

Exponential family measurement error models for single-cell CRISPR screens

TIMOTHY BARRY

Dept. of Biostatistics, Harvard T.H. Chan School of Public Health, Boston MA

tbarry@hsph.harvard.edu

KATHRYN ROEDER

Dept. of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, PA

EUGENE KATSEVICH

Dept. of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA

SUMMARY

CRISPR genome engineering and single-cell RNA sequencing have accelerated 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 considerable 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 hundreds of processors on clouds (e.g., Microsoft Azure) and high-performance clusters. Leveraging this infrastructure, we apply GLM-EIV to analyze two recent, large-scale, single-cell CRISPR screen datasets, yielding several new insights.

Key words: CRISPR, single-cell, GLM, mixture 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 (Musunuru *and others*, 2021) and accelerating biological discovery (Przybyla and Gilbert, 2022). Recently, scientists have paired CRISPR genome engineering with single-cell RNA sequencing (Datlinger *and others*, 2017). The resulting assays, known as “single-cell CRISPR screens,” link genetic perturbations in individual cells to changes in gene expression. Single-cell CRISPR screens have enabled breakthrough progress on longstanding challenges in genetics, such as causally mapping genome wide association study (GWAS) variants to target genes at genome-wide scale (Morris *and others*, 2023).

Despite their promise, single-cell CRISPR screens present considerable statistical challenges. One 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. A 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 *and others*, 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 to bring the cost of GLM-EIV down to within about an order of magnitude of the thresholding method. We additionally develop a Docker-containerized application to deploy GLM-EIV at-scale across tens or hundreds of processors on clouds (e.g., Microsoft Azure) and high-performance clusters.

Our analyses indicate that single-cell CRISPR screens fall into two main problem settings: the more challenging “high background contamination” setting and the easier “low background contamination” setting. GLM-EIV outperforms thresholded regression by a considerable margin

in the high background contamination setting; in the low background contamination setting, by contrast, GLM-EIV and thresholded regression perform similarly, provided that accurate guide RNA-to-cell assignments are used within the thresholded regression model. We show that a simplified version of GLM-EIV can be used to obtain these guide RNA-to-cell assignments in the low background contamination setting, thereby neutralizing a tuning parameter that until this point has been challenging to select.

2. ASSAY BACKGROUND

There are several classes of single-cell CRISPR screen assays, each suited to answer a different set of biological questions. In this work we mostly 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. 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 (Morris *and others*, 2023).

High-MOI single-cell CRISPR screens are a promising emerging technology for resolving this challenge (Morris *and others*, 2023; Mostafavi *and others*, 2023). 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 several methods for high-MOI single-cell CRISPR screen analysis (Gasperini *and others*, 2019; Xie *and others*, 2019; Barry *and others*, 2021; Wang, 2021). For example, Gasperini et al. applied negative binomial GLMs (as implemented in the Monocle software; Trapnell *and others* (2014)) to carry out the differential expression analysis described above. Moreover, Xie et al. applied chi-squared-like tests of independence for this purpose. Unfortunately, both of these approaches have limitations: the former can break down when the gene expression model is misspecified, and the latter does not adjust for the presence of technical confounders. In a prior work we introduced introduced SCEPTRE, a custom implementation of the conditional randomization test (Candès *and others*, 2018; Liu *and others*, 2022) tailored to single-cell CRISPR screen data. SCEPTRE simultaneously adjusts for confounder presence and ensures

robustness to expression model misspecification, thereby overcoming limitations of previous approaches and demonstrating improved sensitivity and specificity on single-cell CRISPR screen data. In this work we tackle a set of analysis challenges complimentary to those addressed by SCEPTR. Most importantly, we seek to account for the fact that the perturbation is measured with noise. Additionally, we seek to *estimate* (with confidence) the effect size of a perturbation on gene expression change, an objective that we did not consider in the original SCEPTR study.

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 not directly applicable. Finally, sequenced gRNAs sometimes map to cells that have not received a perturbation. This phenomenon, which we call “background contamination,” results from errors in the sequencing and alignment processes. 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 $i \in \{1, \dots, n\}$, 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 $z_i \in \mathbb{R}^{d-2}$ be the cell-specific covariates, including sequencing batch, percent mitochondrial reads, etc. (We note that most single-cell CRISPR screens have been carried out on cell lines consisting of a uniform cell type; however, if multiple cell types are present in the data, then cell type could be included as a covariate in the model.) The letters “m,” “g”, and “d” stand for “mRNA,” “gRNA,” and “depth,” respectively.

Building on the work of several previous authors (Robinson and Smyth, 2008; Townes *and others*, 2019; 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

$$m_i | (p_i, z_i, d_i^m) \sim \text{NB}_{s^m}(\mu_i^m); \quad \log(\mu_i^m) = \beta_0^m + \beta_1^m p_i + \gamma_m^T z_i + \log(d_i^m), \quad (3.1)$$

where (i) $\text{NB}_{s^m}(\mu_i^m)$ is a negative binomial distribution with mean μ_i^m and known size parameter s^m ; (ii) $\beta_0^m \in \mathbb{R}$, $\beta_1^m \in \mathbb{R}$, and $\gamma_m \in \mathbb{R}^{d-2}$ are unknown parameters; and (iii) $\log(d_i^m)$ is an offset term. (We note that the “size parameter” is simply the inverse of the negative binomial dispersion parameter; “size parameter” does not refer to library size in this context.) Similarly, we model

the gRNA counts by

$$g_i | (p_i, z_i, d_i^g) \sim \text{NB}_{sg} (\mu_i^g); \quad \log(\mu_i^g) = \beta_0^g + \beta_1^g p_i + \gamma_g^T z_i + \log(d_i^g), \quad (3.2)$$

where μ_i^g , s^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 *and others*, 2017; Hill *and others*, 2018). We model the marginal perturbation as $p_i \sim \text{Bern}(\pi)$, where p_i is an unobserved binary variable indicating presence ($p_i = 1$) or absence ($p_i = 0$) of the perturbation. We restrict π , the probability of perturbation, to the interval $(0, 1/2]$ to ensure that the model is identifiable; this restriction is reasonable given that each perturbation infects only a small fraction of cells. The gRNA intercept term β_0^g controls the ambient level of gRNA expression, i.e. the rate at which gRNA reads are generated in the absence of the perturbation. The perturbation coefficient β_1^g controls the extent to which perturbed and unperturbed cells differentially express the gRNA; the target of inference β_1^m is challenging to estimate when β_1^g is close to zero, as the gRNA distributions of the perturbed and unperturbed cells are hard to differentiate in this region of the problem space. Together, (3.1), (3.2), and the marginal distribution of p_i define the negative binomial GLM-EIV model.

The log-transformed sequencing depth $\log(d_i^m)$ is included as an offset term in (3.1) 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 (3.1) reveals that the mean gene expression μ_i^m of the i th cell is $\exp(\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i) d_i^m$. Because d_i^m is the sequencing depth, $\exp(\beta_0^m + \beta_1^m p_i + \gamma_m^T z_i)$ 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 analyzed two large-scale, high-MOI, single-cell CRISPR screen datasets published by Gasperini *and others* (2019) and Xie *and others* (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 studied thresholding from empirical and theoretical perspectives, highlighting several potential limitations of the approach. In the context of the negative binomial GLM-EIV model introduced above (3.1-3.2), the thresholding method leverages the gRNA counts (3.2) to impute the latent perturbation indicator (3.2), thereby reducing the full data generating process to a single, gene expression model (3.1). We studied Gasperini et al.'s variant of the thresholding method (i.e., thresholded negative binomial regression), as this version of the thresholding method is standard and 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:

$$m_i | (\hat{p}_i, z_i, d_i^m) \sim \text{NB}_{s^m}(\mu_i^m); \quad \log(\mu_i^m) = \beta_0^m + \beta_1^m \hat{p}_i + \gamma_m^T z_i + \log(d_i^m). \quad (4.3)$$

The model (4.3) is equivalent to the model (3.2), but the latent perturbation indicator p_i has been replaced by the imputed perturbation indicator \hat{p}_i .

3. Fit a GLM to (4.3) to obtain an estimate and CI for the target of inference β_1^m .

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).

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 distribution of a gRNA from the Gasperini (Figure 2e) and Xie (Figure 2f) dataset (counts of 0 omitted). The distributions peaked at 1 and then tapered off gradually; there did not exist a sharp boundary that cleanly separated 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 may not imply a clear threshold selection strategy.

Next, we studied the thresholding method from a theoretical perspective, recovering in a simplified Gaussian setting phenomena revealed in the empirical analysis. Due to space constraints we relegate this analysis to Appendix A, but we briefly summarize the main results here. First, we derived an exact expression for the asymptotic relative bias of the thresholding estimator $\hat{\beta}_1^m$. Leveraging this exact expression, we showed that (i) the thresholding estimator strictly underestimates (in absolute value) the true value of β_1^m over all choices of the threshold and over all values of the regression coefficients (an example of *attenuation bias*; Stefanski (2000)); and (ii) the magnitude of the bias decreases monotonically in β_1^q , comporting with the intuition that the problem becomes easier as the gRNA mixture distribution becomes increasingly well-separated. Second, we derived an asymptotically exact bias-variance decomposition for $\hat{\beta}_m$, demonstrating that as the threshold tends to infinity, the bias decreases and the variance increases.

5. GLM-BASED ERRORS-IN-VARIABLES (GLM-EIV)

We introduce the general GLM-EIV model, which generalizes the negative binomial GLM-EIV model (3.1-3.2) 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.

5.1 Model and model properties

The general GLM-EIV model uses an arbitrary GLM to model the gene and gRNA modalities:

$$m_i | (p_i, z_i, o_i^m) \sim f_m(\mu_i^m); \quad r_m(\mu_i^m) = \beta_0^m + \beta_1^m p_i + \gamma_m^T z_i + o_i^m, \quad (5.4)$$

$$g_i | (p_i, z_i, o_i^g) \sim f_g(\mu_i^g); \quad r_g(\mu_i^g) = \beta_0^g + \beta_1^g p_i + \gamma_g^T z_i + o_i^g. \quad (5.5)$$

Here, f_m (resp., f_g) is an exponential family distribution with mean μ_i^m (resp., μ_i^g); r_m and r_g are the link function for the gene and gRNA models, respectively; and o_i^m and o_i^g are 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 $\log(d_i^g)$). Again, we assume that the unobserved perturbation indicator p_i is drawn from a $\text{Bern}(\pi)$ distribution. More model details are available in Appendix B.

The GLM-EIV model can be seen as a generalization of the simple errors-in-variables model (when the predictor is binary); the latter is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i^* + \epsilon_i; \quad x_i = x_i^* + \tau_i, \quad (5.6)$$

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 (5.6) 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 (5.6) or other standard measurement error models.

5.2 Estimation and inference, and computational infrastructure

We derived an EM algorithm (Algorithm 1) to estimate the parameters of the GLM-EIV model.

We briefly introduce some notation. Let $\beta_m = [\beta_0^m, \beta_1^m, \gamma_m]^T$ be the vector of unknown gene model parameters and $\beta_g = [\beta_0^g, \beta_1^g, \gamma_g]^T$ the vector of unknown gRNA model parameters. Let m , g , o^m , and o^g be the vector of gene expressions, gRNA expressions, gene library sizes, and gRNA library sizes. Finally, let X be the observed design matrix; let \tilde{X} be the augmented design matrix that results from concatenating the column of (unobserved) p_i s to X ; and let $\tilde{X}(0)$ (resp., $\tilde{X}(1)$) be the matrix that results from setting all of the p_i s in \tilde{X} to 0 (resp., 1).

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 $i \in \{1, \dots, n\}$ given the current parameter estimates $(\beta_m^{(t)}, \beta_g^{(t)}, \pi^{(t)})$ and observed data (m_i, g_i) is $T_i(1) = \mathbb{P}(p_i = 1 | M_i = m_i, G_i = g_i, \beta_m^{(t)}, \beta_g^{(t)}, \pi^{(t)})$. We can calculate this quantity by applying (i) Bayes rule, (ii) the conditional independence property of M_i and G_i , (iii) the density of M_i and G_i , and (iv) a log-sum-exp-type trick to ensure numerical stability. Next, we produce updated estimates $\pi^{(t+1)}$, $\beta_g^{(t+1)}$, and $\beta_m^{(t+1)}$ of the parameters by maximizing the M step objective function. It turns out that maximizing 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 (B.1) 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 derived an analytic formula for the asymptotic observed information matrix using Louis's The-

Algorithm 1 EM algorithm for GLM-EIV model.

Input: Pilot estimates β_m^{curr} , β_g^{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** $i \in \{1, \dots, n\}$ **do** ▷ E step

 - $T_i(1) \leftarrow \mathbb{P}(p_i = 1 | M_i = m_i, G_i = g_i, \beta_m^{\text{curr}}, \beta_g^{\text{curr}}, \pi^{\text{curr}})$
 - $T_i(0) \leftarrow 1 - T_i(1)$

- end for**
- $\pi^{\text{curr}} \leftarrow (1/n) \sum_{i=1}^n T_i(1)$ ▷ M step
- $w \leftarrow [T_1(0), T_2(0), \dots, T_n(0), T_1(1), T_2(1), \dots, T_n(1)]^T$
- 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 $(\hat{\beta}_m, \hat{\beta}_g, \hat{\pi})$

orem (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.

A downside of 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 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” on top of one another. Briefly, we first obtain pilot estimates for the nuisance parameters β_0^m , γ_m , β_0^g , and γ_g by regressing the gene and gRNA expression vectors onto the observed design matrix X ; the resulting estimates are close to the full GLM-EIV model maximum likelihood estimates because the probability of perturbation is small. Next, we obtain pilot estimates for π and the perturbation effect parameters β_1^m and β_1^g by estimating a simplified, “reduced” GLM-EIV model; this second step does not require fitting any GLMs. (See

Appendix C for additional details.) Overall, the statistical accelerations reduce the number of GLMs that must be 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, and `ondisc`, an R/C++ package that we developed (in a separate project; preprint forthcoming) to facilitate large-scale computing on single-cell data. `Nextflow` and `ondisc` together enable the construction of highly portable single-cell pipelines: one can analyze data out-of-memory on a laptop or in a distributed fashion across hundreds of processors 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 recycles computation when possible, saving a considerable amount of compute; see Appendix C.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.

5.3 *The gRNA mixture assignment method*

Thus far we have described two methods for estimating the effect of a perturbation on gene expression: the simple thresholding method and the more complex GLM-EIV method. A third approach of intermediate complexity — which we call the “gRNA mixture assignment” approach — is to (i) fit a mixture model to the gRNA count distribution, (ii) use this fitted mixture model to impute perturbation identities onto cells, and then (iii) regress the gene expressions onto the imputed perturbation indicators (as well as the remaining covariates). The gRNA mixture assignment approach enjoys at least two strengths relative to the simpler thresholding approach: the former negates the threshold tuning parameter and can account for variation across cells due to covariates.

Reprogle *and others* (2020) proposed a simple gRNA mixture assignment strategy that involves fitting a Poisson-Gaussian mixture model to the log-transformed gRNA counts and then assigning gRNAs to cells using the posterior perturbation probabilities of the fitted model. (We call this method the Nat. Biotech. 2020 method, representing the journal and year in which the method appeared.) Unfortunately, this method poses several conceptual and practical difficulties. First, it is unclear how the method fits the Poisson component of the mixture distribution to the log-transformed gRNA expressions, as the transformed expressions are not integer-valued. Second, due to recent changes in the Python ecosystem, we and others have had difficulty with installing the Python package upon which the Nat. Biotech. 2020 method relies. (See Appendix D for further discussion of the Nat. Biotech. 2020 method.)

Following Repogle *and others* (2020), we devised an alternate gRNA mixture assignment strategy that is tethered more closely to the data-generating mechanism. For a given gRNA, we regress the gRNA counts onto the (latent) perturbation indicator and covariates (while ignoring the gene expressions; model 5.5). We assign perturbation identities to cells by thresholding the posterior perturbation probabilities of the fitted model at 1/2. The latent variable gRNA model is a subset of the full GLM-EIV model (5.4-5.5). Thus, we used the GLM-EIV EM algorithm to fit the latent variable gRNA model, enabling us to exploit the various techniques that we developed in the context of GLM-EIV for obtaining fast and numerically stable estimates.

6. SIMULATION STUDY

We conducted a comprehensive suite of six simulation studies to compare the empirical performance of GLM-EIV, the thresholding method, and the gRNA mixture assignment method. (We coupled the latter method to standard regression on the imputed perturbation assignments to estimate the perturbation effect size.) We describe one simulation study here and defer the remaining simulation studies to the Appendix G. We generated data on $n = 50,000$ cells from

the GLM-EIV model, setting 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. We generated the gene expression count data from two response distributions: Poisson and negative binomial (size parameter fixed at $s = 20$ for the latter; see simulation study 3 for an exploration of different values of s). We generated the gRNA count data from a Poisson distribution. For each parameter setting (defined by a β_1^g -distribution pair), we synthesized $n_{\text{sim}} = 500$ i.i.d. datasets. Appendix G compares the parameter values used in the simulation study to those estimated from real data.

We applied four methods to the simulated data: “vanilla” GLM-EIV, accelerated GLM-EIV, thresholded regression, and the gRNA mixture assignment method. We used the Bayes-optimal decision boundary for classification as the threshold for the thresholding method (as derived in Section A.12). We ran all methods on the negative binomial data twice: once treating the size parameter s as a known constant and once treating s as unknown. In the latter case we used the `glm.nb` function from the MASS package to estimate s before applying the methods (Ripley and others, 2013). We note that none of the methods accounts for the error in estimating s when computing coefficient standard errors. We display the results of the simulation study in Figure 3. Columns correspond to distributions (i.e., Poisson, NB with known s , and NB with unknown s), and rows correspond to performance metrics (i.e., bias, mean squared error, CI coverage rate (nominal rate 95%), CI width, and method run time). The β_1^g parameter is plotted on the horizontal axis, and the methods are depicted in different colors. (GLM-EIV is masked by accelerated GLM-EIV in several panels).

We found that GLM-EIV outperformed the gRNA mixture method and that the gRNA mixture method outperformed thresholded regression across the metrics of bias, mean squared error, and confidence interval coverage. We reasoned that GLM-EIV outperformed the gRNA mixture 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 reasoned that the gRNA mixture method outperformed thresholded regression because the gRNA mixture method better accounted for heterogeneity across cells due to the covariates. Notably, 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 running time of accelerated GLM-EIV was almost within an order of magnitude of that of the thresholding method. As expected, the confidence interval coverage of the methods degraded somewhat in the negative binomial case under estimated s as opposed to known s , but this difference was not substantial. Appendix G presents additional simulation studies in which we generate data from a Gaussian model, vary β_1^m and s , and assess the performance of the methods on data containing unmeasured covariates and outliers.

7. REAL DATA APPLICATION I: ESTIMATING PERTURBATION EFFECTS ON HIGH-MOI DATA

Leveraging our computational infrastructure, we applied GLM-EIV and the thresholding method to analyze the entire Gasperini and Xie datasets. GLM-EIV ran in under two days on both datasets, using no more than 250 processors and two gigabytes of memory per process. We report only the most important aspects of the analysis and results in the main text; full details are available in Appendix E. 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 $\bar{\hat{\beta}}_g$ and $\bar{\hat{\pi}}$ across the $\hat{\beta}_g$ s and $\hat{\pi}$ s, and (iv) used $\bar{\hat{\beta}}_g$ and $\bar{\hat{\pi}}$ to estimate a dataset-wide Bayes-optimal decision boundary (Section A.12). We repeated this procedure on both datasets, yielding a threshold of 3 for Gasperini and 7 for Xie.

We compared GLM-EIV to thresholded regression on the real data, focusing specifically on the negative control pairs (i.e., gene-perturbation pairs for which the ground truth fold change is known to be 1; Appendix E). 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 gRNA expression $\exp(\hat{\beta}_1^g)$ was unexpectedly large: the 95% CI for this parameter (averaged across pairs) was [4306, 5186] and [300, 316] on the Gasperini and Xie data, respectively. We reasoned that the datasets lay in a region of the parameter space in which thresholding is 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 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 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. REAL DATA APPLICATION II: ASSIGNING PERTURBATIONS TO CELLS ON LOW-MOI DATA

We sought to explore whether the gRNA mixture assignment method that we proposed in Section 5.3 — which is in effect a special case of GLM-EIV — might be an independently useful tool for assigning gRNAs to cells on real single-cell CRISPR screen data. We applied the gRNA mixture assignment method to assign gRNAs to cells on a low multiplicity-of-infection (or MOI) single-cell CRISPR screen of immune cells (Papalex *and others*, 2021). (A low-MOI dataset, in contrast to a high-MOI dataset, is one in which the experimenter has aimed to insert exactly one perturbation into each cell.) We elected to assess the performance of the gRNA mixture assignment method on low-MOI data because the “ground truth” gRNA-to-cell mapping is easier to ascertain in low MOI than in high MOI. The majority of cells in a low-MOI screen contains a single perturbation, while a fraction of cells contains zero or two or more perturbations. Thus, if a given gRNA constitutes a large fraction (say, $> 25\%$) of the gRNA reads in a given cell, we can confidently map that gRNA to that cell. Although not foolproof, this strategy yields a reasonable approximation to the ground truth in low MOI. (There is no analogous strategy for obtaining ground truth gRNA assignments in high MOI, as each cell in high MOI contains many gRNAs, and the number of

gRNAs per cell is indeterminate and variable.)

We used our proposed gRNA mixture assignment method to obtain gRNA-to-cell assignments for each gRNA in the low-MOI dataset (after restricting our attention to the 95% most highly expressed gRNAs). We included the standard technical factors as covariates, including biological replicate. We compared the mixture-model-based gRNA assignments to the ground truth assignments; the latter were obtained in the manner described above. Encouragingly, we found that these two methods produced near-identical results. For example, the mixture model determined that gRNA “CUL3g2” was present in 141 cells (and absent in the rest), while the ground truth method indicated that “CUL3g2” was present in 137 cells (Figure 5a). Treating the ground truth assignments as a reference, we constructed a confusion matrix to assess the classification accuracy of the mixture method assignments on CUL3g2 (Figure 5b). The sensitivity, specificity, and balanced accuracy of the mixture method assignments were high (1.000, 0.9998, and 0.9998, respectively).

We replicated this analysis across the entire set of gRNAs, finding that the mixture method assignments exhibited consistently high concordance with the ground truth assignments as measured by sensitivity, specificity, and balanced accuracy (although there were a few outliers; Figure 5c). We concluded that the mixture assignment method was a statistically principled, fast, and numerically stable strategy for the recapitulating the ground truth assignments with high fidelity. We sought to compare our gRNA mixture assignment method against the Nat. Biotech. 2020 Poisson-Gaussian mixture method. Unfortunately, as discussed elsewhere (Section 5.3 and Appendix D), we were unable to get the Nat. Biotech. 2020 method (or approximations thereof written in R) working. We note that, in contrast to the Nat. Biotech. 2020 method, the proposed method allows for the inclusion of covariates (e.g., library size and batch) and models the gRNA counts directly.

9. DISCUSSION

In this work we studied the problem of estimating the effect sizes of perturbations on changes in gene expression in high-MOI single-cell CRISPR screens, focusing specifically on the challenge that the perturbation is unobserved. We showed through empirical, theoretical, and simulation analyses that the commonly-used thresholding method poses several difficulties: there exist settings (i.e., high background contamination settings) in which thresholding is not a tenable strategy, and in settings in which thresholding *is* a tenable strategy (i.e., low background contamination settings), selecting a good threshold is challenging and consequential. Next, we developed GLM-EIV, a method that jointly models the gene and gRNA modalities to implicitly assign perturbation identities to cells and estimate perturbation effect sizes, thereby overcoming limitations of the thresholding method. GLM-EIV demonstrated significantly improved performance relative to the thresholding method in high background contamination settings on both synthetic and realistic semi-synthetic data.

However, GLM-EIV and the thresholding method demonstrated roughly similar performance on the two real high-MOI datasets that we examined, as the real data exhibited lower background contamination than anticipated. We believe that this is an interesting finding in itself; moreover, future datasets may demonstrate higher levels of background contamination, in which case GLM-EIV could serve as an immediately applicable analytic tool. Finally, the gRNA mixture assignment method, which under the hood exploits the estimation machinery of GLM-EIV, is a statistically principled, numerically stable, fast, and accurate strategy for obtaining gRNA-to-cell assignments on real data; these assignments can be used as input to downstream methods (e.g., negative binomial regression or SCEPTRE; Figure 5d).

We anticipate that GLM-EIV could be applied to other types of multi-modal single-cell data, such as 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 *and others*, 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. The full GLM-EIV model potentially could be applied to analyze low-MOI single-cell CRISPR screen data, but we anticipate that the relative ease of assigning gRNAs to cells in low MOI (as described in section 8) may obviate the need for GLM-EIV in that setting.

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 *and others*, 2006). Additional related work is relayed in Appendix F.

GLM-EIV might be applied to areas beyond genomics, such as psychology. Some 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). Re-

searchers 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.

SOFTWARE, CODE, AND RESULTS

The gRNA-only mixture assignment functionality of GLM-EIV is implemented in our **sceptre** toolkit for single-cell CRISPR screen analysis (github.com/Katsevich-Lab/sceptre). The **sceptre** user manual (timothy-barry.github.io/sceptre-book/sceptre.html) presents a detailed guide on analyzing data using the **sceptre** software, including several sections on assigning gRNAs to cells using the mixture assignment method introduced in this work.

Results are deposited at upenn.box.com/v/glmeiv-files-v1. Github repositories containing manuscript replication code, the **glmeiv** R package, and the cloud/HPC-scale GLM-EIV pipeline are available at github.com/timothy-barry/glmeiv-manuscript, github.com/timothy-barry/glmeiv, and github.com/timothy-barry/glmeiv-pipeline, respectively. Detailed replication instructions are available in the first repository.

ACKNOWLEDGEMENTS

We thank Eric Tchetgen Tchetgen for helpful conversations, Xuran Wang for helping to process the Xie dataset, and Songcheng Dai for helping to deploy the GLM-EIV pipeline on Azure. We additionally thank three anonymous reviewers whose comments considerably improved the manuscript. This work used the Extreme Science and Engineering Discovery Environment (XSEDE; NSF grant ACI-1548562) and the Bridges-2 system (NSF grant ACI-1928147) at the Pittsburgh Supercomputing Center. This work is funded by National Institute of Mental Health (NIMH) grant R01MH123184 and NSF grant DMS-2113072.

REFERENCES

- AIGNER, DENNIS J. (1973). Regression with a binary independent variable subject to errors of observation. *Journal of Econometrics* **1**(1), 49–59.
- BARRY, TIMOTHY and others. (2021). SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis. *Genome Biology*, 1–19.
- CANDÈS, EMMANUEL and others. (2018). Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **80**(3), 551–577.
- CARROLL, RAYMOND J and others. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.
- CHOUDHARY, SAKET AND SATIJA, RAHUL. (2022). Comparison and evaluation of statistical error models for scRNA-seq. *Genome Biology* **23**(1), 1–20.
- DATLINGER, PAUL and others. (2017). Pooled crispr screening with single-cell transcriptome readout. *Nature Methods* **14**(3), 297–301.
- FITZPATRICK, PATRICK. (2009). *Advanced calculus*, Volume 5. American Mathematical Soc.
- GALLAGHER, MICHAEL D. AND CHEN-PLOTKIN, ALICE S. (2018). The post-gwas era: From association to function. *American Journal of Human Genetics* **102**(5), 717–730.
- GASPERINI, MOLLY and others. (2019). A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**(1-2), 377–390.e19.
- GRÜN, BETTINA AND LEISCH, FRIEDRICH. (2008). *Finite Mixtures of Generalized Linear Regression Models*. Heidelberg: Physica-Verlag HD, pp. 205–230.

- HAFEMEISTER, CHRISTOPH AND SATIJA, RAHUL. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biology* **20**(1), 1–15.
- HILL, ANDREW J. and others. (2018). On the design of crispr-based single-cell molecular screens. *Nature Methods* **15**(4), 271–274.
- IBRAHIM, JOSEPH G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association* **85**(411), 765–769.
- LIN, KEVIN Z, LEI, JING AND ROEDER, KATHRYN. (2021). Exponential-family embedding with application to cell developmental trajectories for single-cell rna-seq data. *Journal of the American Statistical Association* **116**(534), 457–470.
- LIU, MOLEI, KATSEVICH, EUGENE, JANSON, LUCAS AND RAMDAS, AADITYA. (2022). Fast and powerful conditional randomization testing via distillation. *Biometrika* **109**(2), 277–293.
- LOUIS, BY THOMAS A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* **44**(2), 226–233.
- MCCULLAGH, P. AND NELDER, J. A. (1990). Generalized Linear Models, 2nd Edn.
- MIMITOU, ELENI P and others. (2021). Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nature biotechnology* **39**(10), 1246–1258.
- MORRIS, JOHN and others. (2023). Discovery of target genes and pathways at gwas loci by pooled single-cell crispr screens. *Science* **380**(6646), eadh7699.
- MOSTAFAVI, HAKHAMANESH, SPENCE, JEFFREY P, NAQVI, SAHIN AND PRITCHARD, JONATHAN K. (2023). Systematic differences in discovery of genetic effects on gene expression and complex traits. *Nature Genetics*, 1–10.

- MUSUNURU, KIRAN *and others*. (2021). In vivo crispr base editing of pcsk9 durably lowers cholesterol in primates. *Nature* **593**(7859), 429–434.
- PAPALEXI, EFTHYMIA *and others*. (2021). Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nature genetics* **53**(3), 322–331.
- PRZYBYLA, LARALYNNE AND GILBERT, LUKE A. (2022). A new era in functional genomics screens. *Nature Reviews Genetics* **23**(2), 89–103.
- REPLOGLE, JOSEPH M *and others*. (2020). Combinatorial single-cell crispr screens by direct guide rna capture and targeted sequencing. *Nature biotechnology* **38**(8), 954–961.
- RIPLEY, BRIAN *and others*. (2013). Package ‘mass’. *Cran r* **538**, 113–120.
- ROBINSON, MARK D AND SMYTH, GORDON K. (2008). Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics* **9**(2), 321–332.
- SARKAR, ABHISHEK AND STEPHENS, MATTHEW. (2021). Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics* **53**(6), 770–777.
- SAVOCÀ, E. (2000). Measurement errors in binary regressors: An application to measuring the effects of specific psychiatric diseases on earnings. *Health Services and Outcomes Research Methodology* **1**(2), 149–164.
- STEFANSKI, L. A. (2000). Measurement Error Models. *Journal of the American Statistical Association* **95**(452), 1353–1358.
- TOWNES, F. WILLIAM *and others*. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology* **20**(1), 1–16.

- TRAPNELL, COLE *and others*. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature Biotechnology* **32**(4), 381–386.
- WANG, LINGFEI. (2021). Single-cell normalization and association testing unifying crispr screen and gene co-expression analyses with normalisr. *Nature communications* **12**(1), 6395.
- XIE, SHIQI *and others*. (2019). Global analysis of enhancer targets reveals convergent enhancer-driven regulatory modules. *Cell Reports* **29**(9), 2570–2578.e5.

10. FIGURES

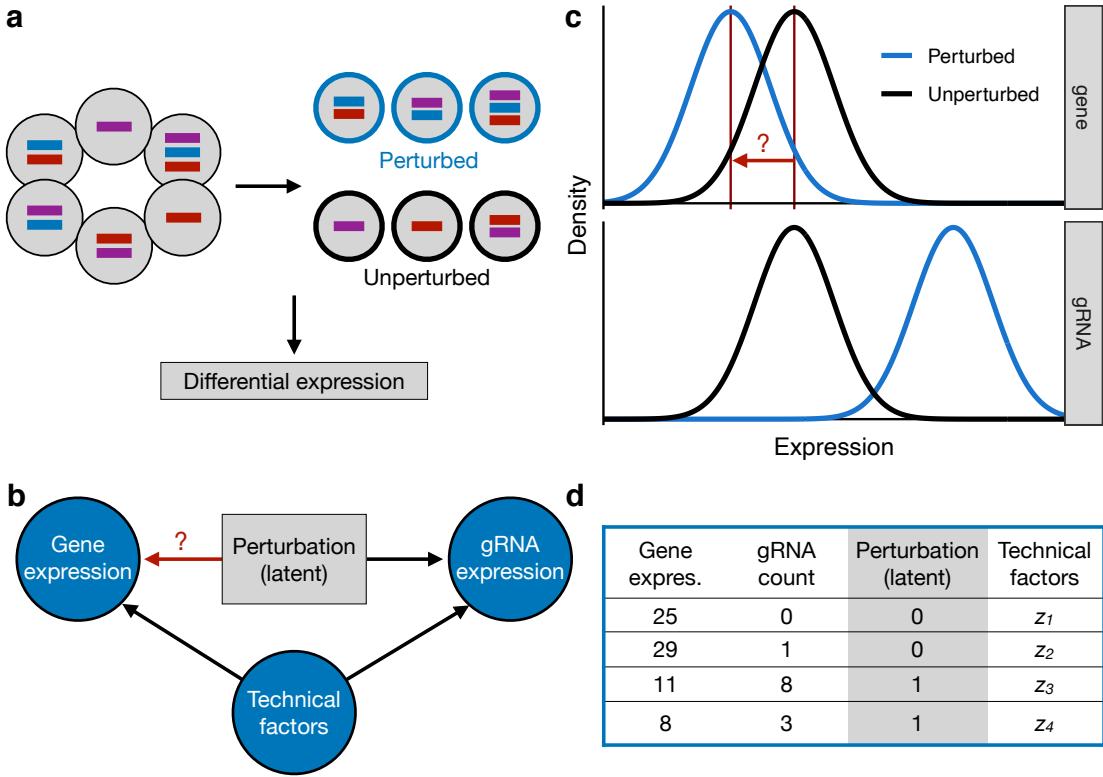


Fig. 1. Experimental design and analysis challenges: **a**, Experimental design. For a given perturbation (e.g., the perturbation indicated in blue), we partition the cells into two groups: perturbed and unperturbed. Next, for a given gene, we conduct a differential expression analysis across the two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

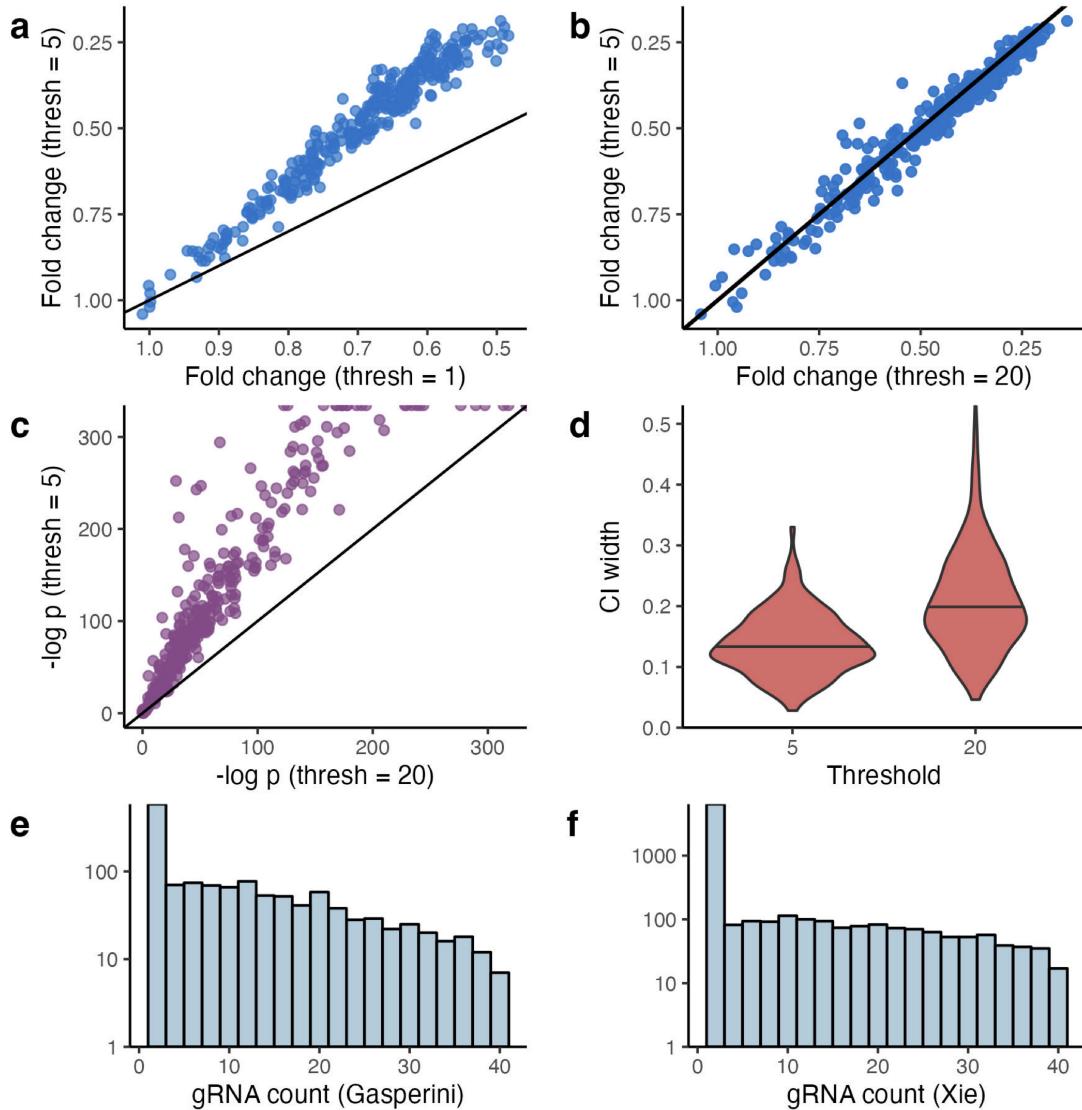


Fig. 2. Empirical challenges of thresholded regression. **a-b**, Estimates for fold change (i.e., $\exp(\beta_1^m)$) produced by threshold = 5 versus threshold = 1 (a) and threshold = 5 versus threshold = 20 (b). The selected threshold substantially impacts the results. **c-d**, p -values (c) and CI widths (d) produced by threshold = 5 versus threshold = 20. The p -values correspond to a test of the null hypothesis $H_0 : \beta_1^m = 0$, i.e., a log fold change in gene expression of zero. A threshold of 5 yields more significant p -values and more confident estimates. **e-f**, Empirical distribution of a gRNA from Gasperini (e) and Xie (f) data (0 counts not shown). These gRNA count distributions do not appear to imply an obvious threshold.

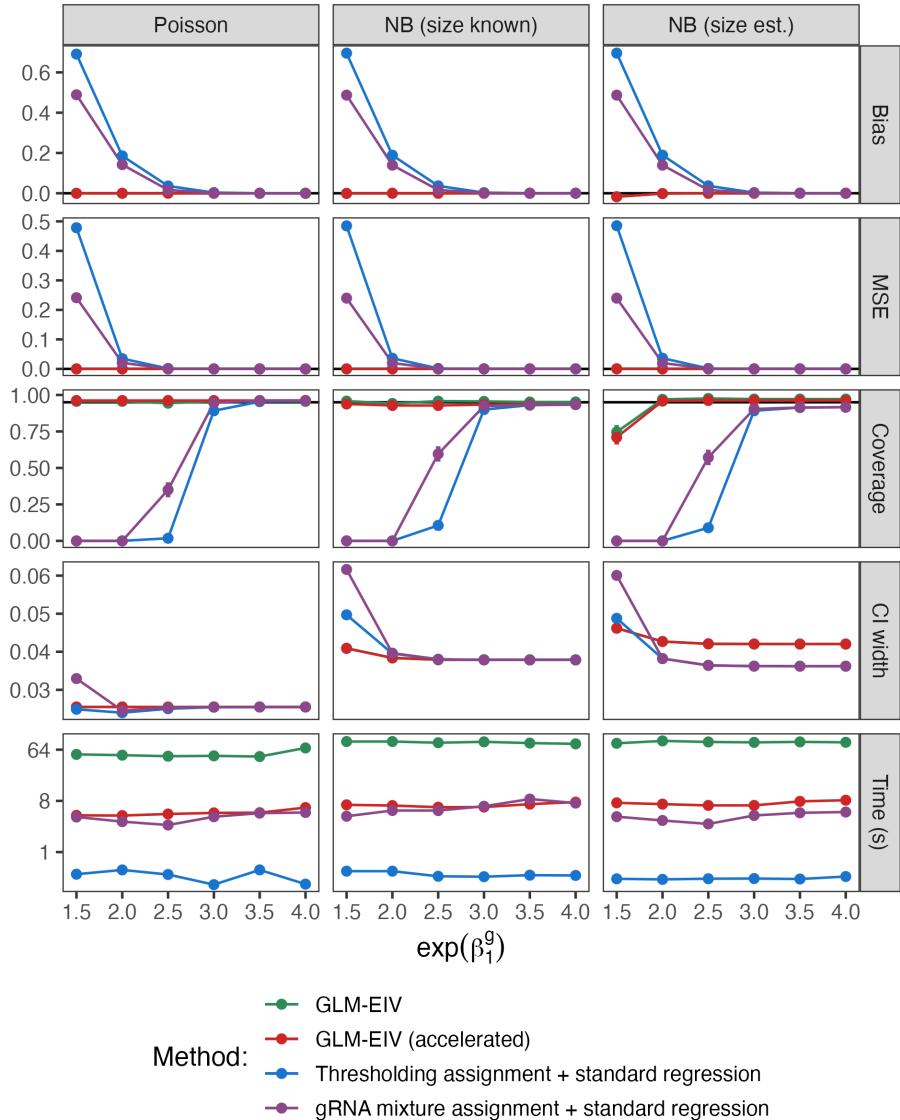


Fig. 3. Simulation study. Columns correspond to distributions (Poisson, NB with known s , NB with estimated s), and rows correspond to metrics (bias, MSE, coverage, CI width, and time). Methods are shown in different colors; GLM-EIV (green) is masked by accelerated GLM-EIV (red) in several panels. Generally, GLM-EIV (both accelerated and non-accelerated versions) outperformed the gRNA-mixture/NB-regression method, which in turn outperformed the thresholding/NB-regression method. The rejection probability (i.e., the probability of rejecting the null hypothesis $H_0 : \beta_1^m = 0$ at level $\alpha = 0.05$) was strictly 1 across methods and parameter settings, likely because the effect size was fairly large.

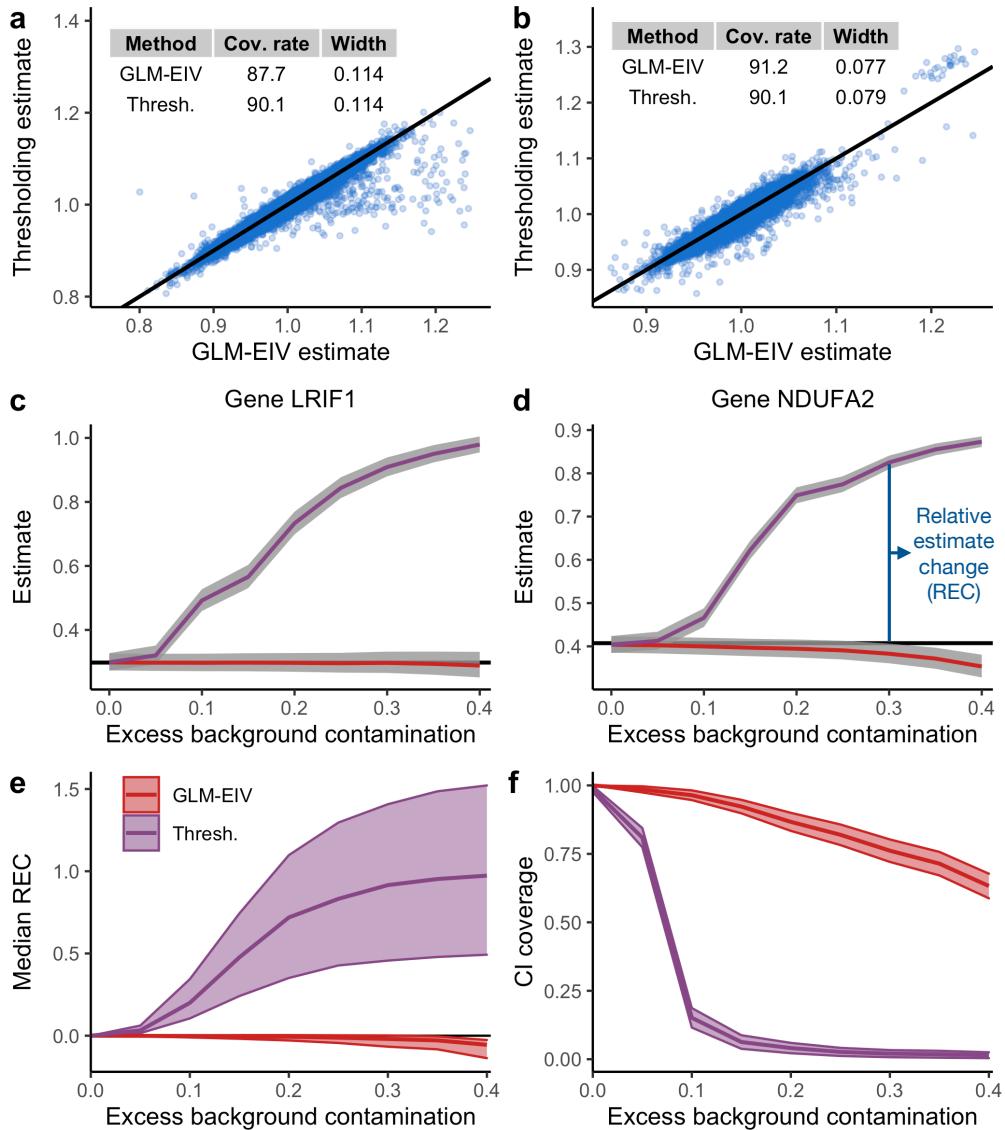


Fig. 4. Applying GLM-EIV to analyze large-scale, high-MOI data. **a-b**, Estimates for fold change produced by GLM-EIV and thresholded regression on Gasperini (**a**) and Xie (**b**) negative control pairs. **c-d**, Estimates produced by GLM-EIV and thresholded regression on two positive control pairs – *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.

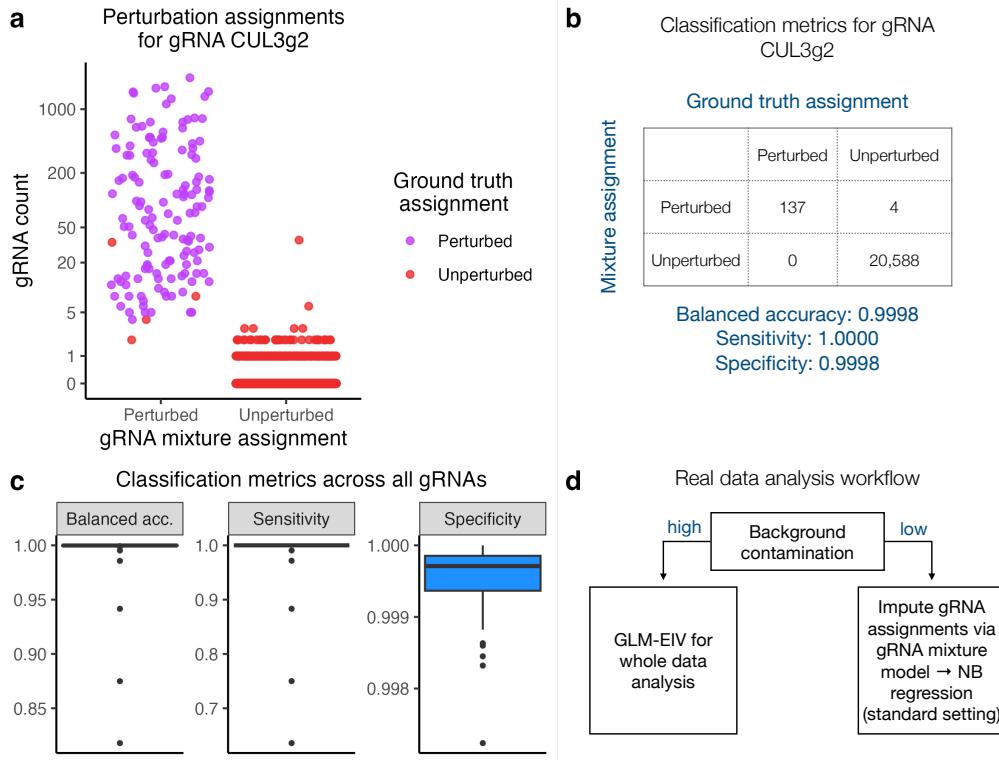


Fig. 5. The gRNA-only mixture assignment functionality of GLM-EIV accurately assigns gRNAs to cells on real low-MOI data. **a**, Each point represents a cell. The position of each cell along the vertical axis indicates the number of gRNA reads (from gRNA “CUL3g2”) observed in that cell. Cells in the left column were classified by the gRNA mixture model as perturbed, while those in the right column were classified as unperturbed. Purple (resp., red) cells were classified by the ground truth method as perturbed (resp., unperturbed). **b**, A confusion matrix comparing the gRNA-to-cell mixture model classifications against the ground truth classifications for gRNA “CUL3g2.” The two sets of classifications were highly concordant, as quantified by balanced accuracy, sensitivity, and specificity metrics. **c**, The balanced accuracy (left), sensitivity (middle), and specificity (right) of the gRNA mixture assignment method across *all* gRNAs. **d**, The proposed data analysis workflow. If the level of background contamination is low, then the gRNA mixture method can be used to impute perturbation identities onto cells, which can then be plugged into downstream analytic tools, such as negative binomial regression or SCEPTRE. On the other hand, if the level of background contamination is high, then the entire GLM-EIV model can be used to analyze the data.

APPENDIX

A. THEORETICAL DETAILS FOR THRESHOLDING ESTIMATOR

We study 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 $(g_1, m_1), \dots, (g_n, m_n)$ on n cells from the following linear model:

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

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. $\hat{\beta}_1^m = [\sum_{i=1}^n (\hat{p}_i - \bar{p})^2]^{-1} [\sum_{i=1}^n (\hat{p}_i - \bar{p})(m_i - \bar{m})]$. We derive the almost sure limit of $\hat{\beta}_1^m$:

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

$$\hat{\beta}_1^m \xrightarrow{a.s.} \beta_1^m \left(\frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} \right) \equiv \beta_1^m \gamma(\beta_1^g, \pi, c, \beta_0^g), \quad (\text{A.2})$$

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

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

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

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 $\pi = 1/2$.) We start with informal proposition statements; we follow up with formal proposition statements below. 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 error 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.

We state five propositions labeled 2 – 6 corresponding to the informal claims above; these propositions are depicted visually in Figure 6.

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

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

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

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

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

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

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

Proposition 5 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.

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

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

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

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

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

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

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

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).

Next, we study the variance of the thresholding estimator, considering a slightly simpler model for this purpose. Suppose the intercepts in (A.1) 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 $\hat{\beta}_m = [\sum_{i=1}^n \hat{p}_i^2]^{-1} [\sum_{i=1}^n \hat{p}_i m_i]$. The following proposition derives the scaled, asymptotic distribution of $\hat{\beta}_m$:

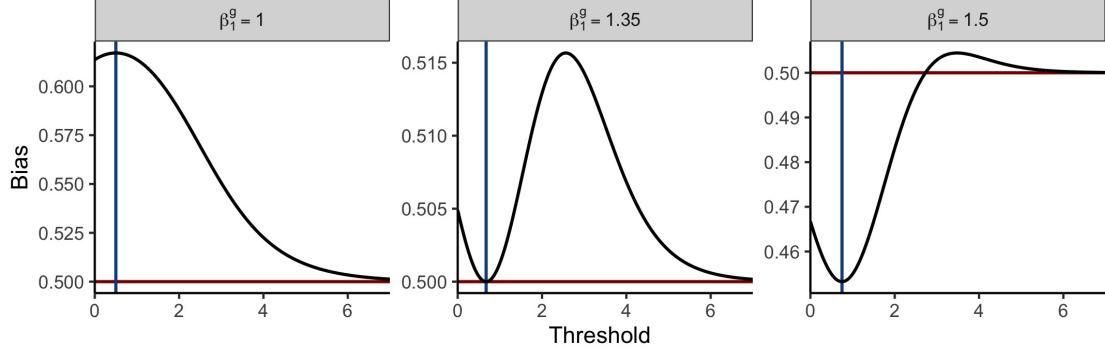


Fig. 6. Bias as a function of threshold. This figure visually depicts Propositions 2-6, which were stated informally above. Asymptotic relative bias is plotted on the vertical axis, and the threshold is plotted on the horizontal axis. Panels correspond to different values of β_1^g . Vertical blue lines indicate the Bayes-optimal decision boundary. Observe that (a) bias is strictly nonzero (proposition 2); (b) bias decreases monotonically in β_1^g (Proposition 3); (c) the Bayes-optimal decision boundary is a critical value of the bias function (Proposition 4), in some cases a maximum and in other cases a minimum; (d) as the threshold tends to infinity, the bias converges to 1/2 (Proposition 5); 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 6). Together, these results illustrate that selecting a good threshold is deceptively challenging.

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

A.1 Organization

The following subsections prove all propositions. Section A.2 introduces some notation. Section A.3 establishes almost sure convergence of the thresholding estimator in the model (A.1), proving Proposition 1. Section A.4 simplifies the expression for the attenuation function γ , and section

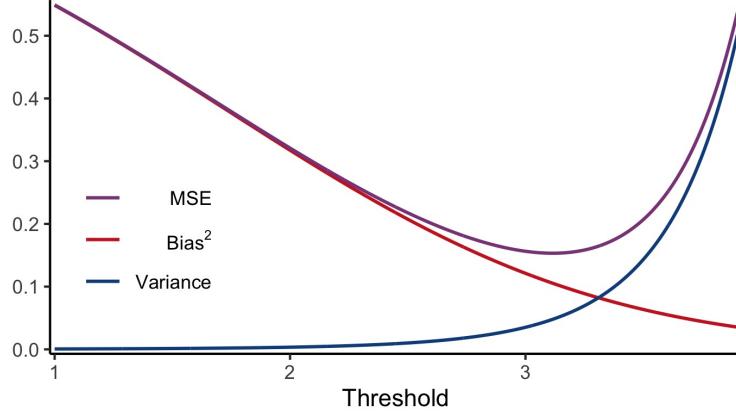


Fig. 7. **Thresholding method bias-variance decomposition.** Bias decreases and variance increases as the threshold tends to infinity. $\beta_1^g = 1$, $\beta_1^m = 1$, and $\pi = 0.1$ in this plot.

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

A.2 Notation

All notation introduced in this subsection (i.e., A.2) pertains to the Gaussian model with intercepts (A.1). Recall that the attenuation function $\gamma : \mathbb{R}^4 \rightarrow \mathbb{R}$ 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

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

Additionally, recall that the asymptotic relative bias function $b : \mathbb{R}^4 \rightarrow \mathbb{R}$ 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 $h : \mathbb{R}^4 \rightarrow \mathbb{R}$ by

$$g(\beta_1^g, c, \pi, \beta_0^g) = (1 - \pi)(\Phi(\beta_0^g + \beta_1^g - c)) - (1 - \pi)(\Phi(\beta_0^g - c)) \quad (\text{A.3})$$

and

$$\begin{aligned} h(\beta_1^g, c, \pi, \beta_0^g) &= [(1 - \pi)(\Phi(\beta_0^g - c)) + \pi(\Phi(\beta_0^g + \beta_1^g - c))] \times \\ &\quad [(1 - \pi)(\Phi(c - \beta_0^g)) + \pi(\Phi(c - \beta_0^g - \beta_1^g))]. \quad (\text{A.4}) \end{aligned}$$

We use $f : \mathbb{R} \rightarrow \mathbb{R}$ 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 $r : \mathbb{R}^p \rightarrow \mathbb{R}$, point $x \in \mathbb{R}^p$, and index $i \in \{1, \dots, p\}$, we use the symbol $D_i r(x)$ to refer to the derivative of the i th argument 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 argument of γ (the argument 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 argument of g (the argument corresponding to c) evaluated at (β_1^g, c, π) .

A.3 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 (A.1). Dividing by n in (A.2), we can express $\hat{\beta}_1^m$ as

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

By weak LLN, $\hat{\beta}_1^m \xrightarrow{P} \text{Cov}(\hat{p}_i, m_i)/\mathbb{V}(\hat{p}_i)$. To compute this quantity, we first compute several simpler quantities:

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

2. Expectation of \hat{p}_i :

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

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

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

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

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

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

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

Combining these expressions, we have that

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

A.4 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

$$\mathbb{E}[\hat{p}_i] = (1 - \pi)\bar{\Phi}(c - \beta_0^g) + \pi\bar{\Phi}(c - \beta_0^g - \beta_1^g) = (1 - \pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c), \quad (\text{A.5})$$

and so

$$\begin{aligned} 1 - \mathbb{E}[\hat{p}_i] &= (1 - \pi) + \pi - \mathbb{E}[\hat{p}_i] = (1 - \pi)(1 - \bar{\Phi}(c - \beta_0^g)) + \pi(1 - \bar{\Phi}(c - \beta_0^g - \beta_1^g)) \\ &= (1 - \pi)\Phi(c - \beta_0^g) + \pi\Phi(c - \beta_0^g - \beta_1^g). \end{aligned} \quad (\text{A.6})$$

Next,

$$\omega = \Phi(\beta_1^g + \beta_0^g - c), \quad (\text{A.7})$$

and so

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

Combining (A.5, A.6, A.7, A.8), we find that

$$\begin{aligned} \gamma(\beta_1^g, c, \pi) &= \frac{\pi(\omega - \mathbb{E}[\hat{p}_i])}{\mathbb{E}[\hat{p}_i](1 - \mathbb{E}[\hat{p}_i])} \\ &= \frac{\pi[(1 - \pi)\Phi(\beta_0^g + \beta_1^g - c) - (1 - \pi)\Phi(\beta_0^g - c)]}{[(1 - \pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c)][(1 - \pi)\Phi(c - \beta_0^g) + \pi\Phi(c - \beta_0^g - \beta_1^g)]}. \end{aligned} \quad (\text{A.9})$$

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

$$\gamma(\beta_1^g, c, 1/2) = \frac{\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)}{[\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)][\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]}. \quad (\text{A.10})$$

Recalling the definitions of g (A.3) and h (A.4), 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 (A.10) is identical to

$$\gamma(\beta_1^g, c, 1/2) = \frac{(4)(1/2)g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)} = \frac{2g(\beta_1^g, c, 1/2)}{4h(\beta_1^g, c, 1/2)}, \quad (\text{A.11})$$

i.e., the numerator and denominator of (A.11) coincide with those of (A.10). We sometimes will use the notation $2 \cdot g$ and $4 \cdot h$ to refer to the numerator and denominator of (A.10), respectively.

A.5 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 the FTC (fundamental theorem of calculus) and the evenness of f , we have that

$$\begin{aligned} D_2g(\beta_1^g, c, \pi) &= -(1 - \pi)f(\beta_0^g + \beta_1^g - c) + (1 - \pi)f(\beta_0^g - c) \\ &= (1 - \pi)f(c - \beta_0^g) - (1 - \pi)f(c - \beta_0^g - \beta_1^g). \end{aligned} \quad (\text{A.12})$$

Second, we have that

$$\begin{aligned} D_2h(\beta_1^g, c, \pi) &= -[(1 - \pi)f(\beta_0^g - c) + \pi f(\beta_0^g + \beta_1^g - c)][(1 - \pi)\Phi(c - \beta_0^g) + \pi\Phi(c - \beta_0^g - \beta_1^g)] \\ &\quad + [(1 - \pi)f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)][(1 - \pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c)] \\ &= [(1 - \pi)f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)] \times \\ &\quad \left[(1 - \pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c) - (1 - \pi)\Phi(c - \beta_0^g) - \pi\Phi(c - \beta_0^g - \beta_1^g) \right]. \end{aligned} \quad (\text{A.13})$$

A.6 Limit of γ in c

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

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

Therefore, we can apply L'Hôpital's rule. We have by (A.12) and (A.13) that

$$\lim_{c \rightarrow \infty} \gamma(\beta_1^g, c, \pi) = \lim_{c \rightarrow \infty} \frac{\pi D_2g(\beta_1^g, c, \pi)}{D_2h(\beta_1^g, c, \pi)}$$

$$\begin{aligned}
&= \lim_{c \rightarrow \infty} \left\{ \frac{(1-\pi)f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)}{\pi(1-\pi)f(c - \beta_0^g) - \pi(1-\pi)f(c - \beta_0^g - \beta_1^g)} \times \right. \\
&\quad \left. \left[(1-\pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c) - (1-\pi)\Phi(c - \beta_0^g) - \pi\Phi(c - \beta_0^g - \beta_1^g) \right] \right\}^{-1}. \quad (\text{A.14})
\end{aligned}$$

We evaluate the two terms in the product (A.14) separately. Dividing by $f(c - \beta_0^g - \beta_1^g) > 0$, we see that

$$\frac{(1-\pi)f(c - \beta_0^g) + \pi f(c - \beta_0^g - \beta_1^g)}{\pi(1-\pi)f(c - \beta_0^g) - \pi(1-\pi)f(c - \beta_0^g - \beta_1^g)} = \frac{\frac{(1-\pi)f(c - \beta_0^g)}{f(c - \beta_0^g - \beta_1^g)} + \pi}{\frac{\pi(1-\pi)f(c - \beta_0^g)}{f(c - \beta_0^g - \beta_1^g)} - \pi(1-\pi)}. \quad (\text{A.15})$$

To evaluate the limit of (A.15), we first evaluate the limit of

$$\begin{aligned}
\frac{f(c - \beta_0^g)}{f(c - \beta_0^g - \beta_1^g)} &= \frac{\exp[-(1/2)(c - \beta_0^g)^2]}{\exp[-(1/2)(c - \beta_0^g - \beta_1^g)^2]} \\
&= \frac{\exp[-(1/2)(c^2 - 2c\beta_0^g + (\beta_0^g)^2)]}{\exp[-(1/2)(c^2 - 2c\beta_0^g - 2c\beta_1^g + (\beta_0^g)^2 + 2(\beta_0^g\beta_1^g) + (\beta_1^g)^2)]} \\
&= \exp[-c^2/2 + c\beta_0^g - (\beta_0^g)^2/2 \\
&\quad + c^2/2 - c\beta_0^g - c\beta_1^g + (\beta_0^g)^2/2 + \beta_0^g\beta_1^g + (\beta_1^g)^2/2] \\
&= \exp[-c\beta_1^g + \beta_0^g\beta_1^g + (\beta_1^g)^2/2] = \exp[\beta_0^g\beta_1^g + (\beta_1^g)^2/2] \exp[-c\beta_1^g]. \quad (\text{A.16})
\end{aligned}$$

Taking the limit in (A.16), we obtain

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

for $\beta_1^g > 0$. We now can evaluate the limit of (A.15):

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

Next, we compute the limit of the other term in the product (A.14):

$$\begin{aligned}
\lim_{c \rightarrow \infty} \left[(1-\pi)\Phi(\beta_0^g - c) + \pi\Phi(\beta_0^g + \beta_1^g - c) \right. \\
\left. - (1-\pi)\Phi(c - \beta_0^g) - \pi\Phi(c - \beta_0^g - \beta_1^g) \right] = -(1-\pi) - \pi = -1. \quad (\text{A.17})
\end{aligned}$$

Combining (A.15) and (A.17), the limit (A.14) evaluates to

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

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

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

A corollary is that $\lim_{c \rightarrow \infty} b(\beta_1^g, c, 1/2) = 1/2$.

A.7 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 (A.11), the quotient rule implies that

$$D_2\gamma(\beta_1^g, c, 1/2) = \frac{D_2[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_2[4h(\beta_1^g, c, 1/2)]}{[4h(\beta_1^g, c, \pi)]^2}. \quad (\text{A.18})$$

We have by (A.12) that

$$D_2[2g(\beta_1^g, c_{\text{bayes}}, 1/2)] = f(\beta_1^g/2) - f(-\beta_1^g/2) = f(\beta_1^g/2) - f(\beta_1^g/2) = 0. \quad (\text{A.19})$$

Similarly, we have by (A.13) that

$$D_2[4h(\beta_1^g, c_{\text{bayes}}, \pi)] = [f(\beta_1^g/2) + f(-\beta_1^g/2)][\Phi(-\beta_1^g/2) + \Phi(\beta_1^g/2) - \Phi(\beta_1^g/2) - \Phi(-\beta_1^g/2)] = 0. \quad (\text{A.20})$$

Plugging in (A.20) and (A.19) to (A.18), 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.8 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 $r : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}$ 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 5 and proven in Section A.6) 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.9 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 A.1 The following inequality holds:

$$\begin{aligned} & [\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)] \cdot [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c) + \Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ & \geq [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)] [\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]. \quad (\text{A.21}) \end{aligned}$$

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

Case 1: $\beta_1^g < 0$. 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 (A.21) is negative.

Turning our attention of the left-hand side of (A.21), we see that

$$\Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1 - \Phi(\beta_0^g + \beta_1^g - c) + \Phi(c - \beta_0^g - \beta_1^g) = 1. \quad (\text{A.22})$$

Additionally, $\Phi(\beta_0^g - c) < 1$ and $\Phi(c - \beta_0^g) > 0$. Combining these facts with (A.22), 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 (A.21) is positive.

The inequality holds for $\beta_1^g < 0$.

Case 2: $\beta_1^g \geq 0$. We will show that the first term on the LHS of (A.21) is greater than the first term on the RHS of (A.21), and likewise that the second term on the LHS is greater than the second term on the RHS, implying the truth of the inequality. Focusing on the first term, the positivity of $\Phi(\beta_0^g - c)$ implies that $\Phi(\beta_0^g - c) \geq -\Phi(\beta_0^g - c)$, and so

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

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

$$\beta_1^g + \beta_0^g - c \geq \beta_0^g - c \implies \Phi(\beta_1^g + \beta_0^g - c) - \Phi(\beta_0^g - c) \geq 0. \quad (\text{A.23})$$

Adding $\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)$ to both sides of (A.23) 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 $\beta_1^g \in \mathbb{R}$. \square

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

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

$$D_1[2g(\beta_1^g, c, 1/2)] = f(\beta_0^g + \beta_1^g - c), \quad (\text{A.24})$$

$$\begin{aligned} D_1[4h(\beta_1^g, c, 1/2)] &= f(\beta_0^g + \beta_1^g - c) [\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ &\quad - f(\beta_0^g + \beta_1^g - c) [\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)]. \end{aligned} \quad (\text{A.25})$$

Proof: Apply FTC and product rule. \square

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 (A.21) and multiplying by $f(\beta_0^g + \beta_1^g - c) > 0$ yields

$$\begin{aligned} &f(\beta_0^g + \beta_1^g - c) [\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)] [\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] \\ &\geq f(\beta_0^g + \beta_1^g - c) [\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)] [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)] \\ &\quad - f(\beta_0^g + \beta_1^g - c) [\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)] [\Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c)]. \end{aligned} \quad (\text{A.26})$$

Next, recall that

$$2g(\beta_1^g, c, 1/2) = \Phi(\beta_0^g + \beta_1^g - c) - \Phi(\beta_0^g - c). \quad (\text{A.27})$$

and

$$4h(\beta_1^g, c, 1/2) = [\Phi(\beta_0^g - c) + \Phi(\beta_0^g + \beta_1^g - c)] [\Phi(c - \beta_0^g) + \Phi(c - \beta_0^g - \beta_1^g)]. \quad (\text{A.28})$$

Substituting (A.24, A.25, A.27, A.28) into (A.26) produces

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

or

$$D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)] \geq 0. \quad (\text{A.29})$$

The quotient rule implies that

$$D_1\gamma(\beta_1^g, c, 1/2) = \frac{D_1[2g(\beta_1^g, c, 1/2)]4h(\beta_1^g, c, 1/2) - 2g(\beta_1^g, c, 1/2)D_1[4h(\beta_1^g, c, 1/2)]}{[4h(\beta_1^g, c, 1/2)]^2}. \quad (\text{A.30})$$

We conclude by (A.29) and (A.30) 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.10 Strict attenuation bias

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

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

Similarly,

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

The function $\gamma(\beta_1^g, c, 1/2, \beta_0^g)$ is monotonically increasing in β_1^g (as stated in Proposition 3 and proven in section A.9). 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.11 Bias-variance decomposition in no-intercept model

We prove the bias-variance decomposition for the no-intercept version of (A.1). 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,

$$\sqrt{n}(\hat{\beta}_m - l) = \frac{(1/\sqrt{n}) \sum_{i=1}^n \hat{p}_i (m_i - l\hat{p}_i)}{(1/n) \sum_{i=1}^n \hat{p}_i^2}. \quad (\text{A.31})$$

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 $\hat{p}_i p_i$: $\mathbb{E}[\hat{p}_i p_i] = \mathbb{E}[\hat{p}_i | p_i = 1] \mathbb{P}[p_i = 1] = \omega\pi$.
3. Expectation of $\hat{p}_i m_i$:

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

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

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

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

$$\mathbb{E}[\hat{p}_i(m_i - l\hat{p}_i)] = \mathbb{E}[\hat{p}_i m_i] - l\mathbb{E}[\hat{p}_i] = \beta_m \omega \pi - \left(\frac{\beta_m \omega \pi}{\zeta(1-\pi) + \omega \pi} \right) [\zeta(1-\pi) + \omega \pi] = 0. \quad (\text{A.32})$$

Additionally,

$$\begin{aligned} \mathbb{V}[\hat{p}_i(m_i - l\hat{p}_i)] &= \mathbb{E}[\hat{p}_i^2(m_i - l\hat{p}_i)^2] - (\mathbb{E}[\hat{p}_i(m_i - l\hat{p}_i)])^2 \\ &= \mathbb{E}[\hat{p}_i m_i^2] - 2l\mathbb{E}[m_i \hat{p}_i] + l^2 \mathbb{E}[\hat{p}_i] = \beta_m^2 \omega \pi + \mathbb{E}[\hat{p}_i] - 2l\beta_m \omega \pi + l^2 \mathbb{E}[\hat{p}_i] \\ &= \beta_m \omega \pi (\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1 + l^2). \end{aligned} \quad (\text{A.33})$$

Therefore, by CLT, (A.32), and (A.33),

$$(1/\sqrt{n}) \sum_{i=1}^n \hat{p}_i(m_i - l\hat{p}_i) \xrightarrow{d} N(0, \beta_m \omega \pi (\beta_m - 2l) + \mathbb{E}[\hat{p}_i](1 + l^2)). \quad (\text{A.34})$$

Next, by weak LLN,

$$(1/n) \sum_{i=1}^n \hat{p}_i^2 = (1/n) \sum_{i=1}^n \hat{p}_i \xrightarrow{P} \mathbb{E}[\hat{p}_i]. \quad (\text{A.35})$$

Finally, by (A.31), (A.34), (A.35), and Slutsky's Theorem,

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

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

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

completing the bias-variance decomposition.

A.12 Bayes-optimal decision boundary for non-Gaussian mixture distributions and GLMs

We report the Bayes-optimal decision boundary for gRNA count distributions that are non-Gaussian. First, consider a simple two-component Poisson mixture model with means μ_0 and μ_1 and mixing probability π :

$$p(k; \mu_0, \mu_1, \pi) = (1 - \pi)f(k; \mu_0) + \pi f(k; \mu_1),$$

where $f(k; \mu) = (\mu^k e^{-\mu})/\mu!$ is a Poisson density. Suppose we draw an observation from this distribution. The Bayes-optimal threshold for classifying the observation as having been drawn from the first or second component is

$$\frac{\mu_0 - \mu_1 + \log(\pi) - \log(1 - \pi)}{\log(\mu_0) - \log(\mu_1)}. \quad (\text{A.36})$$

Next, consider the slightly more complex Poisson mixture GLM:

$$g_i | (p_i, z_i, o_i) \sim \text{Pois}(\mu_i); \quad r(\mu_i) = \beta_0 + \beta_1 p_i + \gamma^T z_i + o_i,$$

where $p_i \sim \text{Bern}(\pi)$ is unobserved. Conditional on the covariates and offset, the mean of the unperturbed component is $\mu_i(1) = r^{-1}(\beta_0 + \gamma^T z_i + o_i)$, and that of the perturbed component is $\mu_i(0) = r^{-1}(\beta_0 + \beta_1 + \gamma^T z_i + o_i)$. The Bayes-optimal threshold is obtained by plugging in $\mu_i(1)$ for μ_1 and $\mu_i(0)$ for μ_0 in (A.36). To obtain a fixed gRNA assignment threshold across cells, we compute the Bayes-optimal decision boundary for each cell and then take the average across cells. The situation is similar for the negative binomial (with known size s) distribution; the Bayes-optimal decision boundary in this case is

$$\frac{s [\log(\mu_0 + s) - \log(\mu_1 + s)] + \log(\pi) - \log(1 - \pi)}{\log(\mu_0(\mu_1 + s)) - \log(\mu_1(\mu_0 + s))}.$$

B. ESTIMATION AND INFERENCE IN THE GLM-EIV MODEL

B.1 *Detailed specification of the model*

We provide a more precise and technical specification of the GLM-EIV model than provided in the main text. 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 i th 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 i th 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 i th observation be $r_m(\mu_i^m) \equiv l_i^m$, where $r_m : \mathbb{R} \rightarrow \mathbb{R}$ is a strictly increasing, differentiable link function. Let $\psi_m : \mathbb{R} \rightarrow \mathbb{R}$ 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 $c_m : \mathbb{R} \rightarrow \mathbb{R}$ be the carrying density of the selected exponential family distribution. The density f_m of m_i conditional on the canonical parameter η_i is $f_m(m_i; \eta_i^m) = \exp \{m_i \eta_i^m - \psi_m(\eta_i^m) + c_m(m_i)\}$. We implicitly set the “scale” term (i.e., the $a(\phi)$ term in McCullagh and Nelder (1990), Eqn. 2.4, p. 28) to unity; this slightly simplified model encompasses the most important distributions for our purposes, including the Poisson, negative binomial, and Gaussian (with unit variance) distributions.

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 $\eta_i^g = ([\psi'_g]^{-1} \circ r_g^{-1})(l_i^g) \equiv h_g(l_i^g)$. 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 $\theta = [\beta_m, \beta_g, \pi]^T \in \mathbb{R}^{2d+1}$.

B.2 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_i s 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 $i \in \{1, \dots, n\}$ 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 $m := [m_1, \dots, m_n]$, 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. $[m, m]^T \equiv [m_1, \dots, m_n, m_1, \dots, m_n]$, and let $[g, g]^T$, $[o^g, o^g]^T$, and $[o^m, o^m]^T$ be defined similarly.

B.3 Log likelihood and estimation

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 model log-likelihood is

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

We see from (B.1) that the GLM-EIV model is equivalent to a two-component mixture of *products* of GLM densities. 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 i th canonical parameter at the t -th iteration of the algorithm of the gene ex-

pression distribution that results from setting p_i to k , i.e. $[\eta_i^m(k)]^{(t)} \equiv h_m(\langle \tilde{x}_i(k), \beta_m^{(t)} \rangle + o_i^m)$. Similarly, let $[\eta_i^g(k)]^{(t)}$ be defined by $[\eta_i^g(k)]^{(t)} \equiv h_g(\langle \tilde{x}_i(k), \beta_g^{(t)} \rangle + o_i^g)$. Next, for $k \in \{0, 1\}$, define $\alpha_i^{(t)}(k)$ by

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

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

$$\begin{aligned}T_i^{(t)}(1) &= \mathbb{P}(P_i = 1 | M_i = m_i, G_i = g_i, \theta^{(t)}) = \frac{\pi^{(t)}(1) \alpha_i^{(t)}(1)}{\sum_{k=0}^1 \pi^{(t)}(k) \alpha_i^{(t)}(k)} \quad (\text{by Bayes rule}) \\ &= \frac{1}{\frac{\pi^{(t)}(0) \alpha_i(0)}{\pi^{(t)}(1) \alpha_i(1)} + 1} = \frac{1}{\exp\left(\log\left(\frac{\pi^{(t)}(0) \alpha_i(0)}{\pi^{(t)}(1) \alpha_i(1)}\right)\right) + 1} = \frac{1}{\exp(q_i^{(t)}) + 1}, \quad (\text{B.2})\end{aligned}$$

where we set

$$q_i^{(t)} := \log\left(\frac{\pi^{(t)}(0) \alpha_i^{(t)}(0)}{\pi^{(t)}(1) \alpha_i^{(t)}(1)}\right). \quad (\text{B.3})$$

Next, we have that

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

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

M step

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

$$\mathcal{L}(\theta; m, g, p) = \sum_{i=1}^n [p_i \log(\pi) + (1 - p_i) \log(1 - \pi)] + \sum_{i=1}^n \log(f_m(m_i; \eta_i^m)) + \sum_{i=1}^n \log(f_g(g_i; \eta_i^g)). \quad (\text{B.4})$$

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

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \sum_{i=1}^n \left[T_i^{(t)}(1) \log(\pi) + T_i^{(t)}(0) \log(1-\pi) \right] \\ &\quad + \sum_{k=0}^1 \sum_{i=1}^n T_i^{(t)}(k) \log [f_m(m_i; \eta_i^m(k))] + \sum_{k=0}^1 \sum_{i=1}^n T_i^{(t)}(k) \log [f_g(g_i; \eta_i^{g,b}(k))]. \end{aligned} \quad (\text{B.5})$$

The three terms of (B.5) 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 (B.5), 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 ,

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

Thus, the maximizer $\pi^{(t+1)}$ of (B.5) in π is $\pi^{(t+1)} = (1/n) \sum_{i=1}^n T_i^{(t)}(1)$. Next, define $w^{(t)} = [T_1^{(t)}(0), \dots, T_n^{(t)}(0), T_1^{(t)}(1), \dots, T_n^{(t)}(1)]^T \in \mathbb{R}^{2n}$. We can view the second term of (B.5) 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 (B.5) 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 (B.5) 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.4 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 $i \in \{1, \dots, n\}$, $j \in \{0, 1\}$, and $\theta = (\pi, \beta_m, \beta_g)$, let $T_i^\theta(j)$ be defined by

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

Let the $n \times n$ matrix $T^\theta(j)$ be given by $T^\theta(j) = \text{diag}\{T_1^\theta(j), \dots, T_n^\theta(j)\}$. Next, define the diagonal $n \times n$ matrices Δ^m , $[\Delta']^m$, V^m , and H^m by

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

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

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

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 (B.1) 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 (B.1) is a complicated mixture. Louis

(1982) showed that $J(\theta; m, g)$ is equivalent to the following quantity:

$$\begin{aligned} J(\theta; m, g) = & -\mathbb{E} [\nabla^2 \mathcal{L}(\theta; m, g, p)|G = g, M = m] \\ & + \mathbb{E} [\nabla \mathcal{L}(\theta; m, g, p)|G = g, M = m] \mathbb{E} [\nabla \mathcal{L}(\theta; m, g, p)|G = g, M = m]^T \\ & - \mathbb{E} [\nabla \mathcal{L}(\theta; m, g, p) \nabla \mathcal{L}(\theta; m, g, p)^T |G = g, M = m]. \quad (\text{B.6}) \end{aligned}$$

The observed information matrix $J(\theta; m, g)$ has dimension $(2d + 1) \times (2d + 1)$. Recall that the complete-data log-likelihood (B.4) 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 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

$$J_\pi(\theta; m, g) = -\mathbb{E} [\nabla_\pi^2 \mathcal{L}(\theta; m, g, p)] + (\mathbb{E} [\nabla_\pi \mathcal{L}(\theta; m, g, p)])^2 - \mathbb{E} [(\nabla_\pi \mathcal{L}(\theta; m, g, p))^2]. \quad (\text{B.7})$$

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

$$\begin{aligned} \nabla_\pi \mathcal{L}(\theta; m, g, p) &= \frac{\partial}{\partial \pi} \left(\sum_{i=1}^n p_i \log(\pi) + \sum_{i=1}^n (1-p_i) \log(1-\pi) \right) \\ &= \frac{\sum_{i=1}^n p_i}{\pi} - \frac{\sum_{i=1}^n (1-p_i)}{1-\pi} = \frac{\sum_{i=1}^n p_i}{\pi} - \frac{n - \sum_{i=1}^n p_i}{1-\pi} = \left(\frac{1}{\pi} + \frac{1}{1-\pi} \right) \sum_{i=1}^n p_i - \frac{n}{1-\pi}. \quad (\text{B.8}) \end{aligned}$$

The second derivative is

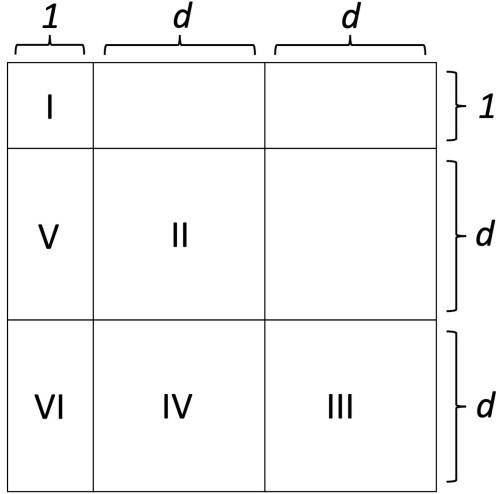


Fig. 8. Block structure of the observed information matrix $J(\theta; m, g) = -\nabla^2 \mathcal{L}(\theta; m, g)$. The matrix is symmetric, and so we only need to compute submatrices I-VI to compute the entire matrix.

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

We compute the expectation of the first term of (B.7):

$$\begin{aligned} \mathbb{E}[-\nabla_\pi^2 \mathcal{L}(\theta; m, g, p)] &= -\mathbb{E} \left[\frac{(\sum_{i=1}^n p_i) - n}{(1 - \pi)^2} - \frac{\sum_{i=1}^n p_i}{\pi^2} \right] \\ &= -\mathbb{E} \left\{ \left[\frac{1}{(1 - \pi)^2} - \frac{1}{\pi^2} \right] \sum_{i=1}^n p_i - \frac{n}{(1 - \pi)^2} \right\} = - \left\{ \left[\frac{1}{(1 - \pi)^2} - \frac{1}{\pi^2} \right] \sum_{i=1}^n T_i^\theta(1) - \frac{n}{(1 - \pi)^2} \right\} \\ &= \left[\frac{1}{\pi^2} - \frac{1}{(1 - \pi)^2} \right] \sum_{i=1}^n T_i^\theta(1) + \frac{n}{(1 - \pi)^2}. \quad (\text{B.9}) \end{aligned}$$

Next, we compute the difference of the second two pieces of (B.7). To this end, define $a \equiv 1/(1 - \pi) + 1/\pi$ and $b \equiv n/(1 - \pi)$. We have that

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

Next,

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

Therefore,

$$\begin{aligned}
& (\mathbb{E}[\nabla_\pi \mathcal{L}(\theta; m, g, p)])^2 - \mathbb{E}[\nabla_\pi \mathcal{L}(\theta; m, g, p)^2] \\
&= a^2 \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[p_i] \mathbb{E}[p_j] - a^2 \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[p_i p_j] = a^2 \left(\sum_{i=1}^n \mathbb{E}[p_i]^2 - \mathbb{E}[p_i^2] \right) \\
&= a^2 \left(\sum_{i=1}^n [T_i^\theta(1)]^2 - T_i^\theta(1) \right) = \left(\frac{1}{(1-\pi)} + \frac{1}{\pi} \right)^2 \left(\sum_{i=1}^n [T_i^\theta(1)]^2 - T_i^\theta(1) \right). \quad (\text{B.10})
\end{aligned}$$

Stringing (B.7), (B.9) and (B.10) together, we obtain

$$\begin{aligned}
J_\pi(\theta; m, g) &= \left[\frac{1}{\pi^2} - \frac{1}{(1-\pi)^2} \right] \sum_{i=1}^n T_i^\theta(1) + \frac{n}{(1-\pi)^2} \\
&\quad + \left(\frac{1}{(1-\pi)} + \frac{1}{\pi} \right)^2 \left(\sum_{i=1}^n [T_i^\theta(1)]^2 - T_i^\theta(1) \right). \quad (\text{B.11})
\end{aligned}$$

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

$$\begin{aligned}
J_{\beta^m}(\theta; m, g) &= -\mathbb{E}[\nabla_{\beta^m}^2 \mathcal{L}(\theta; m, g, p)] \\
&\quad + \mathbb{E}[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)] \mathbb{E}[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)]^T - \mathbb{E}[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p) \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)^T]. \quad (\text{B.12})
\end{aligned}$$

Standard GLM results imply that $-\nabla_{\beta^m}^2 \mathcal{L}(\theta; m, g, p) = \tilde{X}^T (\Delta^m V^m \Delta^m - [\Delta']^m H^m) \tilde{X}$ and $\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p) = \tilde{X}^T \Delta^m s^m$. We compute the first term of (B.12). The (k, l) th entry of this matrix is

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

We therefore have that

$$\mathbb{E} \left[-\nabla_{\beta^m}^2 \mathcal{L}(\theta; m, g, p) \right] = \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s) [\Delta^m(s) V^m(s) \Delta^m(s) - [\Delta']^m(s) H^m(s)] \tilde{X}(s). \quad (\text{B.13})$$

Next, we compute the difference of the last two terms of (B.12). The (k, l) th entry is

$$\begin{aligned} & \left[\mathbb{E} [\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)] \mathbb{E} [\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)]^T \right. \\ & \quad \left. - \mathbb{E} [\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p) \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)^T] \right] [k, l] \\ &= \left[\mathbb{E} [\tilde{X}^T \Delta^m s^m] \mathbb{E} [\tilde{X}^T \Delta^m s^m]^T \right] [k, l] - \mathbb{E} [\tilde{X}^T \Delta^m s^m (s^m)^T \Delta^m \tilde{X}] [k, l] \\ &= \mathbb{E} [\tilde{X}_{[, k]}^T \Delta^m s^m] \mathbb{E} [\tilde{X}_{[, l]}^T \Delta^m s^m] - \mathbb{E} [\tilde{X}_{[, k]}^T \Delta^m s^m (s^m)^T \Delta^m \tilde{X}_{[, l]}] \\ &= \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}[\tilde{x}_{jl} \Delta_j^m s_j^m] - \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}] \\ & \quad - \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 [\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)] \\ & \quad \cdot [\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)] \\ & - \sum_{i=1}^n [\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))^2 \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^m(s) H_i^m(t) T_i^\theta(t) \Delta_i^m(t) H_i^m(t) \tilde{x}_{il}(t) \right] \\ & \quad - \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] \\ & \quad - \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 (B.12) is therefore

$$\begin{aligned} & \mathbb{E} [\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)] \mathbb{E} [\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)]^T - \mathbb{E} [\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p) \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)^T] \\ &= \sum_{s=0}^1 \sum_{t=0}^1 \tilde{X}(s)^T T^\theta(s) \Delta^m(s) H^m(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(t) \\ &\quad - \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s) (\Delta^m(s))^2 (H^m(s))^2 \tilde{X}(s). \end{aligned} \quad (\text{B.14})$$

Combining (B.12), (B.13), (B.14), we find that

$$\begin{aligned} J_{\beta^m}(\theta; m, g) &= \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s) [\Delta^m(s) V^m(s) \Delta^m(s) - [\Delta']^m(s) H^m(s)] \tilde{X}(s) \\ &\quad + \sum_{s=0}^1 \sum_{t=0}^1 \tilde{X}(s)^T T^\theta(s) \Delta^m(s) H^m(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(t) \\ &\quad - \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s) (\Delta^m(s))^2 (H^m(s))^2 \tilde{X}(s). \end{aligned} \quad (\text{B.15})$$

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 (B.15). Substituting g for m in this equation yields

$$\begin{aligned} J_{\beta^g}(\theta; m, g) &= \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s) [\Delta^g(s) V^g(s) \Delta^g(s) - [\Delta']^g(s) H^g(s)] \tilde{X}(s) \\ &\quad + \sum_{s=0}^1 \sum_{t=0}^1 \tilde{X}(s)^T T^\theta(s) \Delta^g(s) H^g(s) T^\theta(t) \Delta^g(t) H^g(t) \tilde{X}(t) \\ &\quad - \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s) (\Delta^g(s))^2 (H^g(s))^2 \tilde{X}(s). \end{aligned} \quad (\text{B.16})$$

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

$$\begin{aligned} J_{(\beta^g, \beta^m)}(\theta; m, g) &= \mathbb{E} [-\nabla_{\beta^g} \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)] \\ &\quad + \mathbb{E} [\nabla_{\beta^g} \mathcal{L}(\theta; m, g, p)] \mathbb{E} [\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)]^T - \mathbb{E} [\nabla_{\beta^g} \mathcal{L}(\theta; m, g, p) \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)^T]. \end{aligned} \quad (\text{B.17})$$

First, we have that

$$\mathbb{E} [-\nabla_{\beta^g} \nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)] = 0, \quad (\text{B.18})$$

as differentiating \mathcal{L} 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 (B.17) is

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

$$-\sum_{s=0}^1 \left[\tilde{X}[k]^T T^\theta(s) \Delta^g(s) H^g(s) \Delta^m(s) H^m(s) \tilde{X}[l](s) \right]. \quad (\text{B.19})$$

Combining (B.17), (B.18), and (B.19) produces

$$\begin{aligned} J_{(\beta^g, \beta^m)}(\theta; m, g) &= \sum_{s=0}^1 \sum_{t=0}^1 \tilde{X}(s)^T T^\theta(s) \Delta^g(s) H^g(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(t) \\ &\quad - \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s) \Delta^g(s) H^g(s) \Delta^m(s) H^m(s) \tilde{X}(s). \end{aligned} \quad (\text{B.20})$$

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

$$\begin{aligned} J_{(\beta^m, \pi)}(\theta; m, g) &= \mathbb{E}[-\nabla_{\beta^m} \nabla_\pi \mathcal{L}(\theta; m, g, p)] \\ &\quad + \mathbb{E}[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p)] \mathbb{E}[\nabla_\pi \mathcal{L}(\theta; m, g, p)]^T - \mathbb{E}[\nabla_{\beta^m} \mathcal{L}(\theta; m, g, p) \nabla_\pi \mathcal{L}(\theta; m, g, p)^T]. \end{aligned} \quad (\text{B.21})$$

We have that

$$\mathbb{E}[-\nabla_{\beta^m} \nabla_\pi \mathcal{L}(\theta; m, g, p)] = 0, \quad (\text{B.22})$$

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 (B.8) that $a \sum_{i=1}^n p_i - b$.

The k th entry of the last two terms of (B.21) is

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

$$\begin{aligned}
&= a \sum_{i=1}^n \mathbb{E}[p_i] \mathbb{E}[\tilde{x}_{ik} \Delta_i^m s_i^m] - a \sum_{i=1}^n \mathbb{E}[p_i \tilde{x}_{ik} \Delta_i^m s_i^m] \\
&= a \sum_{i=1}^n T_i^\theta(1) [T_i^\theta(0) \Delta_i^m(0) s_i^m(0) \tilde{x}_{ik}(0) + T_i^\theta(1) \Delta_i^m(1) s_i^m(1) \tilde{x}_{ik}(1)] - a \sum_{i=1}^n T_i^\theta(1) \Delta_i^m(1) s_i^m(1) \tilde{x}_{ik}(1) \\
&\quad = a \sum_{i=1}^n T_i^\theta(0) T_i^\theta(1) \Delta_i^m(0) H_i^m(0) \tilde{x}_{ik}(0) \\
&\quad + a \sum_{i=1}^n ([T_i^\theta(1)]^2 \Delta_i^m(1) H_i^m(1) - T_i^\theta(1) \Delta_i^m(1) H_i^m(1)) \tilde{x}_{ik}(1) \\
&= a \left[\sum_{i=1}^n T_i^\theta(0) T_i^\theta(1) \Delta_i^m(0) H_i^m(0) \tilde{x}_{ik}(0) + \sum_{i=1}^n T_i^\theta(1) \Delta_i^m(1) H_i^m(1) [T_i^\theta(1) - 1] \tilde{x}_{ik}(1) \right] \\
&= a \left[\sum_{i=1}^n T_i^\theta(0) T_i^\theta(1) \Delta_i^m(0) H_i^m(0) \tilde{x}_{ik}(0) - \sum_{i=1}^n T_i^\theta(0) T_i^\theta(1) \Delta_i^m(1) H_i^m(1) \tilde{x}_{ik}(1) \right] \\
&= a (\tilde{X}(0)[, k]^T w^m(0) - \tilde{X}(1)[, k]^T w^m(1)). \quad (\text{B.23})
\end{aligned}$$

Combining (B.21), (B.22), and (B.23), we conclude that

$$J_{(\beta^m, \pi)}(\theta; m, g, p) = \left(\frac{1}{\pi} + \frac{1}{1-\pi} \right) (\tilde{X}(0)^T w^m(0) - \tilde{X}(1)^T w^m(1)). \quad (\text{B.24})$$

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

$$J_{(\beta^g, \pi)}(\theta; m, g, p) = \left(\frac{1}{\pi} + \frac{1}{1-\pi} \right) (\tilde{X}(0)^T w^g(0) - \tilde{X}(1)^T w^g(1)). \quad (\text{B.25})$$

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

I

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

II

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

$$\begin{aligned}
& + \sum_{s=0}^1 \sum_{t=0}^1 \tilde{X}(s)^T T^\theta(s) \Delta^m(s) H^m(s) T^\theta(t) \Delta^m(t) H^m(t) \tilde{X}(t) \\
& \quad - \sum_{s=0}^1 \tilde{X}(s)^T T^\theta(s) (\Delta^m(s))^2 (H^m(s))^2 \tilde{X}(s).
\end{aligned}$$

III

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

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.5 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 using the `glm` function 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 i th 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.

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

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 i th 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

$$\psi_m''(\eta_i^m) = [\sigma_i^m]^2 \tag{B.26}$$

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

$$r_m^{-1}(t) = \psi_m'(h_m(t)). \tag{B.27}$$

Differentiating (B.27) in t , we find that

$$(r_m^{-1})'(t) = \psi_m''(h_m(t))h'_m(t) \iff h'_m(t) = \frac{(r_m^{-1})'(t)}{\psi_m''(h_m(t))}. \tag{B.28}$$

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 ($s > 0$ fixed), log link |
|---------------------------|-------------------------------------|-------------------------------|--|
| <code>linkinv</code> | x | $\exp(x)$ | $\exp(x)$ |
| <code>variance</code> | x | x | $x + x^2/s$ |
| <code>mu.eta</code> | 1 | x | $\exp(x)$ |
| <code>skewness</code> | 0 | $x^{-1/2}$ | $\frac{2x+s}{\sqrt{sx}\sqrt{x+s}}$ |
| <code>mu.eta.prime</code> | 0 | $\exp(x)$ | $\exp(x)$ |

Finally, plugging in l_i^m for t ,

$$h'_m(l_i) = \frac{(r_m^{-1})'(l_i^m)}{\psi''_m(h_m(l_i^m))} = \frac{(r_m^{-1})'(l_i^m)}{\psi''_m(\eta_i^m)} = \text{ by (B.26)} \frac{(r_m^{-1})'(l_i^m)}{[\sigma_i^m]^2}.$$

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

$$\psi'''_m(\eta_i^m) = \gamma_i^m ([\sigma_i^m]^2)^{3/2}. \quad (\text{B.29})$$

Differentiating (B.28) in t , we obtain

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

Plugging in l_i^m for t , we find that

$$h''_m(l_i^m) = \frac{(r_m^{-1})''(l_i^m) - \psi'''_m(\eta_i^m)[h'_m(l_i^m)]^2}{[\sigma_i^m]^2} = \text{ (by B.29)} \frac{(r_m^{-1})''(l_i^m) - ([\sigma_i^m]^2)^{3/2}(\gamma_i^m)[h'_m(l_i^m)]^2}{[\sigma_i^m]^2}.$$

Finally, the expression for V_i^m follows from (B.26). 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 $[m_1, \dots, m_n]$, $[o_1^m, \dots, o_n^m]$, and $[z_1, \dots, z_n]$; and a family object containing functions `linkinv`, `variance`, `mu.eta`, `mu.eta.prime`, and `skewness`.

```

for  $j \in \{0, 1\}$  do
  for  $i \in \{1, \dots, n\}$  do
    3:    $l_i^m(j) \leftarrow \langle \beta_m, \tilde{x}_i(j) \rangle + o_i^m$ 
     $\mu_i^m(j) \leftarrow \text{linkinv}(l_i^m(j))$ 
     $[\sigma_i^m(j)]^2 \leftarrow \text{variance}(\mu_i^m(j))$ 
    6:    $h'_m(l_i^m(j)) \leftarrow \text{mu.eta}(l_i^m(j))/[\sigma_i^m(j)]^2$ 
     $\gamma_i^m(j) \leftarrow \text{skewness}(\mu_i^m(j))$ 
     $[r_m^{-1}]''(l_i^m(j)) \leftarrow \text{mu.eta.prime}(l_i^m(j))$ 
    9:
     $h''_m(l_i^m(j)) \leftarrow \frac{[r_m^{-1}]''(l_i^m(j)) - [([\sigma_i^m(j)]^2)^{3/2}][\gamma_i^m(j)][h'_m(l_i^m(j))]^2}{[\sigma_i^m(j)]^2}$ 
     $\triangleright$  Assign quantities to matrices
     $\Delta_i^m(j) \leftarrow h'_m(l_i^m(j))$ 
     $[\Delta']_i^m(j) \leftarrow h''(l_i^m(j))$ 
    12:   $V_i^m(j) \leftarrow [\sigma_i^m(j)]^2$ 
     $H_i^m(j) \leftarrow s_i^m(j) \leftarrow m_i - \mu_i^m(j)$ 
  end for
15: end for

```

C. STATISTICAL ACCELERATIONS AND COMPUTING

C.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:

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

We similarly can obtain pilot estimates $[\beta_0^g]^{\text{pilot}}$ and $[\gamma_g^T]^{\text{pilot}}$ by regressing the gRNA counts g onto the technical factors X . We extract the fitted values (on the scale of the linear component) for use in a subsequent step: $\hat{f}_i^k = [\beta_0^k]^{\text{pilot}} + \langle [\gamma_k^T]^{\text{pilot}}, z_i \rangle + o_i^k$, for $k \in \{m, g\}$.

Next, we obtain estimates $[\beta_1^m]^{\text{pilot}}$, $[\beta_1^g]^{\text{pilot}}$, and π^{pilot} for β_1^m , β_1^g , and π by fitting a “reduced” GLM-EIV (Algorithm 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:

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

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 $\hat{f}^m := [\hat{f}_1^m, \dots, \hat{f}_n^m]^T$, gRNA offsets $\hat{f}^g := [\hat{f}_1^g, \dots, \hat{f}_n^g]^T$, and no

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
  2: 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  $GLM_k$ .
  4: for  $i \in \{1, \dots, n\}$  do
    5:    $\hat{f}_i^k \leftarrow [\beta_0^k]^{\text{pilot}} + \langle [\gamma_k^T]^{\text{pilot}}, z_i \rangle + o_i^k$                                  $\triangleright$  untransformed fitted values
  6: end for
end for
  8: return  $([\beta_0^m]^{\text{pilot}}, \hat{f}^m, [\gamma_m^T]^{\text{pilot}}, [\beta_0^g]^{\text{pilot}}, [\gamma_g^T]^{\text{pilot}}, \hat{f}^g)$ 

```

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 C.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).

Algorithm 4 Computing $\pi^{\text{pilot}}, [\beta_1^m]^{\text{pilot}}, [\beta_1^g]^{\text{pilot}}$.

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

```

bestLik  $\leftarrow -\infty$  ▷ Reduced GLM-EIV
2: for  $i \in \{1, \dots, B\}$  do
    Randomly generate starting parameters  $\pi^{\text{curr}}, [\beta_1^m]^{\text{curr}}, [\beta_1^g]^{\text{curr}}$ .
4:   while Not converged do
    for  $i \in \{1, \dots, n\}$  do ▷ E step
         $T_i(1) \leftarrow \mathbb{P}(P_i = 1 | M_i = m_i, G_i = g_i, \pi^{\text{curr}}, [\beta_1^m]^{\text{curr}}, [\beta_1^g]^{\text{curr}})$ 
         $T_i(0) \leftarrow 1 - T_i(1)$ 
    8:   end for
     $\pi^{\text{curr}} \leftarrow (1/n) \sum_{i=1}^n T_i(1)$  ▷ M step
10:    $w \leftarrow [T_1(0), T_2(0), \dots, T_n(0), T_1(1), T_2(1), \dots, T_n(1)]^T$ 
    for  $k \in \{g, m\}$  do
        Fit no-intercept, univariate GLM  $GLM_k$  with predictors  $[0, \underbrace{\dots, 0}_n, \underbrace{1, \dots, 1}_n]$ , re-
        spondes  $[k, k]^T$ , offsets  $[\hat{f}^k, \hat{f}^k]^T$ , and weights  $w$ .
        Set  $[\beta_1^k]^{\text{curr}}$  to fitted coefficient of  $GLM_k$ .
    14:   end for
    Compute log likelihood currLik using  $\pi^{\text{curr}}, [\beta_1^m]^{\text{curr}}$ , and  $[\beta_1^g]^{\text{curr}}$ .
16:   end while
    if currLik > bestLik then
        bestLik  $\leftarrow$  currLik
         $\pi^{\text{pilot}} \leftarrow \pi^{\text{curr}}; [\beta_1^m]^{\text{pilot}} \leftarrow [\beta_1^m]^{\text{curr}}; [\beta_1^g]^{\text{pilot}} \leftarrow [\beta_1^g]^{\text{curr}}$ 
    20:   end if
    end for
22: return  $(\pi^{\text{pilot}}, [\beta_1^m]^{\text{pilot}}, [\beta_1^g]^{\text{pilot}})$ 

```

C.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 $\underbrace{[0, \dots, 0]}_n, \underbrace{[1, \dots, 1]}_n$, offsets $[\hat{f}^m, \hat{f}^m]$, and weights $w = [T_1(0), \dots, T_n(0), T_1(1), \dots, T_n(1)]^T$. Throughout, C denotes a universal constant. The log likelihood of this GLM is

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

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 $o_1, \dots, o_n \sim \mathcal{P}_1$, where \mathcal{P}_1 is a distribution. Let $Y_i|o_i, \dots, Y_n|o_i$ 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 $r : \mathbb{R} \rightarrow \mathbb{R}$ 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 $w_1, \dots, w_n \sim \mathcal{P}_2$ 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 $i \in \{1, \dots, n\}$.) 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 $\psi : \mathbb{R} \rightarrow \mathbb{R}$ and $c : \mathbb{R} \rightarrow \mathbb{R}$, respectively. The canonical

parameter η_i of the i th observation is

$$\eta_i = ([\psi']^{-1} \circ r^{-1})(\beta + o_i) := h(\beta + o_i), \quad (\text{C.2})$$

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

$$\mathcal{L}(\beta; y_i) = \sum_{i=1}^n w_i \log [f(y_i; \eta_i)] = C + \sum_{i=1}^n w_i (y_i\eta_i - \psi(\eta_i)). \quad (\text{C.3})$$

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 (C.3); 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 (C.2),

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

Plugging $\eta_i = h(\beta + o_i) = \beta + o_i$ and $\psi(\eta_i) = (1/2)(\beta + o_i)^2$ into (C.3), 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 (C.2),

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

Plugging $\eta_i = h(\beta + o_i) = \beta + o_i$ and $\psi(\eta_i) = \exp(\beta + o_i)$ into (C.3), we obtain

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

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

$$\hat{\beta}^{\text{pois}} = \log \left(\frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i e^{o_i}} \right). \quad (\text{C.4})$$

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

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

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

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

implying $\delta = [\psi']^{-1}$. By (C.2), we have that

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

Therefore,

$$\eta_i = h(\beta + o_i) = \log \left(\frac{\exp(\beta + o_i)}{s + \exp(\beta + o_i)} \right) = \beta + o_i - \log(s + e^\beta e^{o_i}) = \beta - \log(s + e^\beta e^{o_i}) + C, \quad (\text{C.5})$$

and

$$\begin{aligned} \psi(\eta_i) &= -s \log \left(1 - \frac{\exp(\beta + o_i)}{s + \exp(\beta + o_i)} \right) = -s \log \left(\frac{s}{s + \exp(\beta + o_i)} \right) \\ &= -s \log(s) + s \log[s + \exp(\beta + o_i)] = s \log(s + e^s e^{o_i}) + C. \end{aligned} \quad (\text{C.6})$$

Plugging (C.5) and (C.6) into (C.3), the log-likelihood (up to a constant) is

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

The derivative of \mathcal{L} in β is

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

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

$$e^\beta \sum_{i=1}^n \frac{w_i e^{o_i} (y_i + s)}{e^{\beta+o_i} + s} = \sum_{i=1}^n w_i y_i. \quad (\text{C.7})$$

We cannot solve for β in (C.7) analytically. However, we can derive an asymptotically exact solution. By the law of total expectation,

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

the second equality holds because $\mathbb{E}[y_i | o_i] = \mu_i = e^{\beta+o_i}$. Dividing by n on both sides of (C.7) and rearranging,

$$\beta = \log \left(\frac{(1/n) \sum_{i=1}^n w_i e^{o_i} (y_i + s) / (e^{\beta+o_i} + s)}{(1/n) \sum_{i=1}^n w_i y_i} \right). \quad (\text{C.8})$$

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

$$\hat{\beta}^{\text{NB}} \xrightarrow{P} \log \left(\frac{\mathbb{E}[w_i y_i]}{\mathbb{E}[w_i e^{o_i}]} \right). \quad (\text{C.9})$$

But the Poisson MLE $\hat{\beta}^{\text{Pois}}$ (C.4) 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 (C.1) is an approximate intercept-plus-offset model: $T_1(1), \dots, T_n(1)$ are the weights w_1, \dots, w_n , and $\hat{f}_1^m, \dots, \hat{f}_n^m$ are the offsets o_1, \dots, o_m . Of course, $T_1(1), \dots, T_1(n)$ are in general dependent random variables, as are $\hat{f}_1^m, \dots, \hat{f}_n^m$. $T_i(1)$ 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 (C.1), especially when the number of cells n is large. Additionally, we note that the GLM (C.1) 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.

C.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 embarrassingly parallel and therefore can be massively parallelized.

D. THE NAT. BIOTECH. 2020 METHOD

As described in the main text, the Nat. Biotech. 2020 method (of Replogle *and others* (2020)) fits a Poisson-Gaussian mixture model to the log-2 transformed gRNA counts and then assigns gRNAs to cells based on the posterior perturbation probabilities. If a given cell has a posterior perturbation probability greater than 1/2, then the gRNA is assigned to that cell; otherwise, the

Algorithm 5 Applying GLM-EIV at scale.

```

 $G \leftarrow \{\text{gene}_1, \dots, \text{gene}_{d_g}\}; P \leftarrow \{\text{perturbation}_1, \dots, \text{perturbation}_{d_p}\}$ 

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

```

gRNA is not assigned to that cell. Covariates (including gRNA library size, gene library size, batch, etc.) are not included in the model.

As mentioned in the main text, the Nat. Biotech. 2020 method poses several conceptual and practical challenges. First, the log-2 transformed gRNA counts are not integer-valued. Thus, it is unclear how the Poisson component of the mixture distribution is fitted to the data. Second, the authors of the Nat. Biotech. 2020 method used the Python package `Pomegranate` (github.com/jmschrei/pomegranate; version $\leq 0.14.8$) to implement their method. Unfortunately, due to recent updates to the `Pomegranate` package, we and others have been unable to install version $\leq 0.14.8$ (relevant Github issues: github.com/jmschrei/pomegranate/issues/1052, github.com/jmschrei/pomegranate/issues/1057).

Thus, we attempted to implement the Nat. Biotech. 2020 method ourselves in R using the `flexmix` package, a popular package for mixture modeling. We found that `flexmix` throws an

error when one attempts to fit a Poisson distribution to non-integer data. We therefore considered a modification to the Nat. Biotech. 2020 method in which we fitted a two-component Gaussian mixture to the log-transformed gRNA counts, adding a pseudocount of one to avoid taking the log of zero. Unfortunately, our modified version of the Nat. Biotech. 2020 method did not work well in practice, as it categorized all cells as unperturbed on both the simulated gRNA data (Figure 3) and the low-MOI gRNA data (Figure 5). The default CellRanger method for gRNA assignment — which is based on the Nat. Biotech. 2020 method — uses a two-component Gaussian mixture model (www.10xgenomics.com/support/software/cell-ranger/latest/algorithms-overview/cr-crispr-algorithm). The CellRanger method became open-source shortly before the publication of this paper.

E. DATA ANALYSIS DETAILS

First, we performed quality control and basic pre-processing on both datasets. As is standard in single-cell analysis, we removed cells with a high fraction ($> 8\%$) of mitochondrial reads (Choudhary and Satija, 2022). 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 appeared to be 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 and others (2021) for details on the construction of *cis* and *in silico* negative control pairs on the Xie data. Because the negative control pairs are not expected to exhibit a regulatory relationship, the ground truth fold change in gene expression for these pairs is taken to be unity.

We modeled the gene expression counts using a negative binomial distribution with unknown size parameter s ; we estimated s using the `glm.nb` package. Choudhary and Satija (2022) 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.

We describe in more detail how we conducted the “excess background contamination” analysis. For each positive control pair, we varied excess background contamination over the grid $[0.0, 0.05, 0.1, \dots, 0.4]$. 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 $[\hat{\beta}_1^m]^{(1)}, \dots, [\hat{\beta}_1^m]^{(B)}$. 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.

F. ADDITIONAL RELATED WORK

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 *and others* (2021) and Townes *and others* (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.

G. SIMULATION STUDY DETAILS AND ADDITIONAL SIMULATION STUDIES

G.1 *Main text simulation study parameter values*

We constructed a table (Table 2) that maps each model parameter to its (i) main text simulation study value and (ii) estimated value on real data. We obtained the real-data parameter estimates by applying GLM-EIV to analyze a representative gRNA-gene pair from the Gasperini *and others* (2019) data (namely, gene “ENSG00000213931” paired to positive control gRNA “pos_control_Klannchr1_HS4”). The main difference between the simulation parameter values and real-data parameter values is that the perturbation effect size on gRNA expression (i.e., $\exp(\beta_1^g)$) is smaller in the simulation study than on the real data. This difference has the effect of placing the simulation study into a more challenging region of the problem space.

| Parameter | Simulation value | Estimated real data value | Meaning |
|--------------------|----------------------|---------------------------|--------------------------|
| $\exp(\beta_0^m)$ | 0.01 | 0.02 | Gene model intercept |
| $\exp(\beta_1^m)$ | 0.25 | 0.68 | Gene perturbation effect |
| $\exp(\gamma_1^m)$ | 0.9 | 1.0 | Gene batch effect |
| $\exp(\beta_0^g)$ | $5.0 \cdot 10^{-3}$ | $3.4 \cdot 10^{-6}$ | gRNA model intercept |
| $\exp(\beta_1^g)$ | [1.0, 1.5, ..., 4.0] | 6,200 | gRNA perturbation effect |
| $\exp(\gamma_1^g)$ | 1.1 | 1.05 | gRNA batch effect |
| π | 0.02 | 0.004 | Perturbation probability |

Table 2. A mapping of each model parameter to its (i) main text simulation study value and (ii) estimated value on real data.

G.2 *Additional simulation studies*

We report the results of five additional simulation studies. Study 2 considers Gaussian (as opposed to negative binomial or Poisson) data; study 3 varies the negative binomial size parameter s ;

study 4 varies the effect size of the perturbation on gene expression β_1^m ; and study five (resp., six) considers gRNA (resp., gene) expression data that are contaminated by doublets and an unmeasured covariate. In all simulation studies we deployed the accelerated version of GLM-EIV.

Simulation study 2. In simulation study 2 we modeled the gene and gRNA expressions using a Gaussian distribution with an identity link. We generated data on $n = 50,000$ 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 applied GLM-EIV, thresholded regression, and the gRNA mixture assignment method (coupled to linear regression) to analyze the simulated data. The ranking of the methods was as follows: GLM-EIV (best), gRNA mixture assignment method (intermediate), and thresholding method (worst) (Figure 9).

Simulation study 3. Simulation study 3 was similar to the main text simulation study. The difference is that in simulation study 3, we held fixed $\beta_1^g = \log(2.5)$ while varying the negative binomial size parameter s over the grid $1 = 10^{0/9}, 10^{2/9}, 10^{4/9}, \dots, 10^{16/9}, 10^{18/9} = 100$. We applied the three methods twice: once assuming known s and once under unknown s . All methods demonstrated roughly uniform bias over the grid of s values: the bias of GLM-EIV was near zero, while that of the thresholding method and the gRNA mixture method was about 0.02. As s increased, the CI width of all methods decreased (as the gene expression data became more Poisson-like, causing standard errors to shrink). The confidence interval coverage of the thresholding method and the gRNA mixture method degraded, while that of GLM-EIV remained at the roughly nominal level. The former two methods likely lost coverage because their biased estimates caused the increasingly-narrow confidence intervals to be centered at the wrong location.

The results were broadly similar across known s and unknown s (though slightly better under known s).

Simulation study 4. Simulation study 4 also was similar to the main text simulation study. The difference is that in simulation study 4, we held fixed the perturbation effect size on gRNA expression ($\exp(\beta_1^g) = 2.5$) and varied the perturbation effect size on gene expression $\exp(\beta_1^m)$ over the grid $0.2, 0.3, \dots, 0.9, 1.0$. We applied the three methods to analyze data generated from Poisson, negative binomial (with known s), and negative binomial (with unknown s) gene expression distributions. We observed that as the magnitude of the effect size increased (i.e., as $\exp(\beta_1^m)$ decreased from 1.0 to 0.2), GLM-EIV remained roughly unbiased, while the thresholding method and the gRNA mixture assignment method exhibited increasingly severe attenuation bias. Furthermore, GLM-EIV maintained coverage at the nominal level, while the coverage of the thresholding method and the gRNA mixture assignment method degraded due to the aforementioned attenuation bias. Results were broadly similar (albeit slightly worse) under estimated s than known s .

We additionally plotted the rejection probability, i.e. the probability of rejecting the null hypothesis of $H_0 : \exp(\beta_1^m) = 1$ at level 0.05. When $\exp(\beta_1^m) = 1$ (i.e., when we are under the null hypothesis), the rejection probability (which corresponds to type-I error) should be 0.05, the nominal level. When $\exp(\beta_1^m) < 1$ (i.e., when we are under the alternative hypothesis), the rejection probability (which corresponds to power) should be as large as possible (with a value of 1.0 being optimal). We observed that all methods exhibited a rejection probability of roughly 0.05 under the null hypothesis of $\exp(\beta_1^m) = 1$ and a rejection probability of 1.0 under the alternative hypotheses of $\exp(\beta_1^m) = 0.9, 0.8, \dots, 0.2, 0.1$. In other words, over the grid of values that we examined, each method performed optimally with respect to testing the hypothesis $\exp(\beta_1^m) = 1$. (We note that our goal in the simulation studies was to explore discrepancies in estimation accuracy and confidence interval coverage across methods, but we present type-I error control

and power results for completeness.)

Simulation study 5. In simulation study 5 we applied the methods to analyze data drawn from a distribution that lay outside the GLM-EIV family of distributions. First, we simulated gRNA count data from a poisson GLM with two covariates: batch (modeled as a Bernoulli random variable with probability 1/2) and cell cycle (modeled as a uniform random variable on the interval [0,1]). We treated cell cycle as an unmeasured covariate, i.e. we did not give any of the methods access to cell cycle. Next, we randomly selected 1% of cells and doubled the gRNA count in these cells, thereby simulating the presence of doublets (i.e., droplets that contain two cells) in the data. We simulated the gene expression data from the same negative binomial model that we used in the main text simulation (and so the gene expression model was correctly specified.) For simplicity we assumed that the size parameter $s = 20$ was known. We varied the perturbation effect size on gRNA expression $\exp(\beta_1^g)$ over the grid 1, 2, ..., 7 and the perturbation effect size on gene expression $\exp(\beta_1^m)$ over the grid 0.25, 0.5, 0.75, 1.0.

We applied GLM-EIV, thresholded regression, and the gRNA mixture assignment method to analyze the data. GLM-EIV exhibited generally lower bias, lower mean squared error, and better confidence interval coverage than the other methods. The rightmost panel (i.e, $\exp(\beta_1^m) = 1$) corresponds to the null hypothesis of no perturbation effect on gene expression; the left panels (i.e., $\exp(\beta_1^m) = 0.75, 0.5, 0.25$), by contrast, correspond to alternative hypotheses of varying strength. All methods controlled type-I error at the nominal level of 0.05. GLM-EIV demonstrated equal or greater power than the competing methods.

Simulation study 6. Simulation study 6 was similar to simulation study 5, the difference being that simulation study 6 considered a misspecified gene expression model (while simulation study 5 considered a misspecified gRNA count model). We generated the gene expression data from a negative binomial GLM containing the unmeasured covariate of cell cycle, and we doubled the gene expression count in 1% of randomly selected cells to simulate doublets. We generated

gRNA counts from the same gRNA model that we used in the main text simulation (and so the gRNA count model was correctly specified.) Again, we varied $\exp(\beta_1^g)$ over the grid $1, 2, \dots, 7$ and $\exp(\beta_1^m)$ over the grid $0.25, 0.5, 0.75, 1.0$. We found that GLM-EIV generally performed best: GLM-EIV exhibited lower bias, lower mean squared error, and better confidence interval coverage than the other methods. There was one setting for β_1^g (namely, $\exp(\beta_1^g) = 1.5$) for which GLM-EIV did not control type-I error under the null hypothesis of $\exp(\beta_1^m) = 1$. However, this was an extreme value for β_1^g , and GLM-EIV controlled type-I error under all other values of β_1^g .

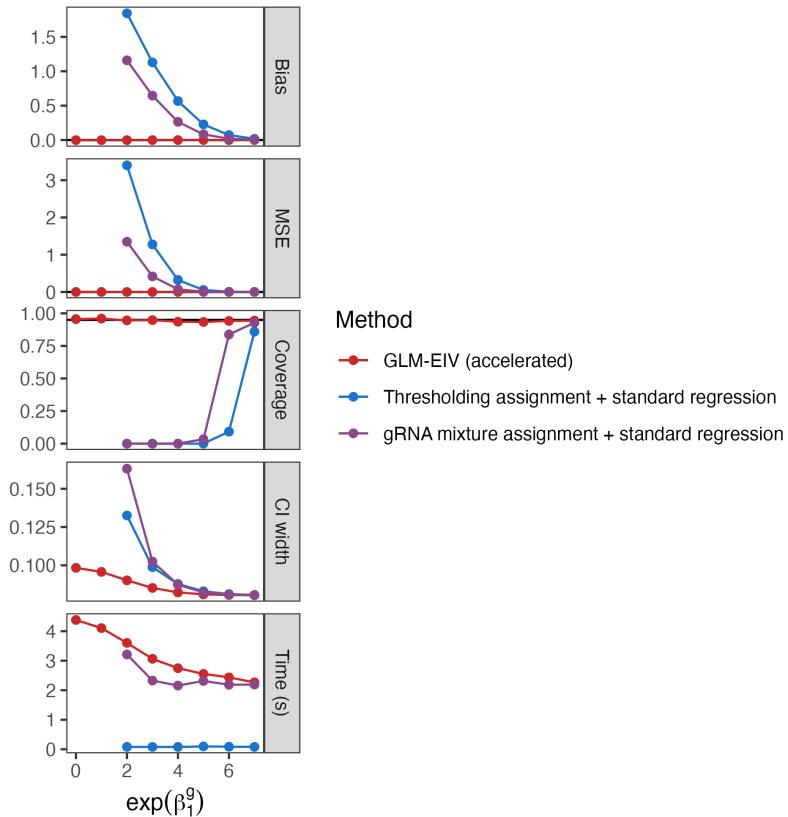


Fig. 9. **Simulation study 2.** Analyzing data generated from a linear Gaussian model. Rejection probability (not plotted) was strictly 1 across methods and parameter settings.

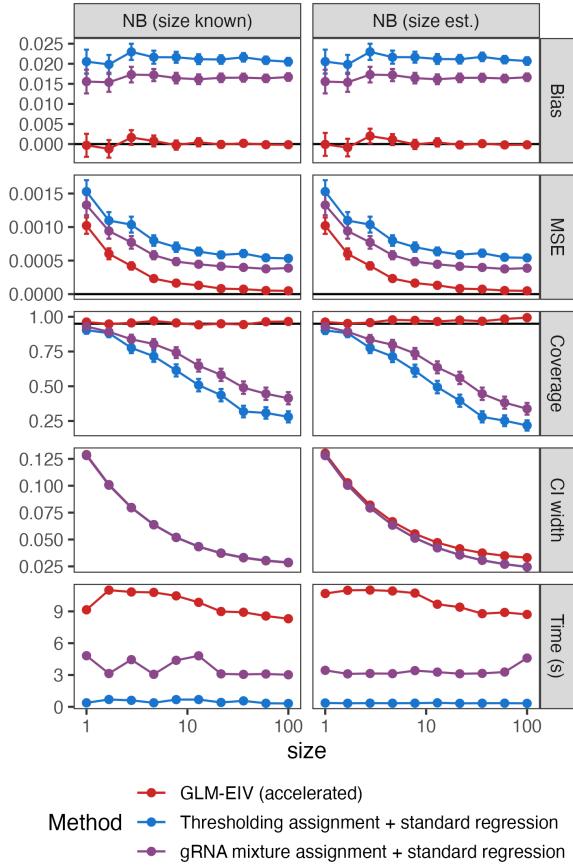


Fig. 10. **Simulation study 3.** Varying the negative binomial size parameter s . Rejection probability (not plotted) was strictly 1 across methods and parameter settings.

REFERENCES

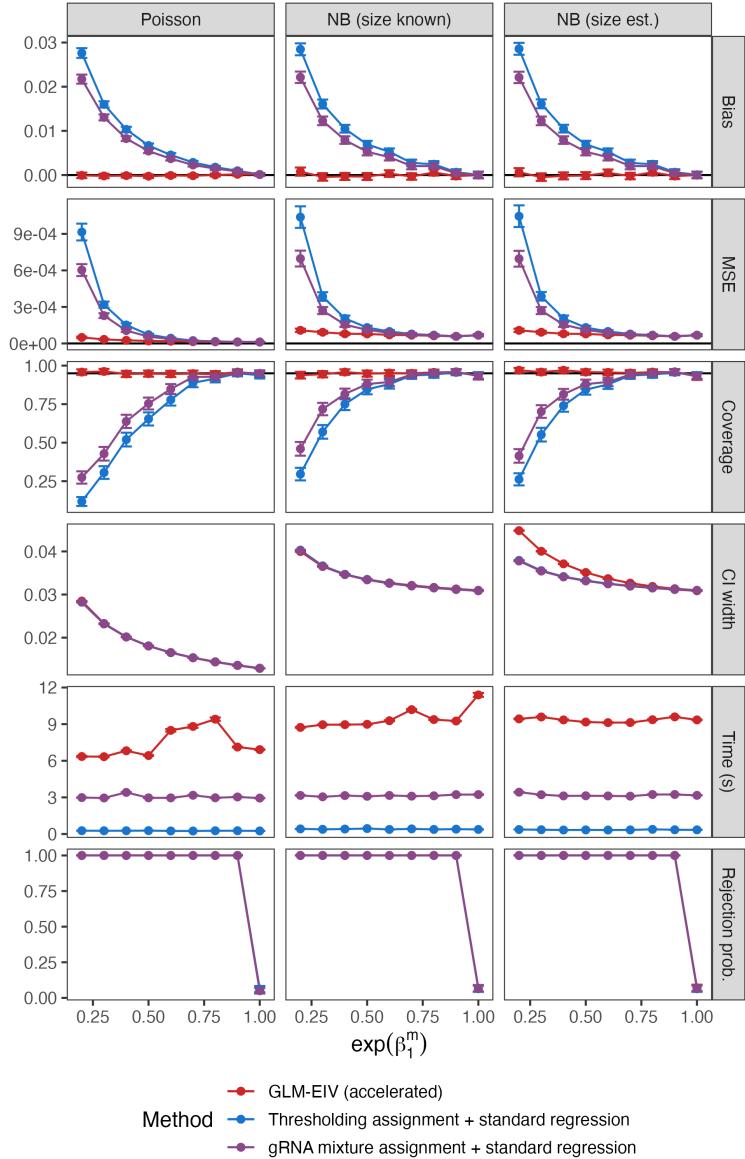


Fig. 11. **Simulation study 4.** Varying the perturbation effect size on gene expression, β_1^m .

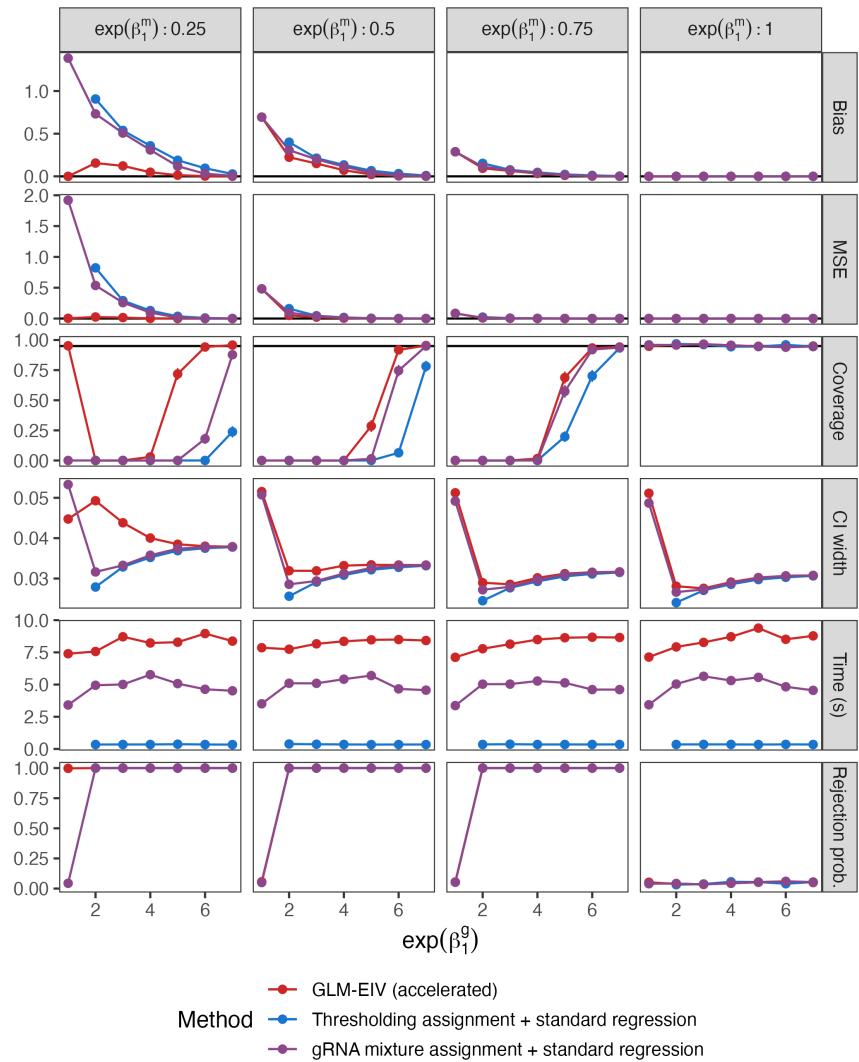


Fig. 12. **Simulation study 5.** Analyzing data using a misspecified gRNA count model.

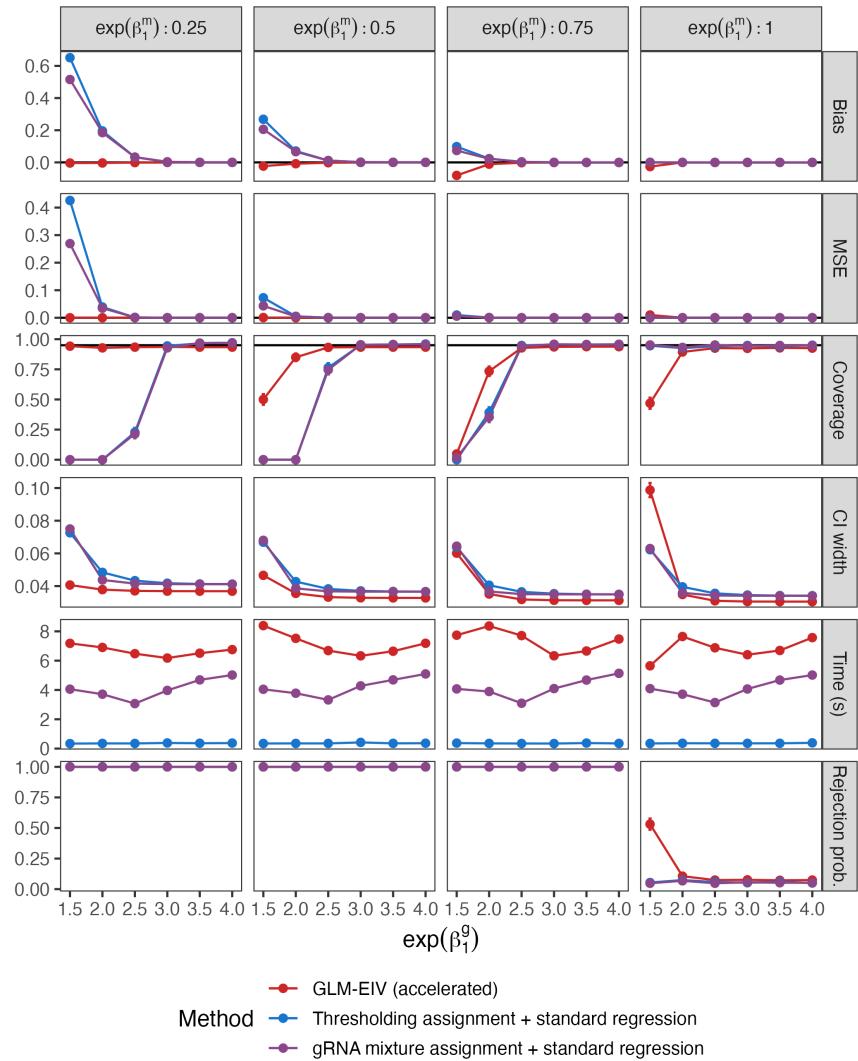


Fig. 13. **Simulation study 6.** Analyzing data using a misspecified gene expression model.