Experimental Designs for Heteroskedastic Variance

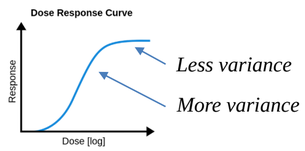
Justin Weltz, Tanner Fiez, Alexander Volfovsky, Eric Laber, Blake Mason, Houssam Nassif, and Lalit Jain
Duke University and Amazon

Introduction

In many problems, there is a set of items, $\mathcal{Z}$, with underlying structure, and the goal is to find which items are best using a set of noisy probes, $\mathcal{X}$. It is natural that some of these probes are noisier than others.

Drug Discovery: $\mathcal{Z} \subset \mathcal{X} \subset \mathbb{R}^d$

$\subset$



$\mathcal{Z}$ — FDA-approved drugs

$\mathcal{X}$ — Experimental drugs

How do we adaptively select probes to measure?

Problem Setup

Given: items $\mathcal{Z} \subset \mathbb{R}^d$, probes $\mathcal{X} \subset \mathbb{R}^d$

Measure: At each time t, observe $y_t = x_t^\top \theta^* + \eta_t$ where

$$\eta_t \sim \mathcal{N}(0, \sigma_t^2) \text{ and } \sigma_t^2 = x_t^\top \Sigma^* x_t,$$

and $\theta^* \in \mathbb{R}^d$ and $\Sigma^* \in \mathbb{R}^{dxd}$ are unknown.

Find: $z^* = \operatorname{argmax}_{z \in \mathcal{Z}} z^T \theta^*$ or $\mathcal{Z}_\alpha = \{ z \in \mathcal{Z} : z^\top \theta^* > \alpha \}$ with $1-\delta$ probability

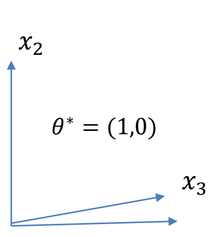
Problem Intuition

Consider a learner that selects a fixed design $\{x_t\}_{t=1}^T$, observes outcomes $\{y_t\}_{t=1}^T$, and constructs the weighted least squares estimator with known heteroskedastic variances, $\hat{\theta}$:

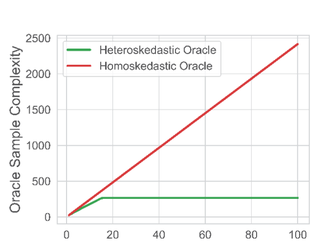
$$\hat{\theta} - \theta^* \sim \mathcal{N}\left\{ \mathbf{0}_d, \left(\sum_{t=1}^T \frac{x_t x_t^\top}{\sigma_t^2} \right)^{-1} \right\}.$$

Goal: Reduce variance of $\hat{\theta}$ in the directions most advantageous for identifying z^* or $\mathcal{Z}_\alpha$.

Benchmark Example: Ignoring heteroskedasticity suffers a multiplicative dependency on $\kappa = \max_x \sigma_x^2 / \min_x \sigma_x^2$.



Knowing the variances changes the design



Learning Heteroskedastic Variances

Goal: Estimate heteroskedastic variances with error bounds that scale favorably in the problem dimension.

Intuition: After Γ samples, we estimate Σ^* with $\widehat{\Sigma_\Gamma}$ using an M-estimation approach and decompose the error as

$$\left|\sigma_x^2 - \widehat{\sigma_x^2}\right| = \left|x^\top(\Sigma^* - \widehat{\Sigma_\Gamma})x\right| < \boldsymbol{A} + \boldsymbol{B} + \boldsymbol{C}.$$

Controlled by…

Algorithm 1: HEAD (Heteroskedasticity Estimation by Adaptive Designs)

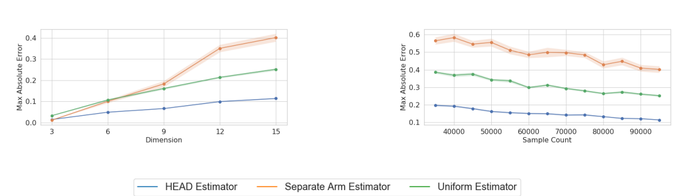
Result: Find $\widehat{\Sigma}_\Gamma$
Input: Arms $\mathcal{X} \in \mathbb{R}^d, \Gamma \in \mathbb{N}$
`//Stage 1: Take half the samples to estimate` θ^*
Determine λ^* according to $\lambda^* = \arg\min_{\lambda \in P_\mathcal{X}} \max_{x \in \mathcal{X}} x^\top \left(\sum_{x' \in \mathcal{X}} \lambda_{x'} x' x'^\top \right)^{-1} x$
Pull arm $x \in \mathcal{X}$ $\lceil \lambda_x^* \Gamma/2 \rceil$ times and collect observations $\{x_t, y_t\}_{t=1}^{\Gamma/2}$
Define $A^* = \sum_{t=1}^{\Gamma/2} x_t x_t^\top$ and $b^* = \sum_{t=1}^{\Gamma/2} x_t y_t$ and estimate $\widehat{\theta}_{\Gamma/2} = A^{*-1} b$
`//Stage 2: Take half the samples to estimate` Σ^* `given` $\widehat{\theta}_{\Gamma/2}$
Determine $\lambda^\dagger \leftarrow \lambda_x^\dagger = \arg\min_{\lambda \in P_\mathcal{X}} \max_{x \in \mathcal{X}} \phi_x^\top \left(\sum_{x' \in \mathcal{X}} \lambda_{x'} \phi_{x'} \phi_{x'}^\top \right)^{-1} \phi_x$, where ϕ_x is a de-duplicated version of $\mathrm{vec}(xx^\top)$
Pull arm $x \in \mathcal{X}$ $\lceil \lambda_x^\dagger \Gamma/2 \rceil$ times and collect observations $\{x_t, y_t\}_{t=\Gamma/2+1}^{\Gamma}$
Let $A^\dagger = \sum_{t=\Gamma/2+1}^{\Gamma} \phi_{x_t} \phi_{x_t}^\top$, $b^\dagger = \sum_{t=\Gamma/2+1}^{\Gamma} \phi_{x_t} \left(y_t - x_t^\top \widehat{\theta}_{\Gamma/2} \right)^2$
Output: $\mathrm{vech}(\widehat{\Sigma}_\Gamma) = A^{\dagger^{-1}} b^\dagger$.

Theorem 3.1. Assume $\Gamma = \Omega\left[\max\left\{\sigma_{max}^2 \log\left(\frac{|\mathcal{X}|}{\delta}\right) d^2, d^2\right\}\right]$. For any $x \in \mathcal{X}$ and $\delta \in (0,1)$, Alg. 1 (HEAD) guarantees the following,

$$\mathbb{P}\left(\left|\sigma_x^2 - \widehat{\sigma_x^2}\right| \le C_{\Gamma,\delta}\right) = 1 - \delta/2 \text{ and } C_{\Gamma,\delta} = \mathcal{O}\left[\left\{\frac{\log\left(\frac{|\mathcal{X}|}{\delta}\right)\sigma_{\max}^2 d^2}{\Gamma}\right\}^{-\frac{1}{2}}\right].$$

Scales with problem dimension, d^2

Empirical Results



HEAD
- optimally allocates samples → outperforms Uniform Est.
- efficiently estimates → outperforms Separate Arm Est.

Best-arm and Level Set Identification

Goal: Efficiently identify $z^* = \operatorname{argmax}_{z \in \mathcal{Z}} z^T \theta^*$, BAI, or $\mathcal{Z}_\alpha = \{ z \in \mathcal{Z} : z^\top \theta^* > \alpha \}$, LS, with $1-\delta$ probability.

Intuition: Use Alg. 1 to **estimate** the heteroskedastic variances. Leverage $\widehat{\sigma_x^2}$ to **minimize** the variance of the weighted least squares estimator in the directions that help identify the objective. **Eliminate** from the set of uncertain items until $\mathcal{Z}_\alpha$ or z^* is identified.

Estimate *Minimize* *Eliminate*

Algorithm 2: (H-RAGE) Heteroskedastic Randomized Adaptive Gap Elimination

Result: Find $z^* := \arg\max_{z \in \mathcal{Z}} z^\top \theta^*$ for BAI or $G_\alpha := \{z \in \mathcal{Z} : z^\top \theta^* > \alpha\}$ for LS
Input: $\mathcal{X} \in \mathbb{R}^d$, $\mathcal{Z} \in \mathbb{R}^d$, confidence $\delta \in (0,1)$, OBJ $\in$ {BAI,LS}, threshold $\alpha \in \mathbb{R}$
Initialize: $\ell \leftarrow 1$, $\mathcal{Z}_1 \leftarrow \mathcal{Z}$, $\mathcal{G}_1 \leftarrow \emptyset$, $\mathcal{B}_1 \leftarrow \emptyset$
`//Variance estimation`
Call Alg. 1 such that $\left|\widehat{\sigma}_x^2 = \min\left\{\max\{x^\top \widehat{\Sigma}_\Gamma x, \sigma_{\min}^2\}, \sigma_{\max}^2\right\} - \sigma_x^2\right| \le \sigma_x^2/2$
while *($|\mathcal{Z}_\ell| > 1$ and OBJ=BAI) or ($|\mathcal{Z}_\ell| > 0$ and OBJ=LS)* **do**
  `//Determine the design`
  Let $\widehat{\lambda}_\ell \in P_\mathcal{X}$ be a minimizer of $q\{\lambda, \mathcal{Y}(\mathcal{Z}_\ell)\}$ if OBJ=BAI and $q(\lambda, \mathcal{Z}_\ell)$ if OBJ=LS where
  $$q(\mathcal{V}) = \inf_{\lambda \in P_\mathcal{X}} q(\lambda; \mathcal{V}) = \inf_{\lambda \in P_\mathcal{X}} \max_{z \in \mathcal{V}} ||z||^2_{\left(\sum_{x \in \mathcal{X}} \widehat{\sigma}_x^{-2} \lambda_x x x^\top\right)^{-1}}$$
  Set $\epsilon_\ell = 2^{-\ell}$, $\tau_\ell = 3\epsilon_\ell^{-2} q(\mathcal{Z}_\ell) \log(8\ell^2|\mathcal{Z}|/\delta)$ `//Determine stepsize`
  Pull arm $x \in \mathcal{X}$ exactly $\lceil \tau_\ell \widehat{\lambda}_{\ell,x} \rceil$ times for n_ℓ samples and collect $\{x_{\ell,i}, y_{\ell,i}\}_{i=1}^{n_\ell}$
  Define $A_\ell := \sum_{i=1}^{n_\ell} \widehat{\sigma}_i^{-2} x_{\ell,i} x_{\ell,i}^\top$, $b_\ell = \sum_{i=1}^{n_\ell} \widehat{\sigma}_i^{-2} x_{\ell,i} y_{\ell,i}$ and construct $\widehat{\theta}_\ell = A_\ell^{-1} b_\ell$
  `//Eliminate arms`
  if *OBJ is BAI* **then**
    $\mathcal{Z}_{\ell+1} \leftarrow \mathcal{Z}_\ell \backslash \{z \in \mathcal{Z}_\ell : \max_{z' \in \mathcal{Z}_\ell} \langle z' - z, \widehat{\theta}_\ell \rangle > \epsilon_\ell\}$
  else
    $\mathcal{G}_{\ell+1} \leftarrow \mathcal{G}_\ell \cup \{z \in \mathcal{Z}_\ell : \langle z, \widehat{\theta}_\ell \rangle - \epsilon_\ell > \alpha\}$
    $\mathcal{B}_{\ell+1} \leftarrow \mathcal{B}_\ell \cup \{z \in \mathcal{Z}_\ell : \langle z, \widehat{\theta}_\ell \rangle + \epsilon_\ell < \alpha\}$
    $\mathcal{Z}_{\ell+1} \leftarrow \mathcal{Z}_\ell \backslash \{\{\mathcal{G}_\ell \setminus \mathcal{G}_{\ell+1}\} \cup \{\mathcal{B}_\ell \setminus \mathcal{B}_{\ell+1}\}\}$
  end
  $\ell \leftarrow \ell + 1$
end
Output: $\mathcal{Z}_\ell$ for BAI or $\mathcal{G}_\ell$ for LS

Matches lower bound up to log factors with an additive dependency on κ

Theorem 4.2. Consider objective, OBJ, of best-arm identification (BAI) or level-set identification (LS). The set returned from Alg. 2 (H-RAGE) achieves OBJ with probability $1-\delta$ at time $\tau = \mathcal{O}\left(\psi_{\mathrm{OBJ}}^* \log(\Delta^{-1}) \left[\log\left(\frac{|\mathcal{Z}|}{\delta}\right) + \log\{\log(\Delta^{-1})\}\right] + \log(\Delta^{-1})\, d^2 + \log\left(\frac{|\mathcal{X}|}{\delta}\right) \kappa^2 d^2\right)$, where ψ_{OBJ}^* is such that $\mathbb{E}(\tau) \ge 2\log(\frac{1}{2.4\delta})\, \psi_{\mathrm{OBJ}}^*$, and Δ is the minimum gap for the objective.

Multivariate Testing Simulation Example

We divide an advertisement into natural locations or features, each of which has different content options.