Tim, Gene, Kathryn

CRISPR genome editing, single-cell sequencing, and exponential family measurement error models

Abstract

CRISPR genome engineering and single-cell sequencing have transformed biological discovery. Single-cell CRISPR screens unite these two technologies, linking genetic perturbations in individual cells to changes in gene expression and illuminating regulatory networks underlying diseases. In this work we study single-cell CRISPR screens from a statistical perspective. First, we demonstrate on real data that a standard method for estimation and inference in single-cell CRISPR screens — "thresholded regression" — exhibits attenuation bias and a bias-variance tradeoff as a function of an intrinsic tuning parameter. We recover these phenomena in precise theoretical terms in an idealized Gaussian setting. Next, we introduce GLM-EIV ("generalized linear model with errors-in-variables"), a new method for single-cell CRISPR screen analysis. GLM-EIV generalizes the classical errors-in-variables model to response distributions and sources of measurement error that are exponential family-distributed, overcoming limitations of thresholded regression. We develop a computational infrastructure to deploy GLM-EIV across hundreds or thousands of processors on clouds (e.g., Microsoft Azure) and high-performance clusters. Leveraging this infrastructure, we apply GLM-EIV to analyze two recent, large-scale, single-cell CRISPR screen datasets, yielding new biological insights.

Contents

1 Introduction

CRISPR is a genome engineering tool that has enabled scientists to precisely edit human and nonhuman genomes, opening the door to new medical therapies [1, 2] and transforming basic biology research [3]. Recently, scientists have paired CRISPR genome engineering with single-cell sequencing [4, 5]. The resulting assays, known as a "single-cell CRISPR screens," link genetic perturbations in individual cells to changes in gene expression, illuminating regulatory networks underlying human diseases and other traits [6].

Despite their promise, single-cell CRISPR screens present substantial statistical challenges. A major difficulty is that CRISPR perturbations are unobservable and assigned stochastically to cells. As a consequence, one cannot know with certainty which cells were perturbed. Instead, one must leverage an indirect, noisy proxy of perturbation presence or absence – namely, transcribed barcode counts – to "guess" which cells were perturbed. Using these imputed perturbation assignments, one can attempt to estimate the effect of the perturbation on gene expression. The standard approach, which we call "thresholded regression" or the "thresholding method," is to assign perturbation identities to cells by simply thresholding the barcode counts.

We study estimation and inference in single-cell CRISPR screens from a statistical perspective, formulating the data generating mechanism using a new class of errors-in-variables (or measurement error) models. We assume

that the response variable y is a GLM of an underlying predictor variable x^*. We do not observe x^* directly; rather, we observe a noisy version x of x^* that itself is a GLM of x^*. The goal of the analysis is to estimate the effect of x^* on y using the observed data (x, y) only. In the context of the biological application, x^*, y, and x are CRISPR perturbations, gene expressions, and barcode counts, respectively.

Our work makes two main contributions. First, we conduct an in-depth study of the thresholding method from empirical and theoretical perspectives. Notably, we demonstrate on real data that the thresholding method exhibits attenuation bias and a bias-variance tradeoff as a function of the selected threshold, and we recover these phenomena in precise mathematical terms in an idealized Gaussian model. Second, we introduce a new method for estimation and inference in single-cell CRISPR screens that accounts for the measurement error inherent in the experiment. The method, called *GLM-EIV* (generalized linear model with errors-in-variables), implicitly estimates the probability that each cell was perturbed, obviating the need to explicitly impute perturbation assignments via thresholding or another heuristic. Theoretical analyses and simulation studies indicate that GLM-EIV outperforms the thresholding method in large regions of the parameter space.

We implement several statistical accelerations (that likely are of independent utility) to bring the cost of GLM-EIV down to within an order of magnitude of the thresholding method. Finally, we develop a computational infrastructure to deploy GLM-EIV at-scale across hundreds or thousands of processors on clouds (e.g., Microsoft Azure) and high-performance clusters. Leveraging this infrastructure, we apply GLM-EIV to analyze two recent, large-scale, high multiplicity-of-infection single-cell CRISPR screen datasets, yielding new biological and statistical insights.

2 Background and analysis challenges

Our focus in this work is on high multiplicity-of-infection (MOI), enhancer-targeting, single-cell CRISPR screens. In this section we cover relevant biological background and motivation.

P1: The human genome consists of genes (segments of DNA that code for proteins), enhancers (segments of DNA that regulate the expression of one or more genes), and other genomic regions. Genome-wide association studies

have revealed that the majority ($> 95\%$) of variants associated with diseases lie outside genes and (very likely) inside enhancers. These noncoding variants contribute to disease by modulating the expression one or more genes, which in turn encode proteins that affect the phenotype. A central open challenge in genetics, therefore, is to link enhancers that harbor disease-associated variants to the genes that they target at genome-wide scale.

P2: High MOI single-cell CRISPR screens are the most promising biotechnology for solving this problem. Describe the experimental protocol here. Explain that we use the terms "barcodes" and "gRNAs" interchangeably, as polyadenylated gRNAs serve as barcodes in CROP-seq. Link to Figure 1.

P3: Single-cell CRISPR screens pose several core analysis challenges. Describe the analysis challenges here: (i) unobserved perturbation; (ii) existence of background reads; (iii) highly discrete count data; (iv) nuisance variables.

3 Related work

Motivated by the challenges of single-cell data, several authors recently have extended statistical models that (implicitly or explicitly) assume Gaussianity and homoscedasticity to a broader class of exponential family distributions. For example, Lin, Lei, and Roeder [7] developed eSVD, an extension of SVD to exponential family and curved Gaussian responses. Unlike SVD, eSVD models the relationship between the mean and variance of a gene's expression level, a phenomenon induced by the countedness of single-cell data [8]. Similarly, Townes et al. [9] proposed GLM-PCA, an extension of PCA that directly models Poisson- or negative binomially-distributed gene expression counts. We see our work as a continuation of this broad effort to "port" common statistical methods and models to single-cell count data. Our focus, however, is on regression rather than dimension reduction: we extend the classical errors-in-variables model to response distributions and sources of measurement error that are exponential family-distributed.

The closest parallels to our work in the statistical methodology literature are Grün & Leisch [10] and Ibrahim [11]. Grün & Leisch considered estimation and inference in a k-component mixture of GLMs. While we prefer to view the GLM-EIV model as an errors-in-variables model, the GLM-EIV model is equivalent to a two-component mixture of *products* of GLM

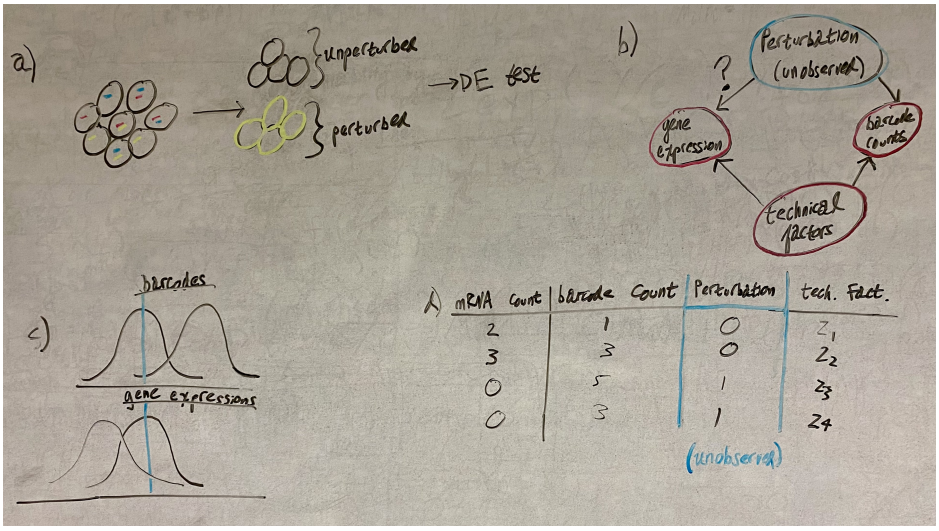

Figure 1: **Overview of experimental design and analysis challenges**: **a,** Experimental design. For a given perturbation (e.g., the perturbation represented in yellow), we partition the cells into two groups: those that received the perturbation, and those that did not receive the perturbation. For a given gene, we conduct a differential expression analysis across the two groups of cells, yielding an estimate of the impact of the given perturbation on the given gene. **b,** DAG representing the variables in the analysis. The perturbation (unobserved) affects both gene expression and gRNA expression; technical factors (e.g., batch, sequencing depth, etc.) act as nuisance variables. The target of inference is the effect of the perturbation on gene expression (denoted with question mark). **c,** Schematic illustrating "background reads." The gRNA modality has a nonzero, "background read" distribution even in the absence of a perturbation, complicating the assignment of perturbations to cells. **d**, Example data for a given perturbation-gene pair. Notice that (i) the perturbations are unobserved, and (ii) the gene and gRNA expression data take the form of discrete counts.

densities. Ibrahim derived a method for fitting GLMs in the presence of missing-at-random covariates. Our method, by contrast, involves fitting two conditionally independent GLMs in the presence of a totally latent covariate. Thus, while Ibrahim is a helpful reference, our estimation and inference tasks are more complex.

The genomics literature has produced several applied methods for linking

perturbations to changes in gene expression in single-cell CRISPR screens: SCEPTRE [12], MIMOSCA [4], and scMAGeCK [13]. These methods in general are focused on hypothesis testing rather than estimation; none, for instance, produces a confidence interval for the effect size of a perturbation on gene expression change. Additionally, two of these methods (MIMOSCA and scMAGeCK) use (possibly penalized) linear models to model gene expressions, thereby disregarding the countedness and sparsity of the data.

4 Thresholding method

In this section we study thresholded regression from empirical and theoretical perspectives, highlighting several fundamental challenges of the method. Let $n \in \mathbb{N}$ be the number of cells assayed in the experiment (typically, $n \approx 100,000 - 250,000$). Consider a single perturbation and a single gene. For cell $i \in \{1, \ldots, n\}$, let $m_i \in \mathbb{N}$ be the number of gene transcripts sequenced in the cell; let $g_i \in \mathbb{N}$ be the number of gRNA transcripts sequenced in the cell; let $l_i^m \in \mathbb{N}$ be the number of gene transcripts sequenced across *all* genes (the library size) in the cell; and finally, let $z_i \in \mathbb{R}^{d-1}$ be the cell-specific vector of technical factors (e.g., sequencing batch, percent mitochondrial reads, etc.). The thresholding method is defined as follows:

1. For a given threshold $c \in \mathbb{N}$, let the imputed perturbation assignment $\hat{p}_i \in \{0, 1\}$ be
$$\begin{cases} \hat{p}_i = 0 \text{ if } g_i \geq c, \\ \hat{p}_i = 1 \text{ if } g_i < c. \end{cases}$$

2. Assume that m_i is related to $\hat{p}_i, l_i$, and z_i through the following GLM:
$$\begin{cases} m_i | (\hat{p}_i, z_i, l_i) \sim \text{NB}_\theta(\mu_i), \\ \log(\mu_i) = \beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i + \log\left(l_i^m\right), \end{cases}$$
where (i) $NB_\theta(\mu_i)$ is a negative binomial distribution with mean μ_i and known size parameter θ; (ii) $\beta_0^m \in \mathbb{R}, \beta_1^m \in \mathbb{R}$, and $\gamma \in \mathbb{R}^{d-1}$ are unknown parameters; and (iii) $\log(l_i^m)$ is an offset term. Fit a GLM to obtain estimates of the parameters.

3. Compute a p-value and confidence interval for the target of inference β_1^m.

The biological interpretation for the target of inference β_1^m as follows: β_1^m is the log-transformed fold change in gene expression in response to the perturbation, removing the effect of the technical factors. Fold change (obtained by exponentiating β_1^m) is the ratio of the mean expression in cells that were perturbed to the mean expression in cells that were not perturbed. $\exp(\beta_1^m) = 1$ indicates no change in mean expression, while $\exp(\beta_1^m) > 1$ and $\exp(\beta_1^m) < 1$ indicate an increase and decrease in expression in response to the perturbation, respectively.

4.1 Empirical analysis

We studied the behavior of the thresholding method on real data, uncovering an attenuation bias effect and a bias-variance tradeoff as a function of the selected threshold. We applied the thresholding method to analyze the set of 381 positive control perturbation-gene pairs in the Gasperini dataset. The positive control pairs consisted of perturbations that targeted gene transcription start sites (TSSs) for inhibition. Repressing the TSS of a given gene decreases its expression; therefore, the positive control pairs *a priori* are expected to exhibit a strong, negative log fold change in expression.

To investigate the sensitivity of the thresholding method to the selected threshold, we deployed the thresholding method using three different choices for the threshold: 1, 5, and 20. We found that the chosen threshold substantially impacted the results (Figure 2a-b). Estimates for log fold change produced by setting the threshold to 1 were smaller in magnitude than those produced by setting the threshold to 5. (Equivalently, estimates for *raw* fold change were closer to the baseline of 1 for threshold = 1; Figure 2a.) Fold change estimates generated by setting the threshold to 5 and setting the threshold to 20 were more concordant; however, the threshold of 20 yielded slightly stronger effect sizes (Figure 2b). These results are consistent with an *attenuation bias* phenomenon: small thresholds tend to underestimate the true parameter value, and this effect diminishes gradually as the threshold increases.

We hypothesized that selecting a large threshold comes with the cost of increasing the variance of the estimator.

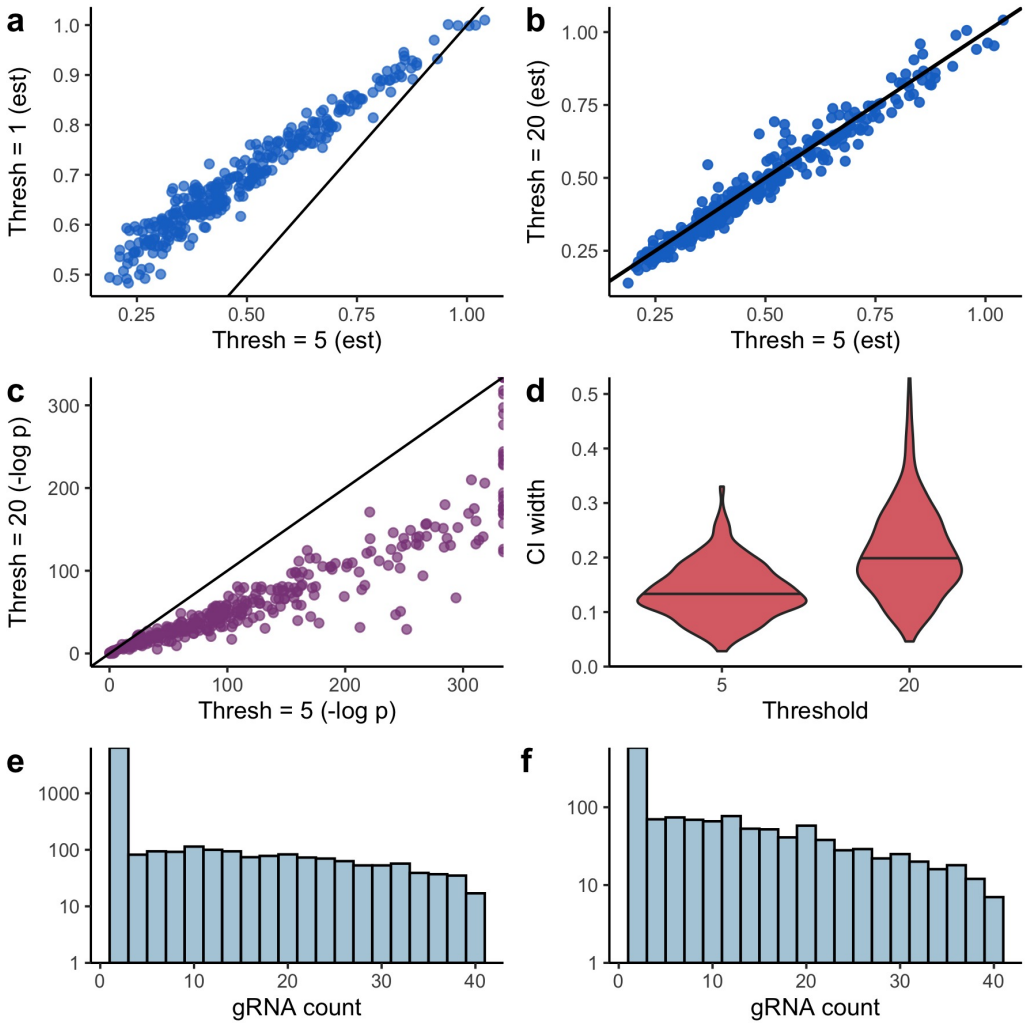

Figure 2: **Empirical challenges of thresholded regression.**

4.2 Theoretical analysis

We study the thresholding method from a theoretical perspective in an idealized Gaussian setting. Suppose that we observe data $(g_1, m_1), \ldots, (g_n, m_n)$ from the following model:

$$\begin{cases} m_i = \beta_0^m + \beta_1^m p_i + \epsilon_i \\ g_i = \beta_0^g + \beta_1^g p_i + \tau_i \\ p_i \sim \text{Bern}(\pi) \\ \epsilon_i, \tau_i \sim N(0,1) \\ p_i \perp\!\!\!\perp \tau_i \perp\!\!\!\perp \epsilon_i. \end{cases} \tag{1}$$

For a given threshold $c \in \mathbb{R}$, the imputed perturbation assignment $\hat{p}_i$ is given by $\hat{p}_i = \mathbb{I}(g_i \geq c)$. The thresholding estimator $\hat{\beta}_1^m$ for β_1^m is

$$\hat{\beta}_1^m = \frac{\sum_{i=1}^n (\hat{p}_i - \overline{\hat{p}})(m_i - \overline{m})}{\sum_{i=1}^n (\hat{p}_i - \overline{\hat{p}})^2}.$$

Proposition 1 *The almost sure limit (as $n \to \infty$) of $\hat{\beta}_1^m$ is*

$$\hat{\beta}_1^m \xrightarrow{a.s.} \beta_1^m \left(\frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} \right), \tag{2}$$

where

$$\begin{cases} \mathbb{E}[\hat{p}_i] = \zeta(1-\pi) + \omega\pi, \\ \omega = \Phi\left(\beta_1^g + \beta_0^g - c\right), \\ \zeta = \Phi\left(\beta_0^g - c\right). \end{cases}$$

Let $\gamma : \mathbb{R}^4 \to \mathbb{R}$ be defined by

$$\gamma(\beta_1^g, \pi, c, \beta_0^g) = \frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])}.$$

We call γ the "attenuation function." Observe that

i. γ does not depend on β_1^m or β_0^m, and

ii. $\hat{\beta}_1^m \xrightarrow{a.s.} [\gamma(\beta_0^g, \beta_1^g, c, \pi)]\beta_1^m$.

Let $b : \mathbb{R}^4 \to \mathbb{R}$ be the asymptotic *relative* bias of $\hat{\beta}_1^m$:

$$\begin{aligned} b(\beta_1^g, \pi, c, \beta_0^g) &= \left(\frac{1}{\beta_1^m}\right) \lim_{n\to\infty} \left(\beta_1^m - \mathbb{E}[\hat{\beta}_1^m]\right) = \left(\frac{1}{\beta_1^m}\right) \left(\beta_1^m - \mathbb{E}\left(\lim_{a.s.} \hat{\beta}_1^m\right)\right) \\ &= \frac{1}{\beta_1^m} \left(\beta_1^m - \gamma(\beta_1^g, \pi, c, \beta_0^g)\beta_1^m\right) = 1 - \gamma(\beta_1^g, \pi, c, \beta_0^g), \end{aligned}$$

where $\lim_{a.s.}$ denotes a.s. convergence. The asymptotic relative bias vanishes when the attenuation function equals 1. (HOW TO MAKE THE ABOVE RIGOROUS?)

Bias as a function of threshold (Panel a)

To investigate the basic question of "What is a good threshold selection strategy?", we study the relationship between the asymptotic relative bias b of $\hat{\beta}_1^m$ and the selected threshold c. For simplicity, we set the perturbation probability π to $1/2$. Let $c_{\text{bayes}} \in \mathbb{R}$ be the Bayes-optimal decision boundary for classifying cells as perturbed or unperturbed, i.e.

$$c_{\text{bayes}} = \arg\min_{c \in \mathbb{R}} \mathbb{P}(\hat{p}_i \neq p_i).$$

Simple algebra shows that $c_{\text{bayes}} = \beta_0^g + (1/2)\beta_1^g$. Below, we give several results for the asymptotic relative bias b of $\hat{\beta}_1^m$. We refer throughout to Figure 3a, which displays plots of asymptotic relative bias versus threshold for different values of β_1^g. We sometimes refer to "asymptotic relative bias" using the shortened term "bias" for succinctness.

- **Proposition 2** *Fix $\pi = 1/2$. For all $(\beta_1^g, c, \beta_0^g) \in \mathbb{R}^3$, the asymptotic relative bias is positive, i.e.*

  $$b(\beta_1^g, 1/2, c, \beta_0^g) > 0.$$

  The thresholding method incurs strict attenuation bias (i.e., it *under*estimates the true effect size) for all choices of the threshold and over all possible values of the model parameters (Figure 3a). Attenuation bias is a common attribute of estimators that ignore measurement in errors-in-variables models [15].

- **Proposition 3** *Fix $\pi = 1/2$. The asymptotic relative bias b decreases monotonically in β_1^g, i.e.*

  $$\frac{\partial b}{\partial(\beta_1^g)}(\beta_1^g, 1/2, c, \beta_0^g) \leq 0.$$

  This result formalizes the intuition that the problem becomes easier as the gRNA mixture distribution becomes increasingly well-separated. To visualize Proposition (3), one can fix a threshold (e.g., $c = 0$) and scan for bias across the panels.

- **Proposition 4** *For $\pi = 1/2$ and given $(\beta_1^g, \beta_0^g) \in \mathbb{R}^2$, the Bayes-optimal decision boundary c_{bayes} is a critical value of the bias function b, i.e.*

$$\frac{\partial b}{\partial c}\left(\beta_1^g, 1/2, c_{bayes}, \beta_0^g\right) = 0.$$

  The Bayes-optimal decision boundary is an optimum (or possibly a saddle point) of the asymptotic relative bias function (Figure 3a, vertical blue lines). Interestingly, c_{bayes} is in some cases a maximizer of the bias (Figure 3a, left) and in other cases a minimizer of the bias (Figure 3a, right).

- **Proposition 5** *Assume without loss of generality that $\beta_1^g > 0$, and fix $\pi = 1/2$. As the threshold c tends to infinity, the asymptotic relative bias b tends to $1/2$, i.e.*

$$\lim_{c \to \infty} b(\beta_1^g, 1/2, c, \beta_0^g) = 1/2.$$

  In other words, we always can set the threshold to a large number and attain a relative bias of $1/2$ (Figure 3a, all panels). This result establishes an upper bound on the bias of thresholded regression (under optimal threshold selection strategy).

- The following proposition compares the two threshold selection strategies introduced above (i.e., large number versus Bayes-optimal decision boundary) head-to-head.

  Proposition 6 *Assume without loss of generality that $\beta_1^g > 0$. For $\beta_1^g \in [0, 2\Phi^{-1}(3/4))$, we have that*

$$b(\beta_1^g, 1/2, c_{bayes}, \beta_0^g) > b(\beta_1^g, 1/2, \infty, \beta_0^g).$$

  For $\beta_1^g = 2\Phi^{-1}(3/4)$, we have that

$$b(\beta_1^g, 1/2, c_{bayes}, \beta_0^g) = b(\beta_1^g, 1/2, \infty, \beta_0^g).$$

  Finally, for $\beta_1^g \in (2\Phi^{-1}(3/4), \infty)$, we have that

$$b(\beta_1^g, 1/2, c_{bayes}, \beta_0^g) < b(\beta_1^g, 1/2, \infty, \beta_0^g).$$

Setting the threshold to a large number yields a smaller bias when β_1^g is small (i.e., $\beta_1^g < 2\Phi^{-1}(3/4) \approx 1.35$; Figure 3a, left); setting the threshold to the Bayes-optimal decision boundary yields a smaller bias when β_1^g is large (i.e., $\beta_1^g > 2\Phi^{-1}(3/4)$; Figure 3a, right); and the two approaches coincide when β_1^g is intermediate (i.e., $\beta_1^g = 2\Phi^{-1}(3/4)$; Figure 3a, middle).

These results are subtle, but we can summarize them as follows. First, selecting a threshold that minimizes the bias is challenging, as there is no rule of thumb that we can apply universally (e.g., "always choose the Bayes-optimal decision boundary" or "always choose a large number") due to the complexity of the bias function. Second, even if we *have* selected a good threshold, we incur nonzero attenuation bias.

Generalizing to $\pi \in [0, 1/2]$ (Panel b)

We generalize the expression for bias when the threshold is large to arbitrary $\pi \in [0, 1/2]$:

Proposition 7 *Assume without loss of generality that $\beta_1^g > 0$. As the threshold c tends to infinity, the asymptotic relative bias b tends to π, i.e.*

$$\lim_{c \to \infty} b(\beta_1^g, \pi, c, \beta_0^g) = \pi.$$

In other words, if the perturbation probability is π, and if we set the threshold to a large number, then the asymptotic relative bias is π (Figure 3b). We can understand this result intuitively by considering an extreme example: when π is very small (e.g., $\pi = 0.01$), almost all cells are unperturbed. Therefore, in selecting a large threshold, we correctly classify nearly all unperturbed cells as unperturbed; on the other hand, the *perturbed* cells that we misclassify as *unperturbed* are swamped in number by the truly unperturbed cells, resulting in a small bias.

Bias-variance tradeoff (Panel c)

Finally, to shed light on limitations of the large threshold selection strategy, we derive an exact bias-variance decomposition for the thresholding

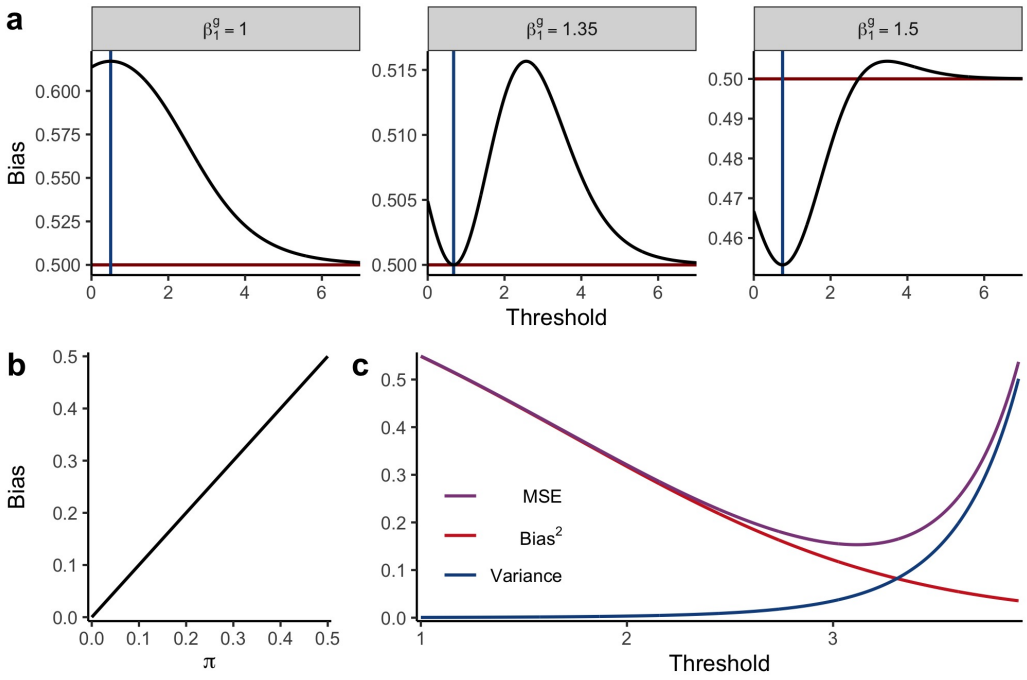

Figure 3: **Theoretical challenges of thresholded regression. a,** Asymptotic relative bias versus threshold for different values of β_1^g. The bias function is highly nonconvex and strictly nonzero. Vertical blue lines, Bayes-optimal decision boundaries. Across all panels, $\beta_0^g = 0$ and $\pi = 1/2$. **b,** Asymptotic relative bias versus π when the threshold is set to a large number. The two quantities coincide exactly. **c,** Bias-variance decomposition for thresholding method in no-intercept model. Bias decreases and variance increases as the threshold tends to infinity. $\beta_1^g = 1, \beta_1^m = 1$, and $\pi = 0.1$.

estimator. We consider a slightly simpler, no-intercept version of (1) for this purpose:

$$\begin{cases} m_i = \beta_m p_i + \epsilon_i \\ g_i = \beta_g p_i + \tau_i \\ p_i \sim \text{Bern}(\pi) \\ \epsilon_i, \tau_i \sim N(0,1) \\ p_i \perp\!\!\!\perp \tau_i \perp\!\!\!\perp \epsilon_i. \end{cases} \tag{3}$$

The thresholding estimator $\hat{\beta}_m$ in the no-intercept case is

$$\hat{\beta}_m = \frac{\sum_{i=1}^n \hat{p}_i m_i}{\sum_{i=1}^n \hat{p}_i^2}. \tag{4}$$

Proposition 8 *The limiting distribution of* $\hat{\beta}_m$ *is*

$$\sqrt{n}(\hat{\beta}_m - l) \xrightarrow{d} N\left(0, \frac{\beta_m \omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1 + l^2)}{\left(\mathbb{E}[\hat{p}_i]\right)^2}\right),$$

where

$$\begin{cases} l = \beta_m \omega\pi / [\zeta(1-\pi) + \omega\pi], \\ \mathbb{E}[\hat{p}_i] = \pi\omega + (1-\pi)\zeta, \\ \omega = \Phi(\beta_g - c), \\ \zeta = \Phi(-c). \end{cases}$$

This result yields an exact bias-variance decomposition for $\hat{\beta}_m$ for large n (Figure 3c). As the threshold tends to infinity, the bias decreases and the variance increases, consistent with the intuition that a large threshold reduces the misclassification rate at the cost of decreasing the "effective sample size." The best strategy for maximizing estimation accuracy (as quantified by mean squared error) is to select a threshold that induces moderate bias. A downside of this approach, however, is that constructing valid confidence intervals becomes more challenging.

5 GLM-EIV

6 Simulation studies

7 Real data analysis

8 Discussion

9 Appendix

9.1 Theoretical details for thresholding estimator

This section contains proofs of the propositions presented Section 4.2, "Theoretical analysis of thresholding estimator." The subsections are organized as follows. Section (9.1.1) introduces some notation. Section (9.1.2) establishes almost sure convergence of the thresholding estimator in the model (1), proving Proposition 1. Section (9.1.3) simplifies the expression for the attenuation function γ, and section (9.1.4) computes derivatives of γ to be used throughout the proofs. Section (9.1.5) establishes the limit in c of γ, proving Proposition 7 and as a corollary Proposition 5. Section (9.1.6) establishes that the Bayes-optimal decision boundary is a critical value of γ, proving Proposition 4, and section (9.1.7) compares the competing threshold selection strategies head-to-head, proving Proposition 6. Section (9.1.8) demonstrates that γ is monotone in β_1^g, proving Proposition 3, and Section (9.1.9) establishes attenuation bias of the thresholding estimator, proving Proposition 2. Finally, Section (9.1.10) derives the bias-variance decomposition of the thresholding estimator in the model (3), proving Proposition 8.

9.1.1 Notation

All notation introduced in this subsection (i.e., 9.1.1) pertains to the Gaussian model with intercepts (1). Recall that the attenuation function $\gamma : \mathbb{R}^4 \to \mathbb{R}$ is defined by

$$\gamma(\beta_1^g, c, \pi, \beta_0^g) = \frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])},$$

where

$$\begin{cases} \mathbb{E}[\hat{p}_i] = \zeta(1-\pi) + \omega\pi, \\ \omega = \Phi\left(\beta_1^g + \beta_0^g - c\right), \\ \zeta = \Phi\left(\beta_0^g - c\right). \end{cases}$$

Additionally, recall that the asymptotic relative bias function $b : \mathbb{R}^4 \to \mathbb{R}$ is

$$b(\beta_1^g, c, \pi, \beta_0^g) = 1 - \gamma(\beta_1^g, c, \pi, \beta_0^g).$$

Next, we define the functions g and $h : \mathbb{R}^4 \to \mathbb{R}$ by

$$g(\beta_1^g, c, \pi, \beta_0^g) = (1-\pi)\left(\Phi(\beta_0^g + \beta_1^g - c)\right) - (1-\pi)\left(\Phi(\beta_0^g - c)\right) \tag{5}$$

and

$$\begin{aligned} h(\beta_1^g, c, \pi, \beta_0^g) = &\left[(1-\pi)\left(\Phi(\beta_0^g - c)\right) + \pi\left(\Phi(\beta_0^g + \beta_1^g - c)\right)\right] \cdot \\ &\left[(1-\pi)\left(\Phi(c - \beta_0^g)\right) + \pi\left(\Phi(c - \beta_0^g - \beta_1^g)\right)\right]. \end{aligned} \tag{6}$$

We use $f : \mathbb{R} \to \mathbb{R}$ to denote the $N(0,1)$ density, and we denote the right-tail probability probability of f by $\bar{\Phi}$, i.e.,

$$\bar{\Phi}(x) = \int_x^\infty f = \Phi(-x).$$

The parameter β_0^g is a given, fixed constant throughout the proofs. Therefore, to minimize notation, we typically use $\gamma(\beta_1^g, c, \pi)$ (resp., $b(\beta_1^g, c, \pi)$, $g(\beta_1^g, c, \pi)$, $h(\beta_1^g, c, \pi)$) to refer to the function γ (resp., b, g, h) evaluated at $(\beta_1^g, c, \pi, \beta_0^g)$. Finally, for a given function $r : \mathbb{R}^p \to \mathbb{R}$, point $x \in \mathbb{R}^p$, and index $i \in \{1, \ldots, p\}$, we use the symbol $D_i r(x)$ to refer to the derivative of the ith component of r evaluated at x (*sensu* [16]). For example, $D_1\gamma(\beta_1^g, c, 1/2)$ is the derivative of the first component of γ (the component corresponding to β_1^g) evaluated at $(\beta_1^g, c, 1/2)$. Likewise, $D_2 g(\beta_1^g, c, \pi)$ is the derivative of the second component of g (the component corresponding to c) evaluated at (β_1^g, c, π).

9.1.2 Almost sure limit of $\hat{\beta}_1^m$

We derive the limit in probability of $\hat{\beta}_1^m$ for the Gaussian model with intercepts (1). Dividing by n in (2), we can express $\hat{\beta}_1^m$ as

$$\hat{\beta}_1^m = \frac{\frac{1}{n}\sum_{i=1}^n (\hat{p}_i - \overline{\hat{p}_i})(m_i - \overline{m})}{\frac{1}{n}\sum_{i=1}^n (\hat{p}_i - \overline{\hat{p}})}.$$

By weak LLN,

$$\hat{\beta}_1^m \xrightarrow{P} \frac{\text{Cov}(\hat{p}_i, m_i)}{\mathbb{V}(\hat{p}_i)}.$$

To compute this quantity, we first compute several simpler quantities:

1. Expectation of m_i: $\mathbb{E}[m_i] = \beta_0^m + \beta_1^m \pi$.

2. Expectation of $\hat{p}_i$:

$$\begin{aligned}
&\mathbb{E}[\hat{p}_i] = \mathbb{P}[\hat{p}_i = 1] = \mathbb{P}[\beta_0^g + \beta_1^g p_i + \tau_i \geq c] = \\
&\text{(By LOTP) } \mathbb{P}[\beta_0^g + \tau_i \geq c]\,\mathbb{P}[p_i = 0] + \mathbb{P}[\beta_0^g + \beta_1^g + \tau_i \geq c]\,\mathbb{P}[p_i = 1] \\
&\quad = \mathbb{P}[\tau_i \geq c - \beta_0^g](1-\pi) + \mathbb{P}[\tau_i \geq c - \beta_1^g - \beta_0^g](\pi) \\
&\quad = \left(\bar{\Phi}(c - \beta_0^g)\right)(1-\pi) + \left(\bar{\Phi}(c - \beta_1^g - \beta_0^g)\right)(\pi) = \\
&\quad\quad \Phi(\beta_0^g - c)(1-\pi) + \Phi(\beta_1^g + \beta_0^g - c)\pi = \zeta(1-\pi) + \omega\pi.
\end{aligned}$$

3. Expectation of $\hat{p}_i p_i$:

$$\mathbb{E}[\hat{p}_i p_i] = \mathbb{E}[\hat{p}_i | p_i = 1]\,\mathbb{P}[p_i = 1] = \mathbb{P}[\beta_0^g + \beta_1^g + \tau_i \geq c]\,\pi = \omega\pi.$$

4. Expectation of $\hat{p}_i m_i$:

$$\begin{aligned}
\mathbb{E}[\hat{p}_i m_i] &= \mathbb{E}[\hat{p}_i(\beta_0^m + \beta_1^m p_i + \epsilon_i)] = \beta_0^m \mathbb{E}[\hat{p}_i] + \beta_1^m \mathbb{E}[\hat{p}_i p_i] + \mathbb{E}[\hat{p}_i \epsilon_i] \\
&= \beta_0^m \mathbb{E}[\hat{p}_i] + \beta_1^m \omega\pi + \mathbb{E}[\hat{p}_i]\mathbb{E}[\epsilon_i] = \beta_0^m \mathbb{E}[\hat{p}_i] + \beta_1^m \omega\pi.
\end{aligned}$$

5. Variance of $\hat{p}_i$: Because $\hat{p}_i$ is binary, we have that $\mathbb{V}[\hat{p}_i] = \mathbb{E}[\hat{p}_i]\,(1 - \mathbb{E}[\hat{p}_i])$.

6. Covariance of $\hat{p}_i, m_i$:

$$\begin{aligned}
\text{Cov}(\hat{p}_i, m_i) &= \mathbb{E}[\hat{p}_i m_i] - \mathbb{E}[\hat{p}_i]\mathbb{E}[m_i] = \beta_0^m \mathbb{E}[\hat{p}_i] + \beta_1^m \omega\pi - \mathbb{E}[\hat{p}_i](\beta_0^m + \beta_1^m \pi) \\
&= \beta_1^m \omega\pi - \mathbb{E}[\hat{p}_i]\beta_1^m \pi = \beta_1^m \pi\,(\omega - \mathbb{E}[\hat{p}_i]).
\end{aligned}$$

Combining these expressions, we have that

$$\hat{\beta}_1^m \xrightarrow{P} \frac{\beta_1^m \pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} = \beta_1^m \gamma(\beta_1^g, c, \pi).$$

9.1.3 Re-expressing γ in a simpler form

We rewrite the attenuation fraction γ in a way that makes it more amenable to theoretical analysis. We leverage the fact that f integrates to unity and is even. We have that

$$\begin{aligned}\mathbb{E}\left[\hat{p}_i\right] &= (1-\pi)\bar{\Phi}(c-\beta_0^g) + \pi\bar{\Phi}(c-\beta_0^g-\beta_1^g) \\ &= (1-\pi)\Phi(\beta_0^g-c) + \pi\Phi(\beta_0^g+\beta_1^g-c), \quad (7)\end{aligned}$$

and so

$$\begin{aligned}1-\mathbb{E}\left[\hat{p}_i\right] = (1-\pi)+\pi-\mathbb{E}[\hat{p}_i] &= (1-\pi)\left(1-\bar{\Phi}(c-\beta_0^g)\right)+\pi\left(1-\bar{\Phi}(c-\beta_0^g-\beta_1^g)\right) \\ &= (1-\pi)\Phi(c-\beta_0^g) + \pi\Phi(c-\beta_0^g-\beta_1^g). \quad (8)\end{aligned}$$

Next,

$$\omega = \Phi(\beta_1^g+\beta_0^g-c), \quad (9)$$

and so

$$\begin{aligned}\omega - \mathbb{E}[\hat{p}_i] &= \Phi(\beta_1^g+\beta_0^g-c) - (1-\pi)\Phi(\beta_0^g-c) - \pi\Phi(\beta_0^g+\beta_1^g-c) \\ &\quad (1-\pi)\Phi(\beta_1^g+\beta_0^g-c) - (1-\pi)\Phi(\beta_0^g-c). \quad (10)\end{aligned}$$

Combining (7, 8, 9, 10), we find that

$$\begin{aligned}\gamma(\beta_1^g, c, \pi) &= \frac{\pi(\omega-\mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1-\mathbb{E}[\hat{p}_i])} \\ &= \frac{\pi\left[(1-\pi)\Phi(\beta_0^g+\beta_1^g-c) - (1-\pi)\Phi(\beta_0^g-c)\right]}{\left[(1-\pi)\Phi(\beta_0^g-c)+\pi\Phi(\beta_0^g+\beta_1^g-c)\right]\left[(1-\pi)\Phi(c-\beta_0^g)+\pi\Phi(c-\beta_0^g-\beta_1^g)\right]}. \quad (11)\end{aligned}$$

As a corollary, when $\pi = 1/2$,

$$\begin{aligned}&\gamma(\beta_1^g, c, 1/2) \\ &\qquad = \frac{\Phi(\beta_0^g+\beta_1^g-c) - \Phi(\beta_0^g-c)}{\left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right]\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right]}. \quad (12)\end{aligned}$$

Recalling the definitions of g (5) and h (6), we can write γ as

$$\gamma(\beta_1^g, c, \pi) = \frac{\pi g(\beta_1^g, c, \pi)}{h(\beta_1^g, c, \pi)}.$$

The special case (12) is identical to

$$\gamma(\beta_1^g, c, 1/2) = \frac{(4)(1/2)g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)} = \frac{2g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)}, \tag{13}$$

i.e., the numerator and denominator of (13) coincide with those of (12). We sometimes will use the notation $2 \cdot g$ and $4 \cdot h$ to refer to the numerator and denominator of (12), respectively.

9.1.4 Derivatives of g and h in c

We compute the derivatives of g and h in c, which we will need to prove subsequent results. First, by FTC and the evenness of f, we have that

$$\begin{aligned} D_2 g(\beta_1^g, c, \pi) &= -(1-\pi) f(\beta_0^g + \beta_1^g - c) + (1-\pi) f(\beta_0^g - c) \\ &= (1-\pi) f(c - \beta_0^g) - (1-\pi) f(c - \beta_0^g - \beta_1^g). \end{aligned} \tag{14}$$

Second, we have that

$$\begin{aligned} & D_2 h(\beta_1^g, c, \pi) \\ &= -[(1-\pi) f(\beta_0^g - c) + \pi f(\beta_0^g + \beta_1^g - c)] \left[(1-\pi)\Phi(c - \beta_0^g) + \pi \Phi(c - \beta_0^g - \beta_1^g)\right] \\ &+ \left[(1-\pi) f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)\right] \left[(1-\pi)\Phi(\beta_0^g - c) + \pi \Phi(\beta_0^g + \beta_1^g - c)\right] \\ &= \left[(1-\pi) f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)\right] \cdot \\ &\quad \Big[(1-\pi)\Phi(\beta_0^g - c) + \pi \Phi(\beta_0^g + \beta_1^g - c) \\ &\qquad - (1-\pi)\Phi(c - \beta_0^g) - \pi \Phi(c - \beta_0^g - \beta_1^g)\Big]. \end{aligned} \tag{15}$$

9.1.5 Limit of γ in c

Assume (without loss of generality) that $\beta_1^g > 0$. We compute $\lim_{c \to \infty} \gamma(\beta_1^g, c, \pi)$. Observe that

$$\lim_{c \to \infty} g(\beta_1^g, c, \pi) = \lim_{c \to \infty} h(\beta_1^g, c, \pi) = 0.$$

Therefore, we can apply L'Hôpital's rule. We have by (14) and (15) that

$$\begin{aligned}\lim_{c\to\infty}\gamma(\beta_1^g,c,\pi) &= \lim_{c\to\infty}\frac{\pi D_2 g(\beta_1^g,c,\pi)}{D_2 h(\beta_1^g,c,\pi)}\\ &= \lim_{c\to\infty}\left\{\frac{(1-\pi)f(c-\beta_0^g)+\pi f(c-\beta_0^g-\beta_1^g)}{\pi(1-\pi)f(c-\beta_0^g)-\pi(1-\pi)f(c-\beta_0^g-\beta_1^g)}\right.\\ &\qquad\left.\cdot\Big[(1-\pi)\Phi(\beta_0^g-c)+\pi\Phi(\beta_0^g+\beta_1^g-c)\right.\\ &\qquad\left.-(1-\pi)\Phi(c-\beta_0^g)-\pi\Phi(c-\beta_0^g-\beta_1^g)\Big]\right\}^{-1}. \quad (16)\end{aligned}$$

We evaluate the two terms in the product (16) separately. Dividing by $f(c-\beta_0^g-\beta_1^g)>0$, we see that

$$\frac{(1-\pi)f(c-\beta_0^g)+\pi f(c-\beta_0^g-\beta_1^g)}{\pi(1-\pi)f(c-\beta_0^g)-\pi(1-\pi)f(c-\beta_0^g-\beta_1^g)} = \frac{\frac{(1-\pi)f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)}+\pi}{\frac{\pi(1-\pi)f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)}-\pi(1-\pi)}. \quad (17)$$

To evaluate the limit of (17), we first evaluate the limit of

$$\begin{aligned}\frac{f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)} &= \frac{\exp\left[-(1/2)(c-\beta_0^g)^2\right]}{\exp\left[-(1/2)(c-\beta_0^g-\beta_1^g)^2\right]}\\ &= \frac{\exp[-(1/2)(c^2-2c\beta_0^g+(\beta_0^g)^2)]}{\exp\left[-(1/2)(c^2-2c\beta_0^g-2c\beta_1^g+(\beta_0^g)^2+2(\beta_0^g\beta_1^g)+(\beta_1^g)^2)\right]}\\ &= \exp\big[-c^2/2+c\beta_0^g-(\beta_0^g)^2/2\\ &\quad +c^2/2-c\beta_0^g-c\beta_1^g+(\beta_0^g)^2/2+\beta_0^g\beta_1^g+(\beta_1^g)^2/2\big]\\ &= \exp[-c\beta_1^g+\beta_0^g\beta_1^g+(\beta_1^g)^2/2] = \exp[\beta_0^g\beta_1^g+(\beta_1^g)^2/2]\exp[-c\beta_1^g]. \quad (18)\end{aligned}$$

Taking the limit in (18), we obtain

$$\lim_{c\to\infty}\frac{f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)} = \exp[\beta_0^g\beta_1^g+(\beta_1^g)^2/2]\lim_{c\to\infty}\exp[-c\beta_1^g] = 0$$

for $\beta_1^g>0$. We now can evaluate the limit of (17):

$$\lim_{c\to\infty}\frac{(1-\pi)f(c-\beta_0^g)+\pi f(c-\beta_0^g-\beta_1^g)}{\pi(1-\pi)f(c-\beta_0^g)-\pi(1-\pi)f(c-\beta_0^g-\beta_1^g)} = \frac{-\pi}{\pi(1-\pi)} = -\frac{1}{1-\pi}.$$

Next, we compute the limit of the other term in the product (16):

$$\lim_{c\to\infty}\Big[(1-\pi)\Phi(\beta_0^g-c)+\pi\Phi(\beta_0^g+\beta_1^g-c) \\ -(1-\pi)\Phi(c-\beta_0^g)-\pi\Phi(c-\beta_0^g-\beta_1^g)\Big]=-(1-\pi)-\pi=-1. \quad (19)$$

Combining (17) and (19), the limit (16) evaluates to

$$\lim_{c\to\infty}\gamma(\beta_1^g,c,\pi)=\left(\frac{1}{1-\pi}\right)^{-1}=1-\pi.$$

It follows that the limit in c of the asymptotic relative bias b is

$$\lim_{c\to\infty}b(\beta_1^g,c,\pi)=1-\lim_{c\to\infty}\gamma(\beta_1^g,c,\pi)=\pi.$$

A corollary is that

$$\lim_{c\to\infty}b(\beta_1^g,c,1/2)=1/2.$$

9.1.6 Bayes-optimal decision boundary as a critical value of γ

Let $c_{\text{bayes}}=\beta_0^g+(1/2)\beta_1^g$. We show that $c=c_{\text{bayes}}$ is a critical value of γ for $\pi=1/2$ and given β_1^g, i.e,

$$D_2\gamma(\beta_1^g,c_{\text{bayes}},1/2)=0.$$

Differentiating (13), the quotient rule implies that

$$D_2\gamma(\beta_1^g,c,1/2)=\frac{D_2[2g(\beta_1^g,c,1/2)]4h(\beta_1^g,c,1/2)-2g(\beta_1^g,c,1/2)D_2[4h(\beta_1^g,c,1/2)]}{[4h(\beta_1^g,c,\pi)]^2}. \quad (20)$$

We have by (14) that

$$D_2[2g(\beta_1^g,c_{\text{bayes}},1/2)]=f(\beta_1^g/2)-f(-\beta_1^g/2)=f(\beta_1^g/2)-f(\beta_1^g/2)=0. \quad (21)$$

Similarly, we have by (15) that

$$D_2[4h(\beta_1^g,c_{\text{bayes}},\pi)]=[f(\beta_1^g/2)+f(-\beta_1^g/2)]\cdot \\ [\Phi(-\beta_1^g/2)+\Phi(\beta_1^g/2)-\Phi(\beta_1^g/2)-\Phi(-\beta_1^g/2)]=0. \quad (22)$$

Plugging in (22) and (21) to (20), we find that

$$D_2[\gamma(\beta_1^g, c_{\text{bayes}}, 1/2)] = 0.$$

Finally, because

$$b(\beta_1^g, c, 1/2) = 1 - \gamma(\beta_1^g, c, 1/2),$$

it follows that

$$D_2[b(\beta_1^g, c_{\text{bayes}}, 1/2)] = -D_2[\gamma(\beta_1^g, c_{\text{bayes}}, 1/2)] = 0.$$

9.1.7 Comparing Bayes-optimal decision boundary and large threshold

We compare the bias produced by setting the threshold to a large number to the bias produced by setting the threshold to the Bayes-optimal decision boundary. Let $r : \mathbb{R}^{\geq 0} \to \mathbb{R}$ be the value of attenuation function evaluated at the Bayes-optimal decision boundary $c_{\text{bayes}} = \beta_0^g + (1/2)\beta_1^g$, i.e.

$$r(\beta_1^g) = \gamma(\beta_1^g, \beta_0^g + (1/2)\beta_1^g, 1/2) = \frac{\Phi(\beta_1^g/2) - \Phi(-\beta_1^g/2)}{[\Phi(-\beta_1^g/2) + \Phi(\beta_1^g/2)]\,[\Phi(\beta_1^g/2) + \Phi(-\beta_1^g/2)]}$$

$$= \frac{\int_{-\beta_1^g/2}^{\beta_1^g/2} f}{[1 - \Phi(\beta_1^g/2) + \Phi(\beta_1^g/2)]\,[\Phi(\beta_1^g/2) + 1 - \Phi(\beta_1^g/2)]} = 2\int_0^{\beta_1^g/2} f = 2\Phi(\beta_1^g/2) - 1.$$

We set r to $1/2$ and solve for β_1^g:

$$r(\beta_1^g) = 1/2 \iff 2\Phi(\beta_1^g/2) - 1 = 1/2$$
$$\iff \Phi(\beta_1^g/2) = 3/4 \iff \beta_1^g = 2\Phi^{-1}(3/4) \approx 1.35.$$

Because r is a strictly increasing function, it follows that $r(\beta_1^g) < 1/2$ for $\beta_1^g < 2\Phi^{-1}(3/4)$ and $r(\beta_1^g) > 1/2$ for $\beta_1^g > 2\Phi^{-1}(3/4)$. Next, because

$$b(\beta_1^g, c_{\text{bayes}}, 1/2) = 1 - \gamma(\beta_1^g, c_{\text{bayes}}, 1/2) = 1 - r(\beta_1^g),$$

we have that $b(\beta_1^g, c_{\text{bayes}}, 1/2) > 1/2$ for $\beta_1^g < 2\Phi^{-1}(3/4)$ and $b(\beta_1^g, c_{\text{bayes}}, 1/2) < 1/2$ for $\beta_1^g > 2\Phi^{-1}(3/4)$. Recall that the bias induced by sending the threshold to infinity (as stated in Proposition 5 and proven in Section 9.1.5) is $1/2$, i.e.

$$b(\beta_1^g, \infty, 1/2) = 1/2.$$

We conclude that $b(\beta_1^g, c_{\text{bayes}}, 1/2) > b(\beta_1^g, \infty, 1/2)$ on $\beta_1^g \in [0, 2\Phi^{-1}(3/4))$; $b(\beta_1^g, c_{\text{bayes}}, 1/2) = b(\beta_1^g, \infty, 1/2)$ for $\beta_1^g = 2\Phi^{-1}(3/4)$; and $b(\beta_1^g, c_{\text{bayes}}, 1/2) < b(\beta_1^g, \infty, 1/2)$ on $\beta_1^g \in (2\Phi^{-1}(3/4), \infty)$.

9.1.8 Monotonicity in β_1^g

We show that γ is monotonically increasing in β_1^g for $\pi = 1/2$ and given threshold c. We begin by stating and proving two lemmas. The first lemma establishes an inequality that will serve as the basis for the proof.

Lemma 1 *The following inequality holds:*

$$\begin{aligned}&[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)] \\ &\quad \cdot [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c) + \Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ &\quad \geq [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)]\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]\,. \quad (23)\end{aligned}$$

Proof: We take cases on the sign on β_1^g.
Case 1: $\beta_g^1 < 0$. Then $\beta_1^g + (\beta^g - c) < (\beta_0^g - c)$, implying $\Phi(\beta_0^g + \beta_1^g - c) < \Phi(\beta_0^g - c)$, or $[\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)] < 0$. Moreover, $[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]$ is positive. Therefore, the right-hand side of (23) is negative.

Turning our attention of the left-hand side of (23), we see that

$$\Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1 - \Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1. \quad (24)$$

Additionally, $\Phi(\beta_0^g - c) < 1$ and $\Phi(c - \beta_0^g) > 0$. Combining these facts with (24), we find that

$$[\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c) + \Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] > 0.$$

Finally, because $[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)] > 0$, the entire left-hand side of (23) is positive. The inequality holds for $\beta_1^g < 0$.
Case 2: $\beta_g^1 \geq 0$. We will show that the first term on the LHS of (23) is greater than the first term on the RHS of (23), and likewise that the second term on the LHS is greater than the second term on the RHS, implying the truth of the inequality. Focusing on the first term, the positivity of $\Phi(\beta_0^g - c)$ implies that

$$\Phi(\beta_0^g - c) \geq -\Phi(\beta_0^g - c),$$

and so

$$\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c) \geq \Phi(\beta_0^g - \beta_1^g - c) - \Phi(\beta_0^g - c).$$

Next, focusing on the second term, $\beta_1^g \geq 0$ implies that

$$\beta_1^g + \beta_0^g - c \geq \beta_0^g - c \implies \Phi(\beta_1^g + \beta_0^g - c) - \Phi(\beta_0^g - c) \geq 0. \quad (25)$$

Adding $\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)$ to both sides of (25) yields

$$\Phi(\beta_1^g+\beta_0^g-c)-\Phi(\beta_0^g-c)+\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g) \geq \Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g).$$

The inequality holds for $\beta_1^g \geq 0$. Combining the cases, the inequality holds for all $\beta_1^g \in \mathbb{R}$. □

The second lemma establishes the derivatives of the functions $2 \cdot g$ and $4 \cdot h$ in β_1^g.

Lemma 2 *The derivatives in* β_1^g *of* $2 \cdot g$ *and* $4 \cdot h$ *are*

$$D_1[2g(\beta_1^g, c, 1/2)] = f(\beta_0^g+\beta_1^g-c) \tag{26}$$

and

$$D_1[4h(\beta_1^g, c, 1/2)] = f(\beta_0^g+\beta_1^g-c)\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right] - f(\beta_0^g+\beta_1^g-c)\left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right]. \tag{27}$$

Proof: Apply FTC and product rule. □

We are ready to prove the monotonicity of γ in β_1^g. Subtracting

$$\left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right]\left[\Phi(\beta_0^g+\beta_1^g-c)-\Phi(\beta_0^g-c)\right]$$

from both sides of (23) and multiplying by $f(\beta_0^g+\beta_1^g-c)>0$ yields

$$\begin{aligned} &f(\beta_0^g+\beta_1^g-c)[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)]\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right] \\ &\geq f(\beta_0^g+\beta_1^g-c)\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right][\Phi(\beta_0^g+\beta_1^g-c)-\Phi(\beta_0^g-c)] \\ &-f(\beta_0^g+\beta_1^g-c)\left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right][\Phi(\beta_0^g+\beta_1^g-c)-\Phi(\beta_0^g-c)]. \end{aligned} \tag{28}$$

Next, recall that

$$2g(\beta_1^g, c, 1/2) = \Phi(\beta_0^g+\beta_1^g-c)-\Phi(\beta_0^g-c). \tag{29}$$

and

$$4h(\beta_1^g, c, 1/2) = \left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right]\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right]. \tag{30}$$

Substituting (26, 27, 29, 30) into (28) produces

$$D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) \geq 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)],$$

or

$$D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)] \geq 0. \quad (31)$$

The quotient rule implies that

$$D_1\gamma(\beta_1^g, c, 1/2) = \frac{D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)]}{[4h(\beta_1^g, c, 1/2)]^2}. \quad (32)$$

We conclude by (31) and (32) that γ is monotonically increasing in β_1^g. Finally, $b(\beta_1^g, c, \pi) = 1 - \gamma(\beta_1^g, c, \pi)$ is monotonically decreasing in β_1^g.

9.1.9 Strict attenuation bias

We begin by computing the limit of γ in β_1^g given $\pi = 1/2$. First,

$$\lim_{\beta_1^g \to \infty} \gamma(\beta_1^g, c, 1/2) = \frac{1 - \Phi(\beta_0^g - c)}{[1 + \Phi(\beta_0^g - c)]\,[\Phi(c - \beta_0^g)]} = \frac{\Phi(c - \beta_0^g)}{[1 + \Phi(\beta_0^g - c)]\,[\Phi(c - \beta_0^g)]} = \frac{1}{1 + \Phi(\beta_0^g - c)} < 1.$$

Similarly,

$$\lim_{\beta_1^g \to -\infty} \gamma(\beta_1^g, c, 1/2) = \frac{-\Phi(\beta_0^g - c)}{[\Phi(\beta_0^g - c)]\,[\Phi(c - \beta_0^g) + 1]} = \frac{-1}{1 + \Phi(c - \beta_0^g)} > -1.$$

The function $\gamma(\beta_1^g, c, 1/2, \beta_0^g)$ is monotonically increasing in β_1^g (as stated in Proposition 3 and proven in section 9.1.8). It follows that

$$-1 < -\frac{1}{1 + \Phi(c - \beta_0^g)} \leq \gamma(\beta_1^g, c, 1/2, \beta_0^g) \leq \frac{1}{1 - \Phi(\beta_0^g - c)} < 1$$

for all $\beta_1^g \in \mathbb{R}$. But β_0^g and c were chosen arbitrarily, and so

$$-1 < \gamma(\beta_1^g, c, 1/2, \beta_0^g) < 1$$

for all $(\beta_1^g, c, \beta_0^g) \in \mathbb{R}^3$. Finally, because $b(\beta_1^g, c, 1/2, \beta_0^g) = 1 - \gamma(\beta_1^g, c, 1/2, \beta_0^g)$, it follows that

$$0 < b(\beta_1^g, c, 1/2, \beta_0^g) < 2$$

for all $(\beta_1^g, c, \beta_0^g) \in \mathbb{R}^3$

9.1.10 Bias-variance decomposition in no-intercept model

We prove the bias-variance decomposition for the no-intercept model (3). Define l (for "limit") by

$$l = \beta_m \left(\frac{\omega\pi}{\zeta(1-\pi) + \omega\pi} \right),$$

where

$$\begin{cases} \omega = \bar{\Phi}(c - \beta_g) = \Phi(\beta_g - c) \\ \zeta = \bar{\Phi}(c) = \Phi(-c). \end{cases}$$

We have that

$$\hat{\beta}_m - l = \frac{\sum_{i=1}^n \hat{p}_i m_i}{\sum_{i=1}^n \hat{p}_i^2} - l = \frac{\sum_{i=1}^n \hat{p}_i m_i}{\sum_{i=1}^n \hat{p}_i^2} - \frac{l \sum_{i=1}^n \hat{p}_i^2}{\sum_{i=1}^n \hat{p}_i^2} = \frac{\sum_{i=1}^n \hat{p}_i (m_i - l\hat{p}_i)}{\sum_{i=1}^n \hat{p}_i^2}.$$

Therefore,

$$\sqrt{n}(\hat{\beta}_m - l) = \frac{(1/\sqrt{n}) \sum_{i=1}^n \hat{p}_i (m_i - l\hat{p}_i)}{(1/n) \sum_{i=1}^n \hat{p}_i^2}. \tag{33}$$

Next, we compute the expectation and variance of $\hat{p}_i(m_i - l\hat{p}_i)$. To do so, we first compute several simpler quantities:

1. Expectation of $\hat{p}_i$:

$$\begin{aligned} \mathbb{E}[\hat{p}_i] = \mathbb{P}(p_i\beta_g + \tau_i \geq c) &= \mathbb{P}(\beta_g + \tau_i \geq c)\pi + \mathbb{P}(\tau_i \geq c)(1-\pi) \\ &= \pi\omega + (1-\pi)\zeta. \end{aligned}$$

2. Expectation of $\hat{p}_i p_i$:

$$\mathbb{E}\left[\hat{p}_i p_i\right] = \mathbb{E}\left[\hat{p}_i | p_i = 1\right] \mathbb{P}\left[p_i = 1\right] = \omega\pi.$$

3. Expectation of $\hat{p}_i m_i$:

$$\begin{aligned} \mathbb{E}[\hat{p}_i m_i] = \mathbb{E}\left[\hat{p}_i(\beta_m p_i + \epsilon_i)\right] &= \mathbb{E}\left[\beta_m \hat{p}_i p_i + \hat{p}_i \epsilon_i\right] \\ &= \beta_m \mathbb{E}\left[\hat{p}_i p_i\right] + \mathbb{E}[\hat{p}_i]\mathbb{E}[\epsilon_i] = \beta_m \omega\pi + 0 = \beta_m \omega\pi. \end{aligned}$$

4. Expectation of $\hat{p}_i m_i^2$:

$$\begin{aligned} &\mathbb{E}\left[\hat{p}_i m_i^2\right] = \mathbb{E}\left[\hat{p}_i(\beta_m p_i + \epsilon_i)^2\right] = \mathbb{E}\left[\hat{p}_i\left(\beta_m^2 p_i^2 + 2\beta_m p_i \epsilon_i + \epsilon_i^2\right)\right] \\ &= \mathbb{E}\left[\hat{p}_i p_i \beta_m^2 + 2\beta_m p_i \hat{p}_i \epsilon_i + \hat{p}_i \epsilon_i^2\right] = \beta_m^2 \mathbb{E}[\hat{p}_i p_i] + 2\beta_m \mathbb{E}[p_i \hat{p}_i]\mathbb{E}[\epsilon_i] + \mathbb{E}[\hat{p}_i]\mathbb{E}[\epsilon_i^2] \\ &= \beta_m^2 \mathbb{E}[\hat{p}_i p_i] + \mathbb{E}[\hat{p}_i] = \beta_m^2 \omega\pi + \mathbb{E}[\hat{p}_i]. \end{aligned}$$

Now, we can compute the expectation and variance of $\hat{p}_i(m_i - l\hat{p}_i)$. First,

$$\begin{aligned}\mathbb{E}\left[\hat{p}_i(m_i - l\hat{p}_i)\right] &= \mathbb{E}[\hat{p}_i m_i] - l\mathbb{E}[\hat{p}_i] \\ &= \beta_m\omega\pi - \left(\frac{\beta_m\omega\pi}{\zeta(1-\pi)+\omega\pi}\right)[\zeta(1-\pi)+\omega\pi] = 0. \quad (34)\end{aligned}$$

Additionally,

$$\begin{aligned}\mathbb{V}\left[\hat{p}_i(m_i - l\hat{p}_i)\right] &= \mathbb{E}\left[\hat{p}_i^2(m_i - l\hat{p}_i)^2\right] - \left(\mathbb{E}\left[\hat{p}_i(m_i - l\hat{p}_i)\right]\right)^2 \\ &= \mathbb{E}\left[\hat{p}_i m_i^2\right] - 2l\mathbb{E}[m_i\hat{p}_i] + l^2\mathbb{E}[\hat{p}_i] = \beta_m^2\omega\pi + \mathbb{E}[\hat{p}_i] - 2l\beta_m\omega\pi + l^2\mathbb{E}[\hat{p}_i] \\ &= \beta_m\omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1+l^2). \quad (35)\end{aligned}$$

Therefore, by CLT, (34), and (35),

$$(1/\sqrt{n})\sum_{i=1}^{n}\hat{p}_i(m_i - l\hat{p}_i) \xrightarrow{d} N\left(0, \beta_m\omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1+l^2)\right). \quad (36)$$

Next, by weak LLN,

$$(1/n)\sum_{i=1}^{n}\hat{p}_i^2 = (1/n)\sum_{i=1}^{n}\hat{p}_i \xrightarrow{P} \mathbb{E}[\hat{p}_i]. \quad (37)$$

Finally, by (33), (36), (37), and Slutsky's Theorem,

$$\sqrt{n}(\hat{\beta}_m - l) \xrightarrow{d} N\left(0, \frac{\beta_m\omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1+l^2)}{\left(\mathbb{E}[\hat{p}_i]\right)^2}\right).$$

Thus, for large $n \in \mathbb{N}$, we have that

$$\begin{cases}\mathbb{E}[\hat{\beta}_m] \approx l, \\ \mathbb{V}[\hat{\beta}_m] \approx \left[\beta_m\omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1+l^2)\right]/[n\mathbb{E}^2[\hat{p}_i]],\end{cases}$$

completing the bias-variance decomposition.

9.2 Derivation of EM algorithm

9.3 Derivation of observed information matrix

9.4 Implementation using R family objects

9.5 Statistical accelerations to GLM-EIV

9.6 Additional simulation results

References

[1] Tanja Rothgangl, Melissa K. Dennis, Paulo J.C. Lin, Rurika Oka, Dominik Witzigmann, Lukas Villiger, Weihong Qi, Martina Hruzova, Lucas Kissling, Daniela Lenggenhager, Costanza Borrelli, Sabina Egli, Nina Frey, Noëlle Bakker, John A. Walker, Anastasia P. Kadina, Denis V. Victorov, Martin Pacesa, Susanne Kreutzer, Zacharias Kontarakis, Andreas Moor, Martin Jinek, Drew Weissman, Markus Stoffel, Ruben van Boxtel, Kevin Holden, Norbert Pardi, Beat Thöny, Johannes Häberle, Ying K. Tam, Sean C. Semple, and Gerald Schwank. In vivo adenine base editing of PCSK9 in macaques reduces LDL cholesterol levels. *Nature Biotechnology*, 39(8):949–957, 2021.

[2] Kiran Musunuru, Alexandra C. Chadwick, Taiji Mizoguchi, Sara P. Garcia, Jamie E. DeNizio, Caroline W. Reiss, Kui Wang, Sowmya Iyer, Chaitali Dutta, Victoria Clendaniel, Michael Amaonye, Aaron Beach, Kathleen Berth, Souvik Biswas, Maurine C. Braun, Huei Mei Chen, Thomas V. Colace, John D. Ganey, Soumyashree A. Gangopadhyay, Ryan Garrity, Lisa N. Kasiewicz, Jennifer Lavoie, James A. Madsen, Yuri Matsumoto, Anne Marie Mazzola, Yusuf S. Nasrullah, Joseph Nneji, Huilan Ren, Athul Sanjeev, Madeleine Shay, Mary R. Stahley, Steven H.Y. Fan, Ying K. Tam, Nicole M. Gaudelli, Giuseppe Ciaramella, Leslie E. Stolz, Padma Malyala, Christopher J. Cheng, Kallanthottathil G. Rajeev, Ellen Rohde, Andrew M. Bellinger, and Sekar Kathiresan. In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates. *Nature*, 593(7859):429–434, 2021.

[3] Laralynne Przybyla and Luke A. Gilbert. A new era in functional genomics screens. *Nature Reviews Genetics*, 0123456789, 2021.

[4] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P. Fulco, Livnat Jerby-Arnon, Nemanja D. Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, Britt Adamson, Thomas M. Norman, Eric S. Lander, Jonathan S. Weissman, Nir Friedman, and Aviv Regev. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, 167(7):1853–1866.e17, 2016.

[5] Paul Datlinger, André F. Rendeiro, Christian Schmidl, Thomas Krausgruber, Peter Traxler, Johanna Klughammer, Linda C. Schuster, Amelie

Kuchler, Donat Alpar, and Christoph Bock. Pooled CRISPR screening with single-cell transcriptome readout. *Nature Methods*, 14(3):297–301, 2017.

[6] John A. Morris, Zharko Daniloski, Júlia Domingo, Timothy Barry, Marcello Ziosi, Dafni A. Glinos, Stephanie Hao, Eleni P. Mimitou, Peter Smibert, Kathryn Roeder, Eugene Katsevich, Tuuli Lappalainen, and Neville E. Sanjana. Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing. *bioRxiv*, page 2021.04.07.438882, 2021.

[7] Kevin Z. Lin, Jing Lei, and Kathryn Roeder. Exponential-Family Embedding With Application to Cell Developmental Trajectories for Single-Cell RNA-Seq Data. *Journal of the American Statistical Association*, 0(0):1–32, 2021.

[8] Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology*, 22(1):1–20, 2021.

[9] F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, 20(1):1–16, 2019.

[10] Bettina Grün and Friedrich Leisch. *Finite Mixtures of Generalized Linear Regression Models*, pages 205–230. Physica-Verlag HD, Heidelberg, 2008.

[11] Joseph G. Ibrahim. Incomplete Data in Generalized Linear Models. *Journal of the American Statistical Association*, 85(411):765–769, 1990.

[12] Timothy Barry, Xuran Wang, John A. Morris, Kathryn Roeder, and Eugene Katsevich. Conditional resampling improves calibration and sensitivity in single-cell CRISPR screen analysis. *bioRxiv*, page 2020.08.13.250092, 2020.

[13] Lin Yang, Yuqing Zhu, Hua Yu, Sitong Chen, Yulan Chu, He Huang, Jin Zhang, and Wei Li. Linking genotypes with multiple phenotypes in single-cell CRISPR screens. *bioRxiv*, page 658146, 2019.

[14] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics*, 53(6):770–777, 2021.

[15] L. A. Stefanski. Measurement Error Models. *Journal of the American Statistical Association*, 95(452):1353–1358, 2000.

[16] Patrick Fitzpatrick. *Advanced calculus*, volume 5. American Mathematical Soc., 2009.