Exponential family measurement error models for single-cell CRISPR screens

Abstract

CRISPR genome engineering and single-cell RNA 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 through theoretical and real data analyses 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, challenging-to-select tuning parameter. To overcome these difficulties, we introduce GLM-EIV ("GLM-based errors-in-variables"), a new method for single-cell CRISPR screen analysis. GLM-EIV extends the classical errors-in-variables model to responses and noisy predictors that are exponential family-distributed and potentially impacted by the same set of confounding variables. 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, demonstrating improved performance in challenging problem settings.

Keywords: genome engineering, GWAS, GLM, latent variable model, cloud computing

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 (Rothgangl et al. 2021; Musunuru et al. 2021) and transforming biological discovery (Przybyla and Gilbert 2021). Recently, scientists have paired CRISPR genome engineering with single-cell RNA sequencing (Dixit et al. 2016; Datlinger et al. 2017). The resulting assays, known as a "single-cell CRISPR screens," link genetic perturbations in individual cells to changes in gene expression, enabling scientists to causally map genome wide association study (GWAS) variants to their target genes at genome-wide scale (Morris et al. 2021).

Despite their promise, single-cell CRISPR screens present substantial statistical challenges. A major difficulty is that the "treatment" – i.e., the presence or absence of a CRISPR perturbation – is assigned randomly to cells and is not directly observable. As a consequence, one cannot know with certainty which cells were perturbed. Instead, one must leverage an indirect, quantitative proxy of perturbation presence or absence to "guess" which cells received a perturbation. This indirect proxy takes the form of a so-called guide RNA count, with higher counts indicating that a cell is more likely to have been perturbed. The standard approach to single-cell CRISPR screen analysis is to impute perturbation assignments onto the cells by simply thresholding the guide RNA counts; using these imputations, one can attempt to estimate the effect of the perturbation on gene expression. We call this standard approach "thresholded regression" or the "thresholding method."

We study estimation and inference in single-cell CRISPR screens from a statistical perspective, formulating the data generating mechanism using a new class of measurement error models. We assume that the response variable y is a GLM of an underlying predictor variable x^* and vector of confounders z. We do not observe x^* directly; rather, we observe a noisy version x of x^* that itself is a GLM of x^* and the same set of confounders z. The goal of the analysis is to estimate the effect of x^* on y using the observed data (x, y, z) only. In the context of the biological application, x^*, x, y, and z are CRISPR perturbations, guide RNA counts, gene expressions, and technical confounders, 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 a simplified Gaussian setting. Second, we introduce a new method, GLM-EIV ("GLM-based errors-in-variables"), for single-cell CRISPR screen analysis. GLM-EIV extends the classical errors-in-variables model (Carroll et al. 2006) to responses and noisy predictors that are exponential family-distributed and potentially impacted by the same set of confounding variables. GLM-EIV thereby implicitly estimates the probability that each cell was perturbed, obviating the need to explicitly impute perturbation assignments via thresholding. We implement several statistical accelerations (that likely 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. We find that in some settings, GLM-EIV outperforms thresholded regression by a considerable margin; in other settings the two methods work best in conjunction, with GLM-EIV providing a statistically principled and empirically effective procedure for selecting the threshold.

2 Assay background

There are several broad classes of single-cell CRISPR screen assays, each suited to answer a different set of biological questions (Gasperini et al. 2019; Datlinger et al. 2021; Mimitou et al. 2019). In this work we focus on high-multiplicity of infection (MOI) single-cell CRISPR screens, which we motivate and describe here. 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 relevance to the current work). GWAS have revealed that the majority ($> 90\%$) of variants associated with diseases lie outside genes and inside enhancers (Gallagher and Chen-Plotkin 2018). These noncoding variants are thought to contribute to disease by modulating the expression of 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 (Gasperini et al. 2020; Morris et al. 2021).

High MOI single-cell CRISPR screens are the most promising biotechnology for solving this challenge. 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 (this work).

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.

The genomics literature has produced a few applied methods for single-cell CRISPR screen analysis (Gasperini et al. 2019; Xie et al. 2019; Barry et al. 2021). Gasperini et al. applied negative binomial GLMs (as implemented in the Monocle software; Trapnell et al. 2014) to carry out the differential expression analysis described above. Xie et al., by contrast, applied chi-squared-like tests of independence for this purpose. Both of these approaches have limitations: the former is not robust to misspecification of the gene expression model, and the latter is unable to correct for the presence of technical confounders. Recently, Barry et al. introduced SCEPTRE, a custom implementation of the conditional randomization test (Candès et al. 2018; Liu et al. 2021) tailored to single-cell CRISPR screen data. SCEPTRE simultaneously adjusts for confounder presence and ensures robustness to expression model misspecification, overcoming limitations of the prior methods and demonstrating state-of-the-art sensitivity and specificity on single-cell CRISPR screen data. In this work we tackle a set of analysis challenges that are complimentary to those addressed by SCEPTRE. Most importantly, we seek to account for the fact that the perturbation is measured with noise, an issue that all available methods (including SCEPTRE) assume away via thresholding. Additionally, we seek to *estimate* (with confidence) the effect size of a perturbation on gene expression change, an objective that is challenging to attain within the nonparametric hypothesis testing framework of SCEPTRE.

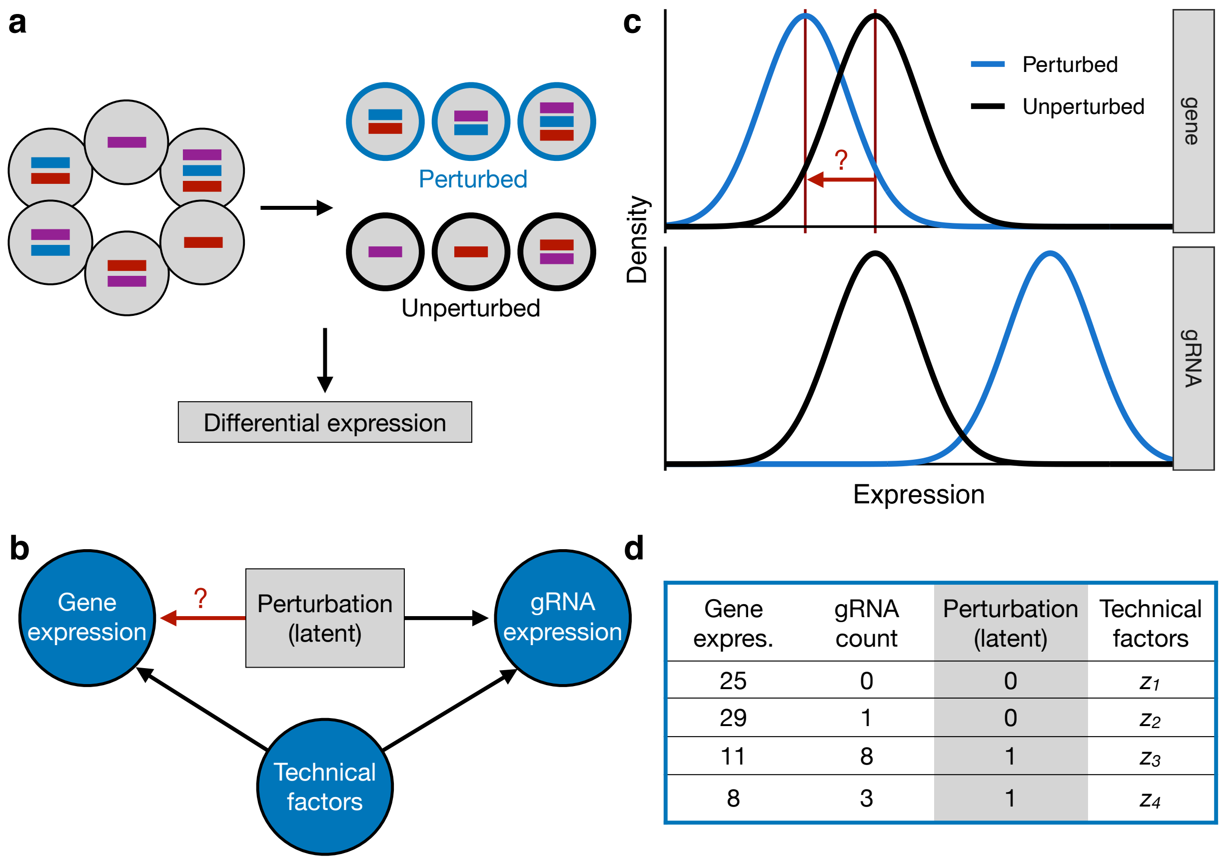

Figure 1: **Experimental design and analysis challenges**: **a,** Experimental design. For a given perturbation (e.g., the perturbation indicated in blue), we partition the cells into two groups: perturbed and unperturbed. Next, for a given gene, we conduct a differential expression analysis across the 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.

3 Analysis challenges and proposed statistical model

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, 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 (Replogle et al. 2020). 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.

We propose to model the single-cell CRISPR screen data-generating process using a pair of GLMs. 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 $p_i \in \{0, 1\}$ indicate perturbation presence or absence; 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 (i.e., the library size or sequencing depth); let d_i^g be the gRNA library size; and finally, let $z_i \in \mathbb{R}^{d-1}$ be the cell-specific technical factors (e.g., sequencing batch, percent mitochondrial reads, etc.) The letters "m," "g", and "d" stand for "mRNA," "gRNA," and "depth," respectively.

Building on the work of several previous authors (Townes et al. 2019; Svensson 2020; Hafemeister and Satija 2019), Sarkar and Stephens (2021) proposed a simple strategy for modeling 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. Sarkar and Stephens' framework enjoys strong theoretical and empirical support; therefore, we generalize their approach to model *both* gene and gRNA modalities in single-cell CRISPR screen experiments. To this end, we assume that the gene expression counts are given by

$$m_i | (p_i, z_i, d_i^m) \sim \text{NB}_{\theta^m}(\mu_i^m); \quad \log(\mu_i^m) = \beta_0^m + \beta_1^m p_i + \gamma_m^T z_i + \log(d_i^m), \tag{1}$$

where (i) $\text{NB}_{\theta^m}(\mu_i^m)$ is a negative binomial distribution with mean μ_i^m 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. Similarly, we model the gRNA counts by

$$g_i | (p_i, z_i, d_i^g) \sim \text{NB}_{\theta^g}(\mu_i^g); \quad \log(\mu_i^g) = \beta_0^g + \beta_1^g p_i + \gamma_g^T z_i + \log(d_i^g), \tag{2}$$

where μ_i^g, θ^g, β_0^g, β_1^g, γ_g, and d_i^g are analogous. We use a negative binomial GLM to model the gRNA counts as well as the gene expressions because the gRNA transcripts are generated via the same biological mechanism as the gene transcripts (Datlinger et al. 2017; Hill et al. 2018). Finally, we model the marginal perturbation probability as

$$p_i \sim \text{Bern}(\pi); \quad \pi \in (0, 1/2], \tag{3}$$

where p_i is unobserved. Together, (1 - 3) define the negative binomial GLM-EIV model.

The log-transformed sequencing depth $\log(d_i^m)$ is included as an offset term in (1) so that $\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i$ can be interpreted as a relative expression. Exponentiating both sides of (1) reveals that the mean gene expression μ_i^m of the ith cell is $\exp\left(\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i\right) d_i^m$. Because d_i^m is the sequencing depth, $\exp\left(\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i\right)$ 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 fold change in expression in response to the perturbation, controlling for the technical factors. Fold change in this context is the ratio of the mean gene expression in perturbed cells to the mean gene expression in unperturbed cells. Hence, $\exp(\beta_1^m) = 1$ (i.e., $\beta_1^m = 0$) indicates no change in expression, whereas $\exp(\beta_1^m) > 1$ (i.e., $\beta_1^m > 0$) and $\exp(\beta_1^m) < 1$ (i.e., $\beta_1^m < 0$) indicate an increase and decrease in expression, respectively.

In this work we analyze two large-scale, high MOI, single-cell CRISPR screen datasets published by Gasperini et al. (2019) and Xie et al. (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 several hundred positive control, gene-targeting perturbations and 50 non-targeting, negative control perturbations to assess method sensitivity and specificity.

4 Analysis of the thresholding method

We study thresholding from empirical and theoretical perspectives, highlighting several limitations of the approach. In the context of the negative bionomial GLM-EIV model introduced above (1-3), the thresholding method leverages the gRNA counts (2) to impute the latent perturbation indicator (3), thereby reducing the full data generating process to a single, gene expression model (1). We study Gasperini et al.'s variant of the thresholding method (i.e., thresholded negative binomial regression), as this version of the thresholding method relates most closely to GLM-EIV. The 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 given by $\hat{p}_i = 0$ if $g_i < c$ and $\hat{p}_i = 1$ otherwise.

2. Assume that m_i is related to $\hat{p}_i, d_i^m$, and z_i through the following GLM:

$$m_i | (\hat{p}_i, z_i, d_i^m) \sim \text{NB}_{\theta^m}(\mu_i^m); \quad \log(\mu_i^m) = \beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i + \log\left(d_i^m\right). \qquad (4)$$

   The model (4) is equivalent to the model (2), but the latent perturbation indicator p_i has been replaced by the imputed perturbation indicator $\hat{p}_i$.

3. Fit a GLM to (4) to obtain an estimate and CI for the target of inference β_1^m.

4.1 Empirical challenges of the thresholding method

To shed light on empirical challenges of the thresholding method, we applied thresholded negative binomial regression to analyze the set of 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 threshold choice, we deployed the method using three different choices for the threshold: 1, 5, and 20. We found that the chosen threshold substantially impacted the results (Figure 2a-b): estimates for fold change produced by threshold = 1 were smaller in magnitude (i.e., closer to the baseline of 1) than those produced by threshold = 5. (Figure 2a.) On the other hand, estimates produced by threshold = 5 and threshold = 20 were more concordant (Figure 2b).

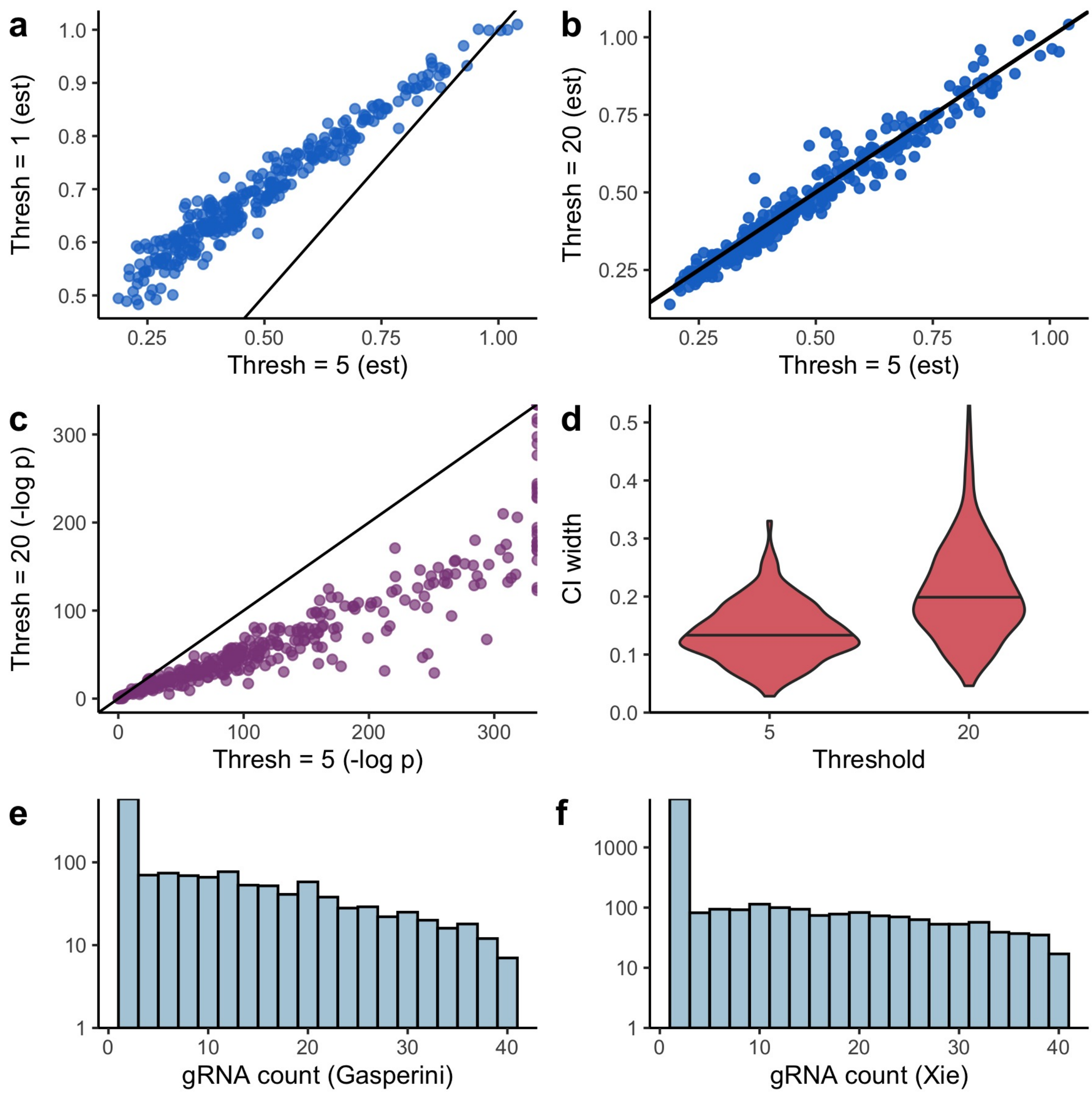

Figure 2: **Empirical challenges of thresholded regression. a-b,** Fold change estimates produced by threshold = 1 versus threshold = 5 (a) and threshold = 20 versus threshold = 5 (b). The selected threshold substantially impacts the results. **c-d,** p-values (c) and CI widths (d) produced by threshold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

We reasoned that thresholded regression systematically underestimated true effect sizes on the positive control pairs, especially for threshold = 1. For a given perturbation, the 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 causes the variance of the estimator to increase. To investigate, we compared p-values and confidence intervals produced by threshold = 5 and threshold = 20 for the target of inference β_1^m. We found that threshold = 5 yielded smaller (i.e., more significant) p-values and narrower confidence intervals than did threshold = 20 (Figures 2c-d). We concluded that the threshold controls a bias-variance tradeoff: as the threshold increases, the bias of the estimator decreases and the 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.

4.2 Theoretical challenges of the thresholding method

Next, we studied the thresholding method from a theoretical perspective, recovering in a simplified Gaussian setting phenomena revealed in the empirical analysis. Suppose we observe gRNA expression and gene expression data $(g_1, m_1), \ldots, (g_n, m_n)$ on n cells from the following linear model:

$$m_i = \beta_0^m + \beta_1^m p_i + \epsilon_i; \quad g_i = \beta_0^g + \beta_1^g p_i + \tau_i; \quad p_i \sim \text{Bern}(\pi); \quad \epsilon_i, \tau_i \sim N(0,1), \tag{5}$$

where p_i, τ_i, and ϵ_i are independent. For a given threshold $c \in \mathbb{R}$, the imputed perturbation assignment $\hat{p}_i$ is $\hat{p}_i = \mathbb{I}(g_i \geq c)$. The thresholding estimator $\hat{\beta}_1^m$ is the OLS solution, i.e. $\hat{\beta}_1^m = \left[\sum_{i=1}^n (\hat{p}_i - \overline{\hat{p}})^2\right]^{-1} \left[\sum_{i=1}^n (\hat{p}_i - \overline{\hat{p}})(m_i - \overline{m})\right]$. We derive the almost sure limit of $\hat{\beta}_1^m$:

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

$$\hat{\beta}_1^m \xrightarrow{a.s.} \beta_1^m \left(\frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} \right) \equiv \beta_1^m \gamma(\beta_1^g, \pi, c, \beta_0^g), \tag{6}$$

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

The function $\gamma : \mathbb{R}^4 \to \mathbb{R}$ does not depend on the gene expression parameters β_1^m or β_0^m. The asymptotic relative bias $b : \mathbb{R}^4 \to \mathbb{R}$ of $\hat{\beta}_1^m$ is given by

$$b(\beta_1^g, \pi, c, \beta_0^g) \equiv \frac{1}{\beta_1^m}\left(\beta_1^m - \lim_{\text{a.s.}} \hat{\beta}_1^m\right) = 1 - \gamma(\beta_1^g, \pi, c, \beta_0^g).$$

Having derived an exact expression for the asymptotic relative bias of $\hat{\beta}_1^m$, we can prove several results about this quantity. We fix π to $1/2$ for simplicity. (In reality, π is smaller, but the relevant statistical phenomena emerge for $\pi = 1/2$.) First, the thresholding estimator strictly underestimates (in absolute value) the true value of β_1^m over all choices of the threshold c and over all values of the regression coefficients (β_0^m, β_1^m) and (β_0^g, β_1^g). This phenomenon, called attenuation bias, is a common attribute of estimators that ignore measurement in errors-in-variables models (Stefanski 2000). Second, the magnitude of the bias decreases monotonically in β_1^g, comporting with the intuition that the problem becomes easier as the gRNA mixture distribution becomes increasingly well-separated. Third, the Bayes-optimal decision boundary $c_{\text{bayes}} \in \mathbb{R}$ (i.e., the most accurate decision boundary for classifying cells) is a critical value of the bias function. Finally, and most subtly, there is no universally applicable rule for selecting a threshold that yields minimal bias: when β_1^g is small, setting the threshold to an arbitrarily large number yields smaller bias than setting the threshold to the Bayes decision boundary; when β_1^g is large, the reverse is true. Appendix A contains detailed proposition statements and proofs of these results.

Next, we studied the variance of the thresholding estimator, considering a slightly simpler model for this purpose. Suppose the intercepts in (5) are fixed at 0 (i.e., $\beta_0^m = \beta_0^g = 0$). For notational simplicity we write $\beta_m = \beta_1^m$ and $\beta_g = \beta_1^g$. The thresholding estimator $\hat{\beta}_m$ is the no-intercept OLS solution $\hat{\beta}_m = \left[\sum_{i=1}^n \hat{p}_i^2\right]^{-1} \left[\sum_{i=1}^n \hat{p}_i m_i\right]$. The following proposition derives the scaled, asymptotic distribution of $\hat{\beta}_m$:

Proposition 2. *The limiting distribution of $\hat{\beta}_m$ is*

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

where

$$l \equiv \beta_m \omega \pi / [\zeta(1-\pi) + \omega\pi]; \quad \mathbb{E}[\hat{p}_i] = \pi\omega + (1-\pi)\zeta; \quad \omega \equiv \Phi(\beta_g - c); \quad \zeta \equiv \Phi(-c).$$

This proposition yields an asymptotically exact bias-variance decomposition for $\hat{\beta}_m$: as the threshold tends to infinity, the bias decreases and the variance increases (Appendix A). Overall, our empirical and theoretical analyses indicate that thresholded regression poses a basic difficulty: selecting a good threshold is challenging, and even if we have selected a threshold that is (in some sense) optimal, the resulting estimator still may suffer from high bias or high variance. These considerations motivate our research question: does modeling the gRNA counts directly, thereby circumventing thresholding, facilitate estimation and inference in single-cell CRISPR screen analysis?

5 GLM-based errors-in-variables (GLM-EIV)

We introduce the general GLM-EIV model, which generalizes the negative binomial GLM-EIV model (1-3) to arbitrary exponential family response distributions and link functions, thereby providing much greater modeling flexibility. We derive efficient methods for estimation and inference in this model and develop a pipeline to deploy the model at-scale. Appendix C develops a parallel methodology in which the gRNA counts are zero-inflated.

5.1 Model

General model This section is more technical than the previous ones; a first-time reader can skip to Section 5.2 without loss of information relevant to the high-level narrative. 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 the log-transformed library sizes (i.e., $\log(d_i^m)$ and $\log(d_i^g)$, respectively).

We use a pair of GLMs 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 $l_i^m \equiv \langle \tilde{x}_i, \beta_m \rangle + o_i^m$. Next, let the mean μ_i^m of the ith observation be $r_m(\mu_i^m) \equiv l_i^m$, 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 $\eta_i^m = ([\psi_m']^{-1} \circ r_m^{-1})\,(l_i^m) \equiv h_m(l_i^m)$. 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 $f_m(m_i; \eta_i^m) = \exp\{m_i \eta_i^m - \psi_m(\eta_i^m) + c_m(m_i)\}$. The function c_m appears as a constant 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, i.e. $l_i^g \equiv \langle \tilde{x}_i, \beta_g \rangle + o_i^g$, $r_g(\mu_i^g) \equiv l_i^g$, and $\eta_i^g = \left([\psi_g']^{-1} \circ r_g^{-1}\right)(l_i^g) \equiv h_g(l_i^g)$. The density f_g of g_i given the canonical parameter is $f_g(m_i; \eta_i^g) = \exp\{g_i \eta_i^g - \psi_g(\eta_i^g) + c_g(g_i)\}$. 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 $\tilde{x}_i(j) \equiv [1, j, z_i]^T$ 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., $l_i^m(j) \equiv \langle \tilde{x}_i(j), \beta_m \rangle + o_i^m$, $\eta_i^m(j) \equiv h_m(l_i^m(j))$, and $\mu_i^m(j) \equiv r_m^{-1}(l_i^m(j))$. Let the corresponding gRNA quantities $l_i^g(j)$, $\eta_i^g(j)$, and $\mu_i^g(j)$ be defined analogously. Next, let $X \in \mathbb{R}^{n \times (d-1)}$ be the observed design matrix, and 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.

$$X \equiv \begin{bmatrix} 1 & z_1 \\ \vdots & \vdots \\ 1 & z_n \end{bmatrix}; \quad \tilde{X} \equiv \begin{bmatrix} 1 & p_1 & z_1 \\ \vdots & \vdots & \vdots \\ 1 & p_n & z_n \end{bmatrix} = \begin{bmatrix} \tilde{x}_1^T \\ \vdots \\ \tilde{x}_n^T \end{bmatrix}.$$

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 $\tilde{X}$, and let $[\tilde{X}(0)^T, \tilde{X}(1)^T]^T$ denote the $\mathbb{R}^{2n \times d}$ matrix that results from vertically concatenating $\tilde{X}(0)$ and $\tilde{X}(1)$. Furthermore, define $m := [m_1, \ldots, m_n]$, and let g, p, o^m, and o^g be defined analogously. Finally, let $[m, m]^T \in \mathbb{R}^{2n}$ be the vector that results from concatenating m to itself, i.e. $[m, m]^T \equiv [m_1, \ldots, m_n, m_1, \ldots, m_n]$, 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

$$f(m_i, g_i, p_i; \theta) = f_p(p_i) f_m(m_i | p_i) f_g(g_i | p_i) = \pi^{p_i}(1 - \pi)^{1 - p_i} f_m(m_i; \eta_i^m) f_g(g_i; \eta^g). \quad (7)$$

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

$$f(m_i, g_i; \theta) = (1 - \pi) f_m(m_i; \eta_i^m(0)) f_g(g_i; \eta_i^g(0)) + \pi f_m(m_i; \eta_i^m(1)) f_g(g_i; \eta_i^g(1)). \quad (8)$$

Therefore, the log-likelihood is

$$\mathcal{L}(\theta; m, g) = \sum_{i=1}^{n} \log \left[(1-\pi) f_m(m_i; \eta_i^m(0)) f_g(g_i; \eta_i^g(0)) + \pi f_m(m_i; \eta_i^m(1)) f_g(g_i; \eta_i^g(1)) \right]. \quad (9)$$

We see from (8) 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 the simple errors-in-variables model (when the predictor is binary); the latter is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i; \quad x_i = x_i^* + \tau_i, \quad (10)$$

where, $x_i^* \sim \text{Bern}(\pi), \epsilon_i, \tau_i \sim N(0,1)$, and ϵ_i,τ_i, and x_i^* are independent. GLM-EIV extends (10) in at least three directions: first, GLM-EIV allows y_i and x_i to follow exponential family (i.e, not just Gaussian) distributions; second, GLM-EIV allows y_i and x_i to be related to x_i^* through arbitrary (i.e., not just linear) link functions; and finally, GLM-EIV allows confounders z_i to impact both x_i and y_i. Therefore, x_i and y_i can be conditionally *dependent* given x_i^*, enabling GLM-EIV to capture more complex dependence relationships between x_i and y_i than is possible in (10) or other standard measurement error models.

5.2 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) in each cell. The membership probability $T_i(1)$ of cell $i \in \{1, \ldots, n\}$ given the current parameter estimates $(\beta_m^{(\text{t})}, \beta_g^{(\text{t})}, \pi^{(\text{t})})$ and observed data (m_i, g_i) is

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

We can calculate this quantity by applying (i) Bayes rule, (ii) the conditional independence property of M_i and G_i, (iii) the density of M_i and G_i, and (iv) a log-sum-exp-type trick to ensure numerical stability. Next, we produce updated estimates $\pi^{(t+1)}$, $\beta_g^{(t+1)}$, and $\beta_m^{(t+1)}$ of the parameters by maximizing the M step objective function. It turns out that maximizing this 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 log likelihood (9) converges (Appendix B). Our EM algorithm is reminiscent of (but distinct from) that of Ibrahim (1990), who also applied weighted GLM solvers to carry out an M step of an EM algorithm.

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 bootstrap. 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 (Louis 1982; Appendix B). Leveraging this analytic formula, we can calculate standard errors quickly, enabling us to perform inference in practice on real, large-scale data.

5.3 Statistical accelerations and computational infrastructure

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 improve chances of 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

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; gRNA expression distribution f_g and link function r_g.

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.

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

  end for

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

simulated data.) We instead devised a strategy to produce a highly accurate pilot estimate 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" (that could be of independent utility for accelerating other single-cell methods) on top of one another; details are deferred to Appendix D. Overall, the statistical accelerations reduce the number of GLMs that we must fit to < 10 in most cases.

Next, we developed a computational infrastructure to apply GLM-EIV to large-scale, single-cell CRISPR screen data. The infrastructure leverages `Nextflow`, a programming language that facilitates building data-intensive pipelines (DI Tommaso et al. 2017), and `ondisc`, an `R/C++` package that we developed (in a separate project) to facilitate large-scale computing on single-cell data (preprint forthcoming). `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 built a Docker-containerized pipeline for deploying GLM-EIV at-scale. The pipeline aggressively recycles computation when possible, saving a considerable amount of compute; see Appendix D.3 for details. Overall, the statistical accelerations and computational infrastructure make the deployment of GLM-EIV to large-scale single-cell CRISPR screen quite feasible.

6 Simulation study

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 using realistic parameter values, 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 problem settings. 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, accelerated GLM-EIV, and thresholded regression. We used the Bayes-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 before applying the methods (Ripley et al. 2013). We display the results of the simulation study in Figure 3. Columns correspond to distributions (i.e., Poisson, NB with known θ, and NB with unknown θ), and 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 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. We additionally found that accelerated GLM-EIV performed as well as vanilla GLM-EIV on all statistical metrics (rows 1-4) despite having substantially lower computational cost (bottom row). In fact, the execution time of accelerated GLM-EIV was almost 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.,

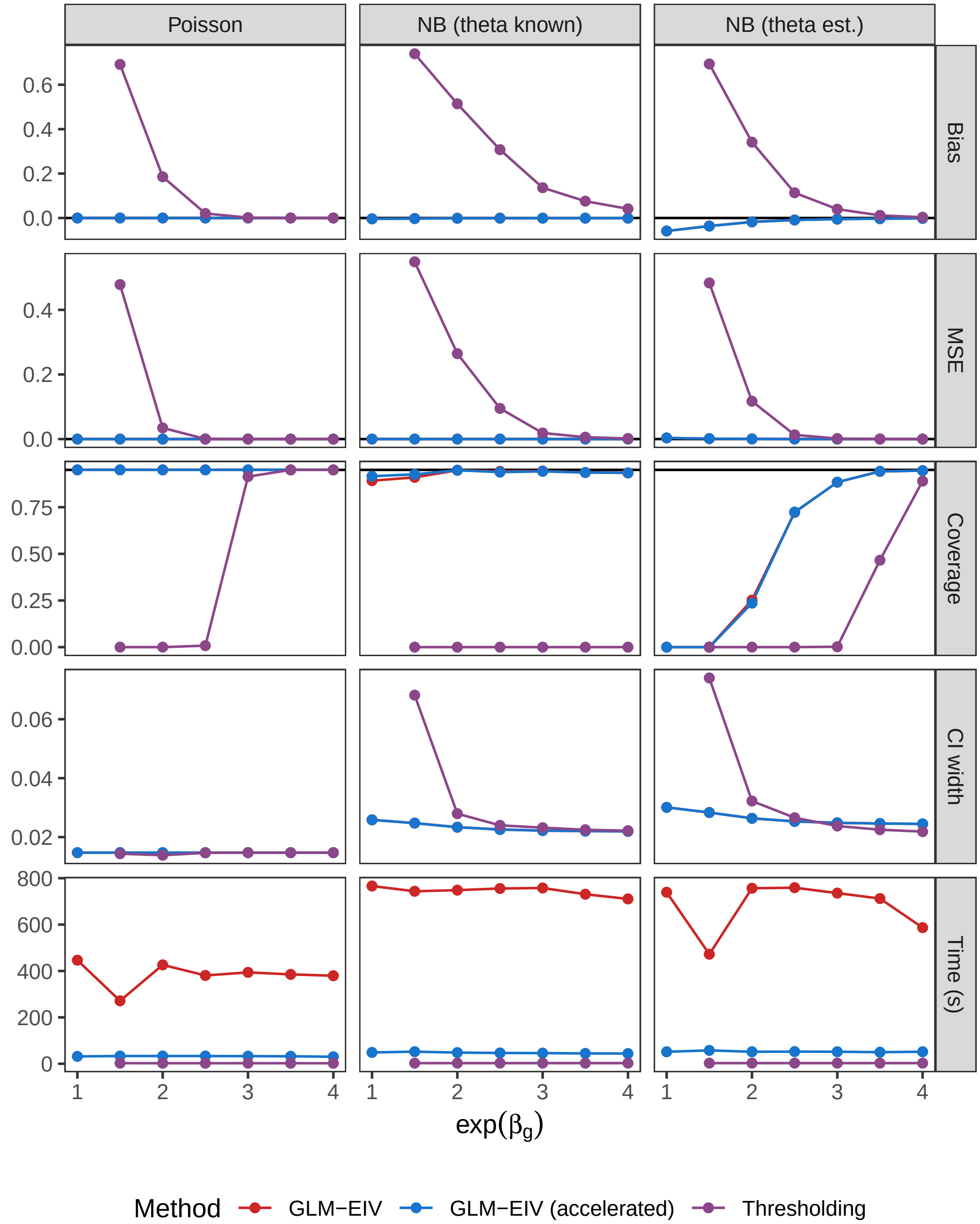

Figure 3: **Simulation study**. Columns correspond to distributions (Poisson, NB with known θ, NB with estimated θ), and rows correspond to metrics (bias, MSE, coverage, CI width, and time). Methods are shown in different colors; GLM-EIV (red) is masked by accelerated GLM-EIV (blue) in several panels. 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).

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 canceled, yielding less biased estimates. GLM-EIV exhibited worse performance under unknown θ than known θ, likely due to poor θ estimation. We note that GLM-EIV and the thresholding method in principle are compatible with *any* θ estimation procedure, including those based on more sophisticated techniques, such as regularization (Hafemeister and Satija 2019). We defer rigorous investigation of the impact of different θ estimation strategies on these methods to future work.

7 Data analysis

Leveraging our computational infrastructure, we applied GLM-EIV and the thresholding method to analyze the entire Gasperini and Xie datasets. We report only the most important aspects of the analysis and results in the main text; full details are available in Appendix F. We set the threshold in the thresholding method to the approximate Bayes-optimal decision boundary, as our theoretical analyses and simulation studies indicated that the Bayes-optimal decision boundary is a good choice for the threshold when the gRNA count distribution is well-separated. Operating under the assumption that the effect of the perturbation on gRNA expression is similar across pairs, we leveraged the fitted GLM-EIV models to approximate the Bayes boundary in the following way: we (i) sampled several hundred gene-perturbation pairs, (ii) extracted the fitted values $\hat{\beta}_g$ and $\hat{\pi}$ from the GLM-EIV models fitted to these pairs, (iii) computed the median $\overline{\hat{\beta}_g}$ and $\overline{\hat{\pi}}$ across the $\hat{\beta}_g$s and $\hat{\pi}$s, and (iv) used $\overline{\hat{\beta}_g}$ and $\overline{\hat{\pi}}$ to estimate a dataset-wide Bayes-optimal decision boundary. We repeated this procedure on both datasets, yielding a threshold of 3 for Gasperini and 7 for Xie. These thresholds were close to the thresholds used in the original publications, which were selected in a more heuristic way.

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 4a-b): estimates, CI coverage rates, and CI widths were concordant. CI coverage rates, which ranged from 87.7%-91.2%, were slightly below the nominal rate of 95%, likely due to mild model misspecification. The estimated effect of the perturbation on gene expression $\exp(\hat{\beta}_1^g)$ was unexpectedly large: the 95% CI for this parameter was $[4306, 5186]$ and $[300, 316]$ on the Gasperini and Xie data, respectively. We reasoned that the datasets lay in an "easy" region of the parameter space, making thresholding a tenable strategy (provided the threshold is selected well). However, this was not obvious *a priori* and may not be the case for other datasets. We note that GLM-EIV produced outlier estimates (defined as estimated fold change < 0.75 or > 1.25) on a small ($< 2.5\%$ on Gasperini, $< 0.05\%$ on Xie) number of pairs consisting of a handful of genes, likely due to non-global EM convergence. These outliers are not plotted in Figures 4a-b but were used to compute the CI coverage reported in the inset tables.

To evaluate performance of GLM-EIV versus thresholding in more challenging settings, we artificially increased the difficulty of the perturbation assignment problem by generating partially-synthetic datasets. First, for a given pair, we sampled gRNA counts directly from the fitted GLM-EIV model. Next, to simulate elevated background contamination, we sampled gRNA counts from a slightly modified version of the fitted model in which we increased the mean gRNA expression of *unperturbed* cells while holding constant the mean gRNA expression of *perturbed* cells. We defined a parameter called "excess background contamination" (normed to take values in $[0, 1]$) to quantify the relative distance between the unperturbed and perturbed gRNA count distributions. We held 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 4c-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 (This result replicated across nearly all pairs; average relative difference 0.003.) We additionally observed that as excess background contamination increased, the performance of thresholded regression degraded considerably while that of GLM-EIV remained stable.

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 4d). Next, we computed the median REC across all positive control pairs (Figure 4e; upper and lower bands indicate the pointwise interquartile range of the 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, letting $\hat{\beta}_1^m$ denote the estimate obtained on the raw data, we computed the CI coverage of $\hat{\beta}_1^m$ as a function of excess contamination. Under the assumption that $\hat{\beta}_1^m$ is close to the true parameter β_1^m, the CI coverage of the former is similar to that of the latter. We computed the

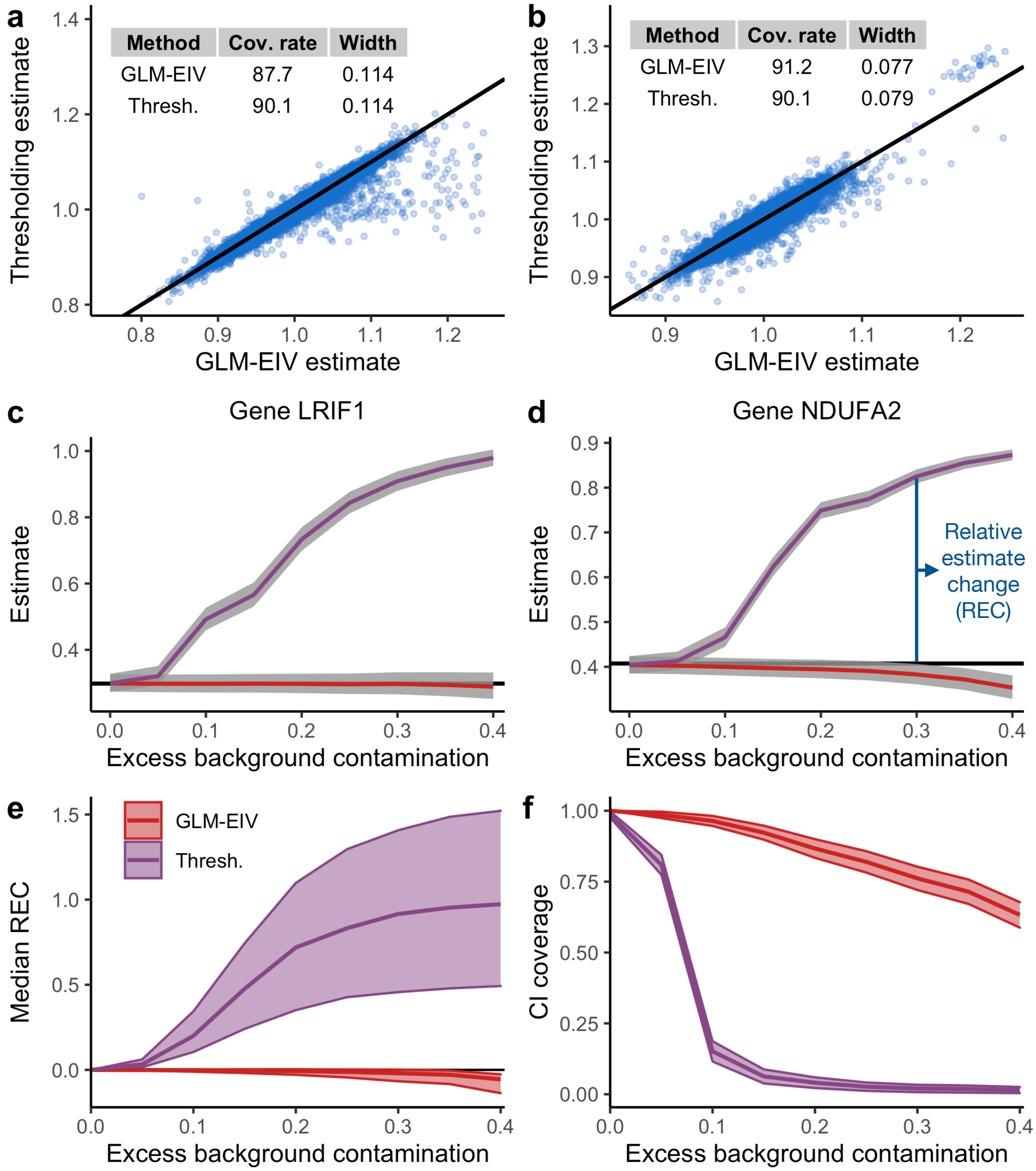

Figure 4: **Data analysis**. **a-b**, Estimates for fold change produced by GLM-EIV and thresholded regression on Gasperini (**a**) and Xie (**b**) negative control pairs. **c-d**, Estimates produced by GLM-EIV and thresholded regression on two positive control pairs – *LRIF1* (**a**) and *NDUFA2* (**b**) – 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. Panels (**c-f**) together illustrate that GLM-EIV demonstrated greater stability than thresholded regression as background contamination increased.

CI coverage of $\hat{\beta}_1^m$ by calculating each individual pair's coverage of $\hat{\beta}_1^m$ (across the Monte Carlo replicates) and then averaging this quantity across all pairs. GLM-EIV exhibited significantly higher CI coverage than thresholded regression as the data became increasingly contaminated (Figure 4f; bands indicate 95% pointwise CIs). Coverage rates were slightly above the nominal level of 95% in some settings because we covered an *estimate* of β_1^m rather than β_1^m itself, leading to mild "overfitting." Nonetheless, this experiment was meaningful to assess the stability of both methods to elevated background contamination.

8 Discussion

In this work we introduced GLM-EIV ("GLM-based errors in variables"), a new model and associated method for single-cell CRISPR screen analysis. GLM-EIV extends the classical errors-in-variables model to responses and noisy predictors that are exponential family-distributed and potentially impacted by the same set of confounding variables. These extensions enable GLM-EIV to resolve novel analysis challenges posed by single-cell CRISPR screens. We demonstrated through simulation studies, real data analyses, and theory that GLM-EIV outperforms thresholded regression by a considerable margin in high background contamination settings. GLM-EIV intuitively achieves this performance gain by leveraging information from *both* modalities (rather than the gRNA modality alone) to assign perturbation identities to cells. On the other hand, in low background contamination settings, GLM-EIV and thresholded regression work best in conjunction, with GLM-EIV providing a statistically principled and empirically effective procedure for selecting the threshold. GLM-EIV thereby neutralizes a tuning parameter that, until this point, has been selected using heuristic procedures, with little confidence that the choice is near optimal. Figure 5 summarizes how we anticipate GLM-EIV being used in practice.

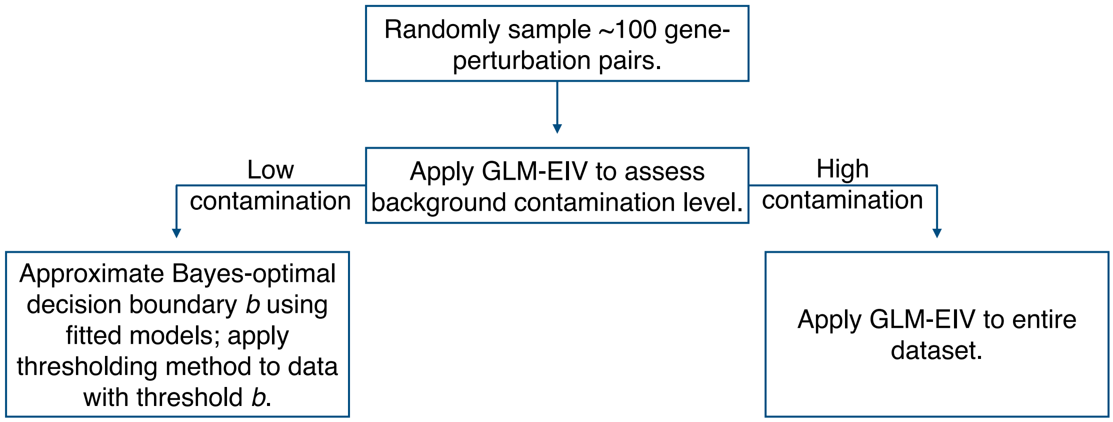

Figure 5: **Use of GLM-EIV in practice**. The decision tree above illustrates how we anticipate GLM-EIV could be used in practice. First, apply GLM-EIV to a set of randomly-sampled gene-perturbation pairs to assess background contamination level (positive control pairs work best for this purpose). If 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.

To our knowledge this is the first single-cell CRISPR screen paper oriented toward a statistical audience. We hope that this work helps to introduce the broader statistics community to an emerging class of functional genomics assays that likely will exert a major impact on biological research in the coming years (Przybyla and Gilbert 2021). Additionally, this is the first work to leverage the `ondisc`-`Nextflow`-HPC/cloud technology stack, a tightly-integrated, user-friendly, and powerful set of tools for large-scale single-cell analysis. We expect this technology stack to be of interest to other single-cell researchers.

We anticipate that GLM-EIV could be applied to other types of single-cell CRISPR screen and multimodal single-cell data. For example, GLM-EIV might be extended (with some effort) to "low-multiplicity of infection" screens (Schraivogel et al. 2020) in which each cell receives one or two perturbations rather than dozens (as is the case in "high multiplicity screens," studied in this work). We also could apply GLM-EIV to analyze multimodal single-cell chromatin accessibility assays. A question of interest in such experiments is whether chromatin state (i.e., closed or open) is associated with the expression of a gene or abundance of a protein (Mimitou et al. 2021). We do not directly observe the chromatin state of a cell; instead, we observe tagged DNA fragments that serve as count-based proxies for whether a given region of chromatin is open or closed. GLM-EIV might be applied in such experiments to aid in the selection of thresholds or to analyze whole datasets.

Several authors working on statistical methods for single-cell data recently have extended models that (implicitly or explicitly) assume Gaussianity and homoscedasticity to a broader class of exponential family distributions. For example, Lin et al. (2021) and Townes et al. (2019) (separately) developed eSVD and GLM-PCA, generalizations of SVD and PCA, respectively, to exponential family response distributions. Unlike their vanilla counterparts, eSVD and GLM-PCA can model gene expression counts directly, improving performance on dimension reduction tasks. We see our work (in part) 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 in several key directions (see above), enabling its direct and natural application to multimodal single-cell data.

The closest parallels to GLM-EIV in the statistical methodology literature are Grün and Leisch (2008) and Ibrahim (1990). Grün and 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. Next, Aigner (1973) and Savoca (2000) proposed measurement error models that consist of unobserved *binary* rather than *continuous* predictors; the latter are more commonly used in measurement error models. GLM-EIV likewise consists of a latent binary predictor, but unlike Aigner and Savoca, GLM-EIV handles a much broader class of exponential family-generated data. Finally, GLM-EIV accounts for a common source of measurement error between the predictor and response, a property not shared by classical measurement error models (Carroll et al. 2006).

GLM-EIV might be applied to areas beyond genomics, such as psychology. Many psychological constructs (e.g., presence or absence of a social media addiction) are latent and can be assessed only through an imperfect proxy (e.g., the number of times one has checked social media). Researchers might use GLM-EIV to regress an outcome variable (e.g., self-reported well-being) onto the latent construct via the imperfect proxy, potentially resolving challenges related to attenuation bias and threshold selection. Applications to psychology and other areas are a topic of future investigation.

9 Acknowledgments and resource availability

[This section is muted to preserve anonymity. The authors have prepared the data, results, and detailed instructions to reproduce the analyses reported in the paper.]

References

Aigner, Dennis J (1973). "Regression with a binary independent variable subject to errors of observation". In: *Journal of Econometrics* 1.1, pp. 49–59.

Barry, Timothy et al. (2021). “SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis”. In: *Genome Biology*, pp. 1–19.

Candès, Emmanuel et al. (2018). “Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection”. In: *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 80.3, pp. 551–577.

Carroll, Raymond J. et al. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall/CRC.

Datlinger, Paul et al. (2017). “Pooled CRISPR screening with single-cell transcriptome readout”. In: *Nature Methods* 14.3, pp. 297–301.

Datlinger, Paul et al. (2021). “Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing”. In: *Nature Methods* 18.6, pp. 635–642.

DI Tommaso, Paolo et al. (2017). “Nextflow enables reproducible computational workflows”. In: *Nature Biotechnology* 35.4, pp. 316–319.

Dixit, Atray et al. (2016). “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”. In: *Cell* 167.7, 1853–1866.e17.

Gallagher, Michael D. and Alice S. Chen-Plotkin (2018). “The Post-GWAS Era: From Association to Function”. In: *American Journal of Human Genetics* 102.5, pp. 717–730.

Gasperini, Molly, Jacob M. Tome, and Jay Shendure (2020). “Towards a comprehensive catalogue of validated and target-linked human enhancers”. In: *Nature Reviews Genetics* 21.5, pp. 292–310.

Gasperini, Molly et al. (2019). “A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens”. In: *Cell* 176.1-2, 377–390.e19.

Grün, Bettina and Friedrich Leisch (2008). "Finite Mixtures of Generalized Linear Regression Models". In: *Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg*. Heidelberg: Physica-Verlag HD, pp. 205–230.

Hafemeister, Christoph and Rahul Satija (2019). "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression". In: *Genome Biology* 20.1, pp. 1–15.

Hill, Andrew J. et al. (2018). "On the design of CRISPR-based single-cell molecular screens". In: *Nature Methods* 15.4, pp. 271–274.

Ibrahim, Joseph G. (1990). "Incomplete Data in Generalized Linear Models". In: *Journal of the American Statistical Association* 85.411, pp. 765–769.

Lin, Kevin Z., Jing Lei, and Kathryn Roeder (2021). "Exponential-Family Embedding With Application to Cell Developmental Trajectories for Single-Cell RNA-Seq Data". In: *Journal of the American Statistical Association* 0.0, pp. 1–32.

Liu, Molei et al. (2021). "Fast and Powerful Conditional Randomization Testing via Distillation". In: *Biometrika*, pp. 1–25.

Louis, By Thomas A (1982). "Finding the Observed Information Matrix when Using the EM Algorithm". In: *Journal of the Royal Statistical Society, Series B* 44.2, pp. 226–233.

Mimitou, Eleni P. et al. (2019). "Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells". In: *Nature Methods* 16.5, pp. 409–412.

Mimitou, Eleni P. et al. (2021). *Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells*. Vol. 39. 10. Springer US, pp. 1246–1258.

Morris, John A. et al. (2021). “Discovery of target genes and pathways of blood trait loci using pooled CRISPR screens and single cell RNA sequencing”. In: *bioRxiv*, p. 2021.04.07.438882.

Musunuru, Kiran et al. (2021). “In vivo CRISPR base editing of PCSK9 durably lowers cholesterol in primates”. In: *Nature* 593.7859, pp. 429–434.

Przybyla, Laralynne and Luke A. Gilbert (2021). “A new era in functional genomics screens”. In: *Nature Reviews Genetics* 0123456789.

Replogle, Joseph M. et al. (2020). “Combinatorial single-cell CRISPR screens by direct guide RNA capture and targeted sequencing”. In: *Nature Biotechnology*.

Ripley, Brian et al. (2013). “Package ‘mass’”. In: *Cran r* 538, pp. 113–120.

Rothgangl, Tanja et al. (2021). “In vivo adenine base editing of PCSK9 in macaques reduces LDL cholesterol levels”. In: *Nature Biotechnology* 39.8, pp. 949–957.

Sarkar, Abhishek and Matthew Stephens (2021). “Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis”. In: *Nature Genetics* 53.6, pp. 770–777.

Savoca, E. (2000). “Measurement errors in binary regressors: An application to measuring the effects of specific psychiatric diseases on earnings”. In: *Health Services and Outcomes Research Methodology* 1.2, pp. 149–164.

Schraivogel, Daniel et al. (2020). “Targeted Perturb-seq enables genome-scale genetic screens in single cells”. In: *Nature Methods* 17.6, pp. 629–635.

Stefanski, L. A. (2000). “Measurement Error Models”. In: *Journal of the American Statistical Association* 95.452, pp. 1353–1358.

Svensson, Valentine (2020). “Droplet scRNA-seq is not zero-inflated”. In: *Nature Biotechnology* 38, pp. 142–150.

Townes, F. William et al. (2019). “Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model”. In: *Genome Biology* 20.1, pp. 1–16.

Trapnell, Cole et al. (2014). “The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells”. In: *Nature biotechnology* 32.4, pp. 381–386.

Xie, Shiqi et al. (2019). “Global Analysis of Enhancer Targets Reveals Convergent Enhancer-Driven Regulatory Modules”. In: *Cell Reports* 29.9, 2570–2578.e5.

Appendices

A Theoretical details for thresholding estimator

A.1 Proposition statements

This subsection contains detailed proposition statements for the informal claims made in Section 4.2. For simplicity we derive results in the case $\pi = 1/2$. In reality π is smaller; however, setting π to 1/2 simplifies the proofs substantially and enables us to recover the most interesting and important statistical phenomena (e.g., attenuation bias, monotonic impact of β_1^g on problem difficulty, etc.). We state state five propositions labeled 3 – 7 corresponding to the informal claims made in Section 4.2; these propositions visually are depicted in Figure 6.

First, 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:

Proposition 3. *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.*

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

Next, the asymptotic relative bias b decreases monotonically in β_1^g :

Proposition 4. *Fix $\pi = 1/2$. The asymptotic relative bias b decreases monotonically in β_1^g, i.e.*

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

Let c_{bayes} denote the Bayes-optimal decision boundary for classifying cells as perturbed or unperturbed, i.e. $c_{\text{bayes}} = (1/2)(\beta_0^g + \beta_1^g)$ for $\pi = 1/2$. We have that c_{bayes} is a critical value of the bias function.

Proposition 5. *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.*

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

Furthermore, as the threshold tends to infinity, the asymptotic relative bias b tends to π.

Proposition 6. *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.*

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

As a corollary, when $\pi = 1/2$, asymptotic relative bias tends to $1/2$ as c tends to infinity. Finally, we compare two threshold selection strategies head-to-head: setting the threshold to an arbitrarily large number, and setting the threshold to the Bayes-optimal decision boundary.

Proposition 7. *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*

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

In other words, 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 7a, 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 7a, right); and the two approaches coincide when β_1^g is intermediate (i.e., $\beta_1^g = 2\Phi^{-1}(3/4)$; Figure 7a, middle).

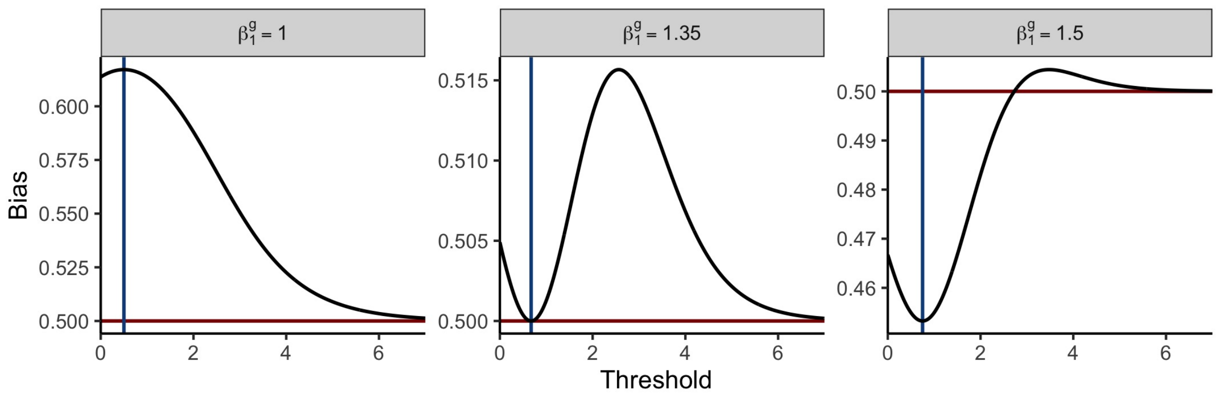

Figure 6: **Bias as a function of threshold**. This figure visually depicts Propositions 3-7, which were stated informally in the main text. Asymptotic relative bias is plotted on the vertical axis, and the threshold is plotted on the horizontal axis. Panels correspond to different values of β_1^g. Vertical blue lines indicate the Bayes-optimal decision boundary. Observe that (a) bias is strictly nonzero (proposition 3); (b) bias decreases monotonically in β_1^g (Proposition 4); (c) the Bayes-optimal decision boundary is a critical value of the bias function (Proposition 5), in some cases a maximum and in other cases a minumum; (d) as the threshold tends to infinity, the bias converges to 1/2 (Proposition 6); and (e) when $\beta_1^g < 1.35$, an arbitrarily large number yields a smaller bias; by contrast, when $\beta_1^g > 1.35$, the Bayes-optimal decision boundary yields a smaller bias (Proposition 7).

A.2 Organization

The following subsections prove all propositions. Section A.3 introduces some notation. Section A.4 establishes almost sure convergence of the thresholding estimator in the model

(5), proving Proposition 1. Section A.5 simplifies the expression for the attenuation function γ, and section A.6 computes derivatives of γ to be used throughout the proofs. Section A.7 establishes the limit in c of γ, proving Proposition 6. Section A.8 establishes that the Bayes-optimal decision boundary is a critical value of γ, proving Proposition 5, and section A.9 compares the competing threshold selection strategies head-to-head, proving Proposition 7. Section A.10 demonstrates that γ is monotone in β_1^g, proving Proposition 4, and Section A.11 establishes attenuation bias of the thresholding estimator, proving Proposition 3. Finally, Section A.12 derives the bias-variance decomposition of the thresholding estimator in the no-intercept version of 5, proving Proposition 2.

A.3 Notation

All notation introduced in this subsection (i.e., A.3) pertains to the Gaussian model with intercepts (5). 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

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

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

$$g(\beta_1^g, c, \pi, \beta_0^g) = (1-\pi)\left(\Phi(\beta_0^g + \beta_1^g - c)\right) - (1-\pi)\left(\Phi(\beta_0^g - c)\right) \tag{11}$$

and

$$h(\beta_1^g, c, \pi, \beta_0^g) = [(1-\pi)\left(\Phi(\beta_0^g - c)\right) + \pi\left(\Phi(\beta_0^g + \beta_1^g - c)\right)] \times$$
$$[(1-\pi)\left(\Phi(c - \beta_0^g)\right) + \pi\left(\Phi(c - \beta_0^g - \beta_1^g)\right)]. \quad (12)$$

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* Fitzpatrick 2009). 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.4 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 (5). Dividing by n in (6), 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, $\hat{\beta}_1^m \xrightarrow{P} \text{Cov}(\hat{p}_i, m_i)/\mathbb{V}\left(\hat{p}_i\right)$. 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 $\hat{p}_i$:

$$\mathbb{E}[\hat{p}_i] = \mathbb{P}\left[\hat{p}_i = 1\right] = \mathbb{P}\left[\beta_0^g + \beta_1^g p_i + \tau_i \geq c\right] =$$

$$\text{(By LOTP) } \mathbb{P}\left[\beta_0^g + \tau_i \geq c\right]\mathbb{P}\left[p_i = 0\right] + \mathbb{P}\left[\beta_0^g + \beta_1^g + \tau_i \geq c\right]\mathbb{P}[p_i = 1]$$

$$= \mathbb{P}\left[\tau_i \geq c - \beta_0^g\right](1-\pi) + \mathbb{P}\left[\tau_i \geq c - \beta_1^g - \beta_0^g\right](\pi)$$

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

$$\Phi(\beta_0^g - c)(1-\pi) + \Phi(\beta_1^g + \beta_0^g - c)\pi = \zeta(1-\pi) + \omega\pi.$$

3. Expectation of $\hat{p}_i p_i$: $\mathbb{E}\left[\hat{p}_i p_i\right] = \mathbb{E}\left[\hat{p}_i | p_i = 1\right]\mathbb{P}\left[p_i = 1\right] = \mathbb{P}\left[\beta_0^g + \beta_1^g + \tau_i \geq c\right]\pi = \omega\pi.$

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

$$\mathbb{E}\left[\hat{p}_i m_i\right] = \mathbb{E}[\hat{p}_i(\beta_0^m + \beta_1^m p_i + \epsilon_i)] = \beta_0^m \mathbb{E}\left[\hat{p}_i\right] + \beta_1^m \mathbb{E}\left[\hat{p}_i p_i\right] + \mathbb{E}[\hat{p}_i \epsilon_i]$$

$$= \beta_0^m \mathbb{E}[\hat{p}_i] + \beta_1^m \omega\pi + \mathbb{E}[\hat{p}_i]\mathbb{E}[\epsilon_i] = \beta_0^m \mathbb{E}[\hat{p}_i] + \beta_1^m \omega\pi.$$

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

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

$$\mathrm{Cov}\left(\hat{p}_i, m_i\right) = \mathbb{E}\left[\hat{p}_i m_i\right] - \mathbb{E}[\hat{p}_i]\mathbb{E}[m_i] = \beta_0^m \mathbb{E}[\hat{p}_i] + \beta_1^m \omega\pi - \mathbb{E}[\hat{p}_i](\beta_0^m + \beta_1^m \pi)$$

$$= \beta_1^m \omega\pi - \mathbb{E}[\hat{p}_i]\beta_1^m \pi = \beta_1^m \pi\left(\omega - \mathbb{E}[\hat{p}_i]\right).$$

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.5 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

$$\mathbb{E}\left[\hat{p}_i\right] = (1-\pi)\bar{\Phi}(c-\beta_0^g) + \pi\bar{\Phi}(c-\beta_0^g-\beta_1^g) = (1-\pi)\Phi(\beta_0^g-c) + \pi\Phi(\beta_0^g+\beta_1^g-c), \quad (13)$$

and so

$$\begin{aligned} 1-\mathbb{E}\left[\hat{p}_i\right] &= (1-\pi)+\pi-\mathbb{E}[\hat{p}_i] = (1-\pi)\left(1-\bar{\Phi}(c-\beta_0^g)\right)+\pi\left(1-\bar{\Phi}(c-\beta_0^g-\beta_1^g)\right) \\ &= (1-\pi)\Phi(c-\beta_0^g)+\pi\Phi(c-\beta_0^g-\beta_1^g). \quad (14) \end{aligned}$$

Next,

$$\omega = \Phi(\beta_1^g+\beta_0^g-c), \quad (15)$$

and so

$$\begin{aligned} \omega-\mathbb{E}[\hat{p}_i] &= \Phi(\beta_1^g+\beta_0^g-c)-(1-\pi)\Phi(\beta_0^g-c)-\pi\Phi(\beta_0^g+\beta_1^g-c) \\ & (1-\pi)\Phi(\beta_1^g+\beta_0^g-c)-(1-\pi)\Phi(\beta_0^g-c). \quad (16) \end{aligned}$$

Combining (13, 14, 15, 16), we find that

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

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

$$\gamma(\beta_1^g, c, 1/2) = \frac{\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)}{[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)]\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]}. \tag{18}$$

Recalling the definitions of g (11) and h (12), 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 (18) is identical to

$$\gamma(\beta_1^g, c, 1/2) = \frac{(4)(1/2)g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)} = \frac{2g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)}, \tag{19}$$

i.e., the numerator and denominator of (19) coincide with those of (18). We sometimes will use the notation $2 \cdot g$ and $4 \cdot h$ to refer to the numerator and denominator of (18), respectively.

A.6 Derivatives of g and h in c

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

$$\begin{aligned} D_2 g(\beta_1^g, c, \pi) &= -(1-\pi)f(\beta_0^g + \beta_1^g - c) + (1-\pi)f(\beta_0^g - c) \\ &= (1-\pi)f(c - \beta_0^g) - (1-\pi)f(c - \beta_0^g - \beta_1^g). \end{aligned} \tag{20}$$

Second, we have that

$$
D_2h(\beta_1^g, c, \pi) = -[(1-\pi)f(\beta_0^g - c) + \pi f(\beta_0^g + \beta_1^g - c)]\,[(1-\pi)\Phi(c-\beta_0^g) + \pi\Phi(c-\beta_0^g-\beta_1^g)]
$$

$$
+ [(1-\pi)f(c-\beta_0^g) + \pi f(c-\beta_0^g-\beta_1^g)]\,[(1-\pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c)]
$$

$$
= [(1-\pi)f(c-\beta_0^g) + \pi f(c-\beta_0^g-\beta_1^g)] \times
$$

$$
\Big[(1-\pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c) - (1-\pi)\Phi(c-\beta_0^g) - \pi\Phi(c-\beta_0^g-\beta_1^g)\Big]. \quad (21)
$$

A.7 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 (20) and (21) that

$$
\lim_{c\to\infty} \gamma(\beta_1^g, c, \pi) = \lim_{c\to\infty} \frac{\pi D_2 g(\beta_1^g, c, \pi)}{D_2 h(\beta_1^g, c, \pi)}
$$

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

$$
\Big[(1-\pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c) - (1-\pi)\Phi(c-\beta_0^g) - \pi\Phi(c-\beta_0^g-\beta_1^g)\Big]\Bigg\}^{-1}. \quad (22)
$$

We evaluate the two terms in the product (22) separately. Dividing by $f(c-\beta_0^g-\beta_1^g) > 0$, we see that

$$
\frac{(1-\pi)f(c-\beta_0^g) + \pi f(c-\beta_0^g-\beta_1^g)}{\pi(1-\pi)f(c-\beta_0^g) - \pi(1-\pi)f(c-\beta_0^g-\beta_1^g)} = \frac{\frac{(1-\pi)f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)} + \pi}{\frac{\pi(1-\pi)f(c-\beta_0^g)}{f(c-\beta_0^g-\beta_1^g)} - \pi(1-\pi)}. \quad (23)
$$

To evaluate the limit of (23), we first evaluate the limit of

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

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

$$= \exp\Big[- c^2/2 + c\beta_0^g - (\beta_0^g)^2/2$$

$$+ c^2/2 - c\beta_0^g - c\beta_1^g + (\beta_0^g)^2/2 + \beta_0^g\beta_1^g + (\beta_1^g)^2/2\Big]$$

$$= \exp[-c\beta_1^g + \beta_0^g\beta_1^g + (\beta_1^g)^2/2] = \exp[\beta_0^g\beta_1^g + (\beta_1^g)^2/2]\exp[-c\beta_1^g]. \quad (24)$$

Taking the limit in (24), 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 (23):

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

Next, we compute the limit of the other term in the product (22):

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

$$- (1-\pi)\Phi(c - \beta_0^g) - \pi\Phi(c - \beta_0^g - \beta_1^g)\Big] = -(1-\pi) - \pi = -1. \quad (25)$$

Combining (23) and (25), the limit (22) 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.8 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 (19), the quotient rule implies that

$$D_2\gamma(\beta_1^g, c, 1/2) = \frac{D_2[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_2[4h(\beta_1^g, c, 1/2)]}{[4h(\beta_1^g, c, \pi)]^2}. \quad (26)$$

We have by (20) that

$$D_2[2g(\beta_1^g, c_{\text{bayes}}, 1/2)] = f(\beta_1^g/2) - f(-\beta_1^g/2) = f(\beta_1^g/2) - f(\beta_1^g/2) = 0. \quad (27)$$

Similarly, we have by (21) that

$$D_2[4h(\beta_1^g, c_{\text{bayes}}, \pi)] = [f(\beta_1^g/2)+f(-\beta_1^g/2)]\,[\Phi(-\beta_1^g/2) + \Phi(\beta_1^g/2) - \Phi(\beta_1^g/2) - \Phi(-\beta_1^g/2)] = 0. \quad (28)$$

Plugging in (28) and (27) to (26), 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.9 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.

$$\begin{aligned} r(\beta_1^g) = \gamma(\beta_1^g, \beta_0^g + (1/2)\beta_1^g, 1/2) &= \frac{\Phi(\beta_1^g/2) - \Phi(-\beta_1^g/2)}{[\Phi(-\beta_1^g/2) + \Phi(\beta_1^g/2)]\,[\Phi(\beta_1^g/2) + \Phi(-\beta_1^g/2)]} \\ &= \frac{\int_{-\beta_1^g/2}^{\beta_1^g/2} f}{[1 - \Phi(\beta_1^g/2) + \Phi(\beta_1^g/2)]\,[\Phi(\beta_1^g/2) + 1 - \Phi(\beta_1^g/2)]} = 2\int_0^{\beta_1^g/2} f = 2\Phi(\beta_1^g/2) - 1. \end{aligned}$$

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 6 and proven in Section A.7) 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.10 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:*

$$\begin{aligned}[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)]\cdot[\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c) + \Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ \geq [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)]\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]\,. \quad (29)\end{aligned}$$

Proof: We take cases on the sign on β_1^g.

<u>Case 1</u>: $\beta_g^1 < 0$. 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 (29) is negative.

Turning our attention of the left-hand side of (29), we see that

$$\Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1 - \Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1. \quad (30)$$

Additionally, $\Phi(\beta_0^g - c) < 1$ and $\Phi(c - \beta_0^g) > 0$. Combining these facts with (30), 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 (29) is positive. The inequality holds for $\beta_1^g < 0$.

<u>Case 2</u>: $\beta_g^1 \geq 0$. We will show that the first term on the LHS of (29) is greater than the first term on the RHS of (29), 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

$$\beta_1^g + \beta_0^g - c \geq \beta_0^g - c \implies \Phi(\beta_1^g + \beta_0^g - c) - \Phi(\beta_0^g - c) \geq 0. \tag{31}$$

Adding $\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)$ to both sides of (31) yields

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

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

□

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

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

$$D_1[2g(\beta_1^g, c, 1/2)] = f(\beta_0^g + \beta_1^g - c), \tag{32}$$

$$\begin{aligned} D_1[4h(\beta_1^g, c, 1/2)] = {} & f(\beta_0^g + \beta_1^g - c)\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ & - f(\beta_0^g + \beta_1^g - c)\,[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)]. \end{aligned} \tag{33}$$

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

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

Next, recall that

$$2g(\beta_1^g, c, 1/2) = \Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c). \tag{35}$$

and

$$4h(\beta_1^g, c, 1/2) = [\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)]\,[\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]. \tag{36}$$

Substituting (32, 33, 35, 36) into (34) 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

$$D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)] \geq 0. \tag{37}$$

The quotient rule implies that

$$D_1\gamma(\beta_1^g, c, 1/2) = \frac{D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)]}{[4h(\beta_1^g, c, 1/2)]^2}. \quad (38)$$

We conclude by (37) and (38) 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.11 Strict attenuation bias

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

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

Similarly,

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

The function $\gamma(\beta_1^g, c, 1/2, \beta_0^g)$ is monotonically increasing in β_1^g (as stated in Proposition 4 and proven in section A.10). 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.12 Bias-variance decomposition in no-intercept model

We prove the bias-variance decomposition for the no-intercept version of (5). Define l (for "limit") by

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

where

$$\omega = \bar{\Phi}(c - \beta_g) = \Phi(\beta_g - c); \quad \zeta = \bar{\Phi}(c) = \Phi(-c).$$

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,

$$\sqrt{n}(\hat{\beta}_m - l) = \frac{(1/\sqrt{n}) \sum_{i=1}^n \hat{p}_i (m_i - l\hat{p}_i)}{(1/n) \sum_{i=1}^n \hat{p}_i^2}. \tag{39}$$

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,

$$\mathbb{E}\left[\hat{p}_i(m_i - l\hat{p}_i)\right] = \mathbb{E}[\hat{p}_i m_i] - l\mathbb{E}[\hat{p}_i] = \beta_m \omega\pi - \left(\frac{\beta_m \omega\pi}{\zeta(1-\pi) + \omega\pi}\right)\left[\zeta(1-\pi) + \omega\pi\right] = 0. \quad (40)$$

Additionally,

$$\mathbb{V}\left[\hat{p}_i(m_i - l\hat{p}_i)\right] = \mathbb{E}\left[\hat{p}_i^2(m_i - l\hat{p}_i)^2\right] - \left(\mathbb{E}\left[\hat{p}_i(m_i - l\hat{p}_i)\right]\right)^2$$
$$= \mathbb{E}\left[\hat{p}_i m_i^2\right] - 2l\mathbb{E}[m_i \hat{p}_i] + l^2 \mathbb{E}[\hat{p}_i] = \beta_m^2 \omega\pi + \mathbb{E}[\hat{p}_i] - 2l\beta_m \omega\pi + l^2 \mathbb{E}[\hat{p}_i]$$
$$= \beta_m \omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1 + l^2). \quad (41)$$

Therefore, by CLT, (40), and (41),

$$(1/\sqrt{n})\sum_{i=1}^{n} \hat{p}_i(m_i - l\hat{p}_i) \xrightarrow{d} N\left(0, \beta_m \omega\pi(\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1 + l^2)\right). \quad (42)$$

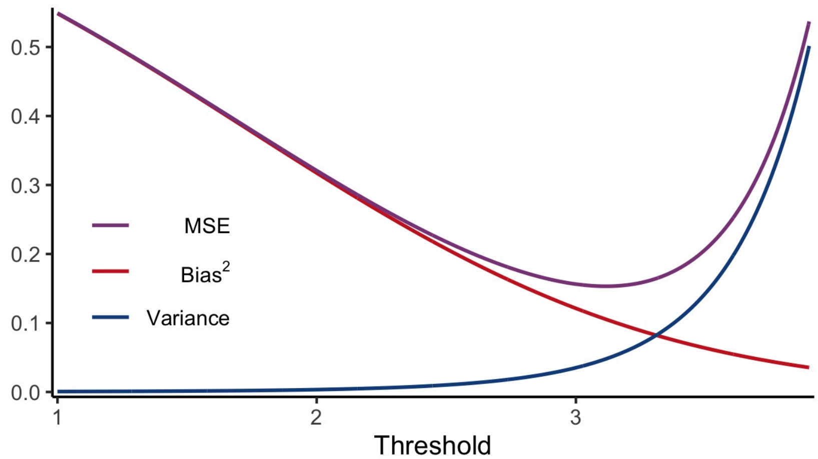

Figure 7: **Thresholding method bias-variance decomposition.** Bias decreases and variance increases as the threshold tends to infinity. $\beta_1^g = 1, \beta_1^m = 1$, and $\pi = 0.1$ in this plot.

Next, by weak LLN,

$$(1/n)\sum_{i=1}^{n} \hat{p}_i^2 = (1/n)\sum_{i=1}^{n} \hat{p}_i \xrightarrow{P} \mathbb{E}[\hat{p}_i]. \tag{43}$$

Finally, by (39), (42), (43), 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

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

completing the bias-variance decomposition. Figure 7 plots the bias-variance decomposition as a function of the threshold.

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. $[\eta_i^m(k)]^{(t)} \equiv h_m\left(\langle \tilde{x}_i(k), \beta_m^{(t)} \rangle + o_i^m\right)$. Similarly, let $[\eta_i^g(k)]^{(t)}$ be defined by $[\eta_i^g(k)]^{(t)} \equiv h_g\left(\langle \tilde{x}_i(k), \beta_g^{(t)} \rangle + o_i^g\right)$. Next, for $k \in \{0,1\}$, define $\alpha_i^{(t)}(k)$ by

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

Finally, let $\pi^{(t)}(1) \equiv \pi^{(t)} = \mathbb{P}\left(P_i = 1 | \theta^{(t)}\right)$ and $\pi^{(t)}(0) \equiv 1 - \pi^{(t)} = \mathbb{P}\left(P_i = 0 | \theta^{(t)}\right)$. The ith membership probability $T_i^{(t)}(1)$ is

$$
\begin{aligned}
T_i^{(t)}(1) = \mathbb{P}(P_i = 1 | M_i = m_i, G_i = g_i, \theta^{(t)}) &= \frac{\pi^{(t)}(1)\alpha_i^{(t)}(1)}{\sum_{k=0}^{1} \pi^{(t)}(k)\alpha_i^{(t)}(k)} \text{ (by Bayes rule)} \\
&= \frac{1}{\frac{\pi^{(t)}(0)\alpha_i(0)}{\pi^{(t)}(1)\alpha_i(1)} + 1} = \frac{1}{\exp\left(\log\left(\frac{\pi^{(t)}(0)\alpha_i(0)}{\pi^{(t)}(1)\alpha_i(1)}\right)\right) + 1} = \frac{1}{\exp\left(q_i^{(t)}\right) + 1}, \quad (44)
\end{aligned}
$$

where we set

$$q_i^{(t)} := \log \left(\frac{\pi^{(t)}(0)\alpha_i^{(t)}(0)}{\pi^{(t)}(1)\alpha_i^{(t)}(1)} \right). \tag{45}$$

Next, we have that

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

We therefore conclude that $T_i^{(t)} = 1/\left(\exp\left(q_i^{(t)}\right) + 1\right)$, which is easily computable.

M step

The complete-data log-likelihood of the GLM-EIV model is

$$\mathcal{L}(\theta; m, g, p) = \sum_{i=1}^{n} [p_i \log(\pi) + (1 - p_i)\log(1 - \pi)] + \sum_{i=1}^{n} \log\left(f_m(m_i; \eta_i^m)\right) + \sum_{i=1}^{n} \log\left(f_g(g_i; \eta_i^g)\right). \tag{46}$$

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

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\left[T_i^{(t)}(1)\log(\pi) + T_i^{(t)}(0)\log(1-\pi)\right] \\ + \sum_{k=0}^{1}\sum_{i=1}^{n} T_i^{(t)}(k)\log\left[f_m(m_i; \eta_i^m(k))\right] + \sum_{k=0}^{1}\sum_{i=1}^{n} T_i^{(t)}(k)\log\left[f_g(g_i; \eta_i^{g,b}(k))\right]. \tag{47}$$

The three terms of (47) 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 (47), 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 (47) 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 (47) 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 $[\tilde{X}(0)^T, \tilde{X}(1)^T]^T$. Therefore, the maximizer $\beta_m^{(t+1)}$ of the second term of (47) 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 (47) 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 $[\tilde{X}(0)^T, \tilde{X}(1)^T]^T$.

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 Δ^g, $[\Delta']^g$, V^g, and H^g analogously. These matrices are *unobserved*, as they depend on $\{p_1, \ldots, p_n\}$. Next, 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 $s^m(j), w^m(j) \in \mathbb{R}^n$ by

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

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 log likelihood (9) 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 (9) is a complicated mixture. Louis (1982) showed that $J(\theta; m, g)$ is equivalent to the following quantity:

$$\begin{aligned} J(\theta; m, g) = -\mathbb{E}\left[\nabla^2 \mathcal{L}(\theta; m, g, p) | G = g, M = m\right] \\ + \mathbb{E}\left[\nabla \mathcal{L}(\theta; m, g, p) | G = g, M = m\right] \mathbb{E}\left[\nabla \mathcal{L}(\theta; m, g, p) | G = g, M = m\right]^T \\ - \mathbb{E}\left[\nabla \mathcal{L}(\theta; m, g, p) \nabla \mathcal{L}(\theta; m, g, p)^T | G = g, M = m\right]. \quad (48) \end{aligned}$$

The observed information matrix $J(\theta; m, g)$ has dimension $(2d+1) \times (2d+1)$. Recall that the complete-data log-likelihood (46) 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 8; 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

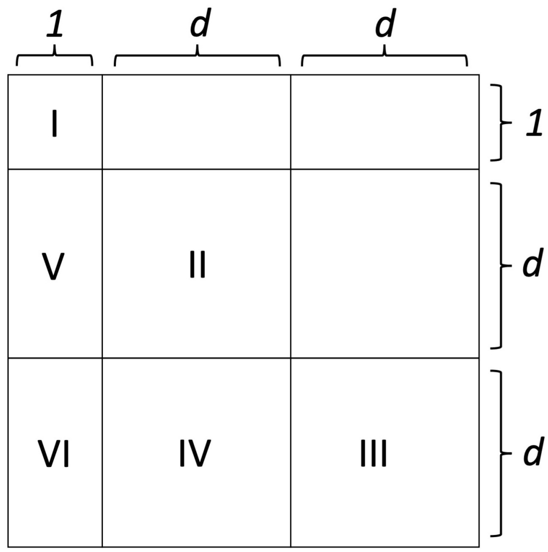

Figure 8: Block structure of the observed information matrix $J(\theta; m, g) = -\nabla^2 \mathcal{L}(\theta; m, g)$. The matrix is symmetric, and so we only need to compute submatrices I-VI to compute the entire matrix.

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

$$J_\pi(\theta; m, g) = -\mathbb{E}\left[\nabla_\pi^2 \mathcal{L}(\theta; m, g, p)\right] + \left(\mathbb{E}\left[\nabla_\pi \mathcal{L}(\theta; m, g, p)\right]\right)^2 - \mathbb{E}\left[\left(\nabla_\pi \mathcal{L}(\theta; m, g, p)\right)^2\right]. \quad (49)$$

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

$$\nabla_\pi \mathcal{L}(\theta; m, g, p) = \frac{\partial}{\partial \pi}\left(\sum_{i=1}^n p_i \log(\pi) + \sum_{i=1}^n (1 - p_i) \log(1 - \pi)\right)$$
$$= \frac{\sum_{i=1}^n p_i}{\pi} - \frac{\sum_{i=1}^n (1 - p_i)}{1 - \pi} = \frac{\sum_{i=1}^n p_i}{\pi} - \frac{n - \sum_{i=1}^n p_i}{1 - \pi} = \left(\frac{1}{\pi} + \frac{1}{1 - \pi}\right) \sum_{i=1}^n p_i - \frac{n}{1 - \pi}. \quad (50)$$

The second derivative is

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

We compute the expectation of the first term of (49):

$$\begin{aligned}
&\mathbb{E}\left[-\nabla^2_\pi \mathcal{L}(\theta; m, g, p)\right] = -\mathbb{E}\left[\frac{(\sum_{i=1}^n p_i) - n}{(1-\pi)^2} - \frac{\sum_{i=1}^n p_i}{\pi^2}\right] \\
&= -\mathbb{E}\left\{\left[\frac{1}{(1-\pi)^2} - \frac{1}{\pi^2}\right]\sum_{i=1}^n p_i - \frac{n}{(1-\pi)^2}\right\} = -\left\{\left[\frac{1}{(1-\pi)^2} - \frac{1}{\pi^2}\right]\sum_{i=1}^n T_i^\theta(1) - \frac{n}{(1-\pi)^2}\right\} \\
&= \left[\frac{1}{\pi^2} - \frac{1}{(1-\pi)^2}\right]\sum_{i=1}^n T_i^\theta(1) + \frac{n}{(1-\pi)^2}. \quad (51)
\end{aligned}$$

Next, we compute the difference of the second two pieces of (49). To this end, define $a \equiv 1/(1-\pi) + 1/\pi$ and $b \equiv n/(1-\pi)$. We have that

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

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,

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

$$= a^2 \left(\sum_{i=1}^{n} [T_i^\theta(1)]^2 - T_i^\theta(1) \right) = \left(\frac{1}{(1-\pi)} + \frac{1}{\pi} \right)^2 \left(\sum_{i=1}^{n} [T_i^\theta(1)]^2 - T_i^\theta(1) \right). \quad (52)$$

Stringing (49), (51) and (52) together, we obtain

$$J_\pi(\theta; m, g) = \left[\frac{1}{\pi^2} - \frac{1}{(1-\pi)^2} \right] \sum_{i=1}^{n} T_i^\theta(1) + \frac{n}{(1-\pi)^2} + \left(\frac{1}{(1-\pi)} + \frac{1}{\pi} \right)^2 \left(\sum_{i=1}^{n} [T_i^\theta(1)]^2 - T_i^\theta(1) \right). \quad (53)$$

Submatrix II

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

$$J_{\beta^m}(\theta; m, g) = -\mathbb{E}\left[\nabla^2_{\beta^m} \mathcal{L}(\theta; m, g, p)\right] + \mathbb{E}\left[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)\right] \mathbb{E}\left[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)\right]^T - \mathbb{E}\left[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p) \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)^T\right]. \quad (54)$$

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 (54). The (k, l)th entry of this matrix is

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

$$= \sum_{s=0}^{1} \tilde{X}(s)[,k]^T T^\theta(s) \left[\Delta^m(s) V^m(s) \Delta^m(s) - [\Delta']^m(s) H^m(s)\right] \tilde{X}(s)[,l].$$

We therefore have that

$$\mathbb{E}\left[-\nabla^2_{\beta^m}\mathcal{L}(\theta; m, g, p)\right] \; = \; \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) \left[\Delta^m(s) V^m(s) \Delta^m(s) - [\Delta']^m(s) H^m(s)\right] \tilde{X}(s). \tag{55}$$

Next, we compute the difference of the last two terms of (54). The (k,l)th entry is

$$\begin{aligned}
&\Bigg[\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\qquad - \mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta; m, g, p) \nabla_{\beta^m}\mathcal{L}(\theta; m, g, p)^T\right]\Bigg][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] \\
&= \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\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]
\end{aligned}$$

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

The sum of the last two terms on the right-hand side of (54) is therefore

$$
\begin{aligned}
&\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)\right]^T - \mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)^T\right] \\
&\qquad = \sum_{s=0}^{1}\sum_{t=0}^{1}\tilde{X}(s)^T T^{\theta}(s)\Delta^m(s)H^m(s)T^{\theta}(t)\Delta^m(t)H^m(t)\tilde{X}(t) \\
&\qquad\qquad - \sum_{s=0}^{1}\tilde{X}(s)^T T^{\theta}(s)(\Delta^m(s))^2(H^m(s))^2\tilde{X}(s). \quad (56)
\end{aligned}
$$

Combining (54), (55), (56), we find that

$$
\begin{aligned}
J_{\beta^m}(\theta;m,g) &= \sum_{s=0}^{1}\tilde{X}(s)^T T^{\theta}(s)\left[\Delta^m(s)V^m(s)\Delta^m(s) - [\Delta']^m(s)H^m(s)\right]\tilde{X}(s) \\
&\quad + \sum_{s=0}^{1}\sum_{t=0}^{1}\tilde{X}(s)^T T^{\theta}(s)\Delta^m(s)H^m(s)T^{\theta}(t)\Delta^m(t)H^m(t)\tilde{X}(t) \\
&\qquad - \sum_{s=0}^{1}\tilde{X}(s)^T T^{\theta}(s)(\Delta^m(s))^2(H^m(s))^2\tilde{X}(s). \quad (57)
\end{aligned}
$$

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 (57). Substituting g for m in this equation yields

$$J_{\beta^g}(\theta; m, g) = \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) \left[\Delta^g(s) V^g(s) \Delta^g(s) - [\Delta']^g(s) H^g(s)\right] \tilde{X}(s)$$
$$+ \sum_{s=0}^{1} \sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s) \Delta^g(s) H^g(s) T^\theta(t) \Delta^g(t) H^g(t) \tilde{X}(t)$$
$$- \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s) (\Delta^g(s))^2 (H^g(s))^2 \tilde{X}(s). \quad (58)$$

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

$$J_{(\beta^g, \beta^m)}(\theta; m, g) = \mathbb{E}\left[-\nabla_{\beta^g} \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)\right]$$
$$+ \mathbb{E}\left[\nabla_{\beta^g} \mathcal{L}(\theta; m, g, p)\right] \mathbb{E}\left[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)\right]^T - \mathbb{E}\left[\nabla_{\beta^g} \mathcal{L}(\theta; m, g, p) \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)^T\right]. \quad (59)$$

First, we have that

$$\mathbb{E}\left[-\nabla_{\beta^g} \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)\right] = 0, \quad (60)$$

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 (59) is

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

$$
\begin{aligned}
&= \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] \\
&\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] \\
&\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 + \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 (61)
\end{aligned}
$$

Combining (59), (60), and (61) produces

$$
\begin{aligned}
J_{(\beta^g,\beta^m)}(\theta;m,g) &= \sum_{s=0}^{1}\sum_{t=0}^{1} \tilde{X}(s)^T T^\theta(s)\Delta^g(s)H^g(s)T^\theta(t)\Delta^m(t)H^m(t)\tilde{X}(t) \\
&\qquad - \sum_{s=0}^{1} \tilde{X}(s)^T T^\theta(s)\Delta^g(s)H^g(s)\Delta^m(s)H^m(s)\tilde{X}(s). \quad (62)
\end{aligned}
$$

Submatrix V

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

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

We have that

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

as β^m and π separate in the log likelihood. Next, set $a \equiv 1/\pi + 1/(1-\pi)$ and $b \equiv n/(1-\pi)$. Recall from GLM theory that $\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p) = \tilde{X}^T\Delta^m s^m$ and from (50) that $a\sum_{i=1}^n p_i - b$. The kth entry of the last two terms of (63) is

$$\begin{aligned}
&\mathbb{E}\left[\nabla_\pi\mathcal{L}(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)[k]\right] - \mathbb{E}\left[\nabla_\pi\mathcal{L}(\theta;m,g,p)\nabla_{\beta^m}\mathcal{L}(\theta;m,g,p)[k]\right] \\
&= \left(\mathbb{E}\left[a\sum_{i=1}^n p_i - b\right]\right)\left(\mathbb{E}\left[\tilde{X}[,k]^T\Delta^m s^m\right]\right) - \mathbb{E}\left[\left(a\sum_{i=1}^n p_i - b\right)\tilde{X}[,k]^T\Delta^m s^m\right] \\
&= \left(a\sum_{i=1}^n \mathbb{E}[p_i] - b\right)\left(\sum_{j=1}^n \mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m]\right) - \mathbb{E}\left[\left(a\sum_{i=1}^n p_i - b\right)\left(\sum_{j=1}^n \tilde{x}_{jk}\Delta_j^m s_j^m\right)\right] \\
&= a\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[p_i]\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m] - b\sum_{j=1}^n \mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m] \\
&\quad - \left[a\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[p_i\tilde{x}_{jk}\Delta_j^m s_j^m] - b\sum_{j=1}^n \mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m]\right] \\
&= a\sum_{i=1}^n\sum_{j=1}^n \mathbb{E}[p_i]\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m] - a\sum_{i\neq j}\mathbb{E}[p_i]\mathbb{E}[\tilde{x}_{jk}\Delta_j^m s_j^m] - a\sum_{i=1}^n \mathbb{E}[p_i\tilde{x}_{ik}\Delta_i^m s_i^m] \\
&= a\sum_{i=1}^n \mathbb{E}[p_i]\mathbb{E}[\tilde{x}_{ik}\Delta_i^m s_i^m] - a\sum_{i=1}^n \mathbb{E}[p_i\tilde{x}_{ik}\Delta_i^m s_i^m] \\
&= a\sum_{i=1}^n T_i^\theta(1)[T_i^\theta(0)\Delta_i^m(0)s_i^m(0)\tilde{x}_{ik}(0) + T_i^\theta(1)\Delta_i^m(1)s_i^m(1)\tilde{x}_{ik}(1)] - a\sum_{i=1}^n T_i^\theta(1)\Delta_i^m(1)s_i^m(1)\tilde{x}_{ik}(1)
\end{aligned}$$

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

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

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

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

$$= a\left(\tilde{X}(0)[,k]^T w^m(0) - \tilde{X}(1)[,k]^T w^m(1)\right). \quad (65)$$

Combining (63), (64), and (65), we conclude that

$$J_{(\beta^m,\pi)}(\theta; m, g, p) = \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right)\left(\tilde{X}(0)^T w^m(0) - \tilde{X}(1)^T w^m(1)\right). \quad (66)$$

Submatrix VI

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

$$J_{(\beta^g,\pi)}(\theta; m, g, p) = \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right)\left(\tilde{X}(0)^T w^g(0) - \tilde{X}(1)^T w^g(1)\right). \quad (67)$$

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

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

III

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

IV

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

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. The procedure that we propose for this purpose is general, but for concreteness, we describe how to implement this procedure in R by extending base family objects. We implicitly condition on p_i, z_i^m, and o_i^m.

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 of the ith example and returns its 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, `mu.eta.prime` is the second derivative of the inverse link function $[r_m^{-1}]''$. Algorithm 2 computes the matrices $\Delta^m(j)$, $[\Delta']^m(j)$, $V^m(j)$, and $H^m(j)$ 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

$$\psi''_m(\eta_i^m) = [\sigma_i^m]^2 \tag{68}$$

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

$$r_m^{-1}(t) = \psi'_m(h_m(t)). \tag{69}$$

Differentiating (69) in t, we find that

$$(r_m^{-1})'(t) = \psi''_m(h_m(t))h'_m(t) \iff h'_m(t) = \frac{(r_m^{-1})'(t)}{\psi''_m(h_m(t))}. \tag{70}$$

Finally, plugging in l_i^m for t,

$$h'_m(l_i) = \frac{(r_m^{-1})'(l_i^m)}{\psi''(h_m(l_i^m))} = \frac{(r_m^{-1})'(l_i^m)}{\psi''_m(\eta_i^m)} = \text{ by (68) } \frac{(r_m^{-1})'(l_i^m)}{[\sigma_i^m]^2}.$$

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

$$\psi'''_m(\eta_i^m) = \gamma_i^m([\sigma_i^m]^2)^{3/2}. \tag{71}$$

Differentiating (70) in t, we obtain

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

Plugging in l_i^m for t, we find that

$$h_m''(l_i^m) = \frac{(r_m^{-1})''(l_i^m) - \psi_m'''(\eta_i^m)[h_m'(l_i^m)]^2}{[\sigma_i^m]^2} = \text{ (by 71) } \frac{(r_m^{-1})''(l_i^m) - ([\sigma_i^m]^2)^{3/2}(\gamma_i^m)[h_m'(l_i^m)]^2}{[\sigma_i^m]^2}.$$

Finally, the expression for V_i^m follows from (68). 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).

Algorithm 2 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

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

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

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

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

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

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

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

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

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. 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

$$\begin{aligned} Q_z(\theta|\theta^{(t)}) = \sum_{i=1}^n \left[T_i^{(t)}(1)\log(\pi) + T_i^{(t)}(0)\log(1-\pi)\right] + \sum_{i=1}^n \sum_{j=0}^1 T_i^{(t)}(j) \log\left[f_m(m_i; \eta_i^m(j))\right] \\ + \sum_{i=1}^n T_i^{(t)}(1)\left[\log(f_g(g_i; \eta_i^{g,z}))\right] + C. \quad (72) \end{aligned}$$

The three terms of (72) 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 $J_z(\theta; m, g) = -\nabla^2 \mathcal{L}_z(\theta; m, g)$.

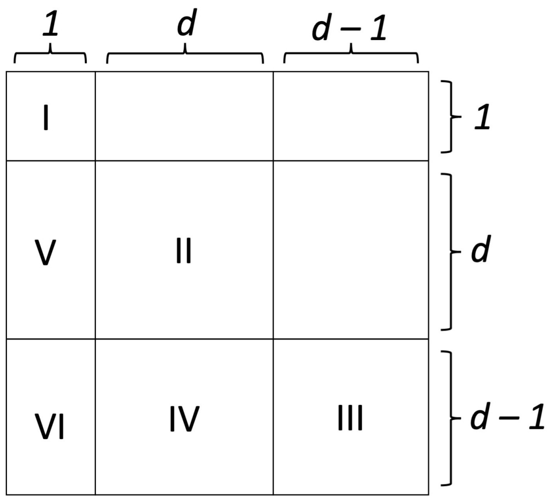

Figure 9: Block structure of the observed information matrix $J_z(\theta; m, g) = -\nabla^2 \mathcal{L}_z(\theta; m, g)$ for the zero-inflated model. Submatrices I, II, and VI are the same as in the background read model; therefore, we only need to compute submatrices III, VI, and V.

Louis's theorem implies that

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

The matrix $J_z(\theta; m, g)$ has dimension $d \times d$ and consists of nine submatrices (Figure 9). 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.

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

$$J_{\beta_{(g,z)}}(\theta;m,g) = -\mathbb{E}\left[\nabla^2_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\right]+\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\right]\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\right]^T$$
$$- \mathbb{E}\left[\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\nabla_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)^T\right]. \quad (73)$$

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 (73). 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,

$$- \mathbb{E}\left[\nabla^2_{\beta_{(g,z)}}\mathcal{L}_z(\theta;m,g,p)\right] = \mathbb{E}\left[X^T P(\Delta^{(g,z)}V^{(g,z)}\Delta^{(g,z)} - (\Delta')^{(g,z)}H^{(g,z)}\right]$$
$$= X^T T^\theta(1)(\Delta^{(g,z)}V^{(g,z)}\Delta^{(g,z)} - (\Delta')^{(g,z)}H^{(g,z)})X. \quad (74)$$

Next, we compute the difference of the last two terms of (73). The (k,l)th entry of this matrix is

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

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

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

$$= \sum_{i=1}^{n} x_{ik} T_i^{\theta}(1)^2 (\Delta_i^{(g,z)})^2 (H_i^{(g,z)})^2 x_{il} - \sum_{i=1}^{n} x_{ik} T_i^{\theta}(1) (\Delta_i^{(g,z)})^2 (H_i^{(g,z)})^2 x_{il}$$

$$= X[,k]^T T^{\theta}(1)^2 (\Delta^{(g,z)})^2 (H^{(g,z)})^2 X[,l] - X[,k]^T T^{\theta}(1) (\Delta^{(g,z)})^2 (H^{(g,z)})^2 X[,l]$$

Therefore, we have that

$$\mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)\right] \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)\right]^T - \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) \nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)^T\right]$$

$$= X^T T^{\theta}(1)^2 (\Delta^{(g,z)})^2 (H^{(g,z)})^2 X - X^T T^{\theta}(1) (\Delta^{(g,z)})^2 (H^{(g,z)})^2 X$$

$$= -X^T T^{\theta}(1) \left(\Delta^{(g,z)}\right)^2 \left(H^{(g,z)}\right)^2 \left(I - T^{\theta}(1)\right) X. \quad (75)$$

Combining (73), (74), and (75), we conclude that

$$J_{\beta_{(g,z)}} = (\theta; m, g) = X^T T^{\theta}(1) (\Delta^{(g,z)} V^{(g,z)} \Delta^{(g,z)} - (\Delta')^{(g,z)} H^{(g,z)}) X$$

$$- X^T T^{\theta}(1) \left(\Delta^{(g,z)}\right)^2 \left(H^{(g,z)}\right)^2 \left(I - T^{\theta}(1)\right) X. \quad (76)$$

Submatrix IV (zero-inflated)

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

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

$$+ \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p)\right] \mathbb{E}\left[\nabla_{\beta_m} \mathcal{L}_z(\theta; m, g, p)\right]^T - \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) \nabla_{\beta_m} \mathcal{L}_z(\theta; m, g, p)\right]^T\right]. \quad (77)$$

First, we have that

$$-\mathbb{E}\left[\nabla_{\beta_{(g,z)}}\nabla_{\beta_m}\mathcal{L}_z(\theta;m,g,p)\right]=0, \tag{78}$$

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 (77). 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^TP\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^TP\Delta^{(g,z)}s^{(g,z)}(s^m)^T\Delta^m\tilde{X}\right][k,l]\\
&=\left[\mathbb{E}\left[X[,k]^TP\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]^TP\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)\\
&=\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}]\\
&=\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-\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}]\\
&=\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-\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]\\
&=\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-\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}$$

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

$$- X[,k]^T \Delta^{(g,z)} H^{(g,z)} T^\theta(1) \Delta^m(1) H^m(1) \tilde{X}[,l]. \quad (79)$$

Combining (73), (74), and (75) yields

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

$$- X^T \Delta^{(g,z)} H^{(g,z)} T^\theta(1) \Delta^m(1) H^m(1) \tilde{X}(1). \quad (80)$$

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

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

$$- \mathbb{E}\left[\nabla_{\beta_{(g,z)}} \mathcal{L}_z(\theta; m, g, p) \nabla_\pi \mathcal{L}_z(\theta; m, g, p) \right]. \quad (81)$$

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 $a = 1/\pi + 1/(1-\pi)$ and $b = n/(1-\pi)$. We have that

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

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 (81). The kth entry of this vector is

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

$$= \left(\mathbb{E}\left[a\sum_{i=1}^{n} p_i - b\right]\right)\left(\mathbb{E}\left[X[,k]^T P\Delta^{(g,z)}s^{(g,z)}\right]\right) - \mathbb{E}\left[\left(a\sum_{i=1}^{n} p_i - b\right) X[,k]^T P\Delta^{(g,z)}s^{(g,z)}\right]$$

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

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

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

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

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

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

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

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

$$= -a\sum_{i=1}^{n} x_{ik}T_i(0)T_i^{\theta}(1)\Delta_i^{(g,z)}H_i^{(g,z)} = -aX[,k]^T w^{(g,z)}. \quad (83)$$

Combining (81), (82), and (83), we conclude that

$$J_{(\beta_{(g,z)},\pi)}(\theta; m, g) = -\left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) X^T w^{(g,z)}. \quad (84)$$

D Statistical accelerations and computing

D.1 Statistical accelerations

We describe in detail the procedure for obtaining the pilot parameter estimates $(\pi^{\text{pilot}}, \beta_m^{\text{pilot}}, \beta_g^{\text{pilot}})$. This procedure consists of two subroutines, which we label Algorithm 3 and Algorithm

4. The first step (Algorithm 3) is to obtain good parameter estimates for $[\beta_0^m, \gamma_m]^T$ and $[\beta_0^g, \gamma_g]^T$ via 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:

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

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

Algorithm 3 Computing $[\beta_0^m]^{\text{pilot}}$, $[\gamma_m^T]^{\text{pilot}}$, $[\beta_0^g]^{\text{pilot}}$, and $[\gamma_g^T]^{\text{pilot}}$.

Input: Data m, g, o^m, o^g, and X; gene expression distribution f_m and link function r_m; gRNA expression distribution f_g and link function r_g; 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.
  Set $[\beta_0^k]^{\text{pilot}}$ and $[\gamma_k^T]^{\text{pilot}}$ to the fitted coefficients of GLM_k.
  for $i \in \{1, \dots, n\}$ **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
return $([\beta_0^m]^{\text{pilot}}, \hat{f}^m, [\gamma_m^T]^{\text{pilot}}, [\beta_0^g]^{\text{pilot}}, [\gamma_g^T]^{\text{pilot}}, \hat{f}^g)$

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 4). 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 4 Computing π^{pilot}, $[\beta_1^m]^{\text{pilot}}$, $[\beta_1^m]^{\text{pilot}}$.

Input: Data m, g; fitted offsets $\hat{f}^m$, $\hat{f}^g$.

  `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 $(\pi^{\text{pilot}}, [\beta_1^m]^{\text{pilot}}, [\beta_1^g]^{\text{pilot}})$

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. 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 section D.2; the latter formula is asymptotically exact). 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 method for obtaining the full set of pilot parameter estimates requires fitting only two GLMs (via IRLS).

D.2 Intercept-plus-offset models

A key step in the algorithm for computing the pilot parameter estimates (Algorithm 4) is to fit a weighted, no-intercept, univariate GLM with nonzero offset terms and a binary predictor variable. 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

$$\begin{aligned}\mathcal{L}(\beta_1; m) = \sum_{i=1}^n T_i(0) f_m(m_i; h_m(\beta_1 + \hat{f}_i^m)) + \sum_{i=1}^n T_i(1) f_m(m_i; h_m(\hat{f}_i^m)) \\ = \sum_{i=1}^n T_i(1) f_m(m_i; h_m(\beta_1 + \hat{f}_i^m)) + C. \quad (85)\end{aligned}$$

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.

General formulation 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

$$\eta_i = ([\psi']^{-1} \circ r^{-1})(\beta + o_i) := h(\beta + o_i), \tag{86}$$

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

$$\mathcal{L}(\beta; y_i) = \sum_{i=1}^{n} w_i \log \left[f(y_i; \eta_i) \right] = C + \sum_{i=1}^{n} w_i (y_i \eta_i - \psi(\eta_i)). \tag{87}$$

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 (87); 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 (86),

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

Plugging $\eta_i = h(\beta + o_i) = \beta + o_i$ and $\psi(\eta_i) = (1/2)(\beta + o_i)^2$ into (87), 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 (86),

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

Plugging $\eta_i = h(\beta + o_i) = \beta + o_i$ and $\psi(\eta_i) = \exp(\beta + o_i)$ into (87), 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

$$\hat{\beta}^{\text{pois}} = \log \left(\frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i e^{o_i}} \right). \tag{88}$$

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 (86), 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,

$$\eta_i = h(\beta+o_i) = \log\left(\frac{\exp(\beta+o_i)}{\theta+\exp(\beta+o_i)}\right) = \beta+o_i-\log\left(\theta+e^{\beta}e^{o_i}\right) = \beta-\log\left(\theta+e^{\beta}e^{o_i}\right)+C, \tag{89}$$

and

$$\begin{aligned}\psi(\eta_i) = -\theta\log\left(1-\frac{\exp(\beta+o_i)}{\theta+\exp(\beta+o_i)}\right) &= -\theta\log\left(\frac{\theta}{\theta+\exp(\beta+o_i)}\right)\\ &= -\theta\log(\theta)+\theta\log[\theta+\exp(\beta+o_i)] = \theta\log(\theta+e^{\beta}e^{o_i})+C. \end{aligned}\tag{90}$$

Plugging (89) and (90) into (87), the log-likelihood (up to a constant) is

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

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

$$e^{\beta}\sum_{i=1}^{n}\frac{w_i e^{o_i}(y_i+\theta)}{e^{\beta}e^{o_i}+\theta} = \sum_{i=1}^{n} w_i y_i. \tag{91}$$

We cannot solve for β in (91) 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 (91) and rearranging,

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

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

$$\hat{\beta}^{\text{NB}} \xrightarrow{P} \log\left(\frac{\mathbb{E}[w_i y_i]}{\mathbb{E}[w_i e^{o_i}]}\right). \tag{93}$$

But the Poisson MLE $\hat{\beta}^{\text{Pois}}$ (88) converges in probability to the same limit:

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

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

Application to GLM-EIV The GLM that we seek to estimate (85) 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 (85), especially when the number of cells n is large. Additionally, we note that the GLM (85) is fitted as a subroutine of the algorithm for producing pilot parameter estimates (Algorithm 4). 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.

D.3 Computing

We describe in detail the at-scale GLM-EIV pipeline. 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" Algorithm 3. 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 (by running a reduced GLM-EIV), fitting the GLM-EIV model (Algorithm 1), and performing inference. The three loops in Algorithm 5 are embarassingly parallel and therefore can be massively parallelized.

Algorithm 5 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 3) 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 3) 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 $[\beta_1^m]^{\text{pilot}}, [\beta_1^g]^{\text{pilot}}, \pi^{\text{pilot}}$ by fitting a reduced GLM-EIV (Algorithm 4).
  Run GLM-EIV using the pilot parameters (Algorithm 1).
end for

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 10).

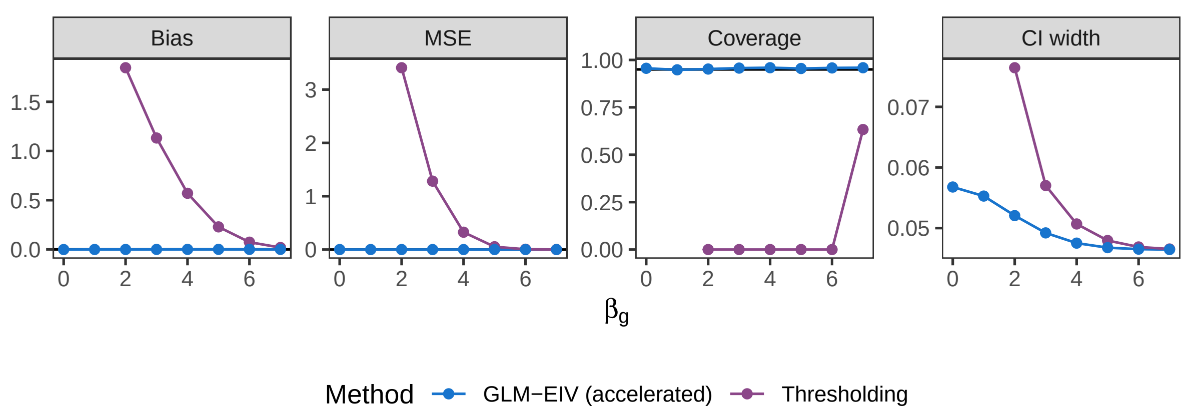

Figure 10: Additional simulation results on Gaussian data. GLM-EIV (accelerated) outperformed the thresholding method on bias, mean squared error, confidence interval coverage rate, and confidence interval width metrics.

F Data analysis details

First, we performed quality control on both datasets. As is standard in single-cell analysis, we removed cells with a high fraction ($> 8\%$) of mitochondrial reads (Choudhary and Satija 2021). We additionally excluded genes that were expressed in fewer than 10% of cells or that had a mean expression level of less than 1. We excluded cells in the Gasperini dataset with gene transcript UMI or gRNA counts below the 5th percentile or above the 95th percentile to reduce the effect of outliers. We did not repeat this latter quality control step on the Xie data because the Xie data were less noisy. The quality-controlled Gasperini and Xie datasets contained $n = 170,645$ (resp. $n = 101,508$) cells, $2,079$ (resp. $1,030$) genes, and $6,598$ (resp. 516) distinct perturbations.

The Gasperini dataset came with $17,028$ candidate *cis* pairs, $97,818$ negative control pairs, and 322 positive control pairs. The *cis* pairs consisted of genes paired to nearby enhancers with unknown regulatory effects. The negative control pairs consisted of non-targeting gRNAs paired to genes. The positive control pairs are described in the main text. The Xie data did not come with either *cis*, negative control, or positive control pairs. Therefore, we constructed a set of 681 candidate *cis* pairs by pairing perturbations to nearby genes, and we constructed a set of $50,000$ *in silico* negative control by pairing perturbations to genes on different chromosomes. See the *Methods* section of Barry et al. (2021) for details on the construction of *cis* and *in silico* negative control pairs on the Xie data.

We modeled the gene expression counts using a negative binomial distributions with unknown size parameter θ; we estimated θ using the glm.nb package. Choudhary and Satija (2021) report that Poisson models accurately capture highly sparse single-cell data.

Although Choudhary and Satija did not investigate the application of Poisson models gRNA data specifically, we modeled the gRNA counts using Poisson distributions, as the gRNA modality exhibited greater sparsity than the gene modality.

We applied GLM-EIV and the thresholding method to analyze the entire set of pairs in both datasets. We did not report results on the candidate *cis* pairs in the text because we do not know the ground truth for these pairs, making them less useful for method assessment. We focused our attention instead on the negative control pairs in both datasets and the positive control pairs in the Gasperini dataset (Figures 2 and 4).

We describe in more detail how we conducted the "excess background contamination" analysis (Figure 4, panels c-f). For each positive control pair, we varied excess background contamination over the grid $[0.0, 0.05, 0.1, \ldots, 0.4]$. For a given level of excess background contamination, we generated $B = 50$ synthetic gRNA datasets, holding fixed the raw gene expressions, covariates, library sizes, and fitted perturbation probabilities. We fitted GLM-EIV and the thresholding method to the data, yielding estimates $[\hat{\beta}_1^m]^{(1)}, \ldots, [\hat{\beta}_1^m]^{(B)}$. Next, we averaged over the $[\hat{\beta}_1^m]^{(i)}$s to obtain the mean estimate for a given pair and level of background contamination, and we calculated the REC using these mean estimates.

References

Barry, Timothy et al. (2021). "SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis". In: *Genome Biology*, pp. 1–19.

Choudhary, Saket and Rahul Satija (2021). "Comparison and evaluation of statistical error models for scRNA-seq". In: *bioRxiv* 8, p. 2021.07.07.451498.

Fitzpatrick, Patrick (2009). *Advanced calculus*. Vol. 5. American Mathematical Soc.

Louis, By Thomas A (1982). “Finding the Observed Information Matrix when Using the EM Algorithm”. In: *Journal of the Royal Statistical Society, Series B* 44.2, pp. 226–233.