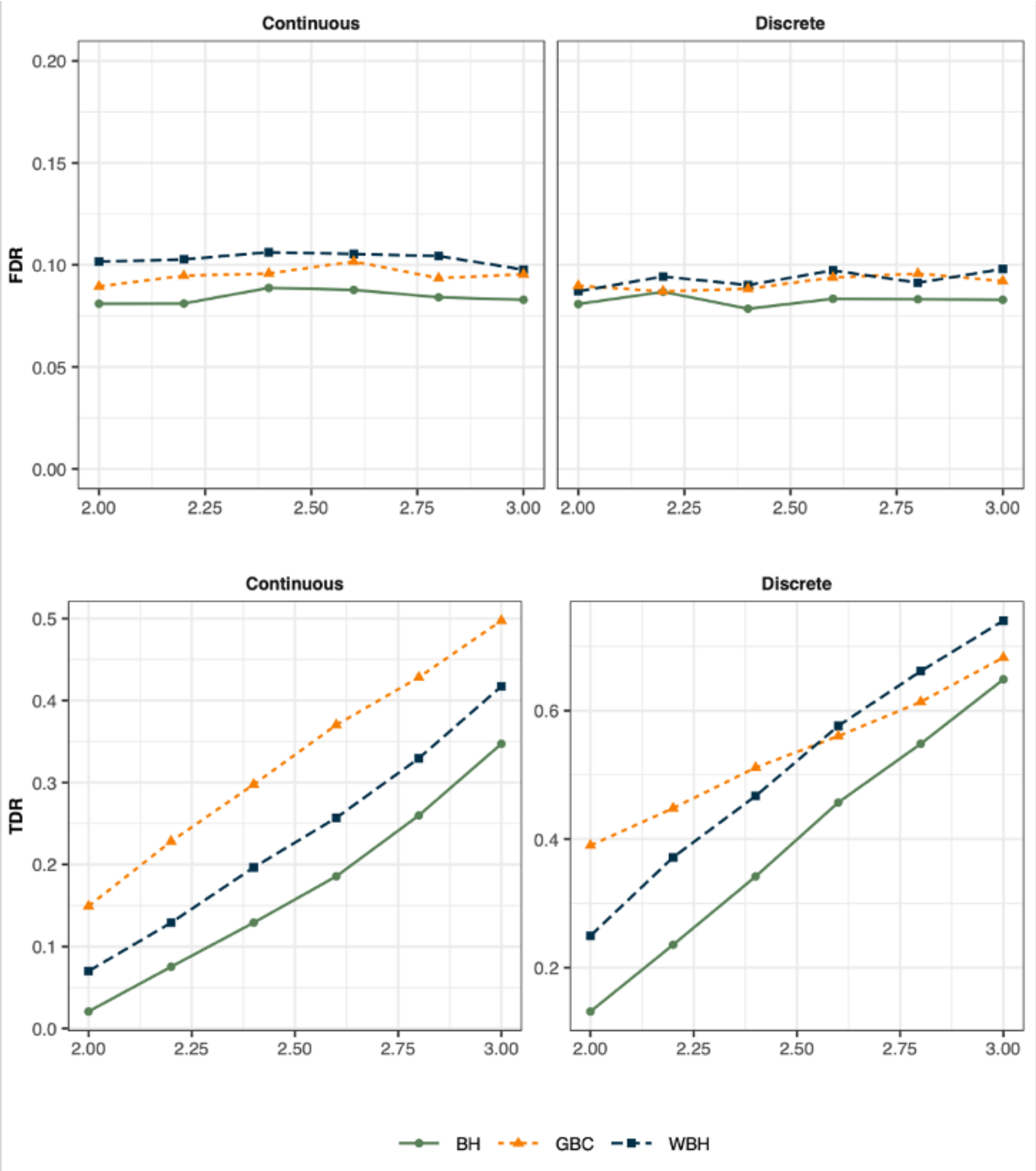
辅助信息：相关的领域知识，例如从同一数据集中精心构造的辅助序列 (Liu, 2014; Cai et al., 2019) 和来自二手数据源的外部协变量或先前数据，其通常会在许多研究与主数据集一起提供，应用中反映出这些数据具有丰富的可利用信息。

$$\theta_i \stackrel{\text{iid}}{\sim} \text{Bernoulli}(\pi_{s_i}), \quad X_i \stackrel{\text{iid}}{\sim} (1 - \theta_i)f_0 + \theta_i f_{1,s_i}, \quad \forall i \in \mathcal{M}$$

$$\text{Clfdr}_{s_i}(X_i) = \mathbb{P}(\theta_i = 0 | X_i, s_i) = \frac{(1 - \pi_{s_i})f_0(X_i)}{(1 - \pi_{s_i})f_0(X_i) + \pi_{s_i}f_{1,s_i}(X_i)}.$$

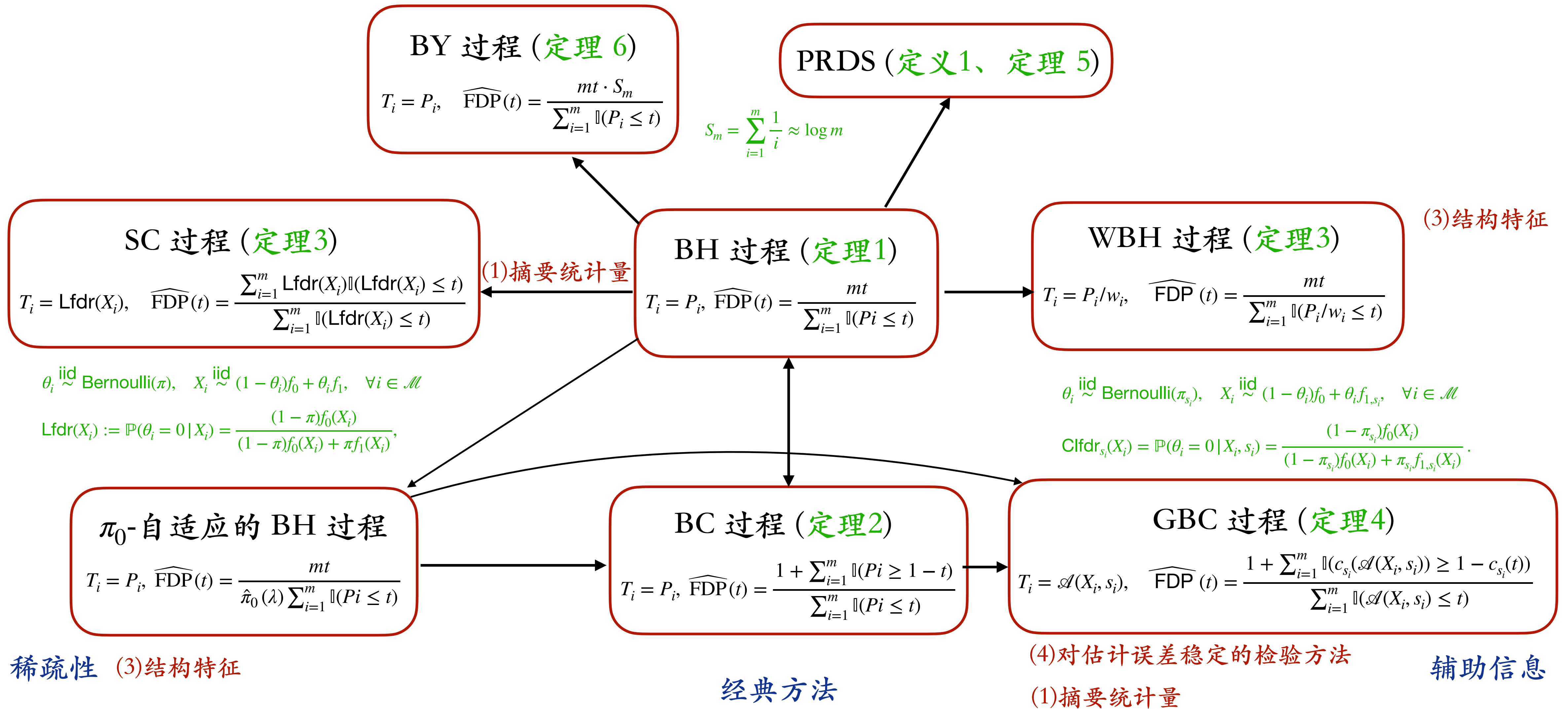
令 $\mathcal{A} : \mathbb{R} \times \mathcal{S} \mapsto \mathbb{R}$ 为一个评估函数用于近似 Clfdr，即 $\widehat{\text{Clfdr}}_s(X) = \mathcal{A}(X, s)$ ，而 $c_s(t) = \mathbb{P}(\mathcal{A}(X, s) \leq t | H_0, s)$ 为在零假设下的条件 CDF。

2.3 研究问题： 辅助信息



非独立性 (2)弱假设下的检验

统一框架: 排序(T_i)-估计($\widehat{\text{FDP}}(t)$)-阈值设定



参考文献

- [1] Pe'er, I., Yelensky, R., Altshuler, D., and Daly, M. J. (2008). Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 32(4):381–385.
- [2] Harvey, C. R. and Liu, Y. (2015). Backtesting. *The Journal of Portfolio Management*, 42(1):13–28
- [3] Benjamini, Y., Krieger, A., and Yekutieli, D. (2006). Adaptive linear step-up false discovery rate controlling procedures. *Biometrika*, 93(3):491–507
- [4] Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association*, 102(478):495–506.
- [5] Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *Journal of the American Statistical Association*, 102(479):901–912.
- [6] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188
- [7] Ramdas, A., Barber, R. F., Wainwright, M. J., and Jordan, M. I. (2017a). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*.
- [8] Zhang, X. and Chen, J. (2022). Covariate adaptive false discovery rate control with applications to omics-wide multiple testing. *Journal of the American Statistical Association*, 117(537):411–427