A generalized errors-in-variables model, with application to single-cell CRISPR screens

Tim Barry[1], Eugene Katsevich[2], Kathyrn Roeder[1]

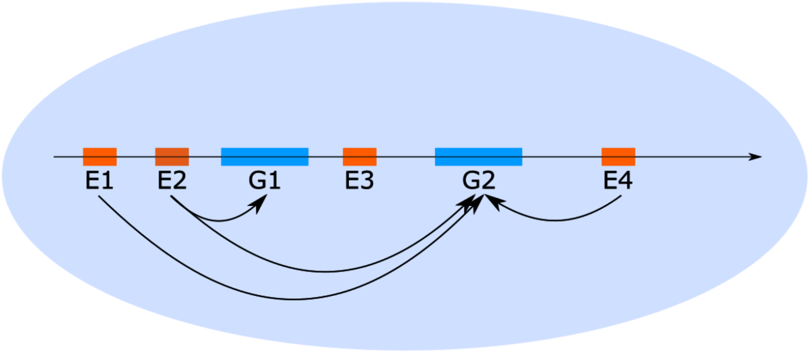
[1]CMU Statistics and Data Science
[2]Wharton Statistics and Data Science

September 2021

Overview

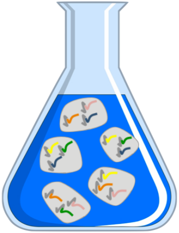
- **Background**
- Analysis objective and challenges
- Existing approach
- Proposed method
- Simulation results
- Real data results

Single-cell CRISPR screens are a powerful technology for mapping the regulatory wiring of the genome.



Single-cell CRISPR screens entail sequencing gRNAs and mRNAs in individual cells.

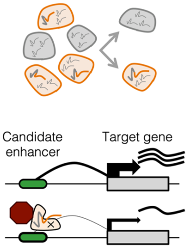
Perturb cells with gRNAs



Sequence single cells

For each cell, measure:
1.gRNAs
2.gene expression

Test for differential expression



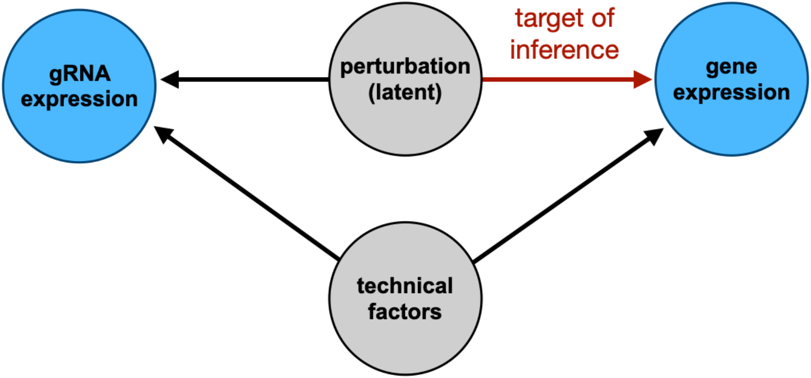
Overview

- Background
- **Analysis Challenges**
- Existing approach
- Proposed method
- Simulation results
- Real data results

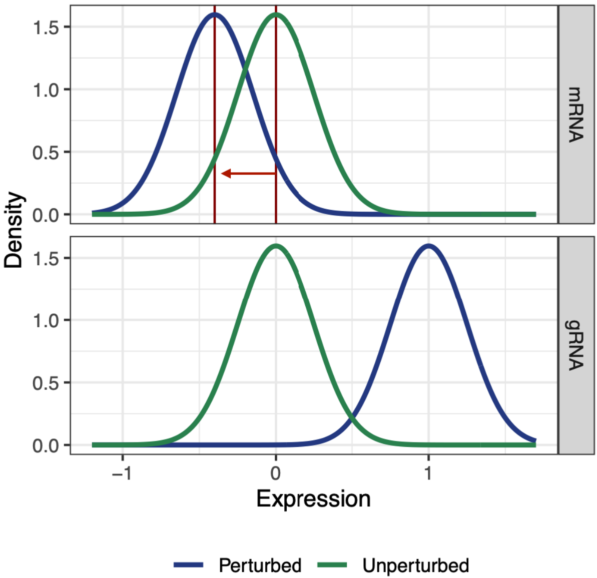
There are several challenges to the analysis of single-cell CRISPR screen data:

1. The perturbation is unobserved.
2. Technical factors, such as batch and sequencing depth, explain variability in mRNA and gRNA counts.
3. Unperturbed cells exhibit “background gRNA reads.” [Schraivogel et al., 2020]
4. The mRNA and gRNA expression data are highly discrete.

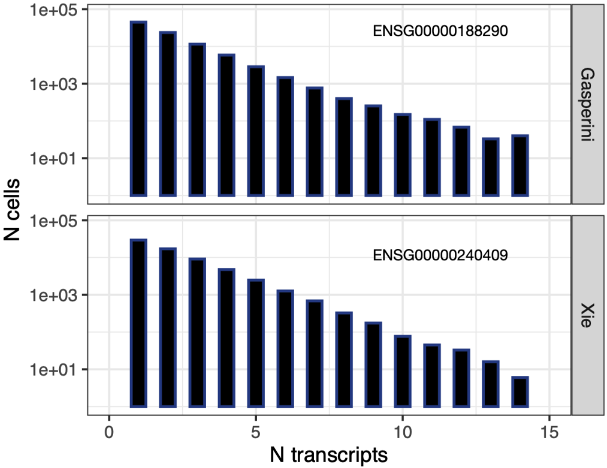
(1) The perturbation is unobserved, and (2) technical factors are present.



(3) Unperturbed cells exhibit background gRNA reads.



(4) The count data are highly discrete.



Overview

- Background
- Analysis Challenges
- **Existing approach**
- Proposed method
- Simulation results
- Real data results

Data and notation

- Observe $n \approx 100,000 - 250,000$ cells.
- Consider a given mRNA and gRNA of interest.
- For cell $i \in \{1, \ldots, n\}$, let
    - $p_i \in \{0, 1\}$ indicate whether a perturbation occurred.
    - $m_i \in \mathbb{N}$ be the mRNA count.
    - $g_i \in \mathbb{N}$ be the gRNA count.
    - $l_i^m \in \mathbb{N}$ be the mRNA library size.
    - $z_i \in \mathbb{R}^{d-1}$ be a vector of technical factors, possibly including an intercept term.
- We measure $\approx 5,000$ genes, $\approx 500 - 5,000$ gRNAs

The “thresholding method”

1. For given threshold $c \in \mathbb{N}$, estimate p_i by

$$\begin{cases} \hat{p}_i = 0 \text{ if } g_i \geq c, \\ \hat{p}_i = 1 \text{ if } g_i < c \end{cases} .$$

2. Fit the regression model [Sarkar and Stephens, 2021]

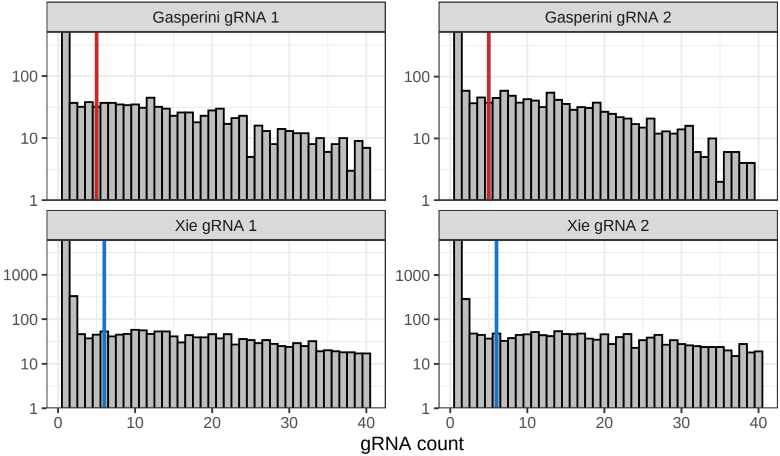
$$m_i \big| \left(z_i, l_i^m\right) \sim \mathrm{NB}_\theta(\mu_i),$$

where $\theta > 0$ is the NB size parameter, and

$$\log\left(\mu_i\right) = \beta_m \hat{p}_i + \gamma_m^T z_i + \log\left(l_i^m\right).$$

3. Obtain an estimate $\hat{\beta}_m$ of β_m and compute a CI for β_m.

Problem 1: There is no clear location in the data at which to draw the threshold.



Problem 2: Thresholding can lead to attenuation bias.

As a simple example, suppose

$$\begin{cases} p_1, \ldots, p_n \sim \mathrm{Bern}(\pi) \\ y_i = \beta_m p_i + \epsilon_i \\ x_i = \beta_g p_i + \tau_i, \end{cases}$$

where $\epsilon_i \perp\!\!\!\perp \tau_i$ and $\epsilon_i, \tau_i \sim N(0, 1)$. We observe

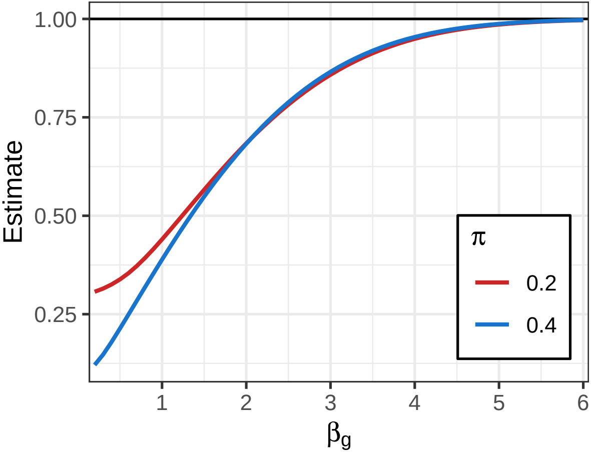
$$\{(x_1, y_1), \ldots, (x_n, y_n)\},$$

and we want to estimate β_m using the thresholding method. Assume β_g and π are known. We select the threshold c so as to minimize the misclassification rate, i.e.,

$$c = \operatorname*{arg\,min}_{c \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{P}_i \neq P_i).$$

The thresholding method shows clear attenuation bias.

- We set β_m to 1.



Research question

Does modeling the gRNA count distribution directly (thereby bypassing thresholding) improve estimation and inference?

Overview

- Background
- Analysis Challenges
- Existing approach
- **Proposed method**
- Simulation results
- Real data results

A bit more notation

For cell $i \in \{1, \ldots, n\}$, let l_i^g be the gRNA library size.

We extend the parametric model of [Sarkar and Stephens, 2021] to model both mRNA *and* gRNA counts.

1. **mRNA**: $m_i|(z_i, l_i^m) \sim \mathrm{NB}_\theta(\mu_i^m)$, where
$$\log\left(\mu_i^m\right) = \beta_m p_i + \gamma_m^T z_i + \log(l_i^m).$$
2. **gRNA**: $g_i|(z_i, l_i^g) \sim \mathrm{NB}_\theta(\mu_i^g)$, where
$$\log(\mu_i^g) = \beta_g p_i + \gamma_g^T z_i + \log(l_i^g)$$
3. **Perturbation**: $p_i \sim \mathrm{Bern}(\pi)$, where $\pi \in [0, 1/2)$. p_i is latent.

Generalizing the NB model to arbitrary exponential family response distribution and link function yields the "GLM-EIV" (GLM errors-in-variables) model.

This extension is important, because authors have used

- Negative binomial, [Choudhary and Satija, 2021]
- Poisson, [Schraivogel et al., 2020]
- and Gaussian [Lin et al., 2021]

distributions to model single-cell data.

Generalizing the NB model to arbitrary exponential family response distribution and link function yields the "GLM-EIV" (GLM errors-in-variables) model.

1. **mRNA density**:

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

2. **gRNA density**:

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

3. **Perturbation density**:

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

Generalizing the NB model to arbitrary exponential family response distribution and link function yields the "GLM-EIV" (GLM errors-in-variables) model.

Consider the mRNA model.

- Let $g_m : \mathbb{R} \to \mathbb{R}$ be the link function, i.e.

$$g_m(\mu_i) = \beta_m p_i + \gamma_m^T z_i + \log(l_i^m).$$

- The canonical parameter for the ith cell, η_i^m, is given by

$$\eta_i^m = \left[\psi_m'\right]^{-1}(\mu_i) = \left[\psi_m'\right]^{-1}\left(g_m^{-1}\left(\beta_m p_i + \gamma_m^T z_i + \log(l_i^m)\right)\right).$$

- Thus, the model is defined by (i) the cumulant-generating function ψ_m, and (ii) the link function g_m.

We derive an EM algorithm to fit the model.

E step:

- Compute membership probabilities $T_1, \dots, T_n$ using the model.

M step:

- Augment count vectors $m \rightarrow [m, m], g \rightarrow [g, g]$.
- Augment offset vectors $l_m \rightarrow [l_m, l_m], l_g \rightarrow [l_g, l_g]$.
- Augment covariate matrix $Z \rightarrow [Z, Z]$; append column of 1s and 0s for perturbation indicators.
- Fit weighted GLM to both modalities using membership probabilities $[T_1, \dots, T_n, 1 - T_1, \dots, 1 - T_n]$ as weights.

We use statistical tricks to produce an accurate pilot estimate of the parameters, enabling us to run the EM algorithm using only one restart.

- Naive approach (random parameter initialization):

$$(15 \text{ EM restarts})\left(\frac{20 \text{ iterations}}{\text{EM restart}}\right)\left(\frac{2 \text{ GLMs}}{\text{iteration}}\right) \approx 600 \text{ GLMs}.$$

- GLM-EIV approach:

$$(1 \text{ EM restart})\left(\frac{3 \text{ iterations}}{\text{EM restart}}\right)\left(\frac{2 \text{ GLMs}}{\text{iteration}}\right) \approx 6 \text{ GLMs}.$$

We derive an analytic expression for the observed information matrix to enable fast inference (CIs, *p*-values).

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

We develop a pipeline to deploy the method across hundreds or thousands of processors on HPC and cloud.

ondisc

nextflow

Method summary

- We extend the parametric model of [Sarkar and Stephens, 2021] to model both mRNA counts and gRNA counts in single-cell CRISPR screen experiments.
  - Arbitrary exponential family distributions and link functions are supported.
- We (i) develop a fast EM algorithm to fit the model, (ii) derive an analytic expression for the observed information matrix, and (iii) implement a computational pipeline to deploy the method on HPC and cloud.
  - These features enable GLM-EIV to scale to large datasets.

Overview

- Background
- Analysis objective and challenges
- Existing approach
- Proposed method
- **Simulation results**
- Real data results

Overview

- Background
- Analysis objective and challenges
- Existing approach
- Proposed method
- Simulation results
- **Real data results**

Real data analysis details

- Quality control
  - Lowly-expressed genes removed
  - Cells with library sizes below 5th percentile or above 95th percentile removed
- mRNA model
  - negative binomial distribution (with log link)
  - size parameter θ estimated from data.
- gRNA model
  - Poisson distribution (with log link) [Choudhary and Satija, 2021]

Choudhary, S. and Satija, R. (2021).
Comparison and evaluation of statistical error models for scRNA-seq.
bioRxiv, (8):2021.07.07.451498.

Gasperini, M., Hill, A. J., McFaline-Figueroa, J. L., Martin, B., Kim, S., Zhang, M. D., Jackson, D., Leith, A., Schreiber, J., Noble, W. S., Trapnell, C., Ahituv, N., and Shendure, J. (2019).
A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens.
Cell, 176(1-2):377–390.e19.

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

Sarkar, A. and Stephens, M. (2021).
Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis.

Nature Genetics, 53(6):770–777.

Schraivogel, D., Gschwind, A. R., Milbank, J. H., Leonce, D. R., Jakob, P., Mathur, L., Korbel, J. O., Merten, C. A., Velten, L., and Steinmetz, L. M. (2020).
Targeted Perturb-seq enables genome-scale genetic screens in single cells.
Nature Methods, 17(6):629–635.

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