Tim, Gene, Kathryn

A class of exponential family measurement error models, with application to CRISPR genome editing and single-cell sequencing

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. Despite their promise, single-cell CRISPR screens 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 ("GLM-based 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 in challenging problem settings. We develop a computational infrastructure to deploy GLM-EIV across tens or hundreds of nodes 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 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 conduct a detailed study of the thresholding method. 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 setting. Second, we introduce a new method, GLM-EIV ("GLM-based 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 thereby 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 possibly are of independent utility) to bring the cost of GLM-EIV down to within about an order of magnitude of the thresholding method. Finally, we develop a Docker-containerized application to deploy GLM-EIV at-scale across tens or hundreds of nodes on clouds (e.g., Microsoft Azure) and high-performance clusters. Leveraging this application, we apply GLM-EIV to analyze

two recent, large-scale, single-cell CRISPR screen datasets.

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 negative binomially-distributed gene expression counts while controlling for technical factors. 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 those addressed by SCEPTRE. Most importantly, we seek to *estimate* (with confidence) the effect size of a perturbation on gene expression change, an objective that is unattainable within the nonparameteric hypothesis testing framework of SCEPTRE.

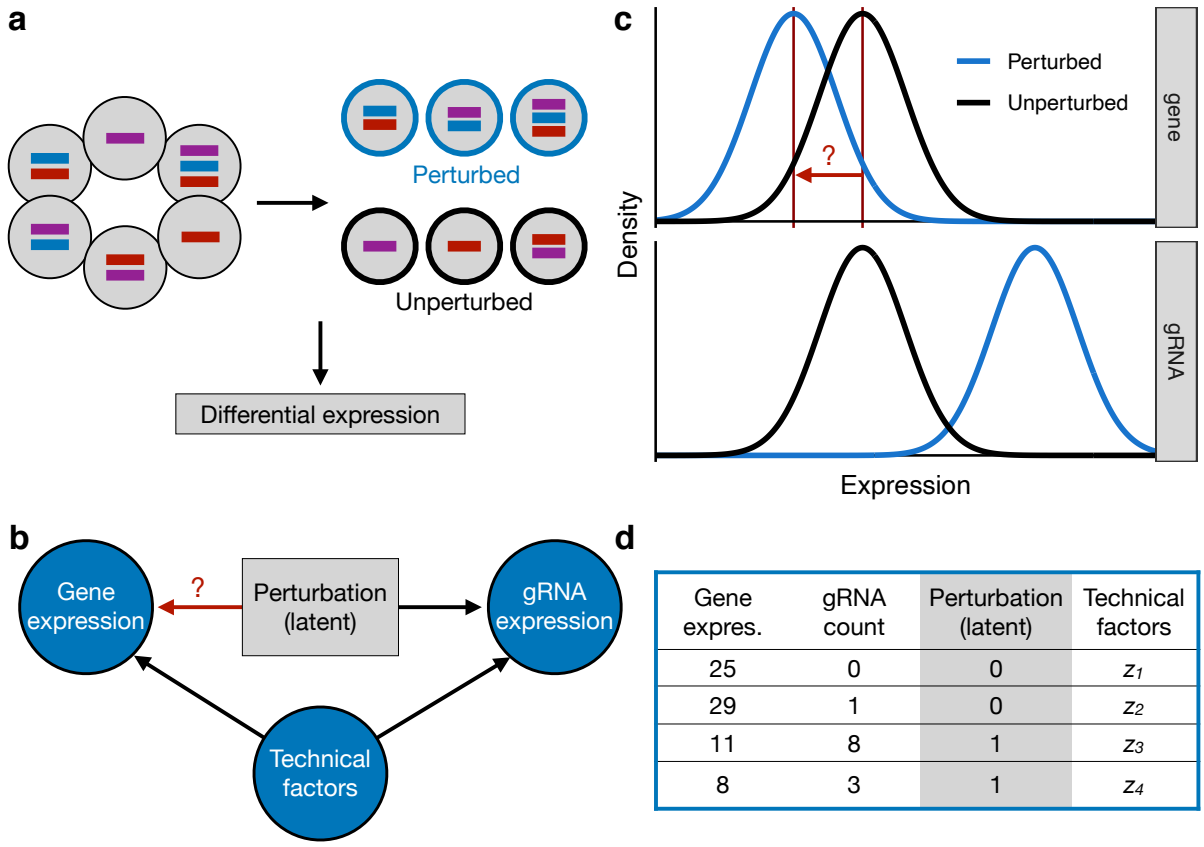
2.2 Assay overview

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 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. Scientists do not know the gene (or genes) through which most noncoding variants exert their effect, 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.



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 confounders, 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 and alignment processes, 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.

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 the measurement of both gene and gRNA expressions and therefore act as confounders (Figure 1b). Third, the gene and gRNA data are sparse, discrete counts. Consequently, classical statistical approaches that assume Gaussianity or homoscedasticity are inapplicable. Finally, and most subtly, sequenced gRNAs sometimes map to cells that have not received a perturbation. This phenomenon, which we "background contamination," 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 is shifted upward 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. in 2019. 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

We study the thresholding method from empirical and theoretical perspectives, highlighting several limitations of the approach. Gasperini and Xie both imputed perturbation assignments onto the cells via thresholding, but they carried out subsequent differential expression analyses in different ways: Gasperini used negative binomial regression, whereas Xie applied nonparametric tests of independence. We study Gasperini's variant of the thresholding method, as it relates more closely to GLM-EIV and (in our view) is the more natural approach.

Let $n \in \mathbb{N}$ be the number of cells 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 $d_i^m \in \mathbb{N}$ be the number of gene transcripts sequenced across *all* genes (the library size or sequencing depth); 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 "d" stand for "mRNA," "gRNA," and "depth," 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 $\hat{p}_i, d_i^m$, and z_i through the following GLM:

$$m_i | (\hat{p}_i, z_i, d_i^m) \sim \text{NB}_{\theta^m}(\mu_i),$$

$$\log(\mu_i) = \beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i + \log\left(d_i^m\right), \tag{1}$$

   where (i) $NB_{\theta^m}(\mu_i)$ 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) $\log(d_i^m)$ is an offset term.

3. Fit a GLM to obtain estimates of the parameters. Compute a p-value and confidence interval for the target of inference β_1^m.

The sequencing depth d_i^m is included 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) yields

$$\mu_i = \exp\left(\beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i\right) d_i^m.$$

The term $\exp\left(\beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i\right)$ is the *fraction* of all transcripts sequenced in the cell produced by the gene under consideration. The target of inference, β_1^m, is the log-transformed fold change in gene expression in response to the perturbation, controlling for the technical factors. Fold change (i.e., $\exp(\beta_1^m)$) is the ratio of mean gene expression in perturbed cells to unperturbed cells: $\exp(\beta_1^m) = 1$ indicates no change in 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 decrease 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

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 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 the thresholding method systematically underestimates effect sizes on the positive control pairs, especially when the threshold is small. 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 obtained 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 versus 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.

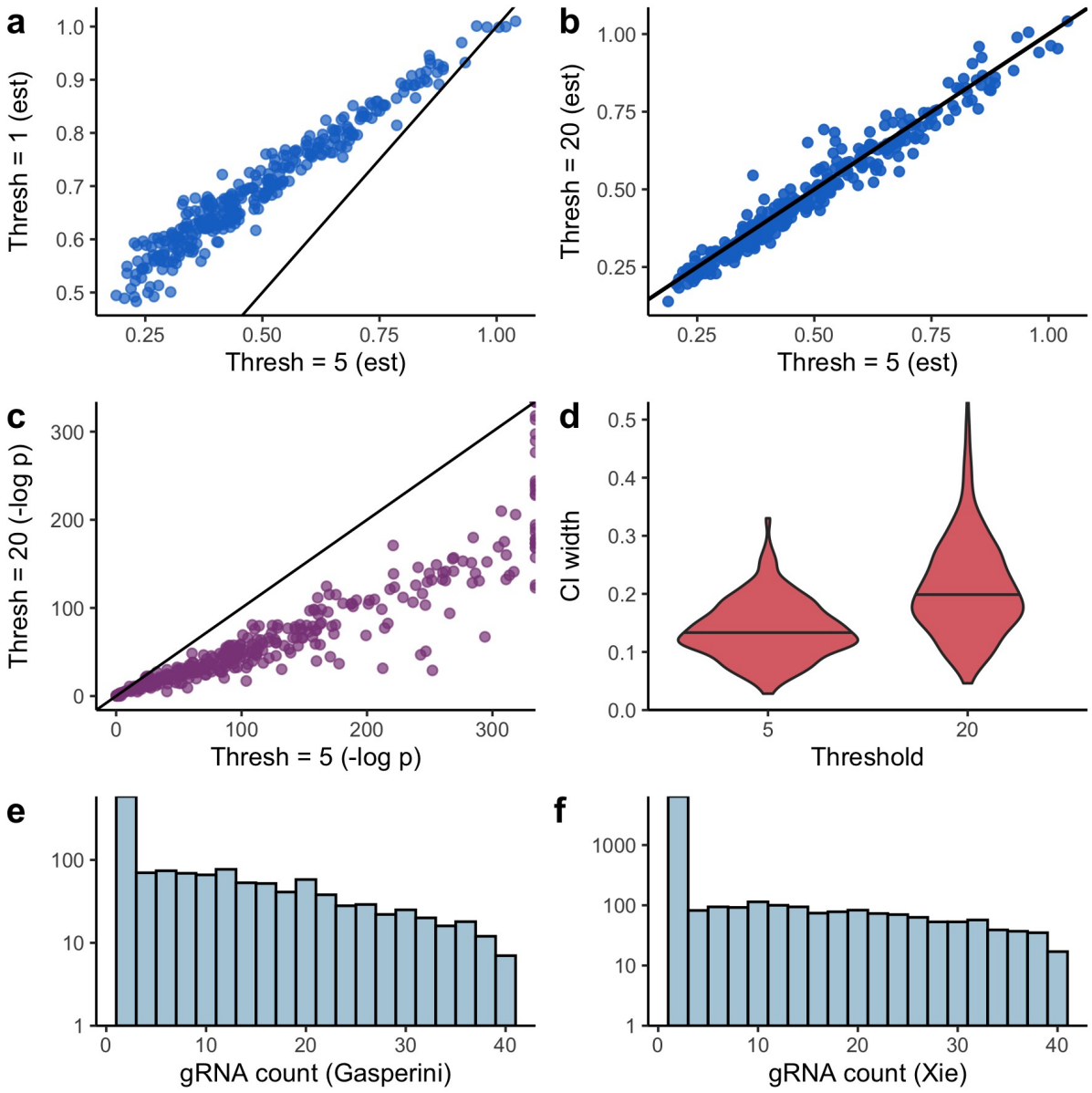


Figure 2: **Empirical challenges of thresholded regression. a-b,** Fold change estimates produced by threshold = 1 versus threshold = 5 (a) and threshold = 20 versus threshold = 5 (b). The selected threshold substantially impacts the results. **c-d,** p-values (c) and CI widths (d) produced by threshold = 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 un-

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

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

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

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.*

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

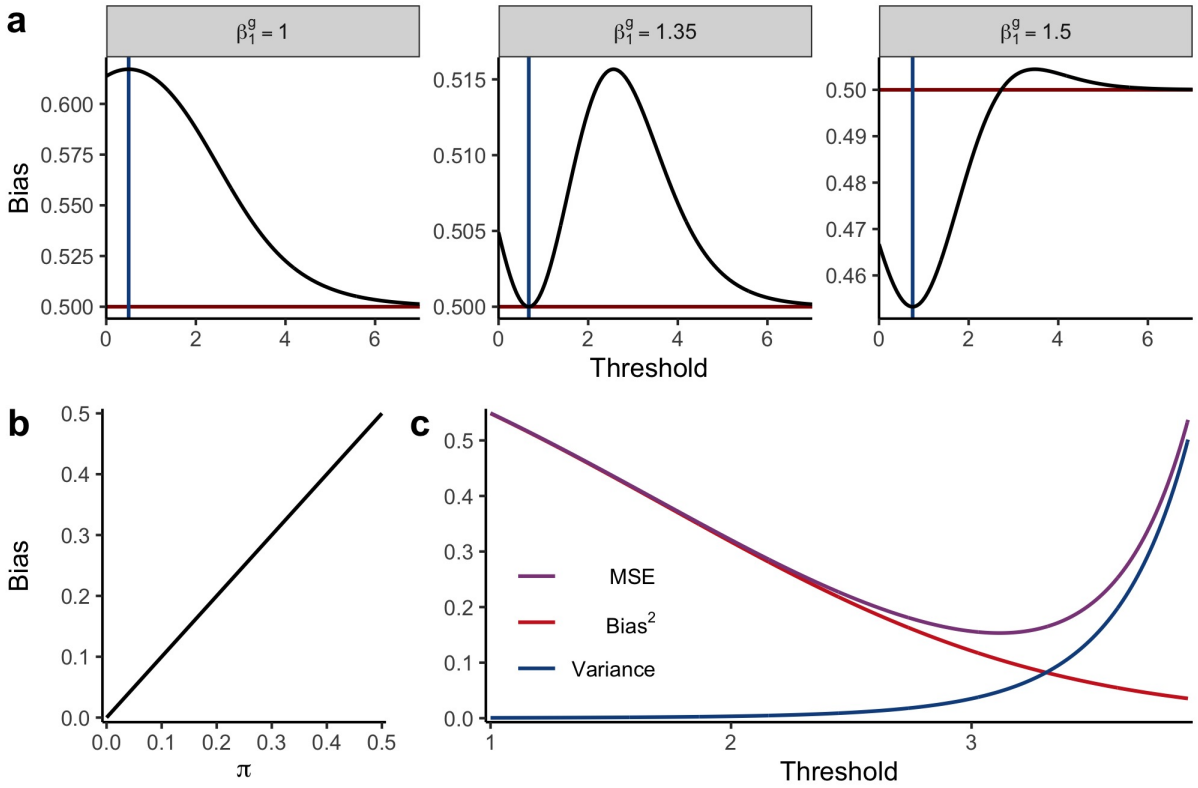


Figure 3: **Theoretical challenges of thresholded regression. a,** Asymptotic relative bias versus threshold for different values of β_1^g. The bias function is highly nonconvex and strictly nonzero. Vertical blue lines, Bayes-optimal decision boundaries. Across all panels, $\beta_0^g = 0$ and $\pi =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

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.

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 GLM-based 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 negative binomial GLMs, 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

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

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

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 $d_i^g \in \mathbb{N}$ 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

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

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

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 the gRNA molecules are transcribed in the cell in the same way as gene transcripts [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.

General 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 $\log(d_i^m)$ and $\log(d_i^g)$, 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 $j \in \{0, 1\}$, let $\tilde{x}_i(j) := [1, j, z_i]^T$ denote the value of $\tilde{x}_i$ that results from setting p_i to j. Next, let $l_i^m(j)$, $\eta_i^m(j)$, and $\mu_i^m(j)$ be the values of l_i^m, η_i^m, and μ_i^m, respectively, that result from setting p_i to j, i.e.,

$$\begin{cases} l_i^m(j) := \langle \tilde{x}_i(j), \beta_m \rangle + o_i^m \\ \eta_i^m(j) := h_m(l_i^m(j)) \\ \mu_i^m(j) = r_m^{-1}(l_i^m(j)). \end{cases}$$

Let the corresponding gRNA quantities $l_i^g(j)$, $\eta_i^g(j)$, and $\mu_i^g(j)$ be defined analogously. Define the observed design matrix $X \in \mathbb{R}^{n \times d-1}$ by

$$X := \begin{bmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix}.$$

Let $\tilde{X} \in \mathbb{R}^{n \times d}$ be the augmented design matrix that results from concatenating the column of (unobserved) p_is to X, i.e.

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

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

$$\begin{bmatrix} \tilde{X}(0) \\ \tilde{X}(1) \end{bmatrix} = [\tilde{X}(0)^T, \tilde{X}(1)^T]^T$$

be the $\mathbb{R}^{2n \times d}$ matrix that results from vertically concatenating $\tilde{X}(0)$ and $\tilde{X}(1)$.

Next, define $m := [m_1, \ldots, m_n]$, and let g, p, o^m, and o^g be defined analogously. Also, let $[m, m]^T \in \mathbb{R}^{2n}$ be the vector that results from concatenating m to itself, i.e.

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

and let $[g, g]^T$, $[o^g, o^g]^T$, and $[o^m, o^m]^T$ be defined similarly.

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

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

The complete-data log-likelihood is

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

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

$$f(m_i, g_i; \theta) = (1-\pi) f_m(m_i; \eta_i^m(0)) f_g(g_i; \eta_i^g(0)) + \pi f_m(m_i; \eta_i^m(1)) f_g(g_i; \eta_i^g(1)). \quad (13)$$

Finally, the log-likelihood is

$$\mathcal{L}(\theta; m, g) = \sum_{i=1}^{n} \log\left[(1-\pi) f_m(m_i; \eta_i^m(0)) f_g(g_i; \eta_i^g(0)) + \pi f_m(m_i; \eta_i^m(1)) f_g(g_i; \eta_i^g(1))\right] \quad (14)$$

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:

$$\begin{cases} y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i \\ x_i = x_i^* + \tau_i, \end{cases} \quad (15)$$

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 (15) is a GLM-EIV model with identity link, Gaussian response, and no covariate terms.

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 $(\beta_m^{(t)}, \beta_g^{(t)}, \pi^{(t)})$ and observed data (m_i, g_i) is

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

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 $\pi^{(t+1)}$ 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 log likelihood (14) converges (see appendix for full details). Our EM algorithm is reminiscent of (but distinct from) that of Ibrahim [11], who also applied weighted GLM solvers to carry out the M step.

Algorithm 1 EM algorithm for GLM-EIV model.

Input: Pilot estimates $\beta_m^{\text{curr}}, \beta_g^{\text{curr}}$, and π^{curr}; data m, g, o^m, o^g, and X; gene expression distribution f_m and link function r_m^{-1}; gRNA expression distribution f_g and link function r_g^{-1}.

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

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

  for $k \in \{g, m\}$ **do**

    Fit a GLM GLM_k with responses $[k, k]^T$, offsets $[o^k, o^k]^T$, weights w, design matrix $[\tilde{X}(0)^T, \tilde{X}(1)^T]^T$, distribution f_k, and link function r_k^{-1}.

    Set β_k^{curr} to the estimated coefficients of GLM_k.

  end for

  Compute marginal likelihood using 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 the EM algorithm is iterative. Therefore, we derive an analytic formula for the asymptotic observed information matrix using Louis's Theorem [27]. 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. The derivation is fairly involved (see appendix); simulation studies confirm the correctness of the results (see Section 5) .

4.3 Statistical accelerations

A downside of the the EM algorithm (Algorithm 1) is that it requires fitting many GLMs. Assuming that we run the algorithm 15 times using randomly-generated pilot estimates (to ensure convergence to the global maximum), and assuming that the algorithm iterates through E and M steps about 10 times per run, we must fit approximately 300 GLMs. (These numbers are based on exploratory applications of the method to real and simulated data.) We devised a strategy (Algorithm 2) to produce a highly accurate pilot estimate $(\pi^{\text{pilot}}, \beta_m^{\text{pilot}}, \beta_g^{\text{pilot}})$ of the true parameters, enabling us to run the algorithm once and converge upon the MLE within a few iterations. The strategy involves layering several statistical "tricks" (some new, some old) on top of one another. We expect these tricks to be independently useful for accelerating other single-cell methods (e.g., SCEPTRE [12]).

The first step (Algorithm 2, lines 2-7) is to obtain good parameter estimates for $[\beta_0^m, \gamma_m]^T$ and $[\beta_0^g, \gamma_g]^T$ using regression. Recall that the underlying gene expression parameter vector β_m is $\beta_m = [\beta_0^m, \beta_1^m, \gamma_m]^T \in \mathbb{R}^d$, where β_0^m is the intercept, β_1^m is the effect of the perturbation, and γ_m^T is the effect of the technical factors. To produce estimates $[\beta_0^m]^{\text{pilot}}$ and $[\gamma_m^T]^{\text{pilot}}$, we regress the gene expressions m onto the technical factors X. The intuition for this procedure is as follows: the probability of perturbation π is very small. Therefore, the true log likelihood is approximately equal to the log likelihood that results from omitting p_i from the model:

$$\begin{aligned}\sum_{i=1}^n f_m(m_i; \eta_i^m) &= \sum_{i=1}^n f_m(m_i; h_m(\beta_0 + \beta_1 p_i + \gamma^T z_i + o_i^m)) \\ &= \underbrace{\sum_{i:p_i=1} f_m(m_i; h_m(\beta_0 + \beta_1 + \gamma^T z_i + o_i^m))}_{\text{few terms}} + \underbrace{\sum_{i:p_i=0} f_m(m_i; h_m(\beta_0 + \gamma^T z_i + o_i^m))}_{\text{many terms}} \\ &\approx \sum_{i=1}^n f_m(m_i; h_m(\beta_0 + \gamma^T z_i + o_i^m)).\end{aligned}$$

We similarly can obtain pilot estimates $[\beta_0^g]^{\text{pilot}}$ and $[\gamma_g^T]^{\text{pilot}}$ by regressing the gRNA counts g onto the technical factors X. We extract the fitted values (on the scale of the linear component) for use in a subsequent step:

$$\hat{f}_i^k = [\beta_0^k]^{\text{pilot}} + \langle [\gamma_k^T]^{\text{pilot}}, z_i \rangle + o_i^k,$$

for $k \in \{m, g\}$.

Next, we obtain estimates $[\beta_1^m]^{\text{pilot}}$, $[\beta_1^g]^{\text{pilot}}$, and π^{pilot} for β_1^m, β_1^g, and π by fitting a "reduced" GLM-EIV (Algorithm 2, lines 8 onward). The log likelihood of the no-intercept, univariate GLM with predictor p_i and offset $\hat{f}_i^m$ is approximately equal to the true log likelihood:

$$\sum_{i=1}^n f_m(m_i; \eta_i^m) = \sum_{i=1}^n f_m(m_i; h_m(\beta_0 + \beta_1 p_i + \gamma^T z_i + o_i^m)) \approx \sum_{i=1}^n f_m(m_i; h_m(\beta_1 p_i + \hat{f}_i^m)).$$

Algorithm 2 Computing pilot parameter estimates.

Input: Data m, g, o^m, o^g, and X; gene expression distribution f_m and link function r_m^{-1}; gRNA expression distribution f_g and link function r_g^{-1}; number of EM starts B.

for $k \in \{m, g\}$ **do**

Fit a GLM GLM_k with responses k, offsets o^k, design matrix X, distribution f_k, and link function r_k^{-1}.

Set $[\beta_0^k]^{\text{pilot}}$ and $[\gamma_k^T]^{\text{pilot}}$ to the fitted coefficients of GLM_k.

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

$\hat{f}_i^k \leftarrow [\beta_0^k]^{\text{pilot}} + \langle [\gamma_k^T]^{\text{pilot}}, z_i \rangle + o_i^k$ ▷ untransformed fitted values

end for

end for

`bestLik` $\leftarrow -\infty$ ▷ Reduced GLM-EIV

for $i \in \{1, \ldots, B\}$ **do**

Randomly generate starting parameters $\pi^{\text{curr}}, [\beta_1^m]^{\text{curr}}, [\beta_1^g]^{\text{curr}}$.

while Not converged **do**

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

$T_i(1) \leftarrow \mathbb{P}(P_i = 1 | M_i = m_i, G_i = g_i, \pi^{\text{curr}}, [\beta_1^g]^{\text{curr}}, [\beta_1^m]^{\text{curr}})$

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

end for

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

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

for $k \in \{g, m\}$ **do**

Fit no-intercept, univariate GLM GLM_k with predictors $[\underbrace{0, \ldots, 0}_{\text{n}}, \underbrace{1, \ldots, 1}_{\text{n}}]$, responses $[k, k]^T$, offsets $[\hat{f}^k, \hat{f}^k]^T$, and weights w.

Set $[\beta_1^k]^{\text{curr}}$ to fitted coefficient of GLM_k.

end for

Compute log likelihood `currLik` using π^{curr},$[\beta_1^m]^{\text{curr}}$, and $[\beta_1^g]^{\text{curr}}$.

end while

if `currLik` $>$ `bestLik` **then**

bestLik $\leftarrow$ currLik

$\pi^{\text{pilot}} \leftarrow \pi^{\text{curr}}$; $[\beta_1^m]^{\text{pilot}} \leftarrow [\beta_1^m]^{\text{curr}}$; $[\beta_1^g]^{\text{pilot}} \leftarrow [\beta_1^g]^{\text{curr}}$

end if

end for

return $(\pi^{\text{pilot}}, [\beta_0^m]^{\text{pilot}}, [\beta_1^m]^{\text{pilot}}, [\gamma_m^T]^{\text{pilot}}, [\beta_0^g]^{\text{pilot}}, [\beta_1^g]^{\text{pilot}}, [\gamma_g^T]^{\text{pilot}})$

Therefore, to estimate β_1^m, β_1^g, and π, we fit a GLM-EIV model with gene expressions m, gRNA counts g, gene offsets $\hat{f}^m := [\hat{f}_1^m, \ldots, \hat{f}_n^m]^T$, gRNA offsets $\hat{f}^g := [\hat{f}_1^g, \ldots, \hat{f}_n^g]^T$, and *no* intercept or covariate terms. Intuitively, we "encode" all information about technical factors, library sizes, and baseline expression levels into $\hat{f}^m$ and $\hat{f}^g$. We run the algorithm $B \approx 15$ times over randomly-selected starting values for β^m, β^g, and π and select the solution with greatest the log likelihood.

The M step of the "reduced" GLM-EIV algorithm requires fitting two no-intercept, univariate GLMs with offsets (Algorithm 2, line 20). We derive analytic formulas for the MLEs of these GLMs in the three most important cases: Gaussian response with identity link, Poisson response with log link, and negative binomial response with log link (see appendix; the latter formula is asymptotically exact, and, to the best of our knowledge, new). Consequently, we do not need to run the relatively slow IRLS procedure to carry out the M step of the "reduced" GLM-EIV algorithm. Overall, the proposed algorithm for obtaining the pilot parameter estimates (Algorithm 2) requires fitting only two GLMs (via IRLS).

4.4 Computing

We developed a computational infrastructure to apply GLM-EIV to large-scale, single-cell CRISPR screen data.[1] The infrastructure leverages Nextflow, a programming language that facilitates building data-intensive pipelines [28], and ondisc[2], an R package that we developed (in a separate project) to facilitate large-scale computing on single-cell data. Nextflow and ondisc together enable the construction of highly portable single-cell pipelines: one can analyze data *out-of-memory* on a laptop or in a *distributed* fashion across tens or hundreds of nodes on a cloud (e.g., Microsoft Azure, Google Cloud) or high-performance cluster.

Leveraging these technologies, we develop a scalable and efficient pipeline for GLM-EIV (Algorithm 3). First, we run a round of "precomputations" on all d_g genes and d_p perturbations. The precomputations involve regressing the gene expressions (or gRNA counts) onto the technical factors, thereby "factoring out" lines 2-7 from Algorithm 2. Next, we run differential expression analyses on the full set of gene-perturbation pairs; for a given pair, this amounts to obtaining the complete set of pilot parameters (Algorithm 2, lines 8 onward), fitting the GLM-EIV model (Algorithm 1), and performing inference. The three loops in Algorithm 3 are embarassingly parallel and therefore can be massively parallelized.

4.5 Zero-inflated model

The GLM-EIV model assumes that unperturbed cells exhibit a nonzero "background read" gRNA count distribution due to errors in the sequencing and alignment processes. This assumption very likely is appropriate for most current single-cell CRISPR screen

[1] Pipeline available at `github.com/timothy-barry/glmeiv-pipeline`.

[2] Package available at `github.com/timothy-barry/ondisc`; preprint forthcoming.

Algorithm 3 Applying GLM-EIV at scale.

$G \leftarrow \{\text{gene}_1, \ldots, \text{gene}_{d_g}\}$; $P \leftarrow \{\text{perturbation}_1, \ldots, \text{perturbation}_{d_p}\}$
for gene $\in G$ **do**
  Run precomputation (Algorithm 2, lines 2-7) on gene.
  Save $\hat{f}^m$, $[\beta_0^m]^{\text{pilot}}$ and $[\gamma_m^T]^{\text{pilot}}$.
end for
for perturbation $\in P$ **do**
  Run precomputation (Algorithm 2, lines 2-7) on perturbation.
  Save $\hat{f}^g$, $[\beta_0^g]^{\text{pilot}}$ and $[\gamma_g^T]^{\text{pilot}}$.
end for
for (gene, perturbation) $\in G \times P$ **do**
  Load $\hat{f}^m, \hat{f}^g$, $[\beta_0^m]^{\text{pilot}}$ $[\gamma_m^T]^{\text{pilot}}$, $[\beta_0^g]^{\text{pilot}}$ and $[\gamma_g^T]^{\text{pilot}}$.
  Compute $[\beta_1^m]^{\text{pilot}}, [\beta_1^g]^{\text{pilot}}, \pi^{\text{pilot}}$ (Algorithm 2, lines 8 onward).
  Run GLM-EIV using the pilot parameters (Algorithm 1).
end for

datasets [21]. However, background contamination may be negligible for some datasets, in which case the gRNA counts of unperturbed cells should be modeled as point masses at zero. We introduce such an extension – which we call the "zero-inflated GLM-EIV model" – in Appendix C. We additionally derive efficient methods for estimation and inference in the zero-inflated model. Given the ubiquity of background contamination in current single-cell CRISPR screen datasets, we focus our attention on the standard GLM-EIV model for the remainder of the main text.

5 Simulation studies

We conducted a simulation study to compare the empirical performance of GLM-EIV to that of the thresholding method. We generated data on $n = 150,000$ cells from the GLM-EIV model (Section 4.1), setting the the target of inference β_1^m to $\log(0.25)$ and the probability of perturbation π to 0.02. $\beta_1^m = \log(0.25)$ represents a decrease in gene expression by a factor of 4, which is a fairly large effect size on the order of what we might observe for a positive control pair. We included "sequencing batch" (modeled as a Bernoulli-distributed variable) as a covariate and sequencing depth (modeled as a Poisson-distributed variable) as an offset. We varied the log-fold change in gRNA expression, β_1^g, over a grid on the interval $[\log(1), \log(4)]$; β_1^g controls problem difficulty, with higher values corresponding to easier problems. Finally, we generated the gene expression and gRNA count data from two response distributions: Poisson and negative binomial (size parameter fixed at $\theta = 20$ for the latter). For each parameter setting (defined by a β_1^g-distribution pair), we synthesized $n_{\text{sim}} = 500$ i.i.d. datasets. Section E presents additional simulation results on Gaussian response distributions.

We applied three methods to the simulated data: "vanilla" GLM-EIV (Algorithm 1), accelerated GLM-EIV (Algorithm 2), and thresholded regression. We used the Bayes-

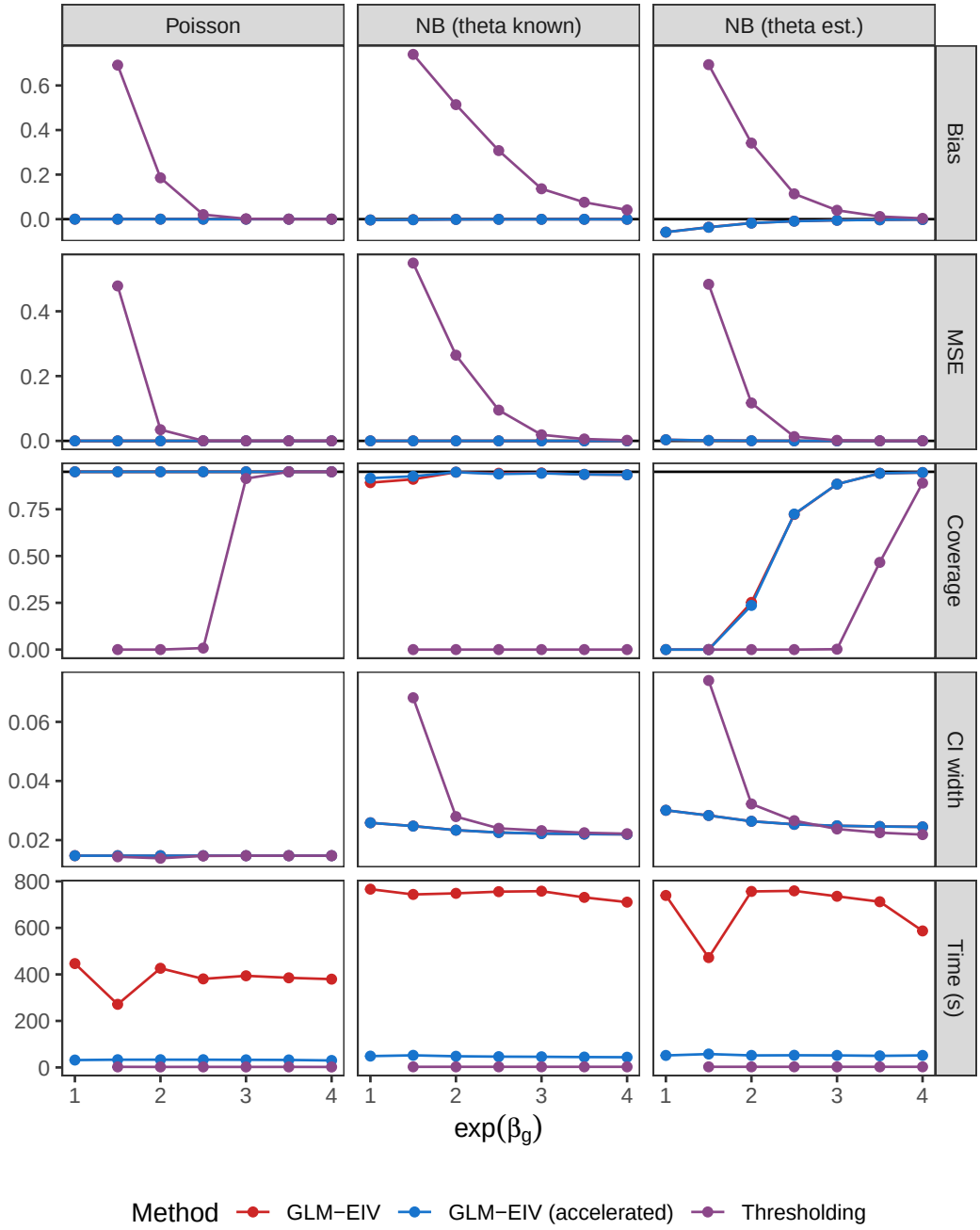


Figure 4: **Simulation study**. Columns correspond to distributions (Poisson, NB with known θ, NB with estimated θ), and rows correspond to metrics (bias, MSE, coverage, CI width, and time). Methods are shown in different colors; GLM-EIV (red) is masked by accelerated GLM-EIV (blue) in several panels. GLM-EIV demonstrated superior statistical performance to the thresholding method on all metrics (rows 1-4). Accelerated GLM-EIV had substantially lower computational cost than "vanilla" GLM-EIV (bottom row) despite demonstrating identical statistical performance (rows 1-4).

optimal decision boundary for classification as the threshold for the thresholding method. We ran all methods on the negative binomial data twice: once treating the size parameter θ as a known constant and once treating θ as unknown. In the latter case we used the `glm.nb` function from the `MASS` package to estimate θ [29]. We display the results of the simulation study in Figure 4. Columns correspond to distributions (i.e., Poisson, negative binomial with known θ, and negative binomial with unknown θ), rows correspond to performance metrics (i.e., bias, mean squared error, CI coverage rate (nominal rate 95%), CI width, and method execution time). The problem difficulty parameter β_1^g is plotted on the horizontal axis, and the methods are depicted in different colors (GLM-EIV = red, accelerated GLM-EIV = blue, and thresholding = purple; GLM-EIV masked by accelerated GLM-EIV in several panels).

First, we observed that GLM-EIV dominated thresholded regression on all statistical metrics: GLM-EIV exhibited lower bias (row 1) and mean squared error (row 2) than thresholded regression; additionally, GLM-EIV had superior confidence interval coverage (row 3) despite having produced generally narrower confidence intervals (row 4). Intuitively, GLM-EIV outperformed the thresholding method because (i) GLM-EIV leveraged information from *both* modalities (rather than the gRNA modality alone) to assign perturbation identities to cells, and (ii) GLM-EIV produced soft rather than hard assignments, capturing the inherent uncertainty in whether a perturbation occurred. Next, we found that accelerated GLM-EIV performed as well as "vanilla" GLM-EIV on the statistical metrics (rows 1-4) despite have substantially lower computational cost (bottom row). In fact, the execution time of accelerated GLM-EIV was approximately within an order of magnitude of that of the thresholding method (bottom row).

Interestingly, thresholded regression exhibited better confidence interval coverage under estimated θ than under known θ (row 3). Estimating θ leads to slight inflation bias (i.e., overestimating the true effect size), whereas, as we showed previously, thresholding leads to attenuation bias (i.e., underestimating the true effect size). These phenomena partially cancel, yielding less biased estimates. We note that GLM-EIV and the thresholding method in principal are compatible with *any* θ estimation procedure, including those based on more sophisticated techniques, such as regularization [24]. We defer rigorous investigation of the impact of different θ estimation strategies on these methods to future work.

6 Data analysis

Leveraging our computational infrastructure (Section 4.4), we applied GLM-EIV to analyze the entire Gasperini and Xie datasets (Figure 5). Our theoretical analyses (Section 3.2) and simulation studies (Section 5) revealed that the Bayes-optimal decision boundary is a good choice for the threshold when the gRNA count distribution is well separated. Therefore, under the assumption that the effect of the perturbation on gRNA expression is similar across pairs, we used the fitted GLM-EIV models to approximate the Bayes-optimal decision boundary on both datasets, yielding a boundary of 3 for Gasperini and

7 for Xie. We then applied the thresholding method to the data, setting the threshold to the Bayes-optimal decision boundary estimated using GLM-EIV.

We compared GLM-EIV to thresholded regression on the real data, focusing specifically on the negative control pairs (i.e., gene-perturbation pairs that, by design, are expected to exhibit a fold change of 1, or no association). We found that GLM-EIV and the thresholding method produced similar results (Figure 5a-b): estimates, CI coverage rates, and CI widths were concordant. We note that GLM-EIV produced outlier estimates (likely due to non-global EM convergence) on a small ($< 2\%$ on Gasperini, $< 0.05\%$ on Xie) number of pairs consisting of a handful of genes (not plotted). The estimated effect of the perturbation on gene expression $\exp(\hat{\beta}_1^g)$ was unexpectedly large (95% CI $[XX, XX]$ and $[300, 316]$ on Gasperini and Xie data, respectively). We concluded that the datasets lay in an "easy" region of the parameter space, making thresholding a tenable strategy (provided the threshold is selected well).

To explore the settings in which GLM-EIV and the thresholding method diverge, we artificially increased the problem difficulty by generating partially-synthetic datasets. To this end, we defined a parameter called "excess background contamination" that controls the extent to which the gRNA count distribution is separated. When excess background contamination is zero, we sample gRNA counts directly from the fitted GLM-EIV model. As excess background contamination increases, we sample gRNA counts from a slightly modified version of the fitted model in which the mean gRNA expression of *unperturbed* cells is increased while the mean gRNA expression of *perturbed* cells is held constant. Finally, when excess background contamination is 1 (the maximum value), we set the mean gRNA expression of unperturbed cells equal to that of perturbed cells and sample gRNA counts from the resulting model. We hold fixed the real-data gene expressions, library sizes, covariates, and fitted perturbation probabilities in all settings.

We generated partially-synthetic data in the above manner for each of the 322 positive control pairs in the Gasperini dataset, varying excess background contamination over the interval $[0, 0.4]$. We then applied GLM-EIV and the thresholding method to analyze the data. We present results on two example pairs (the pair containing gene *LRIF1* and the pair containing gene *NDUFA2*) in Figures 5c-d. We observed that the estimate produced by the methods on the raw data (depicted as a horizontal black line) coincided almost exactly with the estimate produced by the methods on the partially-synthetic data generated by setting excess background contamination to zero (result replicated across nearly all pairs; average relative difference 0.003). This suggested that the fitted models accurately tracked the gRNA count distributions, giving us confidence in our procedure for generating realistic, partially-synthetic datasets. Next, we observed that as excess background contamination increased, the performance of thresholded regression degraded considerably while that of GLM-EIV remained stable on the genes *LRIF1* and *NDUFA2*.

We generalized the above analysis to the entire set of positive control pairs. First, for each pair, we computed the "relative estimate change" (REC) as a function of excess background contamination, defined as the relative difference between the estimate at a given level of excess contamination and zero excess contamination (Figure 5d). Next, we computed the median REC across all positive control pairs (Figure 5e; bands indicate

pointwise 95% CIs for the population median REC). As excess background contamination increased, thresholded regression exhibited severe attenuation bias (as reflected by large median REC values); GLM-EIV, by contrast, remained mostly stable. Finally, taking the estimate obtained on the raw data as "ground truth," we computed the CI coverage rate across pairs as a function of excess contamination (Figure 5f; bands indicate 95% pointwise CIs for population coverage rate). GLM-EIV exhibited significantly higher CI coverage rates than thresholded regression as the data became increasingly contaminated.

7 Discussion

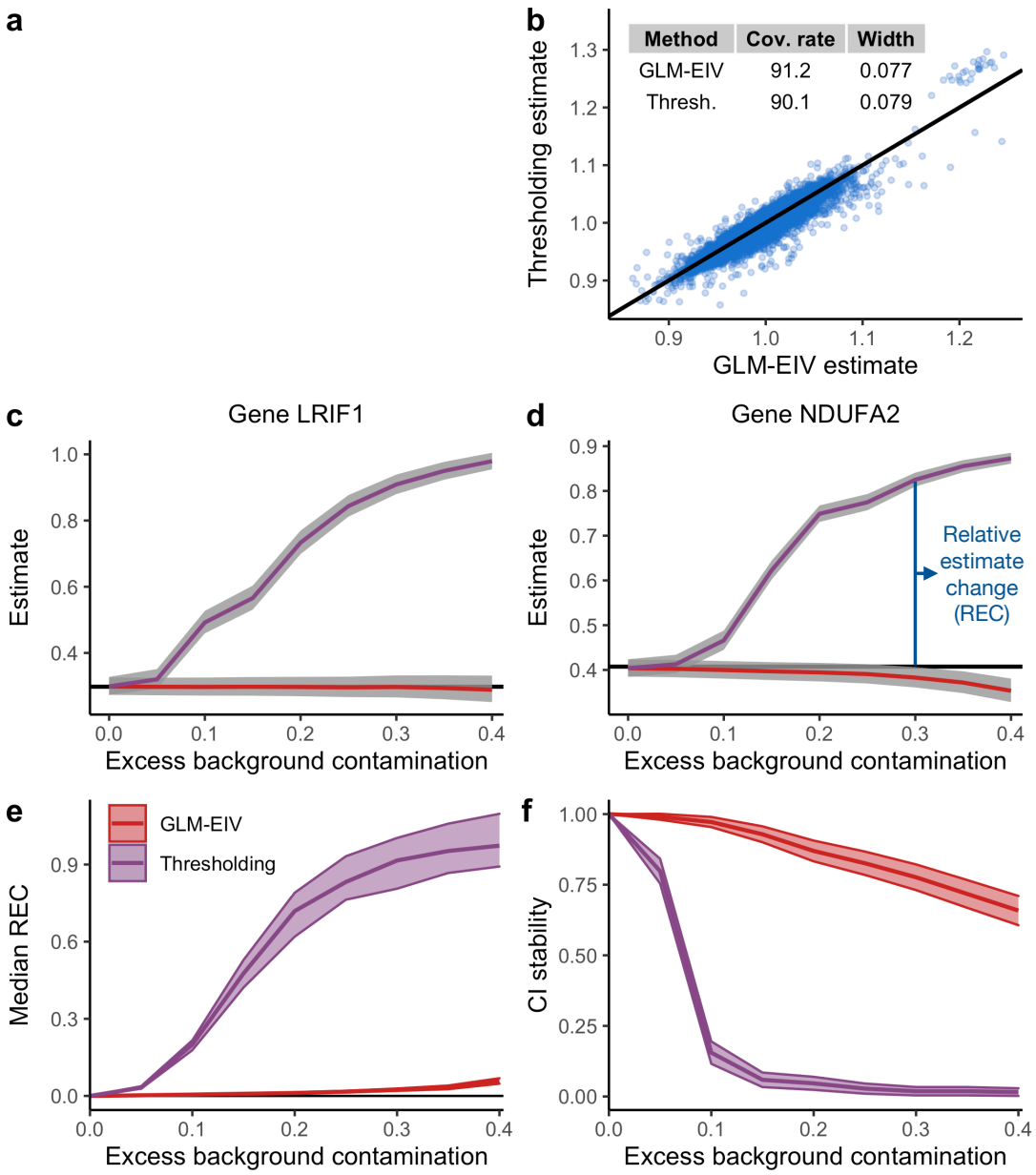


Figure 5: **Data analysis**. **a-b**, Estimates for fold change produced by GLM-EIV and thresholded regression on Gasperini (**a**) and Xie (**b**) negative control pairs. **c-d**, Estimates produced by GLM-EIV and thresholded regression on two positive control pairs – *LRIF1* (**a**) and *NDUFA2* (**b**) – plotted as a function of excess background contamination. Grey bands, 95% CIs for the target of inference outputted by the methods. **e-f**, Median relative estimate change (REC; **e**) and confidence interval coverage rate (**f**) across *all* 322 positive control pairs, plotted as a function of excess background contamination. Colored bands, pointwise 95% CIs for population quantities. Panels (**c-f**) together illustrate that GLM-EIV demonstrated greater stability than thresholded regression as background contamination increased.

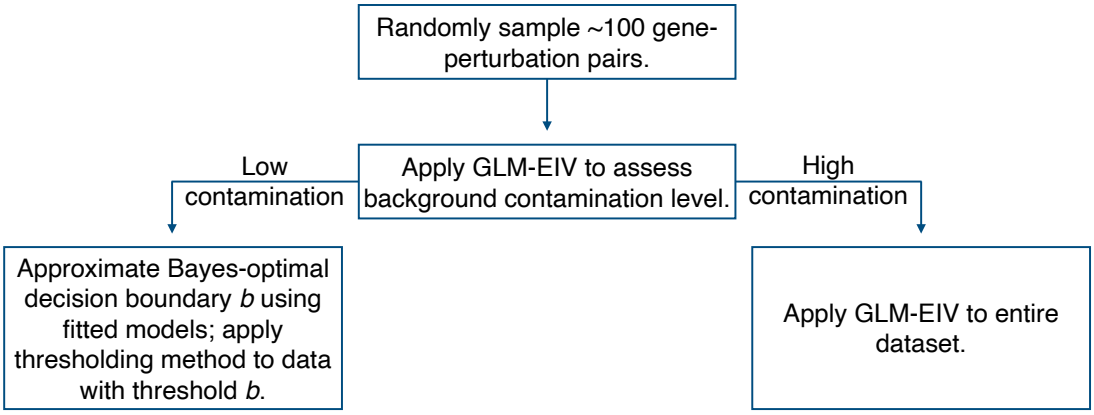


Figure 6: **Use of GLM-EIV in practice**. The decision tree above illustrates how we anticipate GLM-EIV could be used in practice. First, apply GLM-EIV to a set of randomly-sampled gene-perturbation pairs to assess background contamination level (positive control pairs work best for this purpose). If GLM-EIV indicates that background contamination is high (e.g., $\exp(\beta_1^g) \lessapprox 10$), apply GLM-EIV to analyze the entire dataset; otherwise, approximate the Bayes-optimal decision boundary using the fitted GLM-EIV models. Next, apply a thresholding method (e.g., SCEPTRE or thresholded negative binomial regression) to analyze the data, setting the threshold to the estimated Bayes-optimal decision boundary.

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. *Society*, 44(2):226–233, 1982.

[28] Paolo DI Tommaso, Maria Chatzou, Evan W. Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. Nextflow enables reproducible computational workflows. *Nature Biotechnology*, 35(4):316–319, 2017.

[29] Brian Ripley, Bill Venables, Douglas M Bates, Kurt Hornik, Albrecht Gebhardt, David Firth, and Maintainer Brian Ripley. Package 'mass'. *Cran r*, 538:113–120, 2013.

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

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

$$g(\beta_1^g, c, \pi, \beta_0^g) = (1-\pi)\left(\Phi(\beta_0^g + \beta_1^g - c)\right) - (1-\pi)\left(\Phi(\beta_0^g - c)\right) \tag{16}$$

and

$$h(\beta_1^g, c, \pi, \beta_0^g) = \left[(1-\pi)\left(\Phi(\beta_0^g - c)\right) + \pi\left(\Phi(\beta_0^g + \beta_1^g - c)\right)\right] \cdot \left[(1-\pi)\left(\Phi(c - \beta_0^g)\right) + \pi\left(\Phi(c - \beta_0^g - \beta_1^g)\right)\right]. \tag{17}$$

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* [30]). 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,

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

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 $\hat{p}_i$:

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

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

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

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

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

and so

$$1 - \mathbb{E}\left[\hat{p}_i\right] = (1-\pi) + \pi - \mathbb{E}[\hat{p}_i] = (1-\pi)\left(1 - \bar{\Phi}(c - \beta_0^g)\right) + \pi\left(1 - \bar{\Phi}(c - \beta_0^g - \beta_1^g)\right)$$
$$= (1-\pi)\Phi(c - \beta_0^g) + \pi\Phi(c - \beta_0^g - \beta_1^g). \quad (19)$$

Next,

$$\omega = \Phi(\beta_1^g + \beta_0^g - c), \quad (20)$$

and so

$$\omega - \mathbb{E}[\hat{p}_i] = \Phi(\beta_1^g + \beta_0^g - c) - (1-\pi)\Phi(\beta_0^g - c) - \pi\Phi(\beta_0^g + \beta_1^g - c)$$
$$(1-\pi)\Phi(\beta_1^g + \beta_0^g - c) - (1-\pi)\Phi(\beta_0^g - c). \quad (21)$$

Combining (18, 19, 20, 21), we find that

$$\gamma(\beta_1^g, c, \pi) = \frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])}$$
$$= \frac{\pi\left[(1-\pi)\Phi(\beta_0^g + \beta_1^g - c) - (1-\pi)\Phi(\beta_0^g - c)\right]}{\left[(1-\pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c)\right]\left[(1-\pi)\Phi(c - \beta_0^g) + \pi\Phi(c - \beta_0^g - \beta_1^g)\right]}. \quad (22)$$

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

$$\gamma(\beta_1^g, c, 1/2)$$
$$= \frac{\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)}{\left[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)\right]\left[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)\right]}. \quad (23)$$

Recalling the definitions of g (16) and h (17), 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 (23) is identical to

$$\gamma(\beta_1^g, c, 1/2) = \frac{(4)(1/2)g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)} = \frac{2g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)}, \quad (24)$$

i.e., the numerator and denominator of (24) coincide with those of (23). We sometimes will use the notation $2 \cdot g$ and $4 \cdot h$ to refer to the numerator and denominator of (23), 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

$$D_2 g(\beta_1^g, c, \pi) = -(1-\pi) f(\beta_0^g + \beta_1^g - c) + (1-\pi) f(\beta_0^g - c)$$
$$= (1-\pi) f(c - \beta_0^g) - (1-\pi) f(c - \beta_0^g - \beta_1^g). \quad (25)$$

Second, we have that

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

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 (25) and (26) that

$$\begin{aligned} \lim_{c\to\infty} \gamma(\beta_1^g, c, \pi) &= \lim_{c\to\infty} \frac{\pi D_2 g(\beta_1^g, c, \pi)}{D_2 h(\beta_1^g, c, \pi)} \\ &= \lim_{c\to\infty} \Bigg\{ \frac{(1-\pi) f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)}{\pi(1-\pi) f(c - \beta_0^g) - \pi(1-\pi) f(c - \beta_0^g - \beta_1^g)} \\ &\quad \cdot \Big[(1-\pi)\Phi(\beta_0^g - c) + \pi \Phi(\beta_0^g + \beta_1^g - c) \\ &\quad\quad - (1-\pi)\Phi(c - \beta_0^g) - \pi \Phi(c - \beta_0^g - \beta_1^g)\Big] \Bigg\}^{-1}. \quad (27) \end{aligned}$$

We evaluate the two terms in the product (27) separately. Dividing by $f(c - \beta_0^g - \beta_1^g) > 0$, we see that

$$\frac{(1-\pi) f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)}{\pi(1-\pi) f(c - \beta_0^g) - \pi(1-\pi) f(c - \beta_0^g - \beta_1^g)} = \frac{\frac{(1-\pi) f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)} + \pi}{\frac{\pi(1-\pi) f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)} - \pi(1-\pi)}. \quad (28)$$

To evaluate the limit of (28), we first evaluate the limit of

$$\frac{f(c - \beta_0^g)}{f(c - \beta_0^g - \beta_1^g)} = \frac{\exp\left[-(1/2)(c - \beta_0^g)^2\right]}{\exp\left[-(1/2)(c - \beta_0^g - \beta_1^g)^2\right]}$$

$$= \frac{\exp[-(1/2)(c^2 - 2c\beta_0^g + (\beta_0^g)^2)]}{\exp\left[-(1/2)(c^2 - 2c\beta_0^g - 2c\beta_1^g + (\beta_0^g)^2 + 2(\beta_0^g\beta_1^g) + (\beta_1^g)^2)\right]}$$
$$= \exp\big[-c^2/2 + c\beta_0^g - (\beta_0^g)^2/2$$
$$+ c^2/2 - c\beta_0^g - c\beta_1^g + (\beta_0^g)^2/2 + \beta_0^g\beta_1^g + (\beta_1^g)^2/2\big]$$
$$= \exp[-c\beta_1^g + \beta_0^g\beta_1^g + (\beta_1^g)^2/2] = \exp[\beta_0^g\beta_1^g + (\beta_1^g)^2/2]\exp[-c\beta_1^g]. \quad (29)$$

Taking the limit in (29), 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 (28):

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

Next, we compute the limit of the other term in the product (27):

$$\lim_{c\to\infty} \Big[(1-\pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c)$$
$$- (1-\pi)\Phi(c-\beta_0^g) - \pi\Phi(c-\beta_0^g-\beta_1^g)\Big] = -(1-\pi) - \pi = -1. \quad (30)$$

Combining (28) and (30), the limit (27) 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 (24), the quotient rule implies that

$$D_2\gamma(\beta_1^g, c, 1/2) = \frac{D_2[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_2[4h(\beta_1^g, c, 1/2)]}{[4h(\beta_1^g, c, \pi)]^2}. \quad (31)$$

We have by (25) that

$$D_2[2g(\beta_1^g, c_{\text{bayes}}, 1/2)] = f(\beta_1^g/2) - f(-\beta_1^g/2) = f(\beta_1^g/2) - f(\beta_1^g/2) = 0. \tag{32}$$

Similarly, we have by (26) that

$$D_2[4h(\beta_1^g, c_{\text{bayes}}, \pi)] = [f(\beta_1^g/2) + f(-\beta_1^g/2)] \cdot [\Phi(-\beta_1^g/2) + \Phi(\beta_1^g/2) - \Phi(\beta_1^g/2) - \Phi(-\beta_1^g/2)] = 0. \tag{33}$$

Plugging in (33) and (32) to (31), 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:*

$$\begin{aligned}[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)] \\ \cdot [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c) + \Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ \geq [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)]\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]\,. \quad (34)\end{aligned}$$

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 (34) is negative.

Turning our attention of the left-hand side of (34), we see that

$$\Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1 - \Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1. \quad (35)$$

Additionally, $\Phi(\beta_0^g - c) < 1$ and $\Phi(c - \beta_0^g) > 0$. Combining these facts with (35), 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 (34) 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 (34) is greater than the first term on the RHS of (34), 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

$$\beta_1^g + \beta_0^g - c \geq \beta_0^g - c \implies \Phi(\beta_1^g + \beta_0^g - c) - \Phi(\beta_0^g - c) \geq 0. \quad (36)$$

Adding $\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)$ to both sides of (36) yields

$$\begin{aligned}\Phi(\beta_1^g + \beta_0^g - c) - \Phi(\beta_0^g - c) + \Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g) \\ \geq \Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g).\end{aligned}$$

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*

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

and

$$\begin{aligned} D_1[4h(\beta_1^g, c, 1/2)] = f(\beta_0^g + \beta_1^g - c)\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ - f(\beta_0^g + \beta_1^g - c)\,[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)]. \end{aligned} \tag{38}$$

Proof: Apply FTC and product rule. □

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

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

from both sides of (34) and multiplying by $f(\beta_0^g + \beta_1^g - c) > 0$ yields

$$\begin{aligned} & f(\beta_0^g + \beta_1^g - c)[\Phi(\beta_0^g - c) + \Phi\,(\beta_0^g + \beta_1^g - c)]\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ & \quad \geq f(\beta_0^g + \beta_1^g - c)\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)][\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)] \\ & \quad - f(\beta_0^g + \beta_1^g - c)\,[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)][\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)]. \end{aligned} \tag{39}$$

Next, recall that

$$2g(\beta_1^g, c, 1/2) = \Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c). \tag{40}$$

and

$$4h(\beta_1^g, c, 1/2) = [\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)]\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]. \tag{41}$$

Substituting (37, 38, 40, 41) into (39) 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

$$D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)] \geq 0. \tag{42}$$

The quotient rule implies that

$$\begin{aligned} & D_1\gamma(\beta_1^g, c, 1/2) \\ & \qquad = \frac{D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)]}{[4h(\beta_1^g, c, 1/2)]^2}. \end{aligned} \tag{43}$$

We conclude by (42) and (43) 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,

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

Similarly,

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

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,

$$\sqrt{n}(\hat{\beta}_m - l) = \frac{(1/\sqrt{n}) \sum_{i=1}^n \hat{p}_i (m_i - l\hat{p}_i)}{(1/n) \sum_{i=1}^n \hat{p}_i^2}. \tag{44}$$

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 $\hat{p}_i$:

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

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 $\hat{p}_i m_i$:

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

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

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

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

$$\mathbb{E}\left[\hat{p}_i(m_i - l\hat{p}_i)\right] = \mathbb{E}[\hat{p}_i m_i] - l\mathbb{E}[\hat{p}_i] \\ = \beta_m\omega\pi - \left(\frac{\beta_m\omega\pi}{\zeta(1-\pi) + \omega\pi}\right)\left[\zeta(1-\pi) + \omega\pi\right] = 0. \quad (45)$$

Additionally,

$$\mathbb{V}\left[\hat{p}_i(m_i - l\hat{p}_i)\right] = \mathbb{E}\left[\hat{p}_i^2(m_i - l\hat{p}_i)^2\right] - \left(\mathbb{E}\left[\hat{p}_i(m_i - l\hat{p}_i)\right]\right)^2 \\ = \mathbb{E}\left[\hat{p}_i m_i^2\right] - 2l\mathbb{E}[m_i\hat{p}_i] + l^2\mathbb{E}[\hat{p}_i] = \beta_m^2\omega\pi + \mathbb{E}[\hat{p}_i] - 2l\beta_m\omega\pi + l^2\mathbb{E}[\hat{p}_i] \\ = \beta_m\omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1 + l^2). \quad (46)$$

Therefore, by CLT, (45), and (46),

$$(1/\sqrt{n})\sum_{i=1}^{n}\hat{p}_i(m_i - l\hat{p}_i) \xrightarrow{d} N\left(0, \beta_m\omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1 + l^2)\right). \quad (47)$$

Next, by weak LLN,

$$(1/n)\sum_{i=1}^{n}\hat{p}_i^2 = (1/n)\sum_{i=1}^{n}\hat{p}_i \xrightarrow{P} \mathbb{E}[\hat{p}_i]. \quad (48)$$

Finally, by (44), (47), (48), 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

We estimate the parameters of the GLM-EIV model using an EM algorithm.

E step

The E step entails computing the membership probability of each cell. Let $\theta^{(t)} = (\beta_m^{(t)}, \beta_g^{(t)}, \pi^{(t)})$ be the parameter estimate at the t-th iteration of the algorithm. For $k \in \{0, 1\}$, let $[\eta_i^m(k)]^{(t)}$ be the ith canonical parameter at the t-th iteration of the algorithm of the gene expression distribution that results from setting p_i to k, i.e.

$$[\eta_i^m(k)]^{(t)} := h_m\left(\langle \tilde{x}_i(k), \beta_m^{(t)} \rangle + o_i^m\right).$$

Similarly, let $[\eta_i^g(k)]^{(t)}$ be defined by

$$[\eta_i^g(k)]^{(t)} := h_g\left(\langle \tilde{x}_i(k), \beta_g^{(t)} \rangle + o_i^g\right).$$

Next, for $k \in \{0, 1\}$, define $\alpha_i^{(t)}(k)$ by

$$\begin{aligned} \alpha_i^{(t)}(k) &:= \mathbb{P}\left(M_i = m_i, G_i = g_i | P_i = k, \theta^{(t)}\right) \\ &= \mathbb{P}\left(M_i = m_i | P_i = k, \theta^{(t)}\right) \mathbb{P}\left(G_i = g_i | P_i = k, \theta^{(t)}\right) \text{ (because } G_i \perp\!\!\!\perp M_i | P_i) \\ &= f_m\left(m_i; [\eta_i^m(k)]^{(t)}\right) f_g\left(g_i; [\eta_i^g(k)]^{(t)}\right). \end{aligned}$$

Finally, let

$$\begin{cases} \pi^{(t)}(1) := \pi^{(t)} = \mathbb{P}\left(P_i = 1 | \theta^{(t)}\right) \\ \pi^{(t)}(0) := 1 - \pi^{(t)} = \mathbb{P}\left(P_i = 0 | \theta^{(t)}\right). \end{cases}$$

The ith membership probability $T_i^{(t)}(1)$ is

$$T_i^{(t)}(1) = \mathbb{P}(P_i = 1 | M_i = m_i, G_i = g_i, \theta^{(t)}) = \frac{\pi^{(t)}(1)\alpha_i^{(t)}(1)}{\sum_{k=0}^{1} \pi^{(t)}(k)\alpha_i^{(t)}(k)} \text{ (by Bayes rule)}$$
$$= \frac{1}{\frac{\pi^{(t)}(0)\alpha_i(0)}{\pi^{(t)}(1)\alpha_i(1)} + 1} = \frac{1}{\exp\left(\log\left(\frac{\pi^{(t)}(0)\alpha_i(0)}{\pi^{(t)}(1)\alpha_i(1)}\right)\right) + 1} = \frac{1}{\exp\left(q_i^{(t)}\right) + 1}, \quad (49)$$

where we set

$$q_i^{(t)} := \log\left(\frac{\pi^{(t)}(0)\alpha_i^{(t)}(0)}{\pi^{(t)}(1)\alpha_i^{(t)}(1)}\right). \quad (50)$$

Next, we have that

$$q_i^{(t)} = \log\left[\pi^{(t)}(0)\right] + \log\left[f_m\left(m_i; [\eta_i^m(0)]^{(t)}\right)\right] + \log\left[f_g\left(g_i; [\eta_i^g(0)]^{(t)}\right)\right]$$
$$- \log\left[\pi^{(t)}(1)\right] - \log\left[f_m\left(m_i; [\eta_i^m(1)]^{(t)}\right)\right] - \log\left[f_g\left(g_i; [\eta_i^g(1)]^{(t)}\right)\right],$$

which we can compute. We therefore conclude that

$$T_i^{(t)} = \frac{1}{\exp\left(q_i^{(t)}\right) + 1}.$$

M step

Recall that the log-likelihood (12) of the GLM-EIV model is

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

Define $Q(\theta|\theta^{(t)}) = \mathbb{E}_{\left(P|M=m,G=g,\theta^{(t)}\right)}\left[\mathcal{L}(\theta; m, g, p)\right]$. We have that

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n} \left[T_i^{(t)}(1)\log(\pi) + T_i^{(t)}(0)\log(1 - \pi)\right]$$
$$+ \sum_{k=0}^{1}\sum_{i=1}^{n} T_i^{(t)}(k)\log\left[f_m(m_i; \eta_i^m(k))\right] + \sum_{k=0}^{1}\sum_{i=1}^{n} T_i^{(t)}(k)\log\left[f_g(g_i; \eta_i^{g,b}(k))\right]. \quad (51)$$

The three terms of (51) are functions of different parameters: the first is a function of π, the second is a function of β_m, and the third is a function of β_g. Therefore, to find the maximizer $\theta^{(t+1)}$ of (51), we maximize the three terms separately. Differentiating the first term with respect to π, we find that

$$\frac{\partial}{\partial \pi} \sum_{i=1}^{n} \left[T_i^{(t)}(1) \log(\pi) + T_i^{(t)}(0) \log(1-\pi) \right]$$
$$= \frac{\sum_{i=1}^n T_i^{(t)}(1)}{\pi} - \frac{\sum_{i=1}^n T_i^{(t)}(0)}{1-\pi}.$$

Setting the derivative equal to 0 and solving for π,

$$\frac{\sum_{i=1}^n T_i^{(t)}(1)}{\pi} - \frac{\sum_{i=1}^n T_i^{(t)}(0)}{1-\pi} = 0 \iff \sum_{i=1}^{n} T_i^{(t)}(1) - \pi \sum_{i=1}^{n} T_i^{(t)}(1) = \pi \sum_{i=1}^{n} T_i(0)$$
$$\iff \sum_{i=1}^{n} T_i^{(t)}(1) - \pi \sum_{i=1}^{n} T_i^{(t)}(1) = \pi n - \pi \sum_{i=1}^{n} T_i(1) \iff \pi = \frac{\sum_{i=1}^n T_i^{(t)}(1)}{n}.$$

Thus, the maximizer $\pi^{(t+1)}$ of (51) in π is $\pi^{(t+1)} = (1/n) \sum_{i=1}^n T_i^{(t)}(1)$. Next, define $w^{(t)} = [T_1^{(t)}(0), \ldots, T_n^{(t)}(0), T_1^{(t)}(1), \ldots, T_n^{(t)}(1)]^T \in \mathbb{R}^{2n}$. We can view the second term of (51) as the log-likelihood of a GLM – call it $\text{GLM}_m^{(t)}$ – that has exponential family density f_m, link function r_m, responses $[m, m]^T$, offsets $[o^m, o^m]^T$, weights $w^{(t)}$, and design matrix $\begin{bmatrix} \tilde{X}(0) \\ \tilde{X}(1) \end{bmatrix}$. Therefore, the maximizer $\beta_m^{(t+1)}$ of the second term of (51) is the maximizer of $\text{GLM}_m^{(t)}$, which we can compute using the iteratively reweighted least squares (IRLS) procedure, as implemented in R's GLM function. Similarly, the maximizer $\beta_g^{(t+1)}$ of the third term of (51) is the maximizer of the GLM with exponential family density f_g, link function r_g, responses $[g, g]^T$, offsets $[o^g, o^g]^T$, weights $w^{(t)}$, and design matrix $\begin{bmatrix} \tilde{X}(0) \\ \tilde{X}(1) \end{bmatrix}$.

B.2 Inference

We derive the asymptotic observed information matrix of the GLM-EIV log likelihood, enabling us to perform inference on the parameters. First, we define some notation. For $i \in \{1, \ldots, n\}$, $j \in \{0, 1\}$, and $\theta = (\pi, \beta_m, \beta_g)$, let $T_i^\theta(j)$ be defined by

$$T_i^\theta(j) = \mathbb{P}_\theta \left(P_i = j | M_i = m_i, G_i = g_i \right).$$

Let the $n \times n$ matrix $T^\theta(j)$ be given by

$$T^\theta(j) = \text{diag} \left\{ T_1^\theta(j), \ldots, T_n^\theta(j) \right\}.$$

Next, define the diagonal $n \times n$ matrices Δ^m, $[\Delta']^m$, V^m, and H^m by

$$\begin{cases} \Delta^m = \text{diag}\{h_m'(l_1^m), \ldots, h_m'(l_n^m)\} \\ [\Delta']^m = \text{diag}\{h_m''(l_1^m), \ldots, h_m''(l_n^m)\} \\ V^m = \text{diag}\{\psi_m''(\eta_1^m), \ldots, \psi_m''(\eta_n^m)\} \\ H^m = \text{diag}\{m_1 - \mu_1^m, \ldots, m_n - \mu_n^m\}. \end{cases}$$

Define the $n \times n$ matrices $\Delta^g, [\Delta']^g, V^g$, and H^g analogously. These matrices are *unobserved*, as they depend on $\{p_1, \ldots, p_n\}$.

For $j \in \{0, 1\}$, let the diagonal $n \times n$ matrices $\Delta^m(j), [\Delta']^m(j), V^m(j)$, and $H^m(j)$ be given by

$$\begin{cases} \Delta^m(j) = \text{diag}\{h'_m(l_1^m(j)), \ldots, h'_m(l_n^m(j))\} \\ [\Delta']^m(j) = \text{diag}\{h''_m(l_1^m(j)), \ldots, h''_m(l_n^m(j))\} \\ V^m(j) = \text{diag}\{\psi''_m(\eta_1^m(j)), \ldots, \psi''_m(\eta_n^m(j))\} \\ H^m(j) = \text{diag}\{m_1 - \mu_1^m(j), \ldots, m_n - \mu_n^m(j)\}. \end{cases}$$

Define the matrices $\Delta^g(j)$, $[\Delta']^g(j)$, $V^g(j)$, and $H^g(j)$ analogously. Finally, define the vectors $s^m(j)$ and $w^m(j)$ in $\mathbb{R}^n$ by

$$\begin{cases} s^m(j) = [m_1 - \mu_1^m(j), \ldots, m_n - \mu_n^m(j)]^T \\ w^m(j) = [T_1(0)T_1(1)\Delta_1^m(j)H_1^m(j), \ldots, T_n(0)T_n(1)\Delta_n^m(j)H_n^m(j)]^T, \end{cases}$$

and let the vectors $s^g(j)$ and $w^g(j)$ be defined analogously. The quantities $\Delta^m(j), [\Delta']^m(j), V^m(j)$, $H^m(j)$, $s^m(j)$, $w^m(j)$, $\Delta^g(j), [\Delta']^g(j), V^g(j)$, $H^g(j)$, $s^g(j)$, and $w^g(j)$ are all *observed*.

The observed information matrix $J(\theta; m, g)$ evaluated at $\theta = (\pi, \beta_m, \beta_g)$ is the negative Hessian of the marginal log likelihood (14) evaluated at θ, i.e.

$$J(\theta; m, g) = -\nabla^2 \mathcal{L}(\theta; m, g).$$

This quantity, unfortunately, is hard to compute, as the log likelihood (14) is a complicated mixture. Louis [27] showed that $J(\theta; m, g)$ is equivalent to the following quantity:

$$\begin{aligned} J(\theta; m, g) = &-\mathbb{E}\left[\nabla^2 \mathcal{L}(\theta; m, g, p) | G = g, M = m\right] \\ &+ \mathbb{E}\left[\nabla \mathcal{L}(\theta; m, g, p) | G = g, M = m\right] \mathbb{E}\left[\nabla \mathcal{L}(\theta; m, g, p) | G = g, M = m\right]^T \\ &- \mathbb{E}\left[\nabla \mathcal{L}(\theta; m, g, p) \nabla \mathcal{L}(\theta; m, g, p)^T | G = g, M = m\right]. \quad (52) \end{aligned}$$

The observed information matrix $J(\theta; m, g)$ has dimension $(2d + 1) \times (2d + 1)$. Recall that the log-likelihood (12) is the sum of three terms. The first term depends only on π, the second on β_m, and the third on β_g. Therefore, the observed information matrix can be viewed as block matrix consisting of nine submatrices (Figure 7; only six submatrices labelled). Submatrix I depends on π, submatrix II on β_m, submatrix III on β_g, submatrix IV on β_m and β_g, submatrix V on π and β_m, and submatrix VI on π and β_g. We only need to compute these six submatrices to compute the entire matrix, as the matrix is symmetric. The following sections derive formulas for submatrices I-VI. All expectations are understood to be *conditional* on m and g. The notation ∇_v and ∇_v^2 represent the gradient and Hessian, respectively, with respect to the vector v.

Submatrix I

Denote submatrix I by $J_\pi(\theta; m, g)$. The formula for $J_\pi(\theta; m, g)$ is

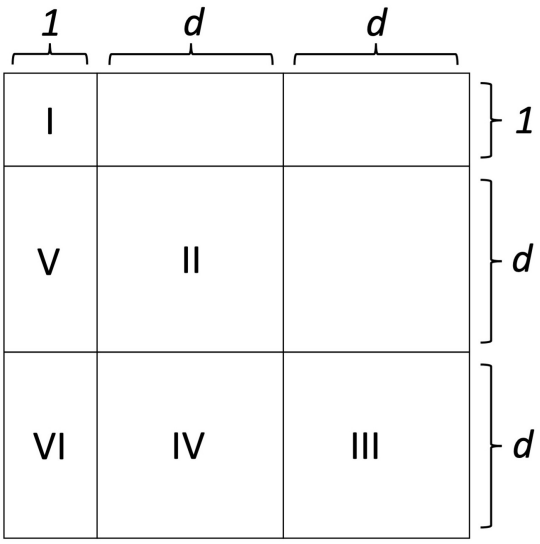


Figure 7: Block structure of the observed information matrix $J(\theta; m, g) = -\nabla^2 \mathcal{L}(\theta; m, g)$. The matrix is symmetric, and so we only need to compute submatrices I-VI to compute the entire matrix.

$$J_\pi(\theta; m, g) = -\mathbb{E}\left[\nabla_\pi^2 \mathcal{L}(\theta; m, g, p)\right] + \left(\mathbb{E}\left[\nabla_\pi \mathcal{L}(\theta; m, g, p)\right]\right)^2 - \mathbb{E}\left[\left(\nabla_\pi \mathcal{L}(\theta; m, g, p)\right)^2\right]. \quad (53)$$

We begin by calculating the first and second derivatives of the log-likelihood $\mathcal{L}$ with respect to π. The first derivative is

$$\begin{aligned}\nabla_\pi \mathcal{L}(\theta; m, g, p) &= \frac{\partial}{\partial \pi}\left(\sum_{i=1}^n p_i \log(\pi) + \sum_{i=1}^n (1 - p_i)\log(1-\pi)\right) \\ &= \frac{\sum_{i=1}^n p_i}{\pi} - \frac{\sum_{i=1}^n (1-p_i)}{1-\pi} = \frac{\sum_{i=1}^n p_i}{\pi} - \frac{n - \sum_{i=1}^n p_i}{1-\pi} \\ &= \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right)\sum_{i=1}^n p_i - \frac{n}{1-\pi}. \quad (54)\end{aligned}$$

The second derivative is

$$\nabla_\pi^2 \mathcal{L}(\theta; m, g, p) \;=\; \frac{\partial^2}{\partial^2 \pi}\left(\frac{\sum_{i=1}^n p_i}{\pi} - \frac{n - \sum_{i=1}^n p_i}{1-\pi}\right) \;=\; \frac{(\sum_{i=1}^n p_i) - n}{(1-\pi)^2} \;-\; \frac{\sum_{i=1}^n p_i}{\pi^2}.$$

We compute the expectation of the first term of (53):

$$\begin{aligned}\mathbb{E}\left[-\nabla_\pi^2 \mathcal{L}(\theta; m, g, p)\right] &= -\mathbb{E}\left[\frac{(\sum_{i=1}^n p_i) - n}{(1-\pi)^2} - \frac{\sum_{i=1}^n p_i}{\pi^2}\right] \\ &= -\mathbb{E}\left\{\left[\frac{1}{(1-\pi)^2} - \frac{1}{\pi^2}\right]\sum_{i=1}^n p_i - \frac{n}{(1-\pi)^2}\right\}\end{aligned}$$

$$= -\left\{ \left[\frac{1}{(1-\pi)^2} - \frac{1}{\pi^2} \right] \sum_{i=1}^{n} T_i^{\theta}(1) - \frac{n}{(1-\pi)^2} \right\}$$
$$= \left[\frac{1}{\pi^2} - \frac{1}{(1-\pi)^2} \right] \sum_{i=1}^{n} T_i^{\theta}(1) + \frac{n}{(1-\pi)^2}. \quad (55)$$

Next, we compute the difference of the second two pieces of (53). To this end, define

$$a := \frac{1}{(1-\pi)} + \frac{1}{\pi}$$

and

$$b := \frac{n}{(1-\pi)}.$$

We have that

$$\mathbb{E}\left[\nabla_{\pi}\mathcal{L}(\theta; m, g, p)^2\right]$$
$$= \mathbb{E}\left[\left(a \sum_{i=1}^{n} p_i - b \right)^2 \right] = \mathbb{E}\left[a^2 \left(\sum_{i=1}^{n} p_i \right)^2 - 2ab \sum_{i=1}^{n} p_i + b^2 \right]$$
$$= a^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[p_i p_j] - 2ab \sum_{i=1}^{n} \mathbb{E}[p_i] + b^2.$$

Next,

$$\left(\mathbb{E}\left[\nabla_{\pi}\mathcal{L}(\theta; m, g, x)\right]\right)^2 = \left(a \sum_{i=1}^{n} \mathbb{E}[p_i] - b \right)^2$$
$$= a^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[p_i] E[p_j] - 2ab \sum_{i=1}^{n} \mathbb{E}[p_i] + b^2.$$

Therefore,

$$\left(\mathbb{E}[\nabla_{\pi}\mathcal{L}(\theta; m, g, p)]\right)^2 - \mathbb{E}\left[\nabla_{\pi}\mathcal{L}(\theta; m, g, p)^2\right]$$
$$= a^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[p_i]\mathbb{E}[p_j] - a^2 \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[p_i p_j] = a^2 \left(\sum_{i=1}^{n} \mathbb{E}[p_i]^2 - \mathbb{E}[p_i^2] \right)$$
$$= a^2 \left(\sum_{i=1}^{n} [T_i^{\theta}(1)]^2 - T_i^{\theta}(1) \right) = \left(\frac{1}{(1-\pi)} + \frac{1}{\pi} \right)^2 \left(\sum_{i=1}^{n} [T_i^{\theta}(1)]^2 - T_i^{\theta}(1) \right). \quad (56)$$

Stringing (53), (55) and (56) together, we obtain

$$J_{\pi}(\theta; m, g) = \left[\frac{1}{\pi^2} - \frac{1}{(1-\pi)^2} \right] \sum_{i=1}^{n} T_i^{\theta}(1) + \frac{n}{(1-\pi)^2}$$
$$+ \left(\frac{1}{(1-\pi)} + \frac{1}{\pi} \right)^2 \left(\sum_{i=1}^{n} [T_i^{\theta}(1)]^2 - T_i^{\theta}(1) \right). \quad (57)$$

Submatrix II

Denote submatrix II by $J_{\beta^m}(\theta; m, g)$. The formula for $J_{\beta^m}(\theta; m, g)$ is

$$\begin{aligned} J_{\beta^m}(\theta; m, g) = &-\mathbb{E}\left[\nabla^2_{\beta^m}\mathcal{L}(\theta; m, g, p)\right] \\ &+ \mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\right]\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\right]^T \\ &- \mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)^T\right]. \quad (58) \end{aligned}$$

Standard GLM results imply that

$$-\nabla^2_{\beta^m}\mathcal{L}(\theta; m, g, p) = \tilde{X}^T(\Delta^m V^m \Delta^m - [\Delta']^m H^m)\tilde{X}$$

and

$$\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p) = \tilde{X}^T\Delta^m s^m.$$

We compute the first term of (58). The (k, l)th entry of this matrix is

$$\begin{aligned} \left(\mathbb{E}\left[-\nabla^2_{\beta^m}\mathcal{L}(\theta; m, g, p)\right]\right)[k, l] &= \mathbb{E}\left\{\tilde{X}[, k]^T(\Delta^m V^m \Delta^m - [\Delta']^m H^m)\tilde{X}[, l]\right\} \\ &= \sum_{i=1}^n \mathbb{E}\left\{\tilde{x}_{i,k}(\Delta_i^m V_i^m \Delta_i^m - [\Delta']_i^m H_i^m)\tilde{x}_{i,l}\right\} \\ &= \sum_{i=1}^n \tilde{x}_{i,k}(0)T_i^\theta(0)[\Delta_i^m(0)V_i^m(0)\Delta_i^m(0) - [\Delta']_i^m(0)H_i^m(0)]\tilde{x}_{i,l}(0) \\ &+ \sum_{i=1}^n \tilde{x}_{i,k}(1)T_i^\theta(1)[\Delta_i^m(1)V_i^m(1)\Delta_i^m(1) - [\Delta']_i^m(1)H_i^m(1)]\tilde{x}_{i,l}(1) \\ &= \sum_{s=0}^1 \tilde{X}(s)[, k]^T T^\theta(s)\left[\Delta^m(s)V^m(s)\Delta^m(s) - [\Delta']^m(s)H^m(s)\right]\tilde{X}(s)[, l]. \end{aligned}$$

We therefore have that

$$\begin{aligned} &\mathbb{E}\left[-\nabla^2_{\beta^m}\mathcal{L}(\theta; m, g, p)\right] \\ &\qquad = \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s)\left[\Delta^m(s)V^m(s)\Delta^m(s) - [\Delta']^m(s)H^m(s)\right]\tilde{X}(s). \quad (59) \end{aligned}$$

Next, we compute the difference of the last two terms of (58). The (k, l)th entry is

$$\begin{aligned} &\Big[\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\right]\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\right]^T \\ &\qquad - \mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)^T\right]\Big][k, l] \\ &= \left[\mathbb{E}\left[\tilde{X}^T\Delta^m s^m\right]\mathbb{E}\left[\tilde{X}^T\Delta^m s^m\right]^T\right][k, l] - \mathbb{E}\left[\tilde{X}^T\Delta^m s^m (s^m)^T\Delta^m\tilde{X}\right][k, l] \end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\tilde{X}[,k]^T \Delta^m s^m\right] \mathbb{E}\left[\tilde{X}[,l]^T \Delta^m s^m\right] - \mathbb{E}\left[\tilde{X}[,k]^T \Delta^m s^m (s^m)^T \Delta^m \tilde{X}[,l]\right] \\
&= \mathbb{E}\left(\sum_{i=1}^{n} \tilde{x}_{ik} \Delta_i^m s_i^m\right) \mathbb{E}\left(\sum_{j=1}^{n} \tilde{x}_{jl} \Delta_j^m s_j^m\right) - \mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{x}_{ik} \Delta_i^m s_i^m s_j^m \Delta_j^m \tilde{x}_{jl}\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[\tilde{x}_{ik} \Delta_i^m s_i^m] \mathbb{E}[\tilde{x}_{jl} \Delta_j^m s_j^m] - \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[\tilde{x}_{ik} \Delta_i^m s_i^m s_j^m \Delta_j^m \tilde{x}_{jl}] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[\tilde{x}_{ik} \Delta_i^m s_i^m] \mathbb{E}\left[\tilde{x}_{jl} \Delta_j^m s_j^m\right] - \sum_{i \neq j} \mathbb{E}[\tilde{x}_{ik} \Delta_i^m s_i^m] \mathbb{E}[s_j^m \Delta_j^m \tilde{x}_{jl}] \\
&\qquad - \sum_{i=1}^{n} \mathbb{E}[\tilde{x}_{ik} \Delta_i^m s_i^m s_i^m \Delta_i^m \tilde{x}_{il}] \\
&= \sum_{i=1}^{n} \mathbb{E}[\tilde{x}_{ik} \Delta_i^m s_i^m] \mathbb{E}[\tilde{x}_{il} \Delta_i^m s_i^m] - \sum_{i=1}^{n} \mathbb{E}[\tilde{x}_{ik} (\Delta_i^m)^2 (H_i^m)^2 \tilde{x}_{il}] \\
&= \sum_{i=1}^{n} \left[\tilde{x}_{ik}(0) \Delta_i^m(0) T_i^\theta(0) H_i^m(0) + \tilde{x}_{ik}(1) \Delta_i^m(1) T_i^\theta(1) H_i^m(1)\right] \\
&\qquad \cdot \left[\tilde{x}_{il}(0) \Delta_i^m(0) T_i^\theta(0) H_i^m(0) + \tilde{x}_{il}(1) \Delta_i^m(1) T_i^\theta(1) H_i^m(1)\right] \\
&- \sum_{i=1}^{n} \left[\tilde{x}_{ik}(0) T_i^\theta(0) (\Delta_i^m(0))^2 (H_i^m(0))^2 \tilde{x}_{il}(0) + \tilde{x}_{ik}(1) T_i^\theta(1) (\Delta_i^m(1))^2 (H_i^m(1))^m \tilde{x}_{il}(1)\right] \\
&= \sum_{s=0}^{1}\sum_{t=0}^{1} \left[\sum_{i=1}^{n} \tilde{x}_{ik}(s) T_i^\theta(s) \Delta_i^m(s) H_i^m(t) T_i^\theta(t) \Delta_i^m(t) H_i^m(t) \tilde{x}_{il}(t)\right] \\
&\qquad - \sum_{s=0}^{1} \left[\sum_{i=1}^{n} \tilde{x}_{ik}(s) T_i^\theta(s) (\Delta_i^m(s))^2 (H_i^m(s))^2 \tilde{x}_{il}(s)\right] \\
&= \sum_{s=0}^{1}\sum_{t=0}^{1} \tilde{X}(s)[,k]^T T^\theta(s) \Delta^m(s) H^m(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(k)[,l] \\
&\qquad - \sum_{s=0}^{1} X(s)[,k]^T T^\theta(s) (\Delta^m(s))^2 (H^m(s))^2 \tilde{X}(s)[,l].
\end{aligned}
$$

The sum of the last two terms on the right-hand side of (58) is therefore

$$
\begin{aligned}
&\mathbb{E}\left[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)\right] \mathbb{E}\left[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)\right]^T \\
&\qquad - \mathbb{E}\left[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p) \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)^T\right] \\
&= \sum_{s=0}^{1}\sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s) \Delta^m(s) H^m(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(t) \\
&\qquad - \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) (\Delta^m(s))^2 (H^m(s))^2 \tilde{X}(s). \quad (60)
\end{aligned}
$$

Combining (58), (59), (60), we find that

$$J_{\beta^m}(\theta; m, g) = \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) \left[\Delta^m(s) V^m(s) \Delta^m(s) - [\Delta']^m(s) H^m(s)\right] \tilde{X}(s) + \sum_{s=0}^{1}\sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s) \Delta^m(s) H^m(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(t) - \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) (\Delta^m(s))^2 (H^m(s))^2 \tilde{X}(s). \quad (61)$$

Submatrix III

Denote submatrix III by $J_{\beta^g}(\theta; m, g)$. The formula for sub-matrix III is similar to that of sub-matrix II (61). Substituting g for m in this equation yields

$$J_{\beta^g}(\theta; m, g) = \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) \left[\Delta^g(s) V^g(s) \Delta^g(s) - [\Delta']^g(s) H^g(s)\right] \tilde{X}(s) + \sum_{s=0}^{1}\sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s) \Delta^g(s) H^g(s) T^\theta(t) \Delta^g(t) H^g(t) \tilde{X}(t) - \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) (\Delta^g(s))^2 (H^g(s))^2 \tilde{X}(s). \quad (62)$$

Submatrix IV

Denote sub-matrix IV by $J_{(\beta^g, \beta^m)}(\theta; m, g)$. The formula for $J_{(\beta^g, \beta^m)}(\theta; m, g)$ is

$$J_{(\beta^g, \beta^m)}(\theta; m, g) = \mathbb{E}\left[-\nabla_{\beta^g}\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\right] + \mathbb{E}\left[\nabla_{\beta^g}\mathcal{L}(\theta; m, g, p)\right]\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\right]^T - \mathbb{E}\left[\nabla_{\beta^g}\mathcal{L}(\theta; m, g, p)\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)^T\right]. \quad (63)$$

First, we have that

$$\mathbb{E}\left[-\nabla_{\beta^g}\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)\right] = 0, \quad (64)$$

as differentiating $\mathcal{L}$ with with respect to β^g yields a vector that is a function of β^g, and differentiating this vector with respect to β^m yields 0. Next, recall from GLM theory that

$$\nabla_{\beta^g}\mathcal{L}(\theta; m, g, p) = \tilde{X}^T \Delta^g s^g$$

and

$$\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p) = \tilde{X}^T \Delta^m s^m.$$

The (k, l)th entry of the last two terms of (63) is

$$
\begin{aligned}
&\Big[\mathbb{E}\left[\nabla_{\beta^g}\mathcal{L}(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)\right]^T \\
&\qquad - \mathbb{E}\left[\nabla_{\beta^g}\mathcal{L}(\theta;m,g,p)\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)^T\right]\Big][k,l] \\
&= \left[\mathbb{E}\left[\tilde{X}^T\Delta^g s^g\right]\mathbb{E}\left[\tilde{X}^T\Delta^m s^m\right]^T\right][k,l] - \mathbb{E}\left[\tilde{X}^T\Delta^g s^g (s^m)^T\Delta^m\tilde{X}\right][k,l] \\
&= \mathbb{E}\left[\tilde{X}[,k]^T\Delta^g s^g\right]\mathbb{E}\left[\tilde{X}[,l]^T\Delta^m s^m\right] - \mathbb{E}\left[\tilde{X}[,k]^T\Delta^g s^g (s^m)^T\Delta^m\tilde{X}[,l]\right] \\
&= \mathbb{E}\left(\sum_{i=1}^n \tilde{x}_{ik}\Delta_i^g s_i^g\right)\mathbb{E}\left(\sum_{j=1}^n \tilde{x}_{jl}\Delta_j^m s_j^m\right) - \mathbb{E}\left(\sum_{i=1}^n\sum_{j=1}^n \tilde{x}_{ik}\Delta_i^g s_i^g s_j^m\Delta_j^m\tilde{x}_{jl}\right) \\
&= \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[\tilde{x}_{ik}\Delta_i^g s_i^g]\mathbb{E}[\tilde{x}_{jl}\Delta_j^m s_j^m] - \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[\tilde{x}_{ik}\Delta_i^g s_i^g s_j^m\Delta_j^m\tilde{x}_{jl}] \\
&= \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[\tilde{x}_{ik}\Delta_i^g s_i^g]\mathbb{E}\left[\tilde{x}_{jl}\Delta_j^m s_j^m\right] - \sum_{i\neq j} \mathbb{E}[\tilde{x}_{ik}\Delta_i^g s_i^g]\mathbb{E}[\tilde{x}_{jl}\Delta_j^m s_j^m] \\
&\qquad - \sum_{i=1}^n \mathbb{E}[\tilde{x}_{ik}\Delta_i^g s_i^g s_i^m\Delta_i^m\tilde{x}_{il}] \\
&= \sum_{i=1}^n \mathbb{E}[\tilde{x}_{ik}\Delta_i^g H_i^g]\mathbb{E}[\tilde{x}_{il}\Delta_i^m H_i^m] - \sum_{i=1}^n \mathbb{E}[\tilde{x}_{ik}H_i^g\Delta_i^g\Delta_i^m H_i^m\tilde{x}_{il}] \\
&= \sum_{i=1}^n \left[\tilde{x}_{ik}(0)\Delta_i^g(0)T_i^\theta(0)H_i^g(0) + \tilde{x}_{ik}(1)\Delta_i^g(1)T_i^\theta(1)H_i^g(1)\right] \\
&\quad \cdot \left[\tilde{x}_{il}(0)\Delta_i^m(0)T_i^\theta(0)H_i^m(0) + \tilde{x}_{il}(1)\Delta_i^m(1)T_i^\theta(1)H_i^m(1)\right] \\
&\qquad - \sum_{i=1}^n [\tilde{x}_{ik}(0)T_i^\theta(0)\Delta_i^g(0)H_i^g(0)\Delta_i^m(0)H_i^m(0)\tilde{x}_{il}(0) \\
&\qquad\quad + \tilde{x}_{ik}(1)T_i^\theta(1)\Delta_i^g(1)H_i^g(1)\Delta_i^m(1)H_i^m(1)\tilde{x}_{il}(1)] \\
&= \sum_{s=0}^1\sum_{t=0}^1 \left[\sum_{i=1}^n \tilde{x}_{ik}(s)T_i^\theta(s)\Delta_i^g(s)H_i^g(s)T_i^\theta(t)\Delta_i^m(t)H_i^m(t)\tilde{x}_{il}(t)\right] \\
&\qquad - \sum_{s=0}^1 \left[\sum_{i=1}^n \tilde{x}_{ik}(s)T_i^\theta(s)\Delta_i^g(s)H_i^g(s)\Delta_i^m(s)H_i^m(s)\tilde{x}_{il}(s)\right] \\
&= \sum_{s=0}^1\sum_{t=0}^1 \left[\tilde{X}(s)[,k]^T T^\theta(s)\Delta^g(s)H^g(s)T^\theta(t)\Delta^m(t)H^m(t)\tilde{X}(t)[,l]\right] \\
&\qquad - \sum_{s=0}^1 \left[\tilde{X}[,k]^T T^\theta(s)\Delta^g(s)H^g(s)\Delta^m(s)H^m(s)\tilde{X}[,l](s)\right]. \quad (65)
\end{aligned}
$$

Combining (63), (64), and (65) produces

$$J_{(\beta^g,\beta^m)}(\theta;m,g) = \sum_{s=0}^{1}\sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s)\Delta^g(s)H^g(s)T^\theta(t)\Delta^m(t)H^m(t)\tilde{X}(t) - \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s)\Delta^g(s)H^g(s)\Delta^m(s)H^m(s)\tilde{X}(s). \quad (66)$$

Submatrix V

Denote submatrix V by $J_{(\beta^m,\pi)}(\theta;m,g)$. The formula for $J_{(\beta^m,\pi)}(\theta;m,g)$ is

$$J_{(\beta^m,\pi)}(\theta;m,g) = \mathbb{E}\left[-\nabla_{\beta^m}\nabla_\pi \mathcal{L}(\theta;m,g,p)\right] + \mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_\pi \mathcal{L}(\theta;m,g,p)\right]^T - \mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)\nabla_\pi \mathcal{L}(\theta;m,g,p)^T\right]. \quad (67)$$

We have that

$$\mathbb{E}\left[-\nabla_{\beta^m}\nabla_\pi \mathcal{L}(\theta;m,g,p)\right] = 0, \quad (68)$$

as β^m and π separate in the log likelihood. Next, set $a := 1/\pi + 1/(1-\pi)$ and $b := n/(1-\pi)$. Recall from GLM theory that

$$\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p) = \tilde{X}^T\Delta^m s^m$$

and from (54) that

$$a\sum_{i=1}^{n} p_i - b.$$

The kth entry of the last two terms of (67) is

$$\begin{aligned}
&\mathbb{E}\left[\nabla_\pi \mathcal{L}(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)[k]\right] - \mathbb{E}\left[\nabla_\pi \mathcal{L}(\theta;m,g,p)\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)[k]\right] \\
&= \left(\mathbb{E}\left[a\sum_{i=1}^{n} p_i - b\right]\right)\left(\mathbb{E}\left[\tilde{X}[,k]^T\Delta^m s^m\right]\right) - \mathbb{E}\left[\left(a\sum_{i=1}^{n} p_i - b\right)\tilde{X}[,k]^T\Delta^m s^m\right] \\
&= \left(a\sum_{i=1}^{n}\mathbb{E}[p_i] - b\right)\left(\sum_{j=1}^{n}\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m]\right) - \mathbb{E}\left[\left(a\sum_{i=1}^{n} p_i - b\right)\left(\sum_{j=1}^{n}\tilde{x}_{jk}\Delta_j^m s_j^m\right)\right] \\
&= a\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[p_i]\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m] - b\sum_{j=1}^{n}\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m] \\
&\quad - \left[a\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[p_i\tilde{x}_{jk}\Delta_j^m s_j^m] - b\sum_{j=1}^{n}\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m]\right] \\
&= a\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{E}[p_i]\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m] - a\sum_{i\neq j}\mathbb{E}[p_i]\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m] - a\sum_{i=1}^{n}\mathbb{E}[p_i\tilde{x}_{ik}\Delta_i^m s_i^m]
\end{aligned}$$

$$= a\sum_{i=1}^{n}\mathbb{E}[p_i]\mathbb{E}[\tilde{x}_{ik}\Delta_i^m s_i^m] - a\sum_{i=1}^{n}\mathbb{E}[p_i\tilde{x}_{ik}\Delta_i^m s_i^m]$$

$$= a\sum_{i=1}^{n}T_i^\theta(1)[T_i^\theta(0)\Delta_i^m(0)s_i^m(0)\tilde{x}_{ik}(0) + T_i^\theta(1)\Delta_i^m(1)s_i^m(1)\tilde{x}_{ik}(1)]$$

$$- a\sum_{i=1}^{n}T_i^\theta(1)\Delta_i^m(1)s_i^m(1)\tilde{x}_{ik}(1)$$

$$= a\sum_{i=1}^{n}T_i^\theta(0)T_i^\theta(1)\Delta_i^m(0)H_i^m(0)\tilde{x}_{ik}(0)$$

$$+ a\sum_{i=1}^{n}\left([T_i^\theta(1)]^2\Delta_i^m(1)H_i^m(1) - T_i^\theta(1)\Delta_i^m(1)H_i^m(1)\right)\tilde{x}_{ik}(1)$$

$$= a\left[\sum_{i=1}^{n}T_i^\theta(0)T_i^\theta(1)\Delta_i^m(0)H_i^m(0)\tilde{x}_{ik}(0) + \sum_{i=1}^{n}T_i^\theta(1)\Delta_i^m(1)H_i^m(1)[T_i^\theta(1)-1]\tilde{x}_{ik}(1)\right]$$

$$= a\left[\sum_{i=1}^{n}T_i^\theta(0)T_i^\theta(1)\Delta_i^m(0)H_i^m(0)\tilde{x}_{ik}(0) - \sum_{i=1}^{n}T_i^\theta(0)T_i^\theta(1)\Delta_i^m(1)H_i^m(1)\tilde{x}_{ik}(1)\right]$$

$$= a\left(\tilde{X}(0)[,k]^T w^m(0) - \tilde{X}(1)[,k]^T w^m(1)\right). \quad (69)$$

Combining (67), (68), and (69), we conclude that

$$J_{(\beta^m,\pi)}(\theta; m, g, p) = \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right)\left(\tilde{X}(0)^T w^m(0) - \tilde{X}(1)^T w^m(1)\right). \quad (70)$$

Submatrix VI

Denote submatrix VI by $J_{(\beta^g,\pi)}(\theta; m, g)$. Calculations similar to those for submatrix V show that

$$J_{(\beta^g,\pi)}(\theta; m, g, p) = \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right)\left(\tilde{X}(0)^T w^g(0) - \tilde{X}(1)^T w^g(1)\right). \quad (71)$$

Combining submatrices

To summarize, the formulas for submatrices I-VI are as follows:

I

$$J_\pi(\theta; m, g) = \left[\frac{1}{\pi^2} - \frac{1}{(1-\pi)^2}\right]\sum_{i=1}^{n}T_i^\theta(1) + \frac{n}{(1-\pi)^2}$$

$$+ \left(\frac{1}{(1-\pi)} + \frac{1}{\pi}\right)^2\left(\sum_{i=1}^{n}[T_i^\theta(1)]^2 - T_i^\theta(1)\right).$$

II

$$J_{\beta^m}(\theta; m, g) = \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) \left[\Delta^m(s) V^m(s) \Delta^m(s) - [\Delta']^m(s) H^m(s)\right] \tilde{X}(s)$$
$$+ \sum_{s=0}^{1} \sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s) \Delta^m(s) H^m(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(t)$$
$$- \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) (\Delta^m(s))^2 (H^m(s))^2 \tilde{X}(s).$$

III

$$J_{\beta^g}(\theta; m, g) = \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) \left[\Delta^g(s) V^g(s) \Delta^g(s) - [\Delta']^g(s) H^g(s)\right] \tilde{X}(s)$$
$$+ \sum_{s=0}^{1} \sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s) \Delta^g(s) H^g(s) T^\theta(t) \Delta^g(t) H^g(t) \tilde{X}(t)$$
$$- \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) (\Delta^g(s))^2 (H^g(s))^2 \tilde{X}(s).$$

IV

$$J_{(\beta^g, \beta^m)}(\theta; m, g) = \sum_{s=0}^{1} \sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s) \Delta^g(s) H^g(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(t)$$
$$- \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) \Delta^g(s) H^g(s) \Delta^m(s) H^m(s) \tilde{X}(s).$$

V

$$J_{(\beta^m, \pi)}(\theta; m, g, p) = \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) \left(\tilde{X}(0)^T w^m(0) - \tilde{X}(1)^T w^m(1)\right).$$

VI

$$J_{(\beta^g, \pi)}(\theta; m, g, p) = \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) \left(\tilde{X}(0)^T w^g(0) - \tilde{X}(1)^T w^g(1)\right).$$

We stitch these pieces together and transpose submatrices IV, V, and VI to produce the whole information matrix $J(\theta; m, g)$. Evaluating this matrix at the EM estimate θ^{EM} and inverting yields the asymptotic covariance matrix, which we can use to compute standard errors.

B.3 Implementation

To evaluate the observed information matrix, we need to compute the matrices $\Delta^m(j)$, $[\Delta']^m(j)$, $V^m(j)$, and $H^m(j)$ and the vectors $s^m(j)$ and $w^m(j)$ for $j \in \{0, 1\}$. We likewise need to compute the analogous gRNA quantities. We describe how to compute these quantities in R by extending base family objects. We implicitly condition on p_i, z_i^m, and o_i^m.

Algorithm 4 Computing the matrices $\Delta^m(j)$, $[\Delta']^m(j)$, $V^m(j)$, $H^m(j)$, and $s^m(j)$ given given β_m.

Input: A coefficient vector β_m; data $[m_1, \ldots, m_n]$, $[o_1^m, \ldots, o_n^m]$, and $[z_1, \ldots, z_n]$; and a family object containing functions `linkinv`, `variance`, `mu.eta`, `mu.eta.prime`, and `skewness`.

for $j \in \{0, 1\}$ **do**
  for $i \in \{1, \ldots, n\}$ **do**
    $l_i^m(j) \leftarrow \langle \beta_m, \tilde{x}_i(j) \rangle + o_i^m$
    $\mu_i^m(j) \leftarrow \texttt{linkinv}(l_i^m(j))$
    $[\sigma_i^m(j)]^2 \leftarrow \texttt{variance}(\mu_i^m(j))$
    $h'_m(l_i^m(j)) \leftarrow \texttt{mu.eta}(l_i^m(j))/[\sigma_i^m(j)]^2$
    $\gamma_i^m(j) \leftarrow \texttt{skewness}(\mu_i^m(j))$
    $[r_m^{-1}]''(l_i^m(j)) \leftarrow \texttt{mu.eta.prime}(l_i^m(j))$

$$h''_m(l_i^m(j)) \leftarrow \frac{[r^{-1}]''(l_i^m(j)) - [([\sigma_i^m(j)]^2)^{3/2}][\gamma_i^m(j)][h'_m(l_i^m(j))]^2}{[\sigma_i^m(j)]^2}$$

▷ Assign quantities to matrices

    $\Delta_i^m(j) \leftarrow h'_m(l_i^m(j))$
    $[\Delta']_i^m(j) \leftarrow h''(l_i^m(j))$
    $V_i^m(j) \leftarrow [\sigma_i^m(j)]^2$
    $H_i^m(j) \leftarrow s_i^m(j) \leftarrow m_i - \mu_i^m(j)$
  end for
end for

An `R` family object contains several functions, including `linkinv`, `variance`, and `mu.eta`. `linkinv` is the inverse link function r_m^{-1}. `variance` takes as an argument the mean μ_i^m and returns the variance $[\sigma_i^m]^2$. `mu.eta` is the derivative of the inverse link function $[r_m^{-1}]'$. We extend the `R` family object by adding two additional functions: `skewness` and `mu.eta.prime`. `skewness` returns the skewness γ_i^m of the distribution as a function of the mean μ_i, i.e.

$$\texttt{skewness}\,(\mu_i) = \mathbb{E}\left[\left(\frac{m_i - \mu_i^m}{\sigma_i^m}\right)^3\right] := \gamma_i^m.$$

Finally, the `mu.eta.prime` is the second derivative of the inverse link function $[r_m^{-1}]''$. Algorithm 4 computes the matrices $\Delta^m(j)$, $[\Delta']^m(j)$, $V^m(j)$, $H^m(j)$, and vector $s^m(j)$

for given β_m and given family object. (The vector $w^m(j)$ can be computed in terms of $\Delta^m(j)$ and $H^m(j)$.) We use $\sigma_i^m(j)$ (resp. $\gamma_i^m(j)$) to refer to the standard deviation (resp. skewness) of the gene expression distribution the ith cell when the perturbation p_i is set to j.

All steps of the algorithm are obvious except the calculation of $h'_m(l_i^m(j))$ (line 6), $h''(l_i^m(j))$ (line 9), and $V_i^m(j)$ (line 12). We omit the (j) notation for compactness. First, we prove the correctness of the expression for $h'_m(l_i^m)$. Recall the basic GLM identities

$$\psi''_m(\eta_i^m) = [\sigma_i^m]^2 \tag{72}$$

and, for all $t \in \mathbb{R}$,

$$r_m^{-1}(t) = \psi'_m(h_m(t)). \tag{73}$$

Differentiating (73) in t, we find that

$$(r_m^{-1})'(t) = \psi''_m(h_m(t))h'_m(t), \tag{74}$$

or

$$h'_m(t) = \frac{(r_m^{-1})'(t)}{\psi''_m(h_m(t))}.$$

Finally, plugging in l_i^m for t,

$$h'_m(l_i) = \frac{(r_m^{-1})'(l_i^m)}{\psi''(h_m(l_i^m))} = \frac{(r_m^{-1})'(l_i^m)}{\psi''_m(\eta_i^m)} = \text{ by (72) } \frac{(r_m^{-1})'(l_i^m)}{[\sigma_i^m]^2}.$$

Next, we prove the correctness for the expression for $h''_m(l_i^m)$. Recall the exponential family identity

$$\psi'''_m(\eta_i^m) = \gamma_i^m([\sigma_i^m]^2)^{3/2}. \tag{75}$$

Differentiating (74) in t, we obtain

$$(r_m^{-1})''(t) = \psi'''_m(h_m(t))[h'_m(t)]^2 + \psi''_m(h_m(t))h''_m(t),$$

or

$$h''_m(t) = \frac{(r_m^{-1})''(t) - \psi'''(h_m(t))[h'_m(t)]^2}{\psi''_m(h_m(t))}.$$

Plugging in l_i^m for t, we find that

$$h''_m(l_i^m) = \frac{(r_m^{-1})''(l_i^m) - \psi'''_m(\eta_i^m)[h'_m(l_i^m)]^2}{[\sigma_i^m]^2}$$
$$= \text{ (by 75) } \frac{(r_m^{-1})''(l_i^m) - ([\sigma_i^m]^2)^{3/2}(\gamma_i^m)[h'_m(l_i^m)]^2}{[\sigma_i^m]^2}.$$

Finally, the expression for V_i^m follows from (72). We can apply a similar algorithm to compute the analogous matrices for the gRNA modality. Table 1 shows the `linkinv`, `variance`, `mu.eta`, `skewness`, and `mu.eta.prime` functions for several common family objects (which are defined by a distribution and link function).

Table 1: `linkinv`, `variance`, `mu.eta`, `skewness`, `mu.eta.prime` for common family objects (i.e., pairs of distributions and link functions).

	Gaussian response, identity link	Poisson response, log link	NB response ($\theta > 0$ fixed), log link
`linkinv`	x	$\exp(x)$	$\exp(x)$
`variance`	x	x	$x + x^2/\theta$
`mu.eta`	1	x	$\exp(x)$
`skewness`	0	$x^{-1/2}$	$\frac{2x+\theta}{\sqrt{\theta x}\sqrt{x+\theta}}$
`mu.eta.prime`	0	$\exp(x)$	$\exp(x)$

C Zero-inflated model

In this section we introduce the "zero-inflated" GLM-EIV model. The zero-inflated GLM-EIV model is appropriate to use when the unperturbed cells do not transcribe *any* gRNA molecules (i.e., when there are no background reads). Let $x_i = [1, z_i]^T \in \mathbb{R}^{d-1}$ be the vector of observed covariates, including an intercept term. (x_i is the same as $\tilde{x}_i$, but with the perturbation indicator p_i removed.) Let $\beta_{g,z} = [\beta_0^g, \gamma_g] \in \mathbb{R}^{d-1}$ be an unknown coefficient vector. ($\beta_{g,z}$ is the same as β_g, but with the perturbation effect β_1^g removed). Let the linear component $l_i^{g,z}$, mean $\mu_i^{g,z}$, and canonical parameter $\eta_i^{g,z}$ of gRNA count distribution of the ith cell be given by

$$\begin{cases} l_i^{g,z} = \langle x_i, \beta_{g,z} \rangle + o_i^g \\ r_g(\mu_i^{g,z}) = l_i^{g,z} \\ \eta_i^{g,z} = ([\psi_g']^{-1} \circ r_g^{-1})(l_i^{g,z}) := h_g(l_i^{g,z}). \end{cases}$$

The density $f_{g,z}$ of gRNA counts in the zero-inflated model is as follows:

$$f_{g,z}(g_i; \eta_i^{g,z}, p_i) = [f_g(g_i; \eta_i^{g,z})]^{p_i} \mathbb{I}(g_i = 0)^{1-p_i}.$$

In other words, when the cell is *perturbed* (i.e., $p_i = 1$), the zero-inflated density $f_{g,z}$ coincides with the background-read density f_g; by contrast, when the cell is *unperturbed* (i.e., $p_i = 0$), the zero-inflated density $f_{g,z}$ is a point mass at zero. The gene expression density f_m and perturbation indicator density f_p are the same across the background read and zero-inflated models. We assume that the gene expression m_i and gRNA count g_i are conditionally independent given the perturbation indicator p_i. The joint density f_z of (m_i, p_i, z_i) is

$$\begin{aligned} f_z(m_i, g_i, p_i) &= f_m(m_i|p_i) f_{g,z}(g_i|p_i) f_p(p_i) \\ &= \pi^{p_i}(1-\pi)^{1-p_i} f_m(m_i; \eta_i^m)[f_g(g_i; \eta_i^{g,z})]^{p_i} \mathbb{I}(g_i = 0)^{1-p_i}. \end{aligned}$$

The log-likelihood $\mathcal{L}_z$ is

$$\mathcal{L}_z(\theta; m, g, p) = \sum_{i=1}^{n} \log\left[\pi^{p_i}(1-\pi)^{1-p_i}\right] + \sum_{i=1}^{n} \log\left[f_m(m_i; \eta_i^m)\right] + \sum_{i=1}^{n} p_i \log\left[f_g(g_i; \eta_i^{g,z})\right] + \sum_{i=1}^{n} (1-p_i) \log\left[\mathbb{I}(g_i = 0)\right],$$

where $\theta = [\pi, \beta_m, \beta_{g,z}]$ is the vector of unknown parameters. Integrating over the unobserved variable p_i, the marginal density f_z of (m_i, g_i) is

$$f_z(m_i, g_i; \theta) = (1-\pi) f_m(m_i; \eta_i^m(0)) \mathbb{I}(g_i = 0) + \pi f_m(m_i; \eta_i^m(1)) f_g(g_i; \eta_i^{g,z}).$$

Finally, the marginal log-likelihood is

$$\mathcal{L}_z(\theta; m_i, g_i) = \sum_{i=1}^{n} \log\left[(1-\pi) f_m(m_i; \eta_i^m(0)) \mathbb{I}(g_i = 0) + \pi f_m(m_i; \eta_i^m(1)) f_g(g_i; \eta_i^{g,z})\right].$$

C.1 Estimation

To estimate the parameters of the zero-inflated GLM-EIV model, we use an EM algorithm similar to Algorithm 1 but with two changes. First, we use a different formula for the ith membership probability at the t-th step of the algorithm $T_i^{(t)}(1)$. (We use $T_i^{(t)}(1)$ to denote the ith membership probability in *both* the background read and zero inflated cases; the difference should be clear from context.) Let $\theta^{(t)} = (\pi^{(t)}, \beta_m^{(t)}, \beta_{g,z}^{(t)})$ be the parameter estimate at the t-th iteration of the algorithm. Arguing in a manner similar to the background read case, we have that

$$T_i^{(t)}(1) = \frac{1}{\exp(q_i^{(t,z)}) + 1},$$

where

$$q_i^{(t,z)} = \log\left(\frac{(1-\pi^{(t)})\mathbb{P}(M_i = m_i | P_i = 0, \theta^{(t)})\mathbb{P}(G_i = g_i | P_i = 0, \theta^{(t)})}{(\pi^{(t)})\mathbb{P}(M_i = m_i | P_i = 1, \theta^{(t)})\mathbb{P}(G_i = g_i | P_i = 1, \theta^{(t)})}\right).$$

The expression for $q_i^{(t,z)}$ is

$$q_i^{(t,z)} = \log\left[1-\pi^{(t)}\right] + \log\left[f_m\left(m_i; [\eta_i^m(0)]^{(t)}\right)\right] + \log\left[\mathbb{I}(g_i = 0)\right] - \log\left[\pi^{(t)}\right] - \log\left[f_m\left(m_i; [\eta_i^m(1)]^{(t)}\right)\right] - \log\left[f_g\left(g_i; [\eta_i^{g,z}]^{(t)}\right)\right],$$

where $[\eta_i^{g,z}]^{(t)} = h_g(\langle x_i, \beta_{g,z}^{(t)} \rangle + o_i^g)$. Notice that if $g_i \geq 1$, then $T_i^{(t)}(1) = 1$. This comports with our intuition that a nonzero gRNA count indicates the presence of a perturbation.

Next, we consider the M step of the EM algorithm, which is similar to the background read case but slightly different. Define $Q_z(\theta | \theta^{(t)}) = \mathbb{E}_{(P|M=m,G=g,\theta^{(t)})}\left[\mathcal{L}_z(\theta; m, g, p)\right]$. We have that

$$Q_z(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\left[T_i^{(t)}(1)\log(\pi) + T_i^{(t)}(0)\log(1-\pi)\right] + \sum_{i=1}^{n}\sum_{j=0}^{1} T_i^{(t)}(j)\log\left[f_m(m_i;\eta_i^m(j))\right] + \sum_{i=1}^{n} T_i^{(t)}(1)\left[\log(f_g(g_i;\eta_i^{g,z}))\right] + C. \quad (76)$$

The three terms of (76) are functions of π, β_m, and $\beta_{g,z}$, respectively. The maximizer $\pi^{(t)}$ and $\beta_m^{(t+1)}$ of the first and second term are the same as in the background read case. The maximizer $\beta_{g,z}^{(t+1)}$ of the third term is the maximizer of the GLM with exponential family density f_g, link function r_g, responses g, weights $T^{(t)}(1)$, design matrix X, offsets o^g.

C.2 Inference

Next, we derive the asymptotic observed information matrix for the zero-inflated model, allowing us to perform inference. Again, let $T^\theta(1) := \text{diag}\{T_1^\theta(1),\ldots,T_n^\theta(1)\}$, but note that $T_i^\theta(1) = \mathbb{P}(P_i = 1|G_i = g_i, M_i = m_i, \theta)$ is computed differently than in the background read case. Define the $n \times n$ matrices $\Delta^{(g,z)}, [\Delta']^{(g,z)}, V^{(g,z)}$, and $H^{(g,z)}$ by

$$\begin{cases} \Delta^{(g,z)} = \text{diag}\{h_g'(l_1^{g,z}),\ldots,h_g'(l_n^{g,z})\} \\ [\Delta']^{(g,z)} = \text{diag}\{h_g''(l_1^{g,z}),\ldots,h_g''(l_n^{g,z})\} \\ V^{(g,z)} = \text{diag}\{\psi_g(\eta_1^{g,z}),\ldots,\psi_g(\eta_n^{g,z})\} \\ H^{(g,z)} = \text{diag}\{m_1-\mu_1^{g,z},\ldots,m_n-\mu_n^{g,z}\}. \end{cases}$$

Also, define the $\mathbb{R}^n$ vectors $s^{(g,z)}$ and $w^{(g,z)}$ by

$$s^{(g,z)} = [g_1-\mu_1^{g,z},\ldots,g_n-\mu_n^{g,z}]^T,$$

and

$$w^{(g,z)} = [T_1^\theta(0)T_1^\theta(1)\Delta_1^{(g,z)}H_1^{(g,z)},\ldots,T_n^\theta(0)T_n^\theta(1)\Delta_n^{(g,z)}H_n^{(g,z)}].$$

These quantities are computable, as they do not depend on the unobserved variables $p_1,\ldots,p_n$. Finally, let the unobserved, $n\times n$ matrix P be defined by $P = \text{diag}\{p_1,\ldots,p_n\}$.

The observed information matrix $J_z(\theta;m,g)$ is given by

$$J_z(\theta;m,g) = -\nabla^2\mathcal{L}_z(\theta;m,g).$$

Louis's theorem implies that

$$J_z(\theta;m,g) = -\mathbb{E}\left[\nabla^2\mathcal{L}_z(\theta;m,g,p)|G=g,M=m\right] + \mathbb{E}\left[\nabla\mathcal{L}_z(\theta;m,g,p)|G=g,M=m\right]\mathbb{E}\left[\nabla\mathcal{L}_z(\theta;m,g,p)|G=g,M=m\right]^T - \mathbb{E}\left[\nabla\mathcal{L}_z(\theta;m,g,p)\nabla\mathcal{L}_z(\theta;m,g,p)^T|G=g,M=m\right].$$

The matrix $J_z(\theta;m,g)$ has dimension $d\times d$ and consists of nine submatrices (Figure 8). Three of these submatrices (i.e., I, II, and V) are the same as the corresponding submatrices in the background read case. We therefore must compute the remaining submatrices (i.e., III, IV, and VI) to compute the entire matrix $J_z(\theta;m,g)$. Again, in the following, all expectations are understood to be conditional on m and g.

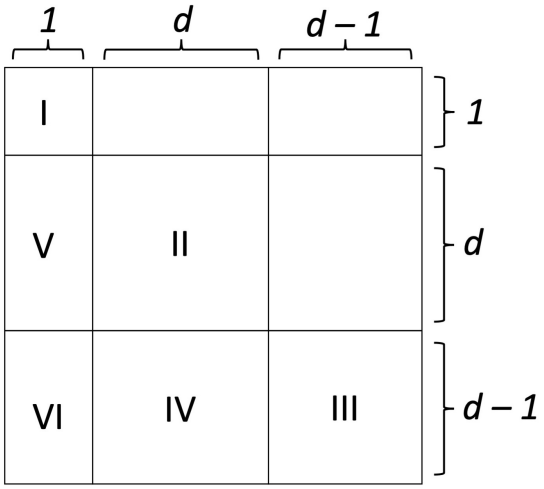


Figure 8: Block structure of the observed information matrix $J_z(\theta; m, g) = -\nabla^2 \mathcal{L}_z(\theta; m, g)$ for the zero-inflated model. Submatrices I, II, and VI are the same as in the background read model; therefore, we only need to compute submatrices III, VI, and V.

Submatrix III (zero-inflated)

Denote submatrix III by $J_{\beta_{(g,z)}}(\theta; m, g)$ The formula for $J_{\beta_{(g,z)}}(\theta; m, g)$ is

$$\begin{aligned} J_{\beta_{(g,z)}}(\theta; m, g) = &-\mathbb{E}\left[\nabla^2_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)\right] \\ &+ \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)\right] \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)\right]^T \\ &- \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) \nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)^T\right]. \quad (77) \end{aligned}$$

GLM theory indicates that

$$-\nabla^2_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) = X^T P(\Delta^{(g,z)} V^{(g,z)} \Delta^{(g,z)} - (\Delta')^{(g,z)} H^{(g,z)}) X$$

and

$$\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) = X^T P \Delta^{(g,z)} s^{(g,z)}.$$

We begin by computing the first term of (77). The only random matrix among X, P, $\Delta^{(g,z)}$, $V^{(g,z)}$, $(\Delta')^{(g,z)}$, and $H^{(g,z)}$ is P. Therefore, by the linearity of expectation,

$$\begin{aligned} -\mathbb{E}\left[\nabla^2_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)\right] &= \mathbb{E}\left[X^T P(\Delta^{(g,z)} V^{(g,z)} \Delta^{(g,z)} - (\Delta')^{(g,z)} H^{(g,z)}\right] \\ &= X^T T^\theta(1)(\Delta^{(g,z)} V^{(g,z)} \Delta^{(g,z)} - (\Delta')^{(g,z)} H^{(g,z)}) X. \quad (78) \end{aligned}$$

Next, we compute the difference of the last two terms of (77). The (k, l)th entry of this matrix is

$$\Bigg[\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_{\beta(g,z)}\mathcal{L}_z(\theta;m,g,p)\right]^T$$
$$-\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)^T\right]\Bigg][k,l]$$
$$=\left[\mathbb{E}\left[X^TP\Delta^{(g,z)}s^{(g,z)}\right]\mathbb{E}\left[X^TP\Delta^{(g,z)}s^{(g,z)}\right]^T\right][k,l]-\mathbb{E}\left[X^TP\Delta s^{(g,z)}(s^{(g,z)})^T\Delta^{(g,z)}PX^T\right][k,l]$$
$$=\mathbb{E}\left[X[,k]^TP\Delta^{(g,z)}s^{(g,z)}\right]\mathbb{E}\left[X[,l]^TP\Delta^{(g,z)}s^{(g,z)}\right]-\mathbb{E}\left[X[,k]^TP\Delta^{(g,z)}s^{(g,z)}(s^{(g,z)})^T\Delta^{(g,z)}PX[,l]\right]$$
$$=\mathbb{E}\left(\sum_{i=1}^n x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}\right)\mathbb{E}\left(\sum_{j=1}^n x_{jl}P_j\Delta_j^{(g,z)}s_j^{(g,z)}\right)$$
$$-\mathbb{E}\left(\sum_{i=1}^n\sum_{j=1}^n x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}s_j^{(g,z)}\Delta_j^{(g,z)}P_jx_{jl}\right)$$
$$=\sum_{i=1}^n\sum_{j=1}^n\mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}]\mathbb{E}[x_{jl}P_j\Delta_j^{(g,z)}s_j^{(g,z)}]-\sum_{i=1}^n\sum_{j=1}^n\mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}s_j^{(g,z)}\Delta_j^{(g,z)}P_jx_{jl}]$$
$$=\sum_{i=1}^n\sum_{j=1}^n\mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}]\mathbb{E}\left[x_{jl}P_j\Delta_j^{(g,z)}s_j^{(g,z)}\right]-\sum_{i\neq j}\mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}]\mathbb{E}[s_j^{(g,z)}P_j\Delta_j^{(g,z)}x_{jl}]$$
$$-\sum_{i=1}^n\mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}s_i^{(g,z)}\Delta_i^{(g,z)}P_ix_{il}]$$
$$=\sum_{i=1}^n\mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}H_i^{(g,z)}]\mathbb{E}[x_{il}P_i\Delta_i^{(g,z)}H_i^{(g,z)}]-\sum_{i=1}^n\mathbb{E}[x_{ik}P_i^2(\Delta_i^{(g,z)})^2(H_i^{(g,z)})^2x_{il}]$$
$$=\sum_{i=1}^n x_{ik}T_i^\theta(1)^2(\Delta_i^{(g,z)})^2(H_i^{(g,z)})^2x_{il}-\sum_{i=1}^n x_{ik}T_i^\theta(1)(\Delta_i^{(g,z)})^2(H_i^{(g,z)})^2x_{il}$$
$$=X[,k]^TT^\theta(1)^2(\Delta^{(g,z)})^2(H^{(g,z)})^2X[,l]-X[,k]^TT^\theta(1)(\Delta^{(g,z)})^2(H^{(g,z)})^2X[,l]$$

Therefore, we have that

$$\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\right]^T-\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)^T\right]$$
$$=X^TT^\theta(1)^2(\Delta^{(g,z)})^2(H^{(g,z)})^2X-X^TT^\theta(1)(\Delta^{(g,z)})^2(H^{(g,z)})^2X$$
$$=-X^TT^\theta(1)\left(\Delta^{(g,z)}\right)^2\left(H^{(g,z)}\right)^2\left(I-T^\theta(1)\right)X. \quad (79)$$

Combining (77), (78), and (79), we conclude that

$$J_{\beta_{(g,z)}}=(\theta;m,g)=X^TT^\theta(1)(\Delta^{(g,z)}V^{(g,z)}\Delta^{(g,z)}-(\Delta')^{(g,z)}H^{(g,z)})X$$
$$-X^TT^\theta(1)\left(\Delta^{(g,z)}\right)^2\left(H^{(g,z)}\right)^2\left(I-T^\theta(1)\right)X. \quad (80)$$

Submatrix IV (zero-inflated)

Denote submatrix IV by $J_{(\beta_{(g,z)},\beta_m)}(\theta;m,g)$. The formula for submatrix IV is

$$J_{(\beta_{(g,z)},\beta_m)}(\theta;m,g) = -\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\nabla_{\beta_m}\mathcal{L}_z(\theta;m,g,p)\right] + \mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_{\beta_m}\mathcal{L}_z(\theta;m,g,p)\right]^T - \mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\nabla_{\beta_m}\mathcal{L}_z(\theta;m,g,p)\right]^T\Big]. \quad (81)$$

First, we have that

$$-\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\nabla_{\beta_m}\mathcal{L}_z(\theta;m,g,p)\right] = 0, \quad (82)$$

as the derivative in β_m of $\mathcal{L}_z(\theta;m,g,p)$ is a function of β_m, and the derivative in $\beta_{(g,z)}$ of this term is 0. Next, we compute the difference of the last two terms of (81). Entry (k,l) of this matrix is

$$\begin{aligned}
&[\mathbb{E}[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)]\mathbb{E}[\nabla_{\beta_m}\mathcal{L}_z(\theta;m,g,p)]^T \\
&\qquad - \mathbb{E}[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\nabla_{\beta_m}\mathcal{L}_z(\theta;m,g,p)^T]][k,l] \\
&= \left[\mathbb{E}\left[X^T P\Delta^{(g,z)}s^{(g,z)}\right]\mathbb{E}\left[\tilde{X}^T\Delta^m s^m\right]^T\right][k,l] - \mathbb{E}\left[X^T P\Delta^{(g,z)}s^{(g,z)}(s^m)^T\Delta^m\tilde{X}\right][k,l] \\
&= \left[\mathbb{E}\left[X[,k]^T P\Delta^{(g,z)}s^{(g,z)}\right]\mathbb{E}\left[\tilde{X}[,l]^T\Delta^m s^m\right]^T\right] - \mathbb{E}\left[X[,k]^T P\Delta^{(g,z)}s^{(g,z)}(s^m)^T\Delta^m\tilde{X}[,l]\right] \\
&= \mathbb{E}\left(\sum_{i=1}^n x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}\right)\mathbb{E}\left(\sum_{j=1}^n \tilde{x}_{jl}\Delta_j^m s_j^m\right) - \mathbb{E}\left(\sum_{i=1}^n\sum_{j=1}^n x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}\Delta_j^m s_j^m\tilde{x}_{jl}\right) \\
&\quad = \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}]\mathbb{E}[\Delta_j^m s_j^m\tilde{x}_{jl}] - \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}\Delta_j^m s_j^m\tilde{x}_{jl}] \\
&\quad = \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}]\mathbb{E}[\Delta_j^m s_j^m\tilde{x}_{jl}] - \sum_{i\neq j} \mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}]\mathbb{E}[\Delta_j^m s_j^m\tilde{x}_{jl}] \\
&\qquad\qquad - \sum_{i=1}^n \mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}s_i^{(g,z)}\Delta_j^m s_j^m\tilde{x}_{jl}] \\
&\quad = \sum_{i=1}^n \mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}H_i^{(g,z)}]\mathbb{E}[\tilde{x}_{il}\Delta_i^m H_i^m] - \sum_{i=1}^n \mathbb{E}[x_{ik}P_i\Delta_i^{(g,z)}H_i^{(g,z)}\Delta_i^m H_i^m\tilde{x}_{il}] \\
&= \sum_{i=1}^n \left[x_{ik}T_i^\theta(1)\Delta_i^{(g,z)}H_i^{(g,z)}\right]\cdot\left[\Delta_i^m(0)T_i^\theta(0)H_i^m(0)\tilde{x}_{il}(0) + \Delta_i^m(1)T_i^\theta(1)H_i^m(1)\tilde{x}_{il}(1)\right] \\
&\qquad\qquad - \sum_{i=1}^n \left[x_{ik}T_i^\theta(1)\Delta_i^{(g,z)}H_i^{(g,z)}\Delta_i^m(1)H_i^m(1)\tilde{x}_{il}(1)\right] \\
&\qquad = \sum_{s=0}^1\sum_{i=1}^n x_{ik}T_i^\theta(1)H_i^{(g,z)}\Delta_i^{(g,z)}T_i^\theta(s)\Delta_i^m(s)H^m(s)\tilde{x}_{il}(s) \\
&\qquad\qquad - \sum_{i=1}^n \left[x_{il}T_i^\theta(1)\Delta_i^{(g,z)}H_i^{(g,z)}\Delta_i^m(1)H_i^m(1)\tilde{x}_{ik}(1)\right]
\end{aligned}$$

$$= \sum_{s=0}^{1} X[,k]^T T^\theta(1) H^{(g,z)} \Delta^{(g,z)} T^\theta(s) \Delta^m(s) H^m(s) \tilde{X}(s)[,l] \\ - X[,k]^T \Delta^{(g,z)} H^{(g,z)} T^\theta(1) \Delta^m(1) H^m(1) \tilde{X}[,l]. \quad (83)$$

Combining (77), (78), and (79) yields

$$J_{(\beta_{(g,z)},\beta_m)}(\theta; m, g) = \left(\sum_{s=0}^{1} X^T T^\theta(1) H^{(g,z)} \Delta^{(g,z)} T^\theta(s) \Delta^m(s) H^m(s) \tilde{X}(s) \right) \\ - X^T \Delta^{(g,z)} H^{(g,z)} T^\theta(1) \Delta^m(1) H^m(1) \tilde{X}(1). \quad (84)$$

Submatrix VI (zero-inflated)

Denote submatrix VI by $J_{(\beta_{(g,z)},\pi)}(\theta; m, g)$. The formula for $J_{(\beta_{(g,z)},\pi)}(\theta; m, g)$ is

$$J_{(\beta_{(g,z)},\pi)}(\theta; m, g) = \mathbb{E}\left[-\nabla_{\beta_{(g,z)}} \nabla_\pi \mathcal{L}_z(\theta; m, g, p) \right] \\ + \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) \right] \mathbb{E}\left[\nabla_\pi \mathcal{L}_z(\theta; m, g, p) \right] \\ - \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) \nabla_\pi \mathcal{L}_z(\theta; m, g, p) \right]. \quad (85)$$

Recall that $\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) = X^T P \Delta^{(g,z)} s^{(g,z)}$ and $\nabla_\pi \mathcal{L}_z(\theta; m, g, p) = a\left(\sum_{i=1}^n p_i\right) - b$, where

$$a = \frac{1}{\pi} + \frac{1}{1-\pi}, \; b = \frac{n}{1-\pi}.$$

We have that

$$\mathbb{E}\left[-\nabla_{\beta_{(g,z)}} \nabla_\pi \mathcal{L}_z(\theta; m, g, p) \right] = 0, \quad (86)$$

as the derivative in π of $\mathcal{L}_z(\theta; m, g, p)$ is a function of π, and the derivative in $\beta_{(g,z)}$ of this term is 0. Next, we compute the difference of the second two terms of (85). The kth entry of this vector is

$$\mathbb{E}\left[\nabla_\pi \mathcal{L}_z(\theta; m, g, p)\right] \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, x)[k] \right] - \mathbb{E}\left[\nabla_\pi \mathcal{L}(\theta; m, g, p) \nabla_{\beta_{(g,z)}} \mathcal{L}(\theta; m, g, p)[k] \right] \\ = \left(\mathbb{E}\left[a \sum_{i=1}^n p_i - b \right] \right) \left(\mathbb{E}\left[X[,k]^T P \Delta^{(g,z)} s^{(g,z)} \right] \right) - \mathbb{E}\left[\left(a \sum_{i=1}^n p_i - b \right) X[,k]^T P \Delta^{(g,z)} s^{(g,z)} \right] \\ = \left(a \sum_{i=1}^n \mathbb{E}[p_i] - b \right) \left(\sum_{j=1}^n \mathbb{E}[x_{jk} p_j \Delta_j^{(g,z)} s_j^{(g,z)}] \right) \\ - \mathbb{E}\left[\left(a \sum_{i=1}^n p_i - b \right) \left(\sum_{j=1}^n \tilde{x}_{jk} p_j \Delta_j^{(g,z)} s_j^{(g,z)} \right) \right] \\ = a \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[p_i] \mathbb{E}[x_{jk} p_j \Delta_j^{(g,z)} s_j^{(g,z)}] - b \sum_{j=1}^n \mathbb{E}[x_{jk} p_j \Delta_j^{(g,z)} s_j^{(g,z)}]$$

$$- \left[a \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[p_i x_{jk} p_j \Delta_j^{(g,z)} s_j^{(g,z)}] - b \sum_{j=1}^{n} \mathbb{E}[x_{jk} p_j \Delta_j^{(g,z)} s_j^{(g,z)}] \right]$$

$$= a \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{E}[p_i] \mathbb{E}[x_{jk} p_j \Delta_j^{(g,z)} s_j^{(g,z)}] - a \sum_{i \neq j} \mathbb{E}[p_i] \mathbb{E}[x_{jk} p_j \Delta_j^{(g,z)} s_j^{(g,z)}] - a \sum_{i=1}^{n} \mathbb{E}[x_{ik} p_i^2 \Delta_i^{(g,z)} s_i^{(g,z)}]$$

$$= a \sum_{i=1}^{n} \mathbb{E}[p_i] \mathbb{E}[x_{ik} p_i \Delta_i^{(g,z)} s_i^{(g,z)}] - a \sum_{i=1}^{n} \mathbb{E}[x_{ik} p_i^2 \Delta_i^{(g,z)} s_i^{(g,z)}]$$

$$= a \sum_{i=1}^{n} T_i^{\theta}(1) x_{ik} T_i^{\theta}(1) \Delta_i^{(g,z)} s_i^{(g,z)} - a \sum_{i=1}^{n} x_{ik} T_i^{\theta}(1) \Delta_i^{(g,z)} s_i^{(g,z)}$$

$$= a \sum_{i=1}^{n} \left(x_{ik} T_i^{\theta}(1)^2 \Delta_i^{(g,z)} s_i^{(g,z)} - x_{ik} T_i^{\theta}(1) \Delta_i^{(g,z)} s_i^{(g,z)} \right) = a \sum_{i=1}^{n} x_{ik} T_i^{\theta}(1) \Delta_i^{(g,z)} s_i^{(g,z)} \left(T_i^{\theta}(1) - 1 \right)$$

$$= -a \sum_{i=1}^{n} x_{ik} T_i(0) T_i^{\theta}(1) \Delta_i^{(g,z)} H_i^{(g,z)} = -a X[,k]^T w^{(g,z)}. \quad (87)$$

Combining (85), (86), and (87), we conclude that

$$J_{(\beta_{(g,z)},\pi)}(\theta; m, g) = - \left(\frac{1}{\pi} + \frac{1}{1-\pi} \right) X^T w^{(g,z)}. \quad (88)$$

D Statistical accelerations

A key step in the algorithm for computing the pilot parameter estimates (Algorithm 2) is to fit a weighted, no-intercept, univariate GLM with nonzero offset terms and a binary predictor variable (line 20). We derive an analytic formula for the MLE of this GLM for three important pairs of response distributions and link functions: Gaussian response with identity link, Poisson response with log link, and negative binomial response with log link. The GLM that we seek to estimate has responses $[m, m]^T$, predictors $[\underbrace{0, \ldots, 0}_{\text{n}}, \underbrace{1, \ldots, 1}_{\text{n}}]$, offsets $[\hat{f}^m, \hat{f}^m]$, and weights $w = [T_1(0), \ldots, T_n(0), T_1(1), \ldots, T_n(1)]^T$. Throughout, C denotes a universal constant. The log likelihood of this GLM is

$$\mathcal{L}(\beta_1; m) = \textstyle\sum_{i=1}^{n} T_i(0) f_m(m_i; h_m(\beta_1 + \hat{f}_i^m)) + \sum_{i=1}^{n} T_i(1) f_m(m_i; h_m(\hat{f}_i^m))$$

$$= \sum_{i=1}^{n} T_i(1) f_m(m_i; h_m(\beta_1 + \hat{f}_i^m)) + C. \quad (89)$$

Thus, finding the MLE $\hat{\beta}_1$ is equivalent to estimating a GLM with intercept β_1, offsets $\hat{f}^m$, weights $T_i(1)$, and *no* covariate terms. We term such a GLM a *intercept-plus-offset* model. Below, we study intercept-plus-offset models in generality.

Intercept-plus-offset models

Let $\beta \in \mathbb{R}$ be an unknown constant. Let $o_1, \ldots, o_n \sim \mathcal{P}_1$, where $\mathcal{P}_1$ is a distribution. Let $Y_i|o_i, \ldots, Y_n|o_i$ be exponential family-distributed random variables with identity sufficient statistic. Suppose the mean μ_i of $Y_i|o_i$ is given by

$$r(\mu_i) = \beta + o_i,$$

where $r : \mathbb{R} \to \mathbb{R}$ is a strictly increasing, differentiable link function. We call this model the *intercept-plus-offset* model.

We derive the (weighted) log likelihood of this model. Let $w_1, \ldots, w_n \sim \mathcal{P}_2$ be weights, where $\mathcal{P}_2$ is a distribution bounded above by 1 and below by 0. (A special case, which corresponds to no weights, is $w_i = 1$ for all $i \in \{1, \ldots, n\}$.) Throughout, we assume that $y_i w_i$ and $\exp(o_i)w_i$ have finite first moment. Suppose the cumulant-generating function and carrying density of the exponential family distribution are $\psi : \mathbb{R} \to \mathbb{R}$ and $c : \mathbb{R} \to \mathbb{R}$, respectively. The canonical parameter η_i of the ith observation is

$$\eta_i = ([\psi']^{-1} \circ r^{-1})(\beta + o_i) := h(\beta + o_i), \tag{90}$$

and the density f of $Y_i|\eta_i$ is

$$f(y_i; \eta_i) = \exp\{y_i\eta_i - \psi(\eta_i) + c(y_i)\}.$$

The weighted log likelihood is

$$\mathcal{L}(\beta; y_i) = \sum_{i=1}^{n} w_i \log\left[f(y_i; \eta_i)\right] = C + \sum_{i=1}^{n} w_i(y_i\eta_i - \psi(\eta_i)). \tag{91}$$

Our goal is to find the weighted MLE $\hat{\beta}$ of β. We consider three important choices for the exponential family distribution and link function. In the first two cases – Gaussian distribution with identity link and Poisson distribution with log link – we find the *finite-sample* maximizer of (91); by contrast, in the third case – negative binomial distribution with log link – we find an *asymptotically exact* maximizer.

Gaussian

First, consider a Gaussian response distribution and identity link function $r(\mu) = \mu$. The cumulant-generating function ψ is $\psi(\eta) = \eta^2/2$, and so, by (90),

$$h(t) = [\psi']^{-1}(r^{-1}(t)) = [\psi']^{-1}(t) = t.$$

Plugging $\eta_i = h(\beta + o_i) = \beta + o_i$ and $\psi(\eta_i) = (1/2)(\beta + o_i)^2$ into (91), we obtain

$$\mathcal{L}(\beta; y) = \sum_{i=1}^{n} w_i(y_i(\beta + o_i) - (\beta + o_i)^2/2).$$

The derivative of this expression in β is

$$\frac{\partial \mathcal{L}(\beta; y)}{\partial \beta} = \sum_{i=1}^{n} w_i(y_i - \beta - o_i) = \sum_{i=1}^{n} w_i(y_i - o_i) - \beta \sum_{i=1}^{n} w_i.$$

Setting this quantity to 0 and solving for β, we find that the MLE $\hat{\beta}^{\text{gauss}}$ is

$$\hat{\beta}^{\text{gauss}} = \frac{\sum_{i=1}^{n} w_i(y_i - o_i)}{\sum_{i=1}^{n} w_i}.$$

Poisson

Next, consider a Poisson response distribution and log link function $r(\mu) = \log(\mu)$. The cumulant-generating function ψ is $\psi(\eta) = e^{\eta}$. Therefore, by (90),

$$h(t) = [\psi']^{-1}(r^{-1}(t)) = [\psi']^{-1}(\exp(t)) = \log(\exp(t)) = t.$$

Plugging $\eta_i = h(\beta + o_i) = \beta + o_i$ and $\psi(\eta_i) = \exp(\beta + o_i)$ into (91), we obtain

$$\mathcal{L}(\beta; y) = \sum_{i=1}^{n} w_i \left(y_i(\beta + o_i) - \exp(\beta + o_i)\right).$$

The derivative of this function in β is

$$\frac{\partial \mathcal{L}(\beta; y)}{\partial \beta} = \sum_{i=1}^{n} w_i y_i - w_i \exp(\beta + o_i) = \sum_{i=1}^{n} w_i y_i - \exp(\beta) \sum_{i=1}^{n} w_i \exp(o_i).$$

Setting to zero and solving for β, we find that the MLE $\hat{\beta}^{\text{pois}}$ is

$$\hat{\beta}^{\text{pois}} = \log\left(\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i e^{o_i}}\right). \tag{92}$$

Negative binomial

Finally, we consider a negative binomial response distribution (with fixed size parameter $\theta > 0$) and log link function $r(\mu) = \log(\mu)$. The cumulant-generating function ψ is $\psi(\eta) = -\theta \log(1 - e^{\eta})$. The derivative ψ' of ψ is

$$\psi'(t) = \theta\left(\frac{e^t}{1 - e^t}\right) = \frac{\theta}{e^{-t} - 1}.$$

Define the function $\delta : \mathbb{R} \to \mathbb{R}$ by $\delta(t) = -\log\left(\theta/t + 1\right)$. We see that

$$\psi'(\delta(t)) = \frac{\theta}{\exp\left(\log(\theta/t + 1)\right) - 1} = t,$$

implying $\delta = [\psi']^{-1}$. By (90), we have that

$$h(t) = [\psi']^{-1}(r^{-1}(t)) = -\log\left(\frac{\theta}{\exp(t)} + 1\right) = \log\left(\frac{\exp(t)}{\theta + \exp(t)}\right).$$

Therefore,

$$\eta_i = h(\beta + o_i) = \log\left(\frac{\exp(\beta + o_i)}{\theta + \exp(\beta + o_i)}\right) = \beta + o_i - \log\left(\theta + e^{\beta} e^{o_i}\right)$$
$$= \beta - \log\left(\theta + e^{\beta} e^{o_i}\right) + C, \quad (93)$$

and

$$\psi(\eta_i) = -\theta \log\left(1 - \frac{\exp(\beta + o_i)}{\theta + \exp(\beta + o_i)}\right) = -\theta \log\left(\frac{\theta}{\theta + \exp(\beta + o_i)}\right)$$
$$= -\theta \log(\theta) + \theta \log[\theta + \exp(\beta + o_i)] = \theta \log(\theta + e^{\beta} e^{o_i}) + C. \quad (94)$$

Plugging (93) and (94) into (91), the log-likelihood (up to a constant) is

$$\mathcal{L}(\beta; y) = \beta \sum_{i=1}^{n} w_i y_i - \sum_{i=1}^{n} w_i y_i \log(\theta + e^{\beta} e^{o_i}) - \theta \sum_{i=1}^{n} w_i \log(\theta + e^{\beta} e^{o_i})$$
$$= \beta \sum_{i=1}^{n} w_i y_i - \sum_{i=1}^{n} (y_i + \theta) w_i \log(\theta + e^{\beta} e^{o_i}).$$

The derivative of $\mathcal{L}$ in β is

$$\frac{\partial \mathcal{L}(\beta; y)}{\partial \beta} = \sum_{i=1}^{n} w_i y_i - \sum_{i=1}^{n} \frac{w_i (y_i + \theta) e^{\beta} e^{o_i}}{\theta + e^{\beta} e^{o_i}}.$$

Setting the derivative to zero, the equation defining the MLE is

$$e^{\beta} \sum_{i=1}^{n} \frac{w_i e^{o_i} (y_i + \theta)}{e^{\beta} e^{o_i} + \theta} = \sum_{i=1}^{n} w_i y_i. \quad (95)$$

We cannot solve for β in (95) analytically. However, we can derive an asymptotically exact solution. By the law of total expectation,

$$\mathbb{E}\left[\frac{w_i e^{o_i}(y_i + \theta)}{e^{\beta + o_i} + \theta}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{w_i e^{o_i}(y_i + \theta)}{e^{\beta + o_i} + \theta}\middle| (o_i, w_i)\right]\right] = \mathbb{E}\left[\frac{w_i e^{o_i}(e^{\beta + o_i} + \theta)}{e^{\beta + o_i} + \theta}\right] = \mathbb{E}[w_i e^{o_i}];$$

the second equality holds because $\mathbb{E}[y_i | o_i] = \mu_i = e^{\beta + o_i}$. Dividing by n on both sides of (95) and rearranging,

$$\beta = \log\left(\frac{(1/n)\sum_{i=1}^{n} w_i e^{o_i}(y_i + \theta)/(e^{\beta} e^{o_i} + \theta)}{(1/n)\sum_{i=1}^{n} w_i y_i.}\right). \quad (96)$$

By weak LLN, the limit (in probability) of the MLE $\hat{\beta}^{\mathrm{NB}}$ is

$$\hat{\beta}^{\mathrm{NB}} \xrightarrow{P} \log\left(\frac{\mathbb{E}[w_i y_i]}{\mathbb{E}[w_i e^{o_i}]}\right). \tag{97}$$

But the Poisson MLE $\hat{\beta}^{\mathrm{Pois}}$ (92) converges in probability to the same limit:

$$\hat{\beta}^{\mathrm{pois}} = \log\left(\frac{(1/n)\sum_{i=1}^{n} w_i y_i}{(1/n)\sum_{i=1}^{n} w_i e^{o_i}}\right) \xrightarrow{P} \log\left(\frac{\mathbb{E}[w_i y_i]}{\mathbb{E}[w_i e^{o_i}]}\right).$$

Therefore, for large n, we can approximate $\hat{\beta}^{\mathrm{NB}}$ by $\hat{\beta}^{\mathrm{pois}}$.

Suppose the log likelihood is unweighted (i.e., $w_i = 1$ for all i). We present an alternate (and simpler, more general) derivation of the approximate formula for the MLE. Let y_i be a random variable with finite first moment and conditional mean $\mathbb{E}[y_i|o_i] = \exp(\beta + o_i)$. By the law of total expectation,

$$\mathbb{E}[y_i] = \mathbb{E}[\mathbb{E}[y_i|o_i]] = \mathbb{E}[\exp(\beta + o_i)] = \exp(\beta)\mathbb{E}(e^{o_i}),$$

or

$$\beta = \log\left(\frac{\mathbb{E}[y_i]}{E[e^{o_i}]}\right). \tag{98}$$

Assuming the MLE is well-behaved, we have that $\hat{\beta} \xrightarrow{P} \beta$. On the other hand, by WLLN and (98),

$$\log\left(\frac{(1/n)\sum_{i=1}^{n} y_i}{(1/n)\sum_{i=1}^{n} e^{o_i}}\right) \xrightarrow{P} \log\left(\frac{\mathbb{E}[y_i]}{\mathbb{E}[e^{o_i}]}\right) = \beta. \tag{99}$$

Therefore, the LHS of (99) approximates the MLE $\hat{\beta}$ for large n.

Application to GLM-EIV

The GLM that we seek to estimate (89) is an approximate intercept-plus-offset model: $T_1(1), \ldots, T_n(1)$ are the weights $w_1, \ldots, w_n$, and $\hat{f}_1^m, \ldots, \hat{f}_n^m$ are the offsets $o_1, \ldots, o_m$. Of course, $T_1(1), \ldots, T_1(n)$ are in general dependent random variables, as are $\hat{f}_1^m, \ldots, \hat{f}_n^m$. $T_i(1)$ depends on m_i and g_i, as well as the final parameter estimate $(\hat{\pi}, \hat{\beta}_m, \hat{\beta}_g)$, which itself is a function of m and g; the situation is similar for the $\hat{f}_i^m$s. In practice, we find that the intercept-plus-offset model is very good approximation to the GLM (89), especially when the number of cells n is large. Additionally, we note that the GLM (89) is fit as a subroutine of the algorithm for producing pilot parameter estimates (Algorithm 2). The quality of the pilot parameter estimates does not affect the validity of the estimation and inference procedures (Algorithm 1), barring issues related to convergence to local optima (which are outside the scope fo the current work).

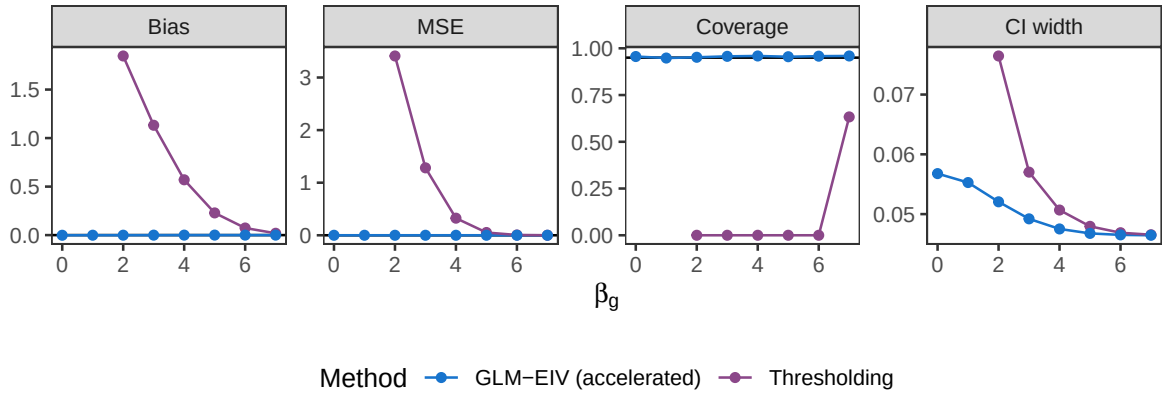


Figure 9: Additional simulation results on Gaussian data. GLM-EIV (accelerated) outperformed the thresholding method on bias, mean squared error, confidence interval coverage rate, and confidence interval width metrics.

E Additional simulation study

We ran an additional simulation study in which we modeled the gene and gRNA expressions using a Gaussian distribution with identity link. We generated data on $n = 150,000$ cells, fixing the target of inference β_1^m to -4 and the probability of perturbation π to 0.05. We included "sequencing batch" (modeled as a Bernoulli-distributed variable) and "sequencing depth" (modeled as a Poisson-distributed variable) as covariates in the model. We did not include sequencing depth as an offset because use of the identity link renders offsets meaningless. We varied β_1^g over a grid on the interval $[0, 7]$. We generated $n_{\text{sim}} = 1,000$ synthetic datasets for each value of β_1^g. We applied accelerated GLM-EIV and thresholded regression to the simulated data. We assessed these methods on the metrics of bias, mean squared error, confidence interval coverage rate, and confidence interval width. We found that accelerated GLM-EIV outperformed the thresholding method: the former method exhibited smaller bias, smaller mean squared error, higher confidence interval coverage rate, and smaller confidence interval width than the latter method (Figure 9).

F Real data analysis details

We performed quality control. We kept cells with percent mitochondrial reads less than 8%. For the Gasperini dataset, we furthermore kept cells with UMI and gRNA counts in the 5th to 95th percentiles to reduce the effect of outliers. (We did not perform this latter filter on the Xie data, as the latter dataset appeared to be less noisy and contained fewer cells.) Next, we kept genes expressed in at least 10% of cells with a mean expression level of at least 1. We did not QC the gRNAs.

The quality-controlled Xie dataset contained $n = 101,508$ cells, $1,030$ genes, and 516

gRNAs. There were $50,000$ *in silico* negative control pairs and 681 candidate *cis* pairs. The quality-controlled Gasperini dataset contained $n = 170,645$ cells, $2,079$ genes, $6,598$ gRNAs. There were $17,028$ candidate *cis* pairs, $97,818$ negative control pairs, and 322 positive control pairs.