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. Single-cell CRISPR screens promise to illuminate regulatory networks underlying diseases but present substantial statistical challenges. 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 **3**

2 Background and analysis challenges **4**

2.1 Related work 4

2.2 Assay overview 5

2.3 Analysis challenges 7

3 Thresholding method **8**
3.1 Empirical challenges of thresholding method 10
3.2 Theoretical challenges of thresholding method 13
3.3 Thresholding method summary 18

4 Generalized linear model with errors in variables **19**
4.1 Model 19
4.2 Estimation and inference 23
4.3 Statistical accelerations 25

5 Simulation studies **25**

6 Real data analysis **25**

7 Discussion **25**

Appendices **25**

A Theoretical details for thresholding estimator **25**
A.1 Notation 25
A.2 Almost sure limit of $\hat{\beta}_1^m$ 27
A.3 Re-expressing γ in a simpler form 28
A.4 Derivatives of g and h in c 29
A.5 Limit of γ in c 29
A.6 Bayes-optimal decision boundary as a critical value of γ 31
A.7 Comparing Bayes-optimal decision boundary and large threshold 32
A.8 Monotonicity in β_1^g 33
A.9 Strict attenuation bias 35
A.10 Bias-variance decomposition in no-intercept model 36

B Estimation and inference in the GLM-EIV model **37**
B.1 Basic model properties 37
B.2 Estimation 38
B.3 Inference 38

C Statistical accelerations **38**

D Additional simulation results **38**
All edits in blue.

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 assigned stochastically to cells and cannot be observed directly. 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 guide RNA 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 guide RNA 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 guide RNA counts, respectively.

Our work makes two main contributions. First, we study 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, GLM-EIV (generalized linear model with errors-in-variables), 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. GLM-EIV 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

2.1 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, a generalization 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 derived a method for estimation and inference in a k-component mixture of GLMs. While we prefer to view GLM-EIV as a generalized errors-in-variables method, the GLM-EIV model is equivalent to a two-component mixture of *products* of GLM

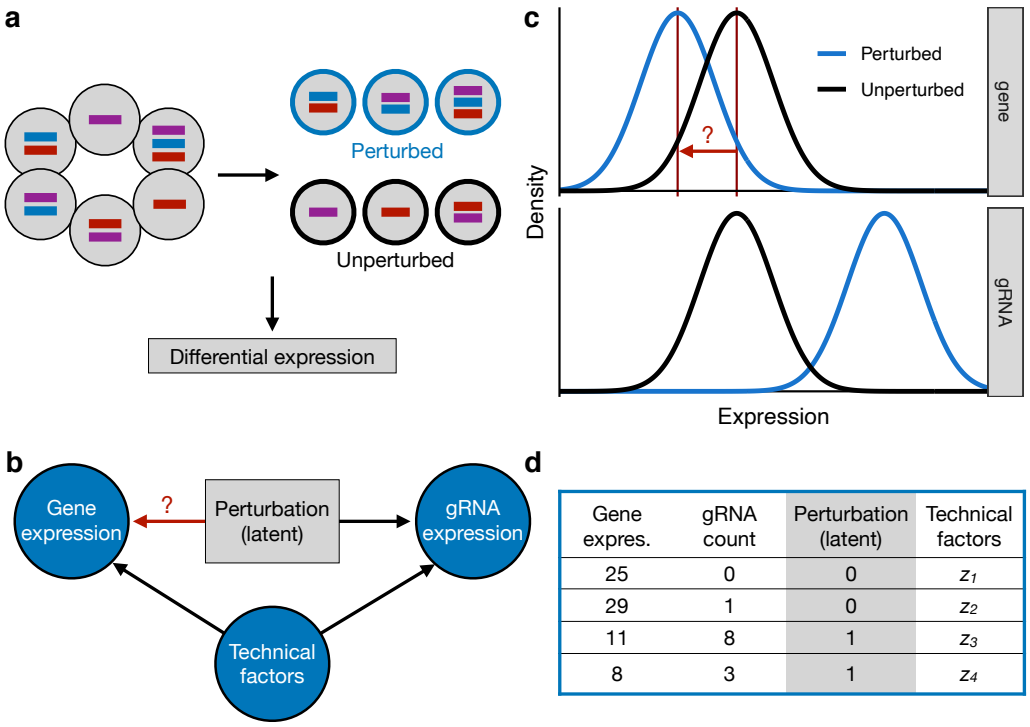
densities. Ibrahim proposed a procedure for fitting GLMs in the presence of missing-at-random covariates. Our method, by contrast, involves fitting two conditionally independent GLMs in the presence of a totally latent covariate. Thus, while Ibrahim and Grün & Leisch are helpful references, our estimation and inference tasks are more complex than theirs.

The genomics literature has produced several applied methods for single-cell CRISPR screen analysis. In a prior work we developed SCEPTRE [12], a custom implementation of the conditional randomization test [13, 14] tailored to single-cell CRISPR screen data. SCEPTRE tests whether a given perturbation is associated with the change in expression of a given gene, adjusting for sources of confounding and ensuring robustness to expression model misspecification. Other applied methods for single-cell CRISPR screen analysis include MIMOSCA [4] and scMAGeCK [15]. These methods, like SCEPTRE, focus on hypothesis testing (rather than estimation), but unlike SCEPTRE, they ignore the countedness of the data and are unable to handle confounders. In this work we tackle a set of analysis challenges that are complimentary to the challenges addressed by SCEPTRE. Most importantly, we seek to *estimate* (with confidence) the effect size of a perturbation on gene expression change, a statistical objective unattainable within the nonparameteric hypothesis testing framework of SCEPTRE.

2.2 Assay overview

There are several broad classes of single-cell CRISPR screen assays, each suited to answer a different set of biological questions [16, 17, 18]. In this work we focus on so-called high-multiplicity of infection (MOI) single-cell CRISPR screens. We expect the ideas that we develop for this assay to apply (with some effort) to other classes of single-cell CRISPR screens as well. In this section we motivate high MOI single-cell screens, overview the experimental protocol, and present relevant analysis challenges.

The human genome consists of genes, enhancers (segments of DNA that regulate the expression of one or more genes), and other genomic elements (that are not of importance to the current discussion). Genome-wide association studies (GWAS) have revealed that the majority ($> 90\%$) of variants associated with diseases lie outside genes and (very likely) inside enhancers [19]. These noncoding variants are thought to contribute to disease by modulating the expression one or more disease-relevant genes. We do not know the gene (or genes) through which most noncoding variants exert their effect,



Gene expres.	gRNA count	Perturbation (latent)	Technical factors
25	0	0	z_1
29	1	0	z_2
11	8	1	z_3
8	3	1	z_4

Figure 1: **Experimental design and analysis challenges**: **a,** Experimental design. For a given perturbation (e.g., the perturbation indicated in blue), we partition the cells into two groups: perturbed and unperturbed. Next, for a given gene, we conduct a differential expression analysis across the two groups, yielding an estimate of the impact of the given perturbation on the given gene. **b,** DAG representing all variables in the system. The perturbation (latent) impacts both gene expression and gRNA expression; technical factors act as nuisance variables, also impacting gene and gRNA expression. The target of estimation is the effect of the perturbation on gene expression. **c,** Schematic illustrating the "background read" phenomenon. Due to errors in the sequencing process, unperturbed cells exhibit a nonzero gRNA count distribution (bottom). The target of estimation is the change in mean gene expression in response to the perturbation (top). **d**, Example data on four cells for a given perturbation-gene pair. Note that (i) the perturbation is unobserved, and (ii) the gene and gRNA data are discrete counts.

limiting the interpretability of GWAS results. A central open challenge in genetics, therefore, is to link enhancers that harbor GWAS variants to the genes that they target at genome-wide scale [20].

The most promising biotechnology for solving this challenge are high MOI single-cell CRISPR screens. High MOI single-cell CRISPR screens combine CRISPR interference (CRISPRi) – a version of CRISPR that represses a targeted region of the genome – with single-cell sequencing. The experimental protocol is as follows. First, the scientist develops a library of several hundred to several thousand CRISPRi perturbations, each designed to target a candidate enhancer for repression. The scientist then cultures tens or hundreds of thousands of cells and delivers the CRISPRi perturbations to these cells. The perturbations assort into the cells randomly, with each cell receiving on average 10-40 distinct perturbations. Conversely, a given perturbation enters about 0.1-2% of cells.

After waiting several days for CRISPRi to take effect, the scientist profiles each cell's transcriptome (i.e., its gene expressions) and the set of perturbations that it received. Finally, the scientist conducts perturbation-to-gene association analyses. Figure 1a depicts this process schematically, with colored bars (blue, red, and purple) representing distinct perturbations. For a given perturbation (e.g., the perturbation represented in blue), the scientist partitions the cells into two groups: those that received the perturbation (top) and those that did not (bottom). Next, for a given gene, the scientist runs a differential expression analysis across the two groups of cells, producing an estimate for the magnitude of the gene expression change in response to the perturbation. If the estimated change in expression is large, the scientist can conclude that the enhancer *targeted* by the perturbation exerts a strong regulatory effect on the gene. This procedure is repeated for a large set of preselected perturbation-gene pairs. The enhancer-by-enhancer approach is valid because the perturbations assort into cells approximately independently of one another.

2.3 Analysis challenges

High MOI single-cell CRISPR screens present several statistical challenges, four of which we highlight here. Throughout, we consider a single perturbation-gene pair. First, the "treatment" variable – i.e., the presence or absence of a perturbation – cannot be directly observed. Instead, perturbed cells transcribe molecules called *guide RNAs* (or *gRNAs*) that serve as indirect proxies

of perturbation presence. We must leverage these gRNAs to impute (explicitly or implicitly) perturbation assignments onto the cells (Figure 1b). Second, "technical factors" – sources of variation that are experimental rather than biological in origin – impact both gene expression and gRNA expression (Figure 1b). Technical factors act as confounders in the measurement process and therefore must be accounted for in differential expression models. Third, the gene and gRNA data are sparse, discrete counts. Therefore, classical statistical approaches that assume Gaussianity or homoscedasticity are inapplicable. Finally, and most subtly, sequenced gRNAs sometimes are mapped to cells that have not received a perturbation. This phenomenon, which we call the "background read" phenomenon, results from errors in the sequencing and alignment processes [21]. The marginal distribution of the gRNA counts is best conceptualized as a mixture model (Figure 1c; Gaussian distributions used for illustration purposes only). Unperturbed and perturbed cells both exhibit nonzero gRNA count distributions, but this distribution overall is greater for perturbed cells. Figure 1d shows example data on four (of possibly tens or hundreds of thousands of) cells. The analysis objective is to leverage the gene expressions and gRNA counts to estimate the effect of the (latent) perturbation on gene expression, accounting for the technical factors.

In this work we analyze two large-scale, high MOI, single-cell CRISPR screen datasets published by Gasperini et al. and Xie et al. Gasperini (resp., Xie) targeted approximately 6,000 (resp., 500) candidate enhancers in a population of approximately 200,000 (resp., 100,000) cells. Gasperini additionally designed 381 positive control, gene-targeting perturbations and 50 non-targeting, negative control perturbations to assess method sensitivity and specificity.

3 Thresholding method

In this section we study thresholded regression from empirical and theoretical perspectives, uncovering several limitations of the method. Gasperini and Xie both imputed perturbation identities onto the cells via thresholding, but they carried out the subsequent differential expression analysis in different ways: Gasperini used negative binomial regression, whereas Xie used nonparametric independence testing. These two strategies pose similar challenges, but we investigate Gasperini's variant of the thresholding method, as it relates most

closely to GLM-EIV.

Let $n \in \mathbb{N}$ be the number of cellxs assayed in the experiment. 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; let $g_i \in \mathbb{N}$ be the number of gRNA transcripts sequenced; let $l_i^m \in \mathbb{N}$ be the number of gene transcripts sequenced across *all* genes (the library size); and finally, let $z_i \in \mathbb{R}^{d-1}$ be the cell-specific technical factors (e.g., sequencing batch, percent mitochondrial reads, etc.) The letters "m," "g", and "l" stand for "mRNA," "gRNA," and "library," respectively. 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
$$\begin{cases} \hat{p}_i = 0 \text{ if } g_i < c, \\ \hat{p}_i = 1 \text{ if } g_i \geq c. \end{cases}$$

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

We include the library size l_i^m as an offset term in (1) so that $\beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i$ can be interpreted as a relative expression: exponentiating both sides of (1), we obtain
$$\mu_i = \exp\left(\beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i\right) l_i^m.$$
We see that $\exp\left(\beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i\right)$ is the *fraction* of all transcripts sequenced in the cell produced by the gene under consideration.

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, controlling for the technical factors. Fold change (obtained by exponentiating β_1^m) is the ratio of the mean gene expression in cells that

were perturbed to the mean gene 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 (accounting for technical factors).

3.1 Empirical challenges of thresholding method

We examined the behavior of the thresholding method on real data and uncovered attenuation bias and bias-variance tradeoff effects. 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 on the positive control data using three different thresholds: 1, 5, and 20. We found that the chosen threshold substantially impacted the results (Figure 2a-b). Estimates for log fold change produced by threshold = 1 were smaller in magnitude than those produced by threshold = 5. (Equivalently, estimates for *raw* fold change were closer to the baseline of 1 for threshold = 1; Figure 2a.) Estimates produced by threshold = 5 and threshold = 20 were more concordant, but threshold = 20 yielded slightly larger effect sizes (Figure 2b).

We reasoned that thresholded regression systematically underestimated effect sizes on the positive control pairs, especially for small thresholds (an example of *attenuation bias*). For a given perturbation, the vast majority ($> 98\%$) of cells are unperturbed. This imbalance leads to an asymmetry: misclassifying *unperturbed* cells as *perturbed* is intuitively "worse" than misclassifying *perturbed* cells as *unperturbed*. Misclassified unperturbed cells contaminate the set of truly perturbed cells, leading to attenuation bias; by contrast, misclassified perturbed cells are swamped in number and "neutralized" by the truly unperturbed cells. Setting the threshold to a large number reduces the unperturbed-to-perturbed misclassification rate, decreasing bias.

We hypothesized, however, that the reduction in bias conferred by selecting a large threshold comes at the cost of increasing the variance of the estimator. To investigate, we compared p-values and confidence intervals produced by threshold = 5 and threshold = 20 for the target of inference β_1^m.

We found that threshold = 5 yielded smaller (i.e., more significant) p-values and narrower confidence intervals than did threshold = 20 (Figure 2c-d). We concluded that the threshold controls a bias-variance tradeoff: as the threshold increases, bias of the estimator decreases and variance increases.

Finally, to determine whether there is an "obvious" location at which to draw the threshold, we examined the empirical gRNA count distributions and checked for bimodality. Figures 2e and 2f display the empirical distribution of a randomly-selected gRNA from the Gasperini and Xie datasets, respectively (counts of 0 omitted). The distributions peak at 1 and then taper off gradually; there does not exist a sharp boundary that cleanly separates the perturbed from the unperturbed cells. Overall, we concluded that the thresholding method faces several challenges: (i) the threshold is a tuning parameter that significantly impacts the results; (ii) the threshold mediates an intrinsic bias-variance tradeoff; and (iii) the gRNA count distributions do not imply a clear threshold selection strategy.

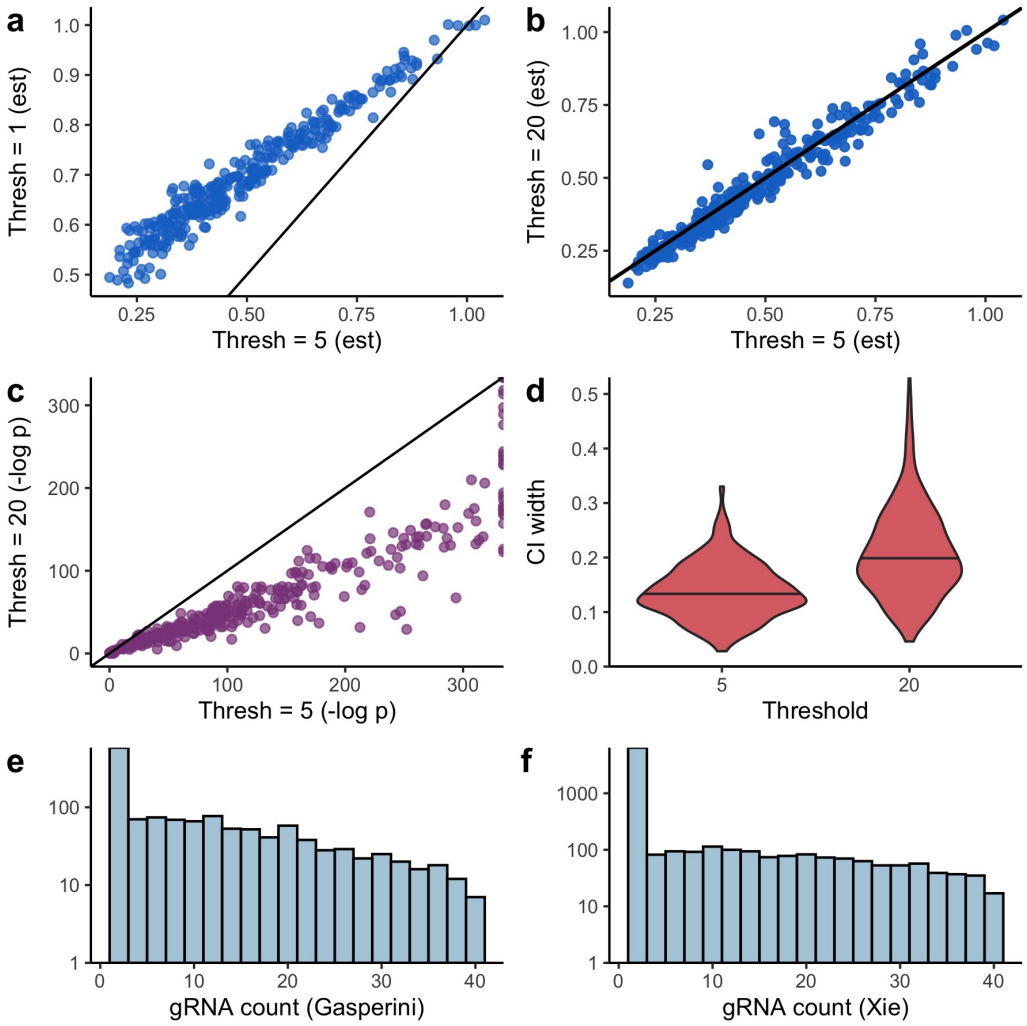


Figure 2: **Empirical challenges of thresholded regression. a-b,** Fold change estimates produced by threshold = 1 versus threshold = 5 (a) and threshold = 20 versus threshold = 5 (b). The selected threshold substantially impacts the results. **c-d,** p-values (c) and CI widths (d) produced by threshold = 20 versus threshold = 5. The latter threshold yields more confident estimates. **e-f**, Empirical distribution of randomly-selected gRNA from Gasperini (e) and Xie (f) data (0 counts not shown). The gRNA data do not appear to imply an obvious threshold selection strategy.

3.2 Theoretical challenges of thresholding method

Next, we study the thresholding method from a theoretical perspective, recovering in precise mathematical terms the attenuation bias and bias-variance tradeoff effects uncovered on real data, as well as several other interesting phenomena. We work in an idealized Gaussian setting. Suppose that we observe gRNA and gene expression data $\{(g_1, m_1), \ldots, (g_n, m_n)\}$ on $n \in \mathbb{N}$ cells from the following model:

$$\begin{cases} m_i = \beta_0^m + \beta_1^m p_i + \epsilon_i \\ g_i = \beta_0^g + \beta_1^g p_i + \tau_i \\ p_i \sim \text{Bern}(\pi) \\ \epsilon_i, \tau_i \sim N(0,1) \\ p_i \perp\!\!\!\perp \tau_i \perp\!\!\!\perp \epsilon_i. \end{cases} \tag{2}$$

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*

$$\hat{\beta}_1^m \xrightarrow{a.s.} \beta_1^m \left(\frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} \right), \tag{3}$$

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 $\hat{\beta}_1^m$:

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

where $\lim_{a.s.}$ denotes a.s. convergence. The asymptotic relative bias vanishes when the attenuation function equals 1.

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 begin by setting 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. SHOULD WE MOVE PROP 4-6 TO THE APPENDIX? THESE PROPS ARE LESS IMPORTANT THAN THE OTHERS AND CONTRIBUTE SOMEWHAT LESS TO THE NARRATIVE.

- **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 [22].

- **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.*

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

  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., $\pi = 0.01$), most cells are unperturbed. Therefore, as discussed in Section 3.1, selecting a large threshold minimizes the unperturbed-to-perturbed misclassification rate, reducing bias.

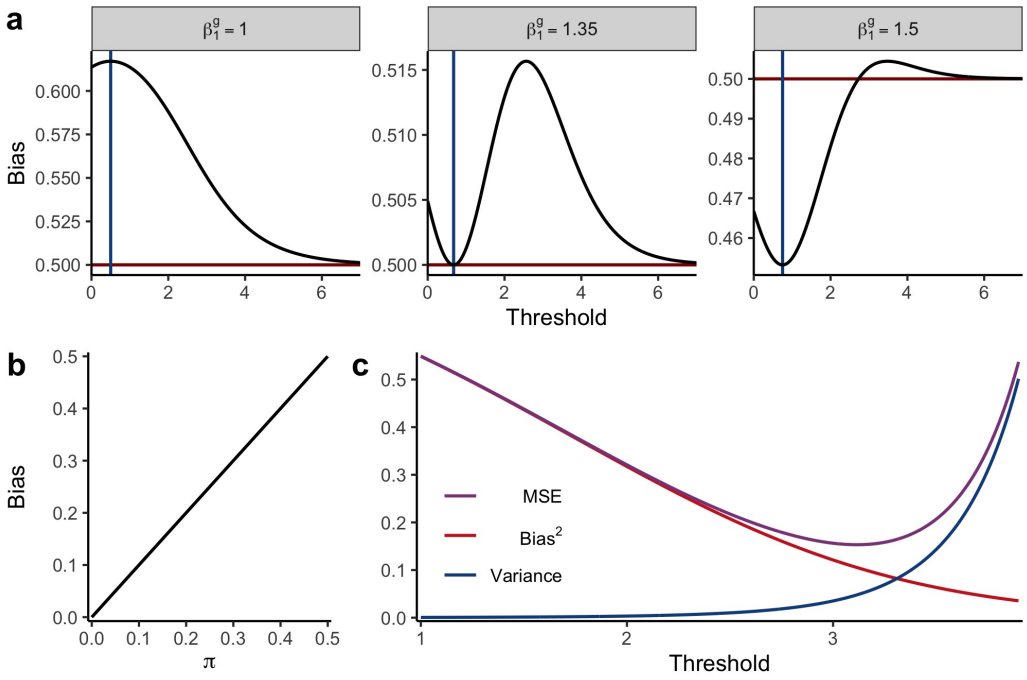


Figure 3: **Theoretical challenges of thresholded regression. a,** Asymptotic relative bias versus threshold for different values of β_1^g. The bias function is highly nonconvex and strictly nonzero. Vertical blue lines, Bayes-optimal decision boundaries. Across all panels, $\beta_0^g = 0$ and $\pi = 1/2$. **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 $\pi = 0.1$.

Bias-variance tradeoff (Panel c)

Finally, to shed light on the costs of selecting a large threshold, we derive an exact bias-variance decomposition for the thresholding estimator. We

consider a slightly simpler, no-intercept version of (2) for this purpose:

$$\begin{cases} m_i = \beta_m p_i + \epsilon_i \\ g_i = \beta_g p_i + \tau_i \\ p_i \sim \text{Bern}(\pi) \\ \epsilon_i, \tau_i \sim N(0,1) \\ p_i \perp\!\!\!\perp \tau_i \perp\!\!\!\perp \epsilon_i. \end{cases} \tag{4}$$

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

$$\hat{\beta}_m = \frac{\sum_{i=1}^n \hat{p}_i m_i}{\sum_{i=1}^n \hat{p}_i^2}. \tag{5}$$

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.

3.3 Thresholding method summary

Empirical and theoretical analyses reveal that the thresholding method poses several challenges: the threshold is a tuning parameter that substantially

impacts the results; strict attenuation bias obtains uniformly over the parameter space and for all choices of the threshold; and there does not exist and obvious threshold selection strategy due to (i) the unimodality of the empirical gRNA count distributions and (ii) the existence of a bias-variance tradeoff mediated by the threshold. These difficulties motivate our core research question: *Does modeling the gRNA count distribution directly, thereby circumventing the need to threshold altogether, lead to simpler, more accurate estimation and inference in single-cell CRISPR screen analysis?* To answer this question, we generalize the classical errors-in-variables model to response distributions and sources of measurement error that are exponential family-distributed.

4 Generalized linear model with errors in variables

In this section we introduce generalized linear model with errors-in-variables (GLM-EIV), derive estimation and inference procedures for the model, and propose several statistical accelerations to reduce the cost of fitting the model.

4.1 Model

Negative binomial model

Building on the work of several previous authors [9, 23, 24], Sarkar and Stephens [25] proposed a simple strategy for modeling for single-cell gene expression data, which, in the framework of the negative binomial GLM, is equivalent to using the log-transformed library size as an offset term (as in (1)). We generalize Sarkar and Stephens' approach to model *both* gene and gRNA modalities. To this end, let the latent variable $p_i \in \{0, 1\}$ indicate whether cell $i \in \{1, \ldots, n\}$ was perturbed. We model the gene expression counts according to

$$m_i | (p_i, z_i, l_i^m) \sim \mathrm{NB}(\mu_i^m), \tag{6}$$

$$\log(\mu_i^m) = \beta_0^m + \beta_1^m p_i + \gamma_m^T z_i + \log(l_i^m), \tag{7}$$

where $\theta^m > 0$ is a known negative binomial size parameter, and $\beta_0^m \in \mathbb{R}, \beta_1^m \in \mathbb{R}$, and $\gamma_m \in \mathbb{R}^{d-2}$ are unknown constants. The model (6) is identical to the thresholding model (1), but the imputed perturbation indicator $\hat{p}_i$ is replaced

by the latent perturbation indicator p_i. Next, let $l_i^g \in \mathbb{N}$ be the number of gRNA transcripts sequenced across *all* gRNAs in cell i (i.e., the gRNA library size). The model for the gRNA counts is

$$g_i|(p_i, z_i, l_i^g) \sim \mathrm{NB}_{\theta^g}\left(\mu_i^g\right), \tag{8}$$

$$\log(\mu_i^g) = \beta_0^g + \beta_1^g p_i + \gamma_g^T z_i + \log(l_i^g), \tag{9}$$

where, similar to above, $\theta^g > 0$ is a known negative binomial size parameter, and $\beta_0^g \in \mathbb{R}, \beta_1^g \in \mathbb{R}, \gamma_g \in \mathbb{R}^{d-2}$ are unknown constants. We use a negative binomial GLM to model the gRNA counts because gRNA molecules are transcribed in the cell in the same way that gene transcripts are [5, 26]. Finally, we model the marginal perturbation probability as

$$p_i \sim \mathrm{Bern}(\pi), \tag{10}$$

where $\pi \in (0, 1/2]$. Together, (6, 7, 8, 9, 10) define the standard GLM-EIV model. The terms $(\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i)$ and $(\beta_0^g + \beta_1^g p_i + \gamma_g^T z_i)$ can be interpreted as relative gene and gRNA expressions, similar to the analogous term in the thresholding model. Likewise, the target of inference β_1^m is the log fold change in gene expression in response to the perturbation, accounting for technical factors.

Full GLM-EIV model

To provide greater modeling flexibility, we generalize the GLM-EIV model to arbitrary exponential family response distributions and link functions. To increase notational compactness, let $\tilde{x}_i = [1, p_i, z_i]^T \in \mathbb{R}^d$ be the vector of covariates (including an intercept term) for the ith cell. (We use the tilde as a reminder that the vector is partially unobserved.) Let $\beta_m = [\beta_0^m, \beta_1^m, \gamma_m]^T \in \mathbb{R}^d$ and $\beta_g = [\beta_0^g, \beta_1^g, \gamma_g]^T \in \mathbb{R}^d$ be the unknown coefficient vectors corresponding to the gene and gRNA expression models, respectively. Finally, let o_i^m and o_i^g be the (possibly zero) offset terms for the gene and gRNA models; in practice, we typically set o_i^m and o_i^g to $\log(l_i^m)$ and $\log(l_i^g)$, respectively.

We use a GLM approach to model the gene and gRNA expressions. Considering first the gene expression model, let the ith linear component l_i^m of the model be

$$l_i^m = \langle \tilde{x}_i, \beta_m \rangle + o_i^m.$$

Let the mean μ_i^m of the ith observation be

$$r_m(\mu_i^m) = l_i^m,$$

where $r_m : \mathbb{R} \to \mathbb{R}$ is a strictly increasing, differentiable link function. Let $\psi_m : \mathbb{R} \to \mathbb{R}$ be the differentiable, cumulant-generating function of the selected exponential family distribution. We can express the canonical parameter η_i^m in terms of ψ_m and r_m by

$$\eta_i^m = \left([\psi_m']^{-1} \circ r_m^{-1}\right)(l_i^m) := h_m(l_i^m).$$

Finally, let $c_m : \mathbb{R} \to \mathbb{R}$ be the carrying density of the selected exponential family distribution. The density f_m of m_i conditional on the the canonical parameter η_i is

$$f_m(m_i; \eta_i^m) = \exp\left\{m_i \eta_i^m - \psi_m(\eta_i^m) + c_m(m_i)\right\}.$$

The function c_m does not appear in the log-likelihood of m_i; therefore, the only functions relevant to inference are ψ_m and r_m.

Let the terms $l_i^g, o_i^g, \mu_i^g, \eta_i^g, \psi_g, r_g, h_g$ and c_g be defined in an analogous way for the gRNA model:

$$\begin{cases} l_i^g = \langle \tilde{x}_i, \beta_g \rangle + o_i^g, \\ r_g(\mu_i^g) = l_i^g, \\ \eta_i^g = \left([\psi_g']^{-1} \circ r_g^{-1}\right)(l_i^g) := h_g(l_i^g). \end{cases}$$

The density f_g of g_i given the canonical parameter is

$$f_g(m_i; \eta_i^g) = \exp\left\{g_i \eta_i^g - \psi_g(\eta_i^g) + c_g(g_i)\right\}.$$

Finally, the unobserved variable p_i is assumed to follow a Bernoulli distribution with mean $\pi \in (0, 1/2]$. Its marginal density f_p is given by

$$f_p(p_i) = \pi^{p_i}(1 - \pi)^{1 - p_i}.$$

The unknown parameters in the model are $\theta = [\beta_m, \beta_g, \pi]^T \in \mathbb{R}^{2d+1}$.

Notation

We briefly introduce notation that we will use throughout. For $k \in \{0, 1\}$, let $\tilde{x}_i(k) := [1, k, z_i]^T$ denote the value of $\tilde{x}_i$ that results from setting p_i to k. Next, let $l_i^m(k) := \langle \tilde{x}_i(k), \beta_m \rangle + o_i^m$ and $\eta_i^m(k) := h_m(l_i^m(k))$, and let the

corresponding gRNA quantities $l_i^g(k)$ and $\eta_i^g(k)$ be defined analogously. Also, let the design matrix $\tilde{X} \in \mathbb{R}^{n \times d}$ be defined by

$$\tilde{X} := \begin{bmatrix} \tilde{x}_1^T \\ \tilde{x}_2^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix} = \begin{bmatrix} 1 & p_1 & z_1 \\ 1 & p_2 & z_2 \\ \vdots & \vdots & \vdots \\ 1 & p_n & z_n \end{bmatrix}.$$

For $k \in \{0, 1\}$, let $\tilde{X}(k) \in \mathbb{R}^{n \times d}$ be the matrix that results from setting p_i to k for all $i \in \{1, \ldots, n\}$ in $\tilde{X}$.

Next, let $[m, m]^T \in \mathbb{R}^{2n}$ denote the vector that results from concatenating $m := [m_1, \ldots, m_n]^T \in \mathbb{R}^n$ to itself, i.e.

$$[m, m]^T := [\underbrace{m_1, m_2 \ldots, m_{n-1}, m_n}_{\text{first copy of } m}, \underbrace{m_1, m_2, \ldots, m_{n-1}, m_n}_{\text{second copy of } m}].$$

Define $[g, g]^T$, $[o^g, o^g]^T$, and $[o^m, o^m]^T$ similarly. Finally, let $\begin{bmatrix} \tilde{X}(0) \\ \tilde{X}(1) \end{bmatrix}$ denote the $\mathbb{R}^{2n \times d}$ matrix that results from vertically concatenating $\tilde{X}(0)$ and $\tilde{X}(1)$.

Log likelihood and model properties

We derive the log-likelihood of the GLM-EIV model. We conduct estimation and inference *conditional* on the library sizes and technical factors l_i^m, l_i^g, and z_i; therefore, we treat these quantities as fixed constants. We assume that the gene expression m_i and gRNA expression g_i are *conditionally independent* given the perturbation p_i. The joint density f of (m_i, g_i, p_i) given θ is

$$f(m_i, g_i, p_i; \theta) = f_m(m_i|p_i) f_g(g_i|p_i) f_p(p_i) = \pi^{p_i}(1-\pi)^{1-p_i} f_m(m_i; \eta_i^m) f_g(g_i; \eta^g). \tag{11}$$

The log-likelihood is

$$\mathcal{L}(\theta; m_i, g_i, p_i) = \sum_{i=1}^{n} \log\left(\pi^{p_i}(1-\pi)^{1-p_i}\right) + \sum_{i=1}^{n} \log\left(f_m(m_i; \eta_i^m)\right) + \sum_{i=1}^{n} \log\left(f_g(g_i; \eta_i^g)\right). \tag{12}$$

Integrating over the unobserved variable p_i, we can write the marginal density f of (m_i, g_i) as

$$f(m_i, g_i; \theta) = (1-\pi) f(m_i; \eta_i^m(0)) f(g_i; \eta_i^g(0)) + \pi f(m_i; \eta_i^m(1)) f(g_i; \eta_i^g(1)). \tag{13}$$

We see from (13) that the GLM-EIV model is equivalent to a two-component mixture of *products* of GLM densities. Additionally, the GLM-EIV model is a generalization of the classical errors-in-variables model (when the predictor is binary). Suppose that we observe data $(x_1, y_1), \ldots, (x_n, y_n)$ from the following model:

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i \\ x_i = x_i^* + \tau_i, \end{cases} \tag{14}$$

where $x_i^* \sim \text{Bern}(\pi), \epsilon_i \sim N(0,1), \tau_i \sim N(0,1)$, and ϵ_i,τ_i, and x_i^* are independent. The model (14) is a special case of the GLM-EIV model. More generally, GLM-EIV allows the use of entire regression functions (that optionally include covariates and use nonlinear links) to model y_i and x_i. Moreover, x_i and y_i need not be Gaussian in the GLM-EIV case.

4.2 Estimation and inference

We derive an EM algorithm (Algorithm 1) to estimate the parameters of the GLM-EIV model. The E step entails computing the membership probability (i.e., the probability of perturbation) of each cell. The membership probability $T_i(1)$ of cell $i \in \{1, \ldots, n\}$ given the current parameter estimates $(\beta_m^{(\text{t})}, \beta_g^{(\text{t})}, \pi^{(\text{t})})$ and observed data (m_i, g_i) is

$$T_i(1) = \mathbb{P}(p_i = 1 | M_i = m_i, G_i = g_i, \beta_m^{(\text{t})}, \beta_g^{(\text{t})}, \pi^{(\text{t})}).$$

We can calculate this quantity by applying (i) Bayes rule, (ii) the conditional independence property of M_i and G_i, (iii) the density of M_i and G_i, and (iv) a log-sum-exp-type trick to ensure numerical stability. Next, we produce updated estimates $\pi^{(t+1)}$, $\beta_g^{(t+1)}$, and $\beta_m^{(t+1)}$ of the parameters by maximizing the M step objective function. It turns out that maximizing the objective function is equivalent to setting $\pi^{(\text{t}+1)}$ to the mean of the current membership probabilities and setting $\beta_g^{(t+1)}$ and $\beta_m^{(t+1)}$ to the fitted coefficients of a GLM weighted by the current membership probabilities (Algorithm 1). We iterate through the E and M steps until the marginal likelihood (13) converges (see appendix for full details). Our EM algorithm is reminiscent of (but distinct

from) that of Ibrahim [11], who also leveraged weighted GLM solvers to carry out an M step.

Algorithm 1 EM algorithm for GLM-EIV model.

Require: Pilot estimates $\beta_m^{\text{curr}}, \beta_g^{\text{curr}}$, and π^{curr}; data $[m_1, \ldots, m_n]$, $[g_1, \ldots, g_n]$, $[o_1^m, \ldots, o_n^m]$, $[o_1^g, \ldots, o_n^g]$, and $[z_1, \ldots, z_n]$.

while Not converged **do**

for $i \in \{1, \ldots, n\}$ **do** ▷ E step

$T_i(1) \leftarrow \mathbb{P}\left(p_i = 1 | M_i = m_i, G_i = g_i, \beta_m^{\text{curr}}, \beta_g^{\text{curr}}, \pi^{\text{curr}}\right)$

$T_i(0) \leftarrow 1 - T_i(1)$

end for

$\pi^{\text{curr}} \leftarrow (1/n) \sum_{i=1}^n T_i(1)$ ▷ M step

$w \leftarrow [T_1(0), T_2(0), \ldots, T_n(0), T_1(1), T_2(1), \ldots, T_n(1)]^T$

Fit GLM with responses $[m, m]^T$, offsets $[o^m, o^m]^T$, weights w, and design matrix $\begin{bmatrix} \tilde{X}(0) \\ \tilde{X}(1) \end{bmatrix}$; set β_m^{curr} to estimated coefficient vector.

Fit GLM with responses $[g, g]^T$, offsets $[o^g, o^g]^T$, weights w, and design matrix $\begin{bmatrix} \tilde{X}(0) \\ \tilde{X}(1) \end{bmatrix}$; set β_g^{curr} to estimated coefficient vector.

Compute marginal likelihood given β_m^{curr}, β_g^{curr}, and π^{curr}.

end while

$\hat{\beta}_m \leftarrow \beta_m^{\text{curr}}$; $\hat{\beta}_g \leftarrow \beta_g^{\text{curr}}$; $\hat{\pi} \leftarrow \pi^{\text{curr}}$.

return $(\hat{\beta}_m, \hat{\beta}_g, \hat{\pi})$

After fitting the model, we perform inference on the estimated parameters. The easiest approach, given the complexity of the log likelihood, would be to run a parametric bootstrap on the fitted model. This strategy, however, is prohibitively slow, as the data are large and we use an EM algorithm (that, in practice, requires multiple starts) to fit the model. Therefore, we derive an analytic formula for the asymptotic observed information matrix using Louis's Theorem [27] (see appendix). Leveraging this analytic formula, we can calculate standard errors (and p-values and confidence intervals) quickly, enabling us to perform inference in practice on real, large-scale data.

4.3 Statistical accelerations

5 Simulation studies

6 Real data analysis

7 Discussion

Appendices

A Theoretical details for thresholding estimator

This section contains proofs of the propositions presented Section 3.2, "Theoretical analysis of thresholding estimator." The subsections are organized as follows. Section (A.1) introduces some notation. Section (A.2) establishes almost sure convergence of the thresholding estimator in the model (2), proving Proposition 1. Section (A.3) simplifies the expression for the attenuation function γ, and section (A.4) computes derivatives of γ to be used throughout the proofs. Section (A.5) establishes the limit in c of γ, proving Proposition 7 and as a corollary Proposition 5. Section (A.6) establishes that the Bayes-optimal decision boundary is a critical value of γ, proving Proposition 4, and section (A.7) compares the competing threshold selection strategies head-to-head, proving Proposition 6. Section (A.8) demonstrates that γ is monotone in β_1^g, proving Proposition 3, and Section (A.9) establishes attenuation bias of the thresholding estimator, proving Proposition 2. Finally, Section (A.10) derives the bias-variance decomposition of the thresholding estimator in the model (4), proving Proposition 8.

A.1 Notation

All notation introduced in this subsection (i.e., A.1) pertains to the Gaussian model with intercepts (2). 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

$$g(\beta_1^g, c, \pi, \beta_0^g) = (1-\pi)\left(\Phi(\beta_0^g + \beta_1^g - c)\right) - (1-\pi)\left(\Phi(\beta_0^g - c)\right) \tag{15}$$

and

$$\begin{aligned} h(\beta_1^g, c, \pi, \beta_0^g) = &\left[(1-\pi)\left(\Phi(\beta_0^g - c)\right) + \pi\left(\Phi(\beta_0^g + \beta_1^g - c)\right)\right] \cdot \\ &\left[(1-\pi)\left(\Phi(c - \beta_0^g)\right) + \pi\left(\Phi(c - \beta_0^g - \beta_1^g)\right)\right]. \end{aligned} \tag{16}$$

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* [28]). 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).

A.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 (2). Dividing by n in (3), 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,

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

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 $\hat{p}_i, m_i$:

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

Combining these expressions, we have that

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

A.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

$$\begin{aligned}\mathbb{E}\left[\hat{p}_i\right] &= (1-\pi)\bar{\Phi}(c-\beta_0^g) + \pi\bar{\Phi}(c-\beta_0^g-\beta_1^g) \\ &= (1-\pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g+\beta_1^g-c), \quad (17)\end{aligned}$$

and so

$$\begin{aligned}1-\mathbb{E}\left[\hat{p}_i\right] = (1-\pi)+\pi-\mathbb{E}[\hat{p}_i] &= (1-\pi)\left(1-\bar{\Phi}(c-\beta_0^g)\right)+\pi\left(1-\bar{\Phi}(c-\beta_0^g-\beta_1^g)\right) \\ &= (1-\pi)\Phi(c-\beta_0^g) + \pi\Phi(c-\beta_0^g-\beta_1^g). \quad (18)\end{aligned}$$

Next,

$$\omega = \Phi(\beta_1^g+\beta_0^g-c), \quad (19)$$

and so

$$\begin{aligned}\omega - \mathbb{E}[\hat{p}_i] &= \Phi(\beta_1^g+\beta_0^g-c) - (1-\pi)\Phi(\beta_0^g-c) - \pi\Phi(\beta_0^g+\beta_1^g-c) \\ &\quad (1-\pi)\Phi(\beta_1^g+\beta_0^g-c) - (1-\pi)\Phi(\beta_0^g-c). \quad (20)\end{aligned}$$

Combining (17, 18, 19, 20), we find that

$$\begin{aligned}\gamma(\beta_1^g, c, \pi) &= \frac{\pi(\omega-\mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1-\mathbb{E}[\hat{p}_i])} \\ &= \frac{\pi\left[(1-\pi)\Phi(\beta_0^g+\beta_1^g-c) - (1-\pi)\Phi(\beta_0^g-c)\right]}{\left[(1-\pi)\Phi(\beta_0^g-c)+\pi\Phi(\beta_0^g+\beta_1^g-c)\right]\left[(1-\pi)\Phi(c-\beta_0^g)+\pi\Phi(c-\beta_0^g-\beta_1^g)\right]}. \quad (21)\end{aligned}$$

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

$$\begin{aligned}&\gamma(\beta_1^g, c, 1/2) \\ &\qquad = \frac{\Phi(\beta_0^g+\beta_1^g-c) - \Phi(\beta_0^g-c)}{\left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right]\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right]}. \quad (22)\end{aligned}$$

Recalling the definitions of g (15) and h (16), 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 (22) is identical to

$$\gamma(\beta_1^g, c, 1/2) = \frac{(4)(1/2)g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)} = \frac{2g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)}, \tag{23}$$

i.e., the numerator and denominator of (23) coincide with those of (22). We sometimes will use the notation $2 \cdot g$ and $4 \cdot h$ to refer to the numerator and denominator of (22), respectively.

A.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

$$\begin{aligned} D_2 g(\beta_1^g, c, \pi) &= -(1-\pi) f(\beta_0^g + \beta_1^g - c) + (1-\pi) f(\beta_0^g - c) \\ &= (1-\pi) f(c - \beta_0^g) - (1-\pi) f(c - \beta_0^g - \beta_1^g). \end{aligned} \tag{24}$$

Second, we have that

$$\begin{aligned} & D_2 h(\beta_1^g, c, \pi) \\ &= -[(1-\pi) f(\beta_0^g - c) + \pi f(\beta_0^g + \beta_1^g - c)] \left[(1-\pi)\Phi(c - \beta_0^g) + \pi \Phi(c - \beta_0^g - \beta_1^g)\right] \\ &+ \left[(1-\pi) f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)\right] \left[(1-\pi)\Phi(\beta_0^g - c) + \pi \Phi(\beta_0^g + \beta_1^g - c)\right] \\ &= \left[(1-\pi) f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)\right] \cdot \\ &\quad \Big[(1-\pi)\Phi(\beta_0^g - c) + \pi \Phi(\beta_0^g + \beta_1^g - c) \\ &\quad - (1-\pi)\Phi(c - \beta_0^g) - \pi \Phi(c - \beta_0^g - \beta_1^g)\Big]. \end{aligned} \tag{25}$$

A.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 (24) and (25) that

$$\begin{aligned}\lim_{c\to\infty}\gamma(\beta_1^g,c,\pi) &= \lim_{c\to\infty}\frac{\pi D_2 g(\beta_1^g,c,\pi)}{D_2 h(\beta_1^g,c,\pi)}\\ &= \lim_{c\to\infty}\left\{\frac{(1-\pi)f(c-\beta_0^g)+\pi f(c-\beta_0^g-\beta_1^g)}{\pi(1-\pi)f(c-\beta_0^g)-\pi(1-\pi)f(c-\beta_0^g-\beta_1^g)}\right.\\ &\qquad\left.\cdot\Big[(1-\pi)\Phi(\beta_0^g-c)+\pi\Phi(\beta_0^g+\beta_1^g-c)\right.\\ &\qquad\left.-(1-\pi)\Phi(c-\beta_0^g)-\pi\Phi(c-\beta_0^g-\beta_1^g)\Big]\right\}^{-1}. \quad (26)\end{aligned}$$

We evaluate the two terms in the product (26) separately. Dividing by $f(c-\beta_0^g-\beta_1^g)>0$, we see that

$$\frac{(1-\pi)f(c-\beta_0^g)+\pi f(c-\beta_0^g-\beta_1^g)}{\pi(1-\pi)f(c-\beta_0^g)-\pi(1-\pi)f(c-\beta_0^g-\beta_1^g)} = \frac{\frac{(1-\pi)f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)}+\pi}{\frac{\pi(1-\pi)f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)}-\pi(1-\pi)}. \quad (27)$$

To evaluate the limit of (27), we first evaluate the limit of

$$\begin{aligned}\frac{f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)} &= \frac{\exp\left[-(1/2)(c-\beta_0^g)^2\right]}{\exp\left[-(1/2)(c-\beta_0^g-\beta_1^g)^2\right]}\\ &= \frac{\exp[-(1/2)(c^2-2c\beta_0^g+(\beta_0^g)^2)]}{\exp\left[-(1/2)(c^2-2c\beta_0^g-2c\beta_1^g+(\beta_0^g)^2+2(\beta_0^g\beta_1^g)+(\beta_1^g)^2)\right]}\\ &= \exp\big[-c^2/2+c\beta_0^g-(\beta_0^g)^2/2\\ &\quad +c^2/2-c\beta_0^g-c\beta_1^g+(\beta_0^g)^2/2+\beta_0^g\beta_1^g+(\beta_1^g)^2/2\big]\\ &= \exp[-c\beta_1^g+\beta_0^g\beta_1^g+(\beta_1^g)^2/2] = \exp[\beta_0^g\beta_1^g+(\beta_1^g)^2/2]\exp[-c\beta_1^g]. \quad (28)\end{aligned}$$

Taking the limit in (28), 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 (27):

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

Next, we compute the limit of the other term in the product (26):

$$\lim_{c\to\infty}\Big[(1-\pi)\Phi(\beta_0^g-c)+\pi\Phi(\beta_0^g+\beta_1^g-c) \\ -(1-\pi)\Phi(c-\beta_0^g)-\pi\Phi(c-\beta_0^g-\beta_1^g)\Big]=-(1-\pi)-\pi=-1. \quad (29)$$

Combining (27) and (29), the limit (26) 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

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

A.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 (23), the quotient rule implies that

$$D_2\gamma(\beta_1^g,c,1/2)=\frac{D_2[2g(\beta_1^g,c,1/2)]4h(\beta_1^g,c,1/2)-2g(\beta_1^g,c,1/2)D_2[4h(\beta_1^g,c,1/2)]}{[4h(\beta_1^g,c,\pi)]^2}. \quad (30)$$

We have by (24) that

$$D_2[2g(\beta_1^g,c_{\text{bayes}},1/2)]=f(\beta_1^g/2)-f(-\beta_1^g/2)=f(\beta_1^g/2)-f(\beta_1^g/2)=0. \quad (31)$$

Similarly, we have by (25) that

$$D_2[4h(\beta_1^g,c_{\text{bayes}},\pi)]=[f(\beta_1^g/2)+f(-\beta_1^g/2)]\cdot \\ [\Phi(-\beta_1^g/2)+\Phi(\beta_1^g/2)-\Phi(\beta_1^g/2)-\Phi(-\beta_1^g/2)]=0. \quad (32)$$

Plugging in (32) and (31) to (30), 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

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

A.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 A.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 $\beta_1^g \in (2\Phi^{-1}(3/4), \infty)$.

A.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:*

$$\begin{aligned}&[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)] \\ &\quad \cdot [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c) + \Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ &\quad \geq [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)]\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]\,. \quad (33)\end{aligned}$$

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 (33) is negative.

Turning our attention of the left-hand side of (33), we see that

$$\Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1 - \Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1. \quad (34)$$

Additionally, $\Phi(\beta_0^g - c) < 1$ and $\Phi(c - \beta_0^g) > 0$. Combining these facts with (34), 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 (33) 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 (33) is greater than the first term on the RHS of (33), 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

$$\beta_1^g + \beta_0^g - c \geq \beta_0^g - c \implies \Phi(\beta_1^g + \beta_0^g - c) - \Phi(\beta_0^g - c) \geq 0. \quad (35)$$

Adding $\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)$ to both sides of (35) 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*

$$D_1[2g(\beta_1^g, c, 1/2)] = f(\beta_0^g+\beta_1^g-c) \tag{36}$$

and

$$D_1[4h(\beta_1^g, c, 1/2)] = f(\beta_0^g+\beta_1^g-c)\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right] - f(\beta_0^g+\beta_1^g-c)\left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right]. \tag{37}$$

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 (33) and multiplying by $f(\beta_0^g+\beta_1^g-c)>0$ yields

$$\begin{aligned} &f(\beta_0^g+\beta_1^g-c)[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)]\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right] \\ &\geq f(\beta_0^g+\beta_1^g-c)\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right][\Phi(\beta_0^g+\beta_1^g-c)-\Phi(\beta_0^g-c)] \\ &-f(\beta_0^g+\beta_1^g-c)\left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right][\Phi(\beta_0^g+\beta_1^g-c)-\Phi(\beta_0^g-c)]. \end{aligned} \tag{38}$$

Next, recall that

$$2g(\beta_1^g, c, 1/2) = \Phi(\beta_0^g+\beta_1^g-c)-\Phi(\beta_0^g-c). \tag{39}$$

and

$$4h(\beta_1^g, c, 1/2) = \left[\Phi(\beta_0^g-c)+\Phi(\beta_0^g+\beta_1^g-c)\right]\left[\Phi(c-\beta_0^g)+\Phi(c-\beta_0^g-\beta_1^g)\right]. \tag{40}$$

Substituting (36, 37, 39, 40) into (38) 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

$$D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)] \geq 0. \quad (41)$$

The quotient rule implies that

$$D_1\gamma(\beta_1^g, c, 1/2) = \frac{D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)]}{[4h(\beta_1^g, c, 1/2)]^2}. \quad (42)$$

We conclude by (41) and (42) 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.

A.9 Strict attenuation bias

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

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

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

A.10 Bias-variance decomposition in no-intercept model

We prove the bias-variance decomposition for the no-intercept model (4). 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,

$$\sqrt{n}(\hat{\beta}_m - l) = \frac{(1/\sqrt{n}) \sum_{i=1}^n \hat{p}_i (m_i - l\hat{p}_i)}{(1/n) \sum_{i=1}^n \hat{p}_i^2}. \tag{43}$$

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,

$$\begin{aligned}\mathbb{E}\left[\hat{p}_i(m_i - l\hat{p}_i)\right] &= \mathbb{E}[\hat{p}_i m_i] - l\mathbb{E}[\hat{p}_i] \\ &= \beta_m\omega\pi - \left(\frac{\beta_m\omega\pi}{\zeta(1-\pi)+\omega\pi}\right)[\zeta(1-\pi)+\omega\pi] = 0. \quad (44)\end{aligned}$$

Additionally,

$$\begin{aligned}\mathbb{V}\left[\hat{p}_i(m_i - l\hat{p}_i)\right] &= \mathbb{E}\left[\hat{p}_i^2(m_i - l\hat{p}_i)^2\right] - \left(\mathbb{E}\left[\hat{p}_i(m_i - l\hat{p}_i)\right]\right)^2 \\ &= \mathbb{E}\left[\hat{p}_i m_i^2\right] - 2l\mathbb{E}[m_i\hat{p}_i] + l^2\mathbb{E}[\hat{p}_i] = \beta_m^2\omega\pi + \mathbb{E}[\hat{p}_i] - 2l\beta_m\omega\pi + l^2\mathbb{E}[\hat{p}_i] \\ &= \beta_m\omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1+l^2). \quad (45)\end{aligned}$$

Therefore, by CLT, (44), and (45),

$$(1/\sqrt{n})\sum_{i=1}^{n}\hat{p}_i(m_i - l\hat{p}_i) \xrightarrow{d} N\left(0, \beta_m\omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1+l^2)\right). \quad (46)$$

Next, by weak LLN,

$$(1/n)\sum_{i=1}^{n}\hat{p}_i^2 = (1/n)\sum_{i=1}^{n}\hat{p}_i \xrightarrow{P} \mathbb{E}[\hat{p}_i]. \quad (47)$$

Finally, by (43), (46), (47), 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.

B Estimation and inference in the GLM-EIV model

B.1 Estimation

B.2 Inference

C Statistical accelerations

D 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] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 80(3):551–577, 2018.

[14] Molei Liu, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas. Fast and Powerful Conditional Randomization Testing via Distillation. *Biometrika*, pages 1–25, 2021.

[15] 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.

[16] Molly Gasperini, Andrew J. Hill, José L. McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D. Zhang, Dana Jackson, Anh Leith, Jacob Schreiber, William S. Noble, Cole Trapnell, Nadav Ahituv, and Jay Shendure. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1-2):377–390.e19, 2019.

[17] Paul Datlinger, André F. Rendeiro, Thorina Boenke, Martin Senekowitsch, Thomas Krausgruber, Daniele Barreca, and Christoph Bock. Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nature Methods*, 18(6):635–642, 2021.

[18] Eleni P. Mimitou, Anthony Cheng, Antonino Montalbano, Stephanie Hao, Marlon Stoeckius, Mateusz Legut, Timothy Roush, Alberto Herrera, Efthymia Papalexi, Zhengqing Ouyang, Rahul Satija, Neville E. Sanjana, Sergei B. Koralov, and Peter Smibert. Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells. *Nature Methods*, 16(5):409–412, 2019.

[19] Michael D. Gallagher and Alice S. Chen-Plotkin. The Post-GWAS Era: From Association to Function. *American Journal of Human Genetics*, 102(5):717–730, 2018.

[20] Molly Gasperini, Jacob M. Tome, and Jay Shendure. Towards a comprehensive catalogue of validated and target-linked human enhancers. *Nature Reviews Genetics*, 21(5):292–310, 2020.

[21] Joseph M. Replogle, Thomas M. Norman, Albert Xu, Jeffrey A. Hussmann, Jin Chen, J. Zachery Cogan, Elliott J. Meer, Jessica M. Terry, Daniel P. Riordan, Niranjan Srinivas, Ian T. Fiddes, Joseph G. Arthur, Luigi J. Alvarado, Katherine A. Pfeiffer, Tarjei S. Mikkelsen, Jonathan S. Weissman, and Britt Adamson. Combinatorial single-cell

CRISPR screens by direct guide RNA capture and targeted sequencing. *Nature Biotechnology*, 2020.

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

[23] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature Biotechnology*, 38:142–150, 2020.

[24] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology*, 20(1):1–15, 2019.

[25] 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.

[26] Andrew J. Hill, José L. McFaline-Figueroa, Lea M. Starita, Molly J. Gasperini, Kenneth A. Matreyek, Jonathan Packer, Dana Jackson, Jay Shendure, and Cole Trapnell. On the design of CRISPR-based single-cell molecular screens. *Nature Methods*, 15(4):271–274, 2018.

[27] By Thomas A Louis. Finding the Observed Information Matrix when Using the EM Algorithm Author (s): Thomas A . Louis Reviewed work (s): Source : Journal of the Royal Statistical Society . Series B (Methodological), Vol . 44 , No . 2 Published by : Blackwell Publishing. *Society*, 44(2):226–233, 2012.

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