

Reduced-rank generalized bilinear models

Kevin S. Kapner^{*1} and Jeffrey W. Miller¹

¹Department of Biostatistics, Harvard T.H. Chan School of Public Health

Abstract

Dimensionality reduction and effect estimation are central tasks in the analysis of high-dimensional data such as in genomics. Generalized bilinear models (GBMs) provide a versatile framework for these tasks, however, the statistical and computational efficiency of GBMs degrades rapidly as the number of sample covariates grows. To address this limitation, we introduce reduced-rank generalized bilinear models (RR-GBMs), which employ a reduced-rank sample coefficient matrix to model the effect of a large number of covariates without requiring an excessive number of parameters. In simulation studies, we find that when the true sample coefficient matrix is reduced rank or close to reduced rank, RR-GBM outperforms the standard full-rank GBM both statistically and computationally, providing more accurate estimates with lower computational burden. We develop a data thinning approach for model selection in the RR-GBM framework, facilitating rank selection. Furthermore, RR-GBM enables a new approach to visualizing the relationships among covariates and among features. We demonstrate the method in an application to Perturb-seq data for pancreatic cancer.

1 Introduction

In single-cell RNA sequencing (scRNA-seq) experiments, researchers often have a large number of sample covariates describing the cells. Some covariates, such as batch, represent effects that one would like to adjust for, since they are not of direct interest. Other covariates, such as experimental condition, represent effects of primary interest that one would like to infer. Modern protocols often produce data across many experimental conditions, for example, Perturb-seq performs scRNA-seq

^{*}Corresponding author, E-mail: kkapner@fas.harvard.edu

on cells with distinct targeted mutations induced by a gene editing technology like CRISPR (Dixit et al., 2016; Adamson et al., 2016; Southard et al., 2025). In these experiments, researchers aim to identify the effect of each perturbation, such as gene loss, expression amplification, or expression reduction, on a phenotype of interest such as survival.

For Gaussian-distributed matrix data, factor analysis and other bilinear models are commonly used to perform dimensionality reduction and effect estimation (Takane and Shibayama, 1991; Tipping and Bishop, 1999). Meanwhile, for non-Gaussian matrix data, generalized bilinear models (GBMs) extend bilinear models by employing a generalized linear model (GLM) for the entries (Choulakian, 1996; Townes et al., 2019; Gabriel, 1998; Miller and Carter, 2020).

This flexible approach models the effect of sample (e.g., cell) covariates, feature (e.g., gene) covariates, and latent factors in an integrated statistical framework. However, the inclusion of sample covariates in a GBM presents a challenge in high-dimensional data, since each additional sample covariate requires estimating a parameter for every feature (Miller and Carter, 2020). Consequently, for genomics data where the number of features can exceed 20,000, very few sample covariates can be included, otherwise the estimation accuracy and computational burden become unacceptable. In the multivariate regression setting—without feature covariates or latent factors—this problem is dealt with using reduced-rank regression (Izenman, 1975; Yee and Hastie, 2003; Camba-Mendez et al., 2003; Davies and Tso, 1982; Braak and Looman, 1994), but these ideas have not yet been explored in the GBM setting.

In this paper, we introduce the reduced-rank generalized bilinear model (RR-GBM) and provide a computationally tractable estimation algorithm that scales to large data sets. Our method estimates a rank-reduced version of the coefficient matrix for the sample covariates, resulting in a significant reduction in the number of parameters in the high-dimensional setting. RR-GBM outperforms the standard full-rank GBM when the true coefficient matrix is approximately reduced-rank or when uninformative covariates are included in the model. The method also provides a way to quantify variable importance and visualize structure among the sample covariate effects via low-dimensional summaries. To select the rank of the sample covariate coefficient matrix as well as the rank of the latent factor matrix, we provide a data-driven model selection method based on data thinning (Neufeld et al., 2024).

The article is organized as follows. In Section 2, we introduce the RR-GBM methodology. Section 3 contains simulation studies illustrating various aspects of the method. Section 4 presents an application to Perturb-seq type data. Section 5 concludes with a brief discussion.

2 Methodology

2.1 Reduced-rank generalized bilinear model

Suppose we have a random data matrix $Y \in \mathbb{R}^{I \times J}$, a matrix of feature covariates $X \in \mathbb{R}^{I \times K}$, and a matrix of sample covariates $Z \in \mathbb{R}^{J \times L}$. Generalized bilinear models (GBMs) assume the form

$$g(\mathbb{E}Y) = XA^T + BZ^T + XCZ^T + U\Sigma V^T \quad (1)$$

where $A \in \mathbb{R}^{J \times K}$, $B \in \mathbb{R}^{I \times L}$, and $C \in \mathbb{R}^{K \times L}$ are matrices of coefficients, $U\Sigma V^T$ is a low-rank matrix of latent effects, and $g(\cdot)$ is a known link function that is applied element-wise; it is assumed that $g(\cdot)$ is a smooth function such that the derivative g' is positive. Here, Y , X , and Z are observed, whereas A , B , C , U , Σ , and V are unknown parameters to be estimated. The interpretation of the dimensions is that I is the number of features, J is the number of samples, K is the number of feature covariates, L is the number of sample covariates, and M is the number of latent factors, which may also need to be estimated. All of the parameters are identifiable under certain constraints. See Miller and Carter (2020) for an in-depth exposition of these models.

We propose to constrain B to be a lower-rank matrix, $B = Q\Lambda R^T$ yielding a *reduced-rank generalized bilinear model* (RR-GBM) in which

$$g(\mathbb{E}Y) = XA^T + Q\Lambda R^T Z^T + XCZ^T + U\Sigma V^T \quad (2)$$

where $Q \in \mathbb{R}^{I \times N}$ and $R \in \mathbb{R}^{L \times N}$ are orthonormal frames, $\Lambda \in \mathbb{R}^{N \times N}$ is diagonal, and N represents the desired rank of B . The rest of the matrices are defined the same way as in the standard GBM. All of the parameters of the RR-GBM are identifiable—including Q , R , and Λ —under natural corresponding constraints, as we discuss in Section 2.2.

The distribution of Y_{ij} is taken to be an exponential dispersion family. More specifically, in this

paper, it is taken to be either the Poisson or the Negative Binomial family.

2.2 Interpretation of the model

The motivation for the RR-GBM is that in the standard GBM, the total number of parameters grows as approximately IL as the number of sample covariates L increases. Since I is typically large, on the order of 10^3 – 10^4 , the computational and statistical efficiency of estimating a full rank B matrix deteriorates quickly as L increases. In contrast, by using a reduced rank matrix $B = Q\Lambda R^T$ in the RR-GBM, the total number of parameters grows as approximately NL as a function of L . Since N is typically small, for instance, $N \leq 10$ or 20 , this yields a massive decrease in the number of parameters that must be estimated. Indeed, in the common use settings described above, the total number of parameters is reduced by a factor of approximately N/L ; see Section S3 for the derivation and exact reduction factor.

The components of the decomposition $B = Q\Lambda R^T$ each reflect a different facet of the structural relationship between the sample covariates and the outcomes. First, the n th column of R represents the effect of each covariate on the latent value of factor n . The n th diagonal entry of Λ represents the overall magnitude of the effect that factor n has on the outcome. The n th column of Q represents the effect that factor n has on each feature (e.g., gene). Putting these pieces together, we obtain a mechanistic representation of how each sample covariate affects the outcome for a given feature in a given sample. The intuition is that the effect of the covariates is mediated by a low-dimensional space of N latent values.

The parameters are not identifiable unless we place appropriate constraints on them. To this end, we assume all of the constraints specified by Miller and Carter (2020), detailed in Section S2 for completeness, as well as additional constraints pertaining to Q , Λ , and R . Specifically, the additional constraints we assume for identifiability of the RR-GBM are:

- (a) Λ is a diagonal matrix such that $\lambda_{11} > \lambda_{22} > \dots > \lambda_{NN} > 0$,
- (b) $Q^T Q = I$ and $R^T R = I$,
- (c) $X^T Q \Lambda R^T = 0$, and
- (d) the first nonzero entry of each column of Q is positive,

where $Q \in \mathbb{R}^{I \times N}$, $\Lambda \in \mathbb{R}^{N \times N}$, $R \in \mathbb{R}^{L \times N}$, and $N < \min\{I, L\}$.

Appropriate selection of the rank of a factorization is a perennial problem, since there is often no clear choice based on fit to the training data, and it is not obvious how to hold out a test set when the data set is a matrix. We present a method for selecting N and M , the ranks of $Q\Lambda R^T$ and $U\Sigma V^T$, respectively, using the data thinning technique of Neufeld et al. (2024) to split the data matrix into independent training and test matrices; see Section 2.5.

2.3 Estimation algorithm for RR-GBM

One approach to fitting the RR-GBM would be to first estimate a full-rank B matrix using an existing GBM algorithm, and then apply the singular value decomposition (SVD) to factorize $B = Q\Lambda R^T$. However, this is not equivalent to enforcing the reduced-rank constraint during estimation by jointly estimating Q , Λ , and R along with the rest of the parameters, which we find yields better performance. In this section, we describe our joint estimation algorithm.

In our proposed algorithm for RR-GBM, the update steps for A , C , Σ , $U\Sigma$, and ΣV^T are the same as those of Miller and Carter (2020). Our algorithm differs only in that we estimate Q , Λ , and R instead of estimating B . The updates involving Q , Λ , and R are done in three sequential steps: (1) updating $Q\Lambda$, (2) updating ΛR^T , and (3) updating Λ .

Briefly, updating $Q\Lambda$ consists of bounded regularized Fisher scoring steps for the product $Q\Lambda$, followed by the enforcement of identifiability constraints, with the updated $Q\Lambda R^T$ coming from a truncated SVD of the updated B matrix. Updating ΛR^T is done in the same way, using steps derived for ΛR^T instead of $Q\Lambda$. One difference, however, is that updating ΛR^T requires the computation of a block matrix of dimension $IL \times IL$ with block sizes of $I \times I$, which can be computationally prohibitive when I and L are large. Thus, to speed up computation of the block Fisher scoring steps for ΛR^T , we use multi-threaded linear algebra routines. Updating Λ consists of a bounded regularized Fisher scoring step. We iteratively cycle through all of the update steps above until a convergence criteria is met or a maximum number of iterations has occurred. See Section S1 for a step-by-step description of the full estimation procedure.

2.4 Visualizing the effect of the sample covariates

An important feature of factorization methods is that they enable one to visualize high-dimensional data in an interpretable way via dimensionality reduction. For instance, principal components analysis (PCA) is often used to visualize clusters and other structures in a collection of high-dimensional data points. Similarly, in RR-GBM, the factorization $B = Q\Lambda R^T$ can roughly be thought of as an application of PCA to the sample covariate coefficient matrix B , and thus, it is amenable to the same visual interpretations.

To illustrate this point, we can visualize the effect of each covariate on each latent factor value by making a heatmap of the estimated ΛR^T matrix; see Figures 5, 6 and 8 for examples. Furthermore, for any two latent factor dimensions, we can make a scatterplot of these effects by plotting the corresponding rows of the estimated ΛR^T matrix; see Figure 8. This is analogous to familiar scatterplots of the first two PC scores, except that each point in the plot corresponds to a covariate rather than a sample; hence, this visualization technique may reveal clusters of covariates or other relationships among covariates in terms of their effect on the outcomes.

Likewise, we can visualize the effect of each latent factor on each feature (e.g., gene) by making a heatmap of the Q matrix, or alternatively, making a scatterplot based on any two columns of Q ; see Figure 8. In such a scatterplot, each point in the plot corresponds to a feature; hence, this may reveal clusters or other relationships among features in terms of how they are affected by the corresponding factors.

2.5 Rank selection using data thinning

In this section, we describe our proposed technique for selecting N , the rank of $B = Q\Lambda R^T$. The same approach can also be used to select the rank of $U\Sigma V^T$. Ideally, one would choose N to minimize the distance between the estimate $\hat{B}$ and the true matrix B , but of course the true B is unknown in practice. For situations with i.i.d. samples, one can split the data into training and test sets, and quantify model performance by fitting on the training set and measuring prediction performance on the test set.

However, it is not appropriate to assume the samples are i.i.d. in our setting. In particular, since

the model involves sample-specific parameters, it is not possible to fit on one set of samples and test on another.

To remedy this problem, we propose using data thinning to split the original data matrix into a training data matrix and test data matrix, both having the same dimensions as the original matrix (Neufeld et al., 2024). Here, we focus on the case of Poisson outcomes, but the technique generalizes to many other common distributions as well. Suppose we have a random data matrix $Y \in \mathbb{R}^{I \times J}$ where $Y_{ij} \sim \text{Poisson}(\mu_{ij})$. For all i and j , given Y_{ij} , sample $Y'_{ij} \sim \text{Binomial}(Y_{ij}, 1/2)$ and set $Y''_{ij} = Y_{ij} - Y'_{ij}$. It can be shown that Y'_{ij} and Y''_{ij} are independent $\text{Poisson}(\mu_{ij}/2)$ random variables, when Y_{ij} is marginalized out (Neufeld et al., 2024). Thus, the matrices Y' and Y'' are independent data matrices with half the mean of the original data matrix. For each candidate rank N , we use Y' to fit the model parameters and then generate our predictions $\hat{\mathbb{E}}Y$ using the covariate values. The RMSE between $\hat{\mathbb{E}}Y$ and Y'' is then obtained. Plotting these RMSE values versus the ranks of the fitted model yields a scree-like plot that helps guide rank selection; see Figure 7.

3 Simulations

In this section, we evaluate the RR-GBM methodology across a range of simulation studies. In each simulation, we randomly generate true parameters for an RR-GBM and simulate a random dataset Y using these true parameters, with either a Poisson or Negative Binomial outcome distribution for Y_{ij} depending on the simulation; see Section S4 for full details.

3.1 Convergence of RR-GBM estimates to true parameter values

To assess the convergence of our parameter estimates to the true values, we performed a simulation study with a range of sample sizes J and sample coefficient matrix ranks N_0 . Specifically, for each $J \in \{500, 1000, 2000, 5000\}$ and $N_0 \in \{1, 2, 3, 4, 5, 6\}$, we simulate Poisson data Y_{ij} (see Section S4), fit the RR-GBM, and compare the estimated parameters to the true parameters. Here, we use $I = 20,000$ features, one latent factor ($M = 1$), and 6 sample covariates ($L = 6$) in both the true and estimated RR-GBMs.

For each dataset, we fit the model using the estimation algorithm described in Section 2.3, setting

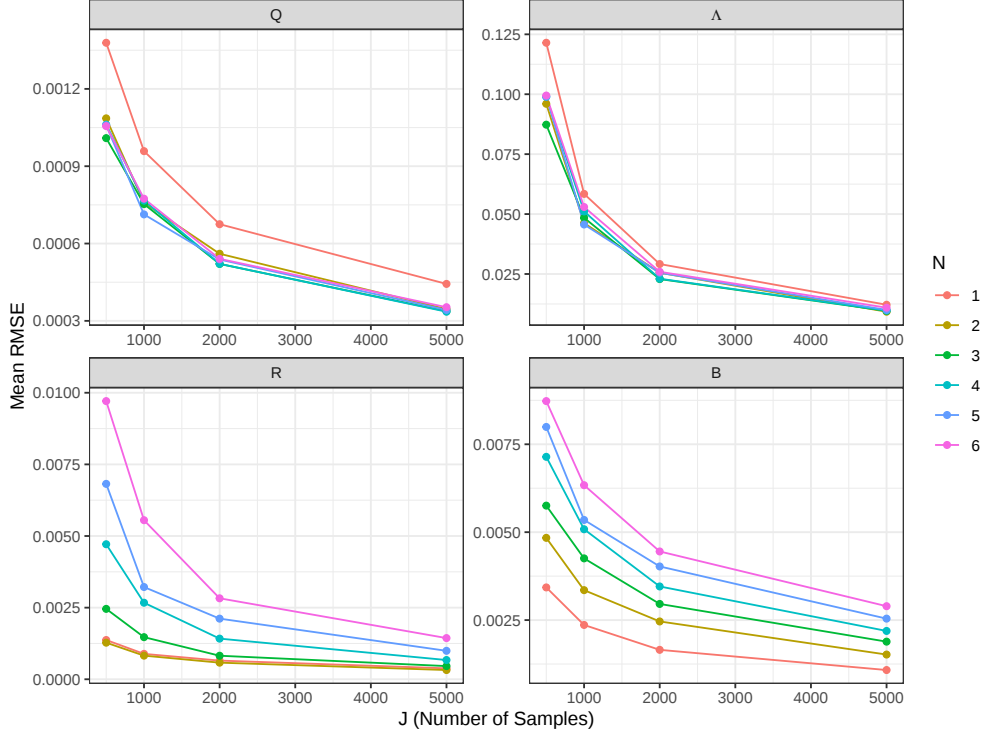


Figure 1: **Convergence of RR-GBM estimates.** RMSE between the true and estimated matrices for Q , Λ , R , and B versus increasing sample size, in the case of Poisson outcomes. In the bottom right plot, $B = Q\Lambda R^T$ and the estimate is $\hat{B} = \hat{Q}\hat{\Lambda}\hat{R}^T$. The true rank N of B was used for fitting the RR-GBM in each case.

the model rank N to the true rank used to generate the data. We evaluate performance by computing the mean RMSE between the true B and estimate $\hat{B}$. In Figure 1, we observe that RMSE decreases as the sample size grows, as expected. We also observe that higher ranks N lead to higher RMSE between $\hat{B}$ and B , due to the larger number of parameters that must be estimated.

3.2 Performance of RR-GBM estimates in the reduced-rank setting

Next, we evaluate the effect of the rank N and number of covariates L on estimation accuracy. We consider two scenarios: (i) the true B matrix is exactly reduced rank, and (ii) the true B matrix is approximately reduced rank. The exactly reduced-rank setting is when the true B matrix has rank $N_0 < L$, which occurs when the effect of the covariates is mediated by N_0 variables that are linear combinations of the covariates. A special case of this would be when only N_0 covariates have nonzero effects. The approximately reduced-rank setting is when the true B matrix can be closely approximated by a rank $N_0 < L$ matrix, which occurs when the effects are predominantly captured

by N_0 factors.

In the full GBM, a total of $(I - K)L$ parameters need to be estimated for B , but in the RR-GBM, only $N(I - K + L - N)$ parameters need to be fit for B . For given I , K , and L with $I \gg \max\{K, N\}$, which is the case for scRNA-seq data, a full GBM requires fitting around L/N times as many parameters for B as a rank N RR-GBM; see Section S3.3.

Thus, we expect that as the number of sample covariates L grows, the benefit of using an RR-GBM will increase.

3.2.1 Exact reduced-rank setting

When the true rank of B is N_0 , we observe that the estimation error for B decreases as we increase N (the rank of B in the fitted RR-GBM), until $N = N_0$ (Figure S1). If a model of rank $N > N_0$ is fit, the estimation procedure terminates without finding a solution. This occurs because when the true B has rank less than N , there is not a way to obtain a unique rank N approximation of B , so the compact SVD does not have a solution. This behavior of the procedure can be used to quickly identify any covariates that have linearly dependent coefficients, that is, covariates that have the same effect on all of the outcomes.

3.2.2 Approximate reduced-rank setting

When the true B matrix is approximately rank N_0 , we observe in Figure 2 that the estimation error for B decreases as we increase N until $N = N_0$, and then the error increases again until $N = L$. The maximum possible value of N is L since there are L covariates; the RR-GBM with $N = L$ is equivalent to the standard full GBM (Miller and Carter, 2020).

Figure 2 illustrates that RR-GBM has better RMSE than the full GBM model when the true B has approximate rank $N_0 \leq N < L$. Indeed, the full GBM corresponds to the model with rank 10 on the x-axis, and we see that the RR-GBM has lower RMSE for every value of N greater or equal to the true approximate rank N_0 . This result can be intuitively understood as coming from the bias-variance tradeoff, since the increased variance due to estimating the parameters of factors with small effects outweighs the reduction in bias-squared due to having the correct full model.

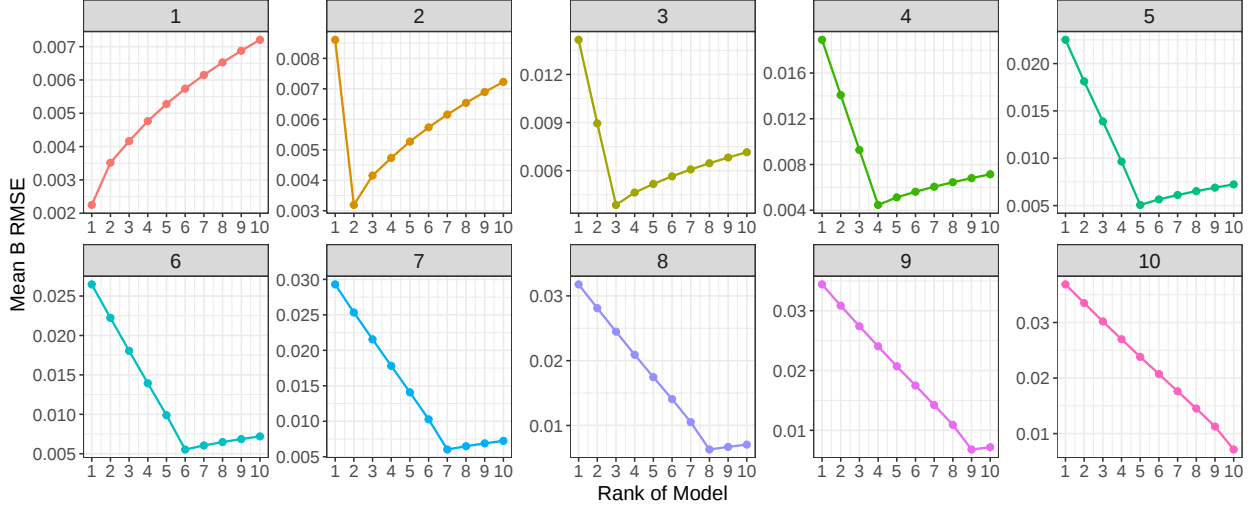


Figure 2: **RR-GBM estimation accuracy as a function of true rank and model rank.** Panels are labeled with their respective true approximate rank of B (corresponding $L - N_0$ singular values are 0.01 times their randomly generated value). The average RMSE (over 100 simulations) between the estimated B (given by $\hat{Q}\hat{\Lambda}\hat{R}^T$) and the true B are displayed. Data was generated using a Negative Binomial outcome distribution and fit with a model with 1 latent factor ($M = 1$).

Even though the true B is full rank, allowing a small amount of bias by ignoring less important factors can reduce the variance in the estimates enough to improve the RMSE.

Figure 3 shows the estimation error as the number of sample covariates L increases, for varying choices of N , when $N_0 = 1$. As expected, normalized RMSE increases with L , however, using an RR-GBM yields much lower error than a full GBM when the true B is approximately reduced rank (Figure 3). Furthermore, the improvement in relative performance increases with L .

3.3 Computation time of RR-GBM compared to full GBM

In Figure 4, we compare the runtime of the RR-GBM and the full GBM, averaged over 100 simulations. As the number of sample covariates L increases, the computation time required to fit the full GBM increases rapidly. Meanwhile, the RR-GBM requires only around twice as much time to fit with $L = 100$ sample covariates as with $L = 10$. Also, we see that these computation times do not depend strongly on the true approximate rank N_0 for $N_0 \ll L$. Here, for the RR-GBM, we set the model rank N equal to the true approximate rank N_0 .

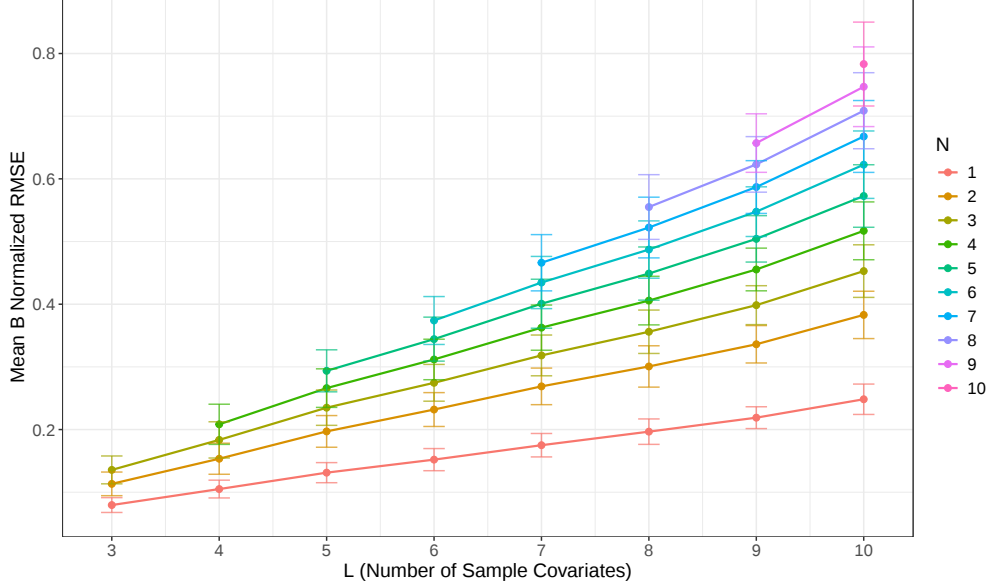


Figure 3: **Performance of RR-GBM estimates as L increases.** Normalized RMSE between the true and estimated B plotted against the number of sample covariates L , for Negative Binomial outcomes with one latent factor ($M = 1$), $J = 2500$, $I = 5000$, and $K = 1$. The approximate rank of B_0 was $N_0 = 1$ for all L . Each point represents the average over 100 simulation runs and the error bars denote standard deviation. Normalized RMSE is defined as $\|\hat{B} - B_0\|_F / \|B_0\|_F$ where $\|\cdot\|_F$ denotes the Frobenius norm. This normalization was performed to adjust for the increasing size of $B_0 \in \mathbb{R}^{I \times L}$ so that performance can be compared across L . We see that $N = 1$ always has the smallest normalized RMSE across increasing L and that normalized RMSE increases monotonically with an increasing fitted N . The top point for each L (corresponding to $N = L$) is equivalent to the full GBM model and always has the highest normalized RMSE, highlighting the benefit of RR-GBM in the reduced rank setting, particularly as the number of sample covariates L grows.

3.4 Visualization of the sample covariate loadings matrix

The decomposition QAR^T yields a method for both visualization of the latent covariate structure and selection of covariates, by plotting the loadings matrix ΛR^T as described in Section 2.4.

We consider two distinct ways of generating the B matrix for simulations. In the first setting, referred to as “Differential Effects”, we initially generate B as described in Section S4, then randomly select a subset of covariates, and for these covariates, we double the entries of B for a random subset of features. Conceptually, in the scRNA-seq setting, this represents the situation where a given sample covariate affects the expression of some percentage of genes. In the second setting, referred to as “Structured Effects”, we generate B using the procedure described in Section S4, except that the R_0 matrix is generated to have a certain column space, as described in Section S4.2.

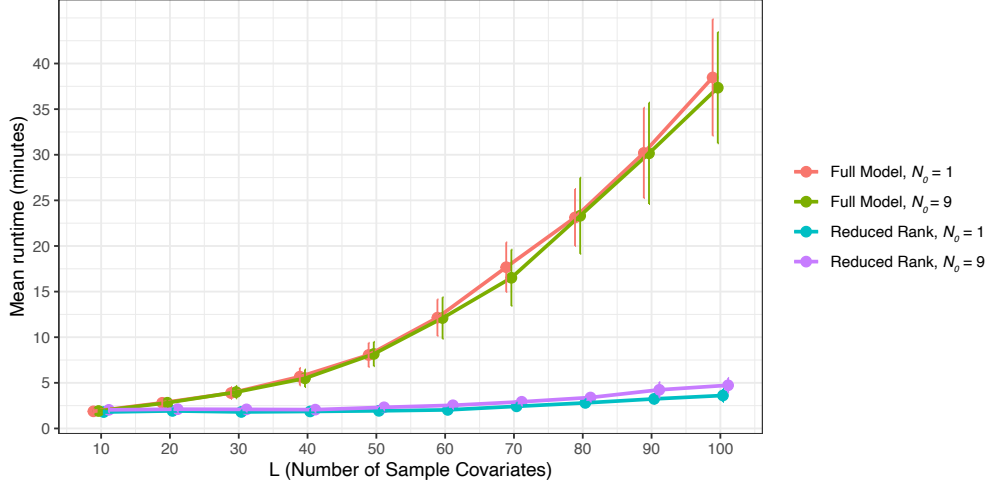


Figure 4: **Computation time of RR-GBM compared to full GBM.** For each $L \in \{10, 20, \dots, 100\}$, Poisson simulated data sets were generated using two settings for the approximate rank of the true B matrix: (i) $N_0 = 1$ and (ii) $N_0 = 9$. Two models were fit to each data set: a full-rank model with $N = L$ and a reduced-rank model with $N = N_0$. Each point shown is the average of 100 simulations with the respective settings, along with $I = 5000$, $J = 2500$, $K = 1$, and $M = 1$. Models were run on a single 48-core CPU HPC cluster with at most 15GB of RAM.

First, in the Differential Effects setting, we consider a simulated dataset in which 5 out of 20 covariates affect the outcome, and their effects are uncorrelated. In Figure 5, we show the true and estimated loadings. We see that the model correctly infers which covariates are relevant, as well as correctly separating their effects into 5 distinct factors.

In this simulation, the magnitude of the loading is related to the percentage of the features that the given covariate influences, with covariate 19 influencing 50% of features for example. Since only the first five factors have non-negligible values in ΛR^T for any covariate, the rank of the model can effectively be reduced to five, resulting in a roughly 75% reduction in the number of parameters that need to be estimated, compared with the full GBM.

Now, consider the Structured Effects setting. When the sample covariates affect the outcome in interrelated ways, multiple sample covariates may contribute to a given latent factor. In Figure 6, we show estimated RR-GBM loadings for two simulation examples with more complex loadings patterns; see Section S4.2.1 for details. These examples demonstrate that the model can distinguish factors with overlapping or even nested sets of sample covariates with nonzero loadings. Thus, not only is RR-GBM able to identify the relevant covariates, it decomposes their overall effect into

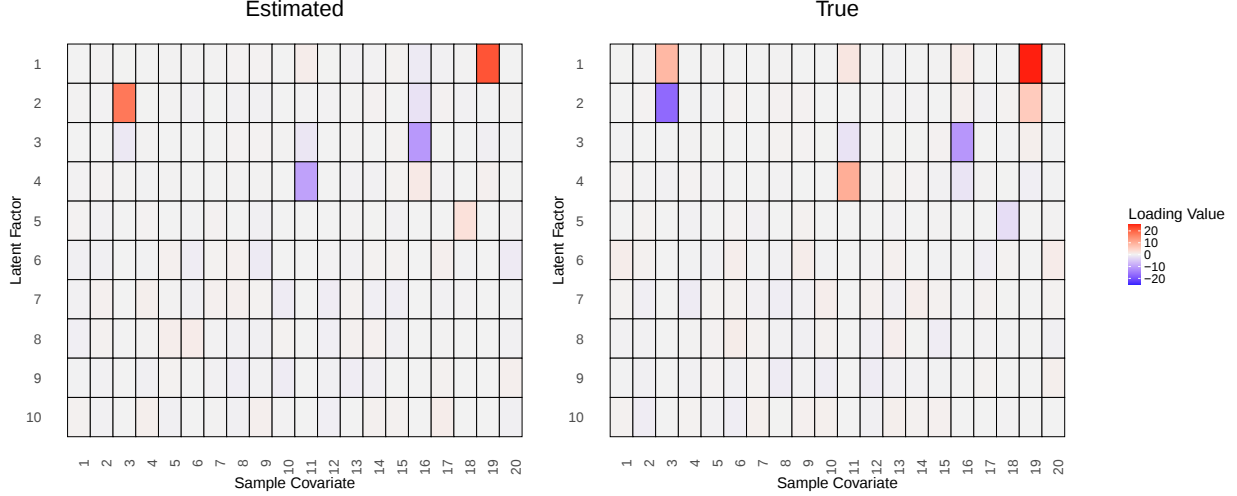


Figure 5: **Visualizing the effect of sample covariates on latent factors in RR-GBM.** The estimated $\hat{\Lambda}\hat{R}^T$ and true $\Lambda_0 R_0^T$ are shown. Data were simulated with 20,000 features, 200 samples, 1 latent factor, and 20 Bernoulli sample covariates with a Negative Binomial outcome distribution. The feature level coefficients for the five sample covariates 18, 11, 16, 3, 19, were altered at random with $\{1\%, 5\%, 10\%, 30\%, 50\%\}$ of feature coefficients being doubled, respectively. The true B_0 matrix was simulated as described in Section S4, before the selected coefficients were altered.

loadings onto shared latent factors which, in turn, act in concert on the outcomes.

3.5 Identification of differential features with respect to each factor

Each column of the Q matrix in the decomposition $Q\Lambda R^T$ represents the effect of the corresponding factor on all of the features. The Q matrix can be visualized using a heatmap, analogously to the visualization of the sample covariate loadings matrix ΛR^T ; see Figure S3 for an example from the Differential Effects simulation. In particular, this allows for the identification of features that differ with respect to each factor. With this interpretation, the magnitude of the entries in the k th column of Q indicate the strength of the differential effects of factor k on the respective features. For instance, in scRNA-seq data, a column of Q may represent the effects of a given gene program on the expression levels of all the genes.

An advantage of the RR-GBM is that it can detect such programs even if the net effect in the B matrix for a given covariate/feature pair is null. More precisely, suppose $b_{il} = 0$ for some covariate l and feature i . Using a full-rank GBM would reveal no relationship between l and i . However, since the factors are shared among all covariates in an RR-GBM, it is possible to infer nonzero contributions q_{in} and r_{ln} in the decomposition $b_{il} = \sum_{n=1}^N \lambda_{nn} q_{in} r_{ln}$. Thus, the RR-GBM can yield

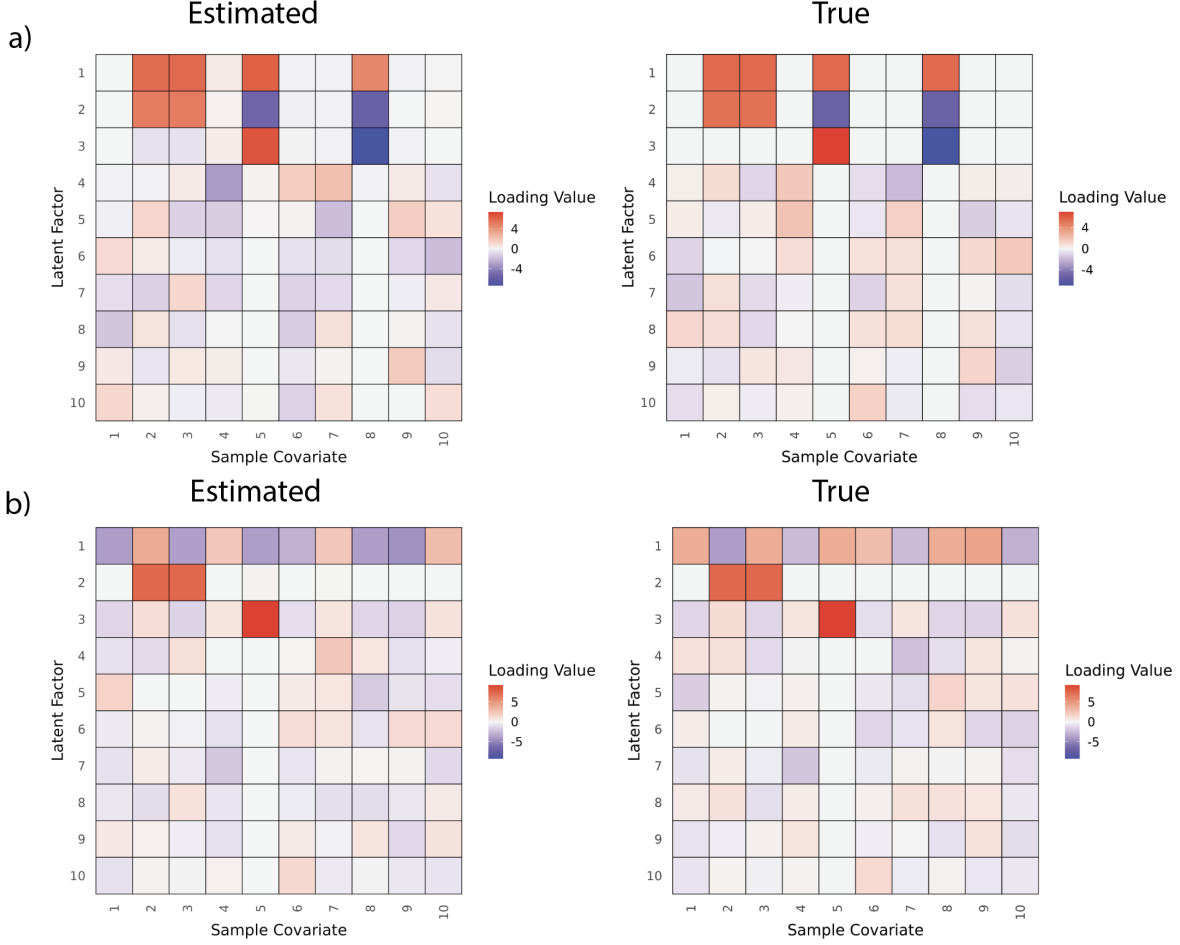


Figure 6: **Visualizing the effect of sample covariates on latent factors in RR-GBM.** Shown are heatmaps of the estimated and true sample covariate loadings in two scenarios, with the estimated matrix $\hat{\Lambda}\hat{R}^T$ on the left and the ground truth ΛR^T on the right. (Scenario a) In the true loadings, four covariates have strong effects on the first three factors, with a nested structure. (Scenario b) In the true loadings, all sample covariates contribute to the first factor, while only covariates two and one contribute to the second and third factors, respectively. In each scenario, $\Lambda_0 R_0^T$ was simulated as described in Section S4.2.1. Negative Binomial outcome data was simulated with 2000 samples, 5000 features, 10 Bernoulli sample covariates, 1 latent factor, and 1 feature covariate. A full rank 10 model was fit to the data using the RR-GBM estimation algorithm.

insights into the effect of covariate l on programs that modulate feature i , even when they cancel out in the aggregate for that covariate. This capability is due to the information sharing implied by the reduced-rank structure.

To recover the usual log-fold change estimates for a given feature and sample covariate pair, we can take the product $\hat{Q}\hat{\Lambda}\hat{R}^T$ to obtain $\hat{B}$ and examine the corresponding entry $\hat{b}_{il}$.

3.6 Empirical rank selection

As shown above, when the true B matrix is exactly or approximately reduced rank, fitting a reduced-rank model results in improved estimates of B compared to using a full GBM. In practice, the rank or approximate rank of B is rarely known and thus it becomes necessary to have a selection method for choosing the rank of B in the fitted model. It is not possible to use a standard cross-validation approach since sample- and gene-specific parameters are fit in the model. To overcome this obstacle, we use the data thinning technique of Neufeld et al. (2024) to create a train and test split of our data and assess prediction performance via RMSE. This procedure creates two independent data matrices, each having the same dimensions as the original data matrix but with half the mean. Thus, we can use one to fit and one to assess performance. The rank of B that yields the smallest test RMSE then provides a data-driven choice of rank to use in the model.

Figure 7 demonstrates this rank selection procedure on data sets generated using true B matrices with approximate ranks ranging from 1 to 10; see Section S4 for details. In each case, we find that the mean test RMSE from 100 simulations is minimized at the true approximate rank of B . This same pattern was observed in all 100 simulations and a version of this figure from one simulation run can be seen in Figure S2.

In addition to selection of rank of B in the reduced-rank regression setting, this method can also be used to select M , the rank of $U\Sigma V^T$, in both the RR-GBM and the full GBM. By performing the same data thinning technique and fitting across a range of M values, this procedure provides an estimate of the number of latent factors in the true data generating process.

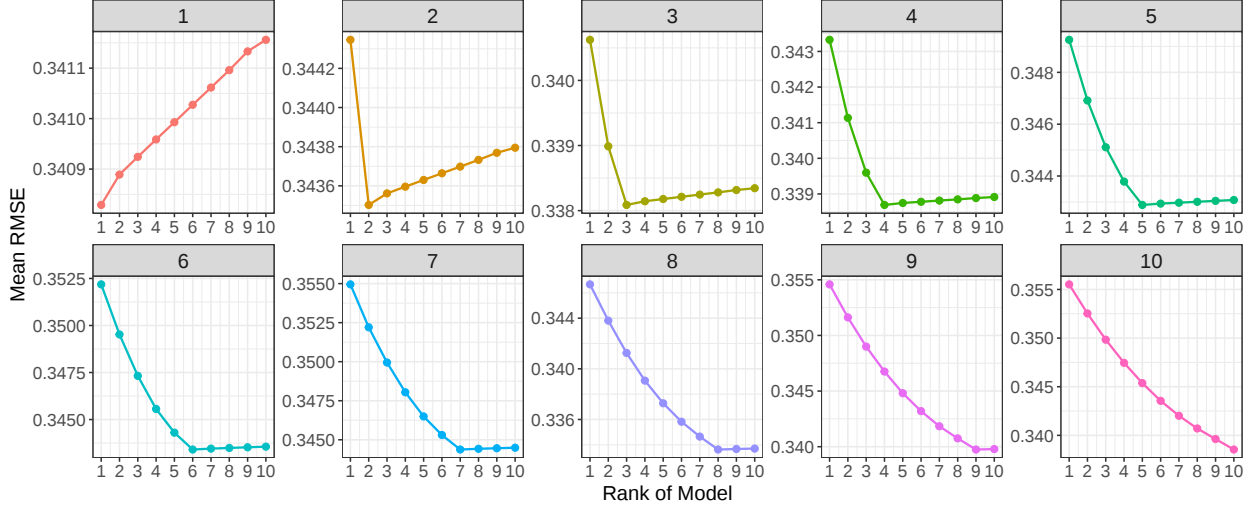


Figure 7: **Empirical selection of the rank in RR-GBM.** Panels are labeled according to the true approximate rank of B , where the corresponding $L - N_0$ singular values are 0.01 times their randomly generated value to create a B with approximate rank N_0 . Each dot indicates the average RMSE (over 100 simulations) between the predicted output values and the true values. Poisson data was generated and fit using a model with 1 latent factor ($M = 1$).

4 Application

The RR-GBM model lends itself to applications in which there are a large number of sample covariates that are potentially important to consider. In particular, this is the case when the primary objective is to determine which of many sample covariates are meaningful and what their effects are. An excellent example is single-cell RNA-sequencing screening, an experimental protocol that has recently been growing in popularity.

In an scRNA-seq screen, a range of perturbations are performed, such as genetic or environmental modifications, and scRNA-seq is used to quantify the resulting transcriptomic profiles. Typically, the experiment is designed such that each cell is perturbed by deleting, amplifying, or inhibiting the activity of one or more genes or exposing the cell to an environmental agent, such as a drug or biologically active compound.

In the gene perturbation setting, the number of genes that are perturbed can range from hundreds to the entire collection of 20,000 genes, and in the environmental perturbation setting, there is no limit to the number of perturbations that can be performed. The objective is to infer which perturbations have significant effects, and how they affect the expression of genes. Thus, from a

statistical perspective, each different perturbation represents a binary sample covariate that equals 0 or 1, indicating whether that perturbation was applied to the given cell.

We refer to Adamson et al. (2016) and Dixit et al. (2016) for background.

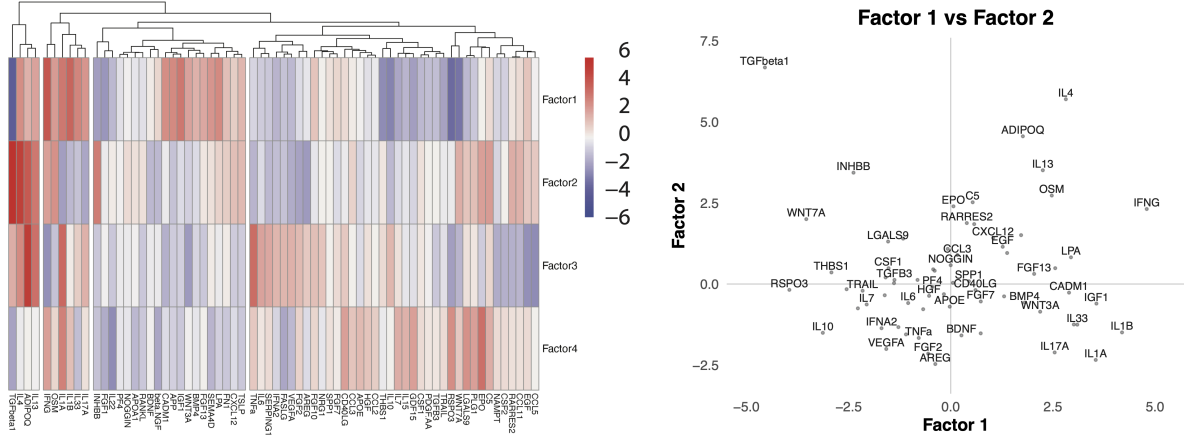
We apply RR-GBM to an scRNA-seq screen data set consisting of the transcriptional responses of pancreatic cancer cells perturbed by a range of tumor microenvironment (TME) protein ligands (Liu et al., 2025). The data consist of counts for $I = 15,876$ genes in $J = 10,881$ cells exposed to $L = 68$ different ligands; see Liu et al. 2025 for experimental design details.

The original study of Liu et al. (2025) analyzed the data using a pipeline in which they first run consensus non-negative matrix factorization (cNMF) to identify gene expression programs (GEPs), filter to GEPs that were variably expressed across cells, and run elastic net regression to infer the effect of each perturbation on the GEPs. For visualization, they compute Pearson residuals, run principal components analysis, use the elbow method to select the number of components, and then run UMAP. In contrast, we simply fit an RR-GBM to the raw count data without any filtering. The columns of Q in the RR-GBM can be interpreted as orthogonal GEPs, aiding the biological interpretation, and the effects of the perturbations on the GEPs are captured by the loadings ΛR^T . Thus, the RR-GBM provides a more direct analytical approach using one coherent statistical model.

Figure 8 summarizes the results using a Negative Binomial RR-GBM with rank $N = 4$, which yields many of the same results as Liu et al. (2025) while using a more unified approach. In Figure 8a, we visualize the relationships among the 68 perturbations via a heatmap of $\hat{\Lambda}\hat{R}^T$ as described in Section 2.4, with the addition of column (ligand) ordering by hierarchical clustering over the 4 latent factors. The ligands with large magnitude loadings in factor 1, such as IFNG, IL1A, and IL1B, are related to inflammatory and growth related signaling; meanwhile, ligands with large magnitude loadings in factor 2, such as TGF β , ADIPOQ, and IL4, are related to immune regulation and tissue remodeling (Lee et al., 2021; Wynn, 2015; Liu et al., 2025).

To investigate the effects of these two factors on gene expression, we visualize the corresponding columns of $\hat{Q}$ in a scatterplot (Figure 8b). Since the sign of each factor is arbitrary, the positive/negative direction of each axis in Figure 8b needs to be interpreted in conjunction with the

a)



b)

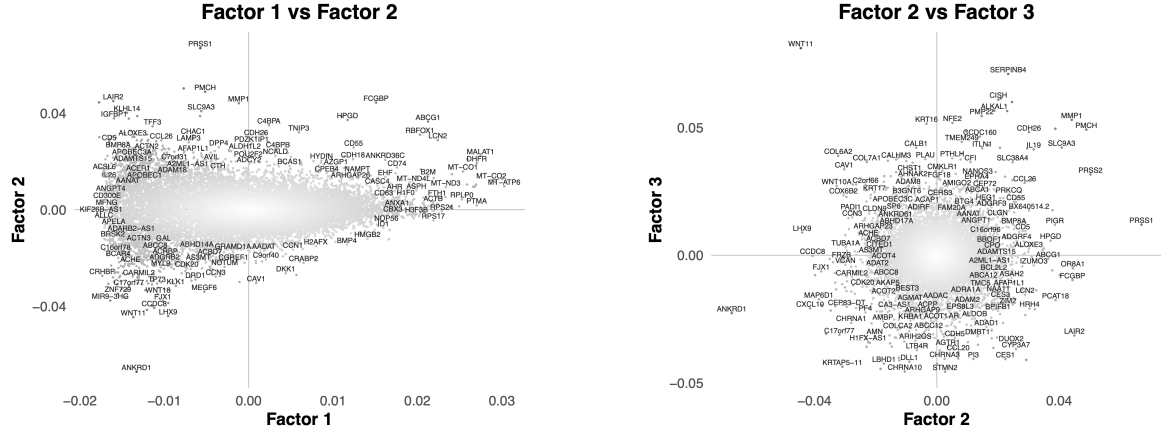


Figure 8: **RR-GBM results on Perturb-seq data for TME ligand experiment.** We fit an RR-GBM with rank $N = 4$ to the (Liu et al., 2025) dataset of a TME phenotypic screen in PDAC organoids. (a) $\hat{\Lambda}\hat{R}^T$ is shown on the left, with hierarchical clustering is applied to the TME ligands (columns). The dendrogram was cut into 4 groups to highlight a potential grouping of ligands. The panel on the right shows an alternate visualization of the top two factors in $\hat{\Lambda}\hat{R}^T$, with factor 2 plotted against factor 1. (b) The figure on the left shows the first and second columns of $\hat{Q}$ plotted against each other to highlight differential genes between them. The figure on the right shows the second and third columns of $\hat{Q}$ plotted against each other.

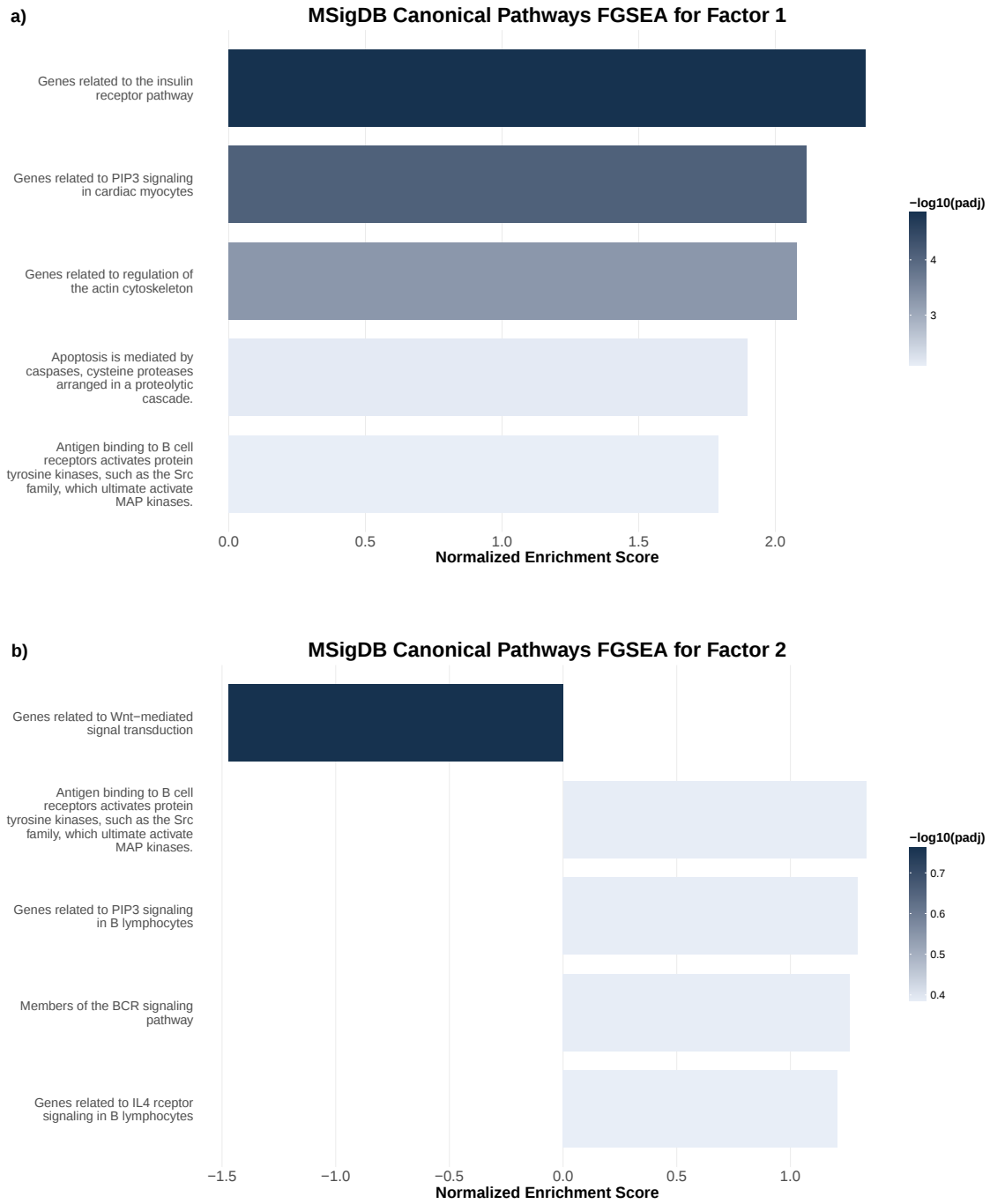


Figure 9: **FGSEA of gene loadings for factors 1 and 2.** The columns of $\hat{Q}$ were used as rankings for the FGSEA algorithm (Korotkevich et al., 2021) to identify relevant pathways that are affected by the corresponding identified ligands. Pathways from the canonical pathways gene set collection from MSigDB were used (Subramanian et al., 2005; Liberzon et al., 2011). (a) The canonical pathways most associated with the gene loadings from factor 1. (b) The canonical pathways most associated with the gene loadings from factor 2.

corresponding row in Figure 8a. For example, we see in Figure 8a that the second cluster of ligands—consisting of IL17A, IL33, IL1B, IL1A, OSM, and IFNG—all have positive loadings for factor 1, which means that genes with a positive factor 1 value in Figure 8b are upregulated in cells exposed to these ligands. To link the results with known biology, we can use standard methods such as enrichment analysis to gain insight into pathways associated with each inferred GEP (Korotkevich et al., 2021). For instance, in Figure 9, we find that notable pathways associated with factor 1 include regulation of the cytoskeleton and PI3K signaling, and notable pathways with factor 2 include down-regulation of Wnt signaling and immune mediated signaling. While experimental validation of these findings are outside the scope of this work, we refer to Liu et al. (2025), where findings related to the PDAC IL-4/IL-13 response program are captured in our factor 1 and the TGF β program is similar to our factor 2. Taken together, these results illustrate how the estimated $\hat{Q}$, $\hat{\Lambda}$, and $\hat{R}$ can be used directly for visualization, analysis, and hypothesis generation in a unified statistical framework without the need for additional pre- or post-processing.

5 Discussion

The RR-GBM makes it possible to apply the GBM framework to data matrices with a large number of sample covariates. Specifically, by constraining the sample covariate parameter matrix B to be low rank, the RR-GBM yields improved estimates at lower computational cost. The resulting factorization of B enables visualization of the relationships among sample covariates in terms of their effects, and identification of biologically meaningful latent factors such as gene expression programs. Our data thinning approach for selecting the best-performing rank of B to use in the RR-GBM is easily conducted in parallel and does not impose any additional computational costs beyond fitting the model a selected number of times.

In future work, the proposed method could also be applied to the feature covariate matrix A , to handle large numbers of feature covariates; by the symmetry of Equation 1, the approach we developed for B can alternatively be used to estimate a reduced-rank A matrix. Additionally, providing uncertainty quantification for the entries of the reduced-rank matrix components would enable confidence interval construction and hypothesis testing on both the latent factor directions (Q) and the associated loadings onto the sample covariates (R). This capability would be useful

for identifying statistically significant genes in a gene expression program or detecting differentially expressed genes between two groups. Another interesting direction would be to apply the iteratively reweighted singular value decomposition (IRSVD) algorithm of Nicol and Miller (2025) to estimate $Q\Lambda R^T$ instead of the bounded regularized Fisher scoring approach used in the present paper, in order to scale up to even larger numbers of features and sample covariates.

Acknowledgments

We thank Rafael Irizarry and Rong Ma for helpful discussions. This work was funded in part by the NIH Genomics Training Grant (T32GM135117). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

- Adamson, B., T. M. Norman, M. Jost, M. Y. Cho, J. K. Nuñez, Y. Chen, J. E. Villalta, L. A. Gilbert, M. A. Horlbeck, M. Y. Hein, R. A. Pak, A. N. Gray, C. A. Gross, A. Dixit, O. Parnas, A. Regev, and J. S. Weissman (2016). A multiplexed single-cell CRISPR screening platform enables systematic dissection of the unfolded protein response. *Cell* 167(7), 1867–1882.e21.
- Blum, A., J. Hopcroft, and R. Kannan (2020). *Foundations of Data Science*. Cambridge University Press.
- Braak, C. J. F. T. and C. W. N. Looman (1994). Biplots in reduced-rank regression. *Biometrical Journal* 36(8), 983–1003.
- Camba-Mendez, G., G. Kapetanios, R. J. Smith, and M. R. Weale (2003). Tests of rank in reduced rank regression models. *Journal of Business & Economic Statistics* 21(1), 145–155.
- Chakraborty, R. and B. C. Vemuri (2019, February). Statistics on the Stiefel manifold: Theory and applications. *The Annals of Statistics* 47(1), 415–438.
- Choulakian, V. (1996). Generalized bilinear models. *Psychometrika* 61(2), 271–283.
- Davies, P. T. and M. K.-S. Tso (1982). Procedures for reduced-rank regression. *Journal of the Royal Statistical Society Series C: Applied Statistics* 31(3), 244–255.

- Dixit, A., O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, B. Adamson, T. M. Norman, E. S. Lander, J. S. Weissman, N. Friedman, and A. Regev (2016). Perturb-Seq: Dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* 167(7), 1853–1866.e17.
- Gabriel, K. R. (1998). Generalised bilinear regression. *Biometrika* 85(3), 689–700.
- Izenman, A. J. (1975). Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis* 5(2), 248–264.
- Korotkevich, G., V. Sukhov, N. Budin, B. Shpak, M. N. Artyomov, and A. Sergushichev (2021). Fast gene set enrichment analysis. *bioRxiv preprint*.
- Lee, B. Y., E. K. J. Hogg, C. R. Below, A. Kononov, A. Blanco-Gomez, F. Heider, J. Xu, C. Hutton, X. Zhang, T. Scheidt, K. Beattie, A. Lamarca, M. McNamara, J. W. Valle, and C. Jørgensen (2021). Heterocellular OSM-OSMR signalling reprograms fibroblasts to promote pancreatic cancer growth and metastasis. *Nature Communications* 12(1), 7336.
- Liberzon, A., A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27(12), 1739–1740.
- Liu, N., W. E. Kattan, B. E. Mead, C. Kummerlowe, T. Cheng, S. Ingabire, J. H. Cheah, C. K. Soule, A. Vrcic, J. K. McIninch, S. Triana, M. Guzman, T. T. Dao, J. M. Peters, K. E. Lowder, L. Crawford, A. P. Amini, P. C. Blainey, W. C. Hahn, B. Cleary, B. Bryson, P. S. Winter, S. Raghavan, and A. K. Shalek (2025). Scalable, compressed phenotypic screening using pooled perturbations. *Nature Biotechnology* 43(8), 1324–1336.
- Marčenko, V. A. and L. A. Pastur (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik* 1(4), 457.
- Miller, J. W. and S. L. Carter (2020). Inference in generalized bilinear models. arXiv:2010.04896.
- Neufeld, A., L. L. Gao, J. Popp, A. Battle, and D. Witten (2024). Inference after latent variable estimation for single-cell RNA sequencing data. *Biostatistics* 25(1), 270–287.

- Nicol, P. B. and J. W. Miller (2025). Model-based dimensionality reduction for single-cell RNA-seq using generalized bilinear models. *Biostatistics* 26(1), kxaf024.
- Southard, K. M., R. C. Ardy, A. Tang, D. D. O’Sullivan, E. Metzner, K. Guruvayurappan, and T. M. Norman (2025). Comprehensive transcription factor perturbations recapitulate fibroblast transcriptional states. *Nature Genetics* 57(9), 2323–2334.
- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* 102(43), 15545–15550.
- Takane, Y. and T. Shibayama (1991). Principal component analysis with external information on both subjects and variables. *Psychometrika* 56(1), 97–120.
- Tipping, M. E. and C. M. Bishop (1999). Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 61(3), 611–622.
- Townes, F. W., S. C. Hicks, M. J. Aryee, and R. A. Irizarry (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology* 20, 1–16. Num Pages: 1-16 Place: London, United Kingdom.
- Tropp, J. A. (2012, August). A comparison principle for functions of a uniformly random subspace. *Probability Theory and Related Fields* 153(3), 759–769.
- Wynn, T. A. (2015). Type 2 cytokines: mechanisms and therapeutic strategies. *Nature Reviews Immunology* 15(5), 271–282.
- Yee, T. W. and T. J. Hastie (2003). Reduced-rank vector generalized linear models. *Statistical Modelling* 3(1), 15–41.

S1 Detailed algorithm for RR-GBM estimation

In this section, we describe our algorithm for estimating the Q , Λ , and R parameter matrices. For details on estimating the rest of the GBM parameters, we refer to Miller and Carter (2020). The following steps are substituted into the algorithm of Miller and Carter (2020) in place of the B estimation steps. We cycle through the steps to update each parameter matrix, iterating until convergence as in Miller and Carter (2020).

We first define non-standard notation, following the conventions used in Miller and Carter (2020). We write $\text{Block}(A_{ij} : i, j \in \{1, \dots, n\})$ to denote the block matrix with block i, j equal to the matrix A_{ij} . For a matrix $A \in \mathbb{R}^{m \times n}$, A_{i*} and A_{*j} denote the diagonal matrices constructed from the i th row and the j th column, respectively. For $A \in \mathbb{R}^{m \times n}$, $\text{vec}(A)$ is defined as $\text{vec}(A) = (a_{11}, a_{21}, \dots, a_{m1}, a_{12}, a_{22}, \dots, a_{m2}, \dots, a_{mn}) \in \mathbb{R}^{mn}$. “Unvectorize” returns the original matrix A that has been vectorized as implemented by $\text{vec}(A)$. X^+ is the Moore–Penrose pseudoinverse of X .

The steps for updating Q , Λ , and R are as follows. The matrices W and E are defined as in Section S5 in Miller and Carter (2020).

- Updating $Q\Lambda$
 1. $\Phi \leftarrow Q\Lambda$
 2. For $i = 1, \dots, I$
 - (a) $\xi \leftarrow (R^T Z^T W_{i*} Z R + \lambda_b I)^{-1} (R^T Z^T E_{i:} - \lambda_b \Phi_{i:})$
 - (b) $\Phi_{i:} \leftarrow \Phi_{i:} + \xi \min\{1, \rho \sqrt{\dim(\Lambda)} / \|\xi\|\}$
 3. $B \leftarrow \Phi R^T$
 4. $D \leftarrow X^+ B$
 5. $B \leftarrow B - X D$
 6. $C \leftarrow C + D$
 7. Run compact SVD of rank N on B yielding $Q\Lambda R^T$
- Updating ΛR^T
 1. $O \leftarrow \Lambda R^T$

2. $F \leftarrow \text{Block}(\sum_{j=1}^J z_{j\ell} z_{j\ell'} (Q^T W_{*j} Q) : \ell, \ell' \in \{1, \dots, L\})$
3. $\xi \leftarrow (F + \lambda_b I)^{-1} (\text{vec}(Q^T E Z) - \lambda_b - \lambda_b \text{vec}(O))$
4. $\text{vec}(O) \leftarrow \text{vec}(O) + \xi \min\{1, \rho \sqrt{\dim(\Lambda)} / \|\xi\|\}$
5. Unvectorize O
6. $B \leftarrow PO$
7. $D \leftarrow X^+ B$
8. $B \leftarrow B - XD$
9. $C \leftarrow C + D$
10. Run compact SVD of rank N on B yielding $Q\Lambda R^T$
- Updating Λ
 1. $F \leftarrow \sum_{i=1}^I (ZRQ_{i*})^T W_{i*} ZRQ_{i*}$
 2. $\xi \leftarrow (F + \lambda_b I)^{-1} (\text{diag}(Q^T E Z R) - \lambda_b \text{diag}(\Lambda))$
 3. $\text{diag}(\Lambda) \leftarrow \text{diag}(\Lambda) + \xi \min\{1, \rho \sqrt{\dim(\Lambda)} / \|\xi\|\}$

S2 Identifiability constraints for RR-GBM

Recall the form of the RR-GBM model in Equation 2:

$$g(\mathbb{E}Y) = XA^T + Q\Lambda R^T Z^T + XCZ^T + U\Sigma V^T.$$

We show that the following constraints ensure identifiability of all the model parameters:

- (a) $X^T X$ and $Z^T Z$ are invertible,
- (b) $X^T Q\Lambda R^T = 0$, $Z^T A = 0$, $X^T U = 0$, and $Z^T V = 0$,
- (c) $U^T U = I$, $V^T V = I$, $Q^T Q = I$, and $R^T R = I$,
- (d) Σ is a diagonal matrix such that $\sigma_{11} > \sigma_{22} > \dots > \sigma_{MM} > 0$,
- (e) the first nonzero entry of each column of U is positive,
- (f) Λ is a diagonal matrix such that $\lambda_{11} > \lambda_{22} > \dots > \lambda_{NN} > 0$, and
- (g) the first nonzero entry of each column of Q is positive.

where $A \in \mathbb{R}^{J \times K}$, $C \in \mathbb{R}^{K \times L}$, $\Sigma \in \mathbb{R}^{M \times M}$, $U \in \mathbb{R}^{I \times M}$, $V \in \mathbb{R}^{J \times M}$, $X \in \mathbb{R}^{I \times K}$, $Z \in \mathbb{R}^{J \times L}$, $Q \in \mathbb{R}^{I \times N}$, $\Lambda \in \mathbb{R}^{N \times N}$, $R \in \mathbb{R}^{L \times N}$, $N < \min\{I, L\}$, and $M < \min\{I, J\}$.

Since $B = Q\Lambda R^T$, the proof of identifiability for A , C , U , Σ , and V is still valid and can be found in Miller and Carter (2020). We now show that Q , Λ , and R are also identifiable given these constraints. If $(A_1, Q_1, \Lambda_1, R_1, C_1, U_1, \Sigma_1, V_1, X, Z)$ and $(A_2, Q_2, \Lambda_2, R_2, C_2, U_2, \Sigma_2, V_2, X, Z)$ satisfy the above constraints and

$$XA_1^T + Q_1\Lambda_1R_1^TZ^T + XC_1Z^T + U_1\Sigma_1V_1^T = XA_2^T + Q_2\Lambda_2R_2^TZ^T + XC_2Z^T + U_2\Sigma_2V_2^T \quad (3)$$

then we know from Miller and Carter (2020) that $A_1 = A_2$, $C_1 = C_2$, $U_1 = U_2$, $\Sigma_1 = \Sigma_2$, and $V_1 = V_2$. Plugging these equations into Equation 3 and canceling yields that

$$Q_1\Lambda_1R_1^TZ^T = Q_2\Lambda_2R_2^TZ^T.$$

Right multiplying by Z and using constraint (a), we have

$$Q_1\Lambda_1R_1^T = Q_2\Lambda_2R_2^T.$$

By the uniqueness properties of the singular value decomposition, constraints (c) and (f) imply that $\Lambda_1 = \Lambda_2$, $Q_1 = Q_2S$, and $R_1^T = SR_2^T$ for a diagonal matrix S of the form $S = \text{Diag}(\pm 1, \dots, \pm 1)$ (Blum et al., 2020). By constraint (g), $S = I$, and hence, $Q_1 = Q_2$ and $R_1 = R_2$. Therefore, by Equation 2, Q , Λ , and R are uniquely determined by $\mathbb{E}Y$ for any given X , Z , M , and N .

S3 Parameter count for B in the GBM and RR-GBM

In this section, we count the number of free parameters in the B matrix under the assumed identifiability constraints in Section S2. We consider both the case of the full GBM and the RR-GBM.

S3.1 Full GBM setting

In the full GBM, the only constraint on B is $X^T B = 0$. Since $X \in \mathbb{R}^{I \times K}$ and $B \in \mathbb{R}^{I \times L}$, the constraint $X^T B = 0$ is equivalent to the requirement that $X^T b_l = 0$ for $l = 1, \dots, L$, where $b_1, \dots, b_L$ are the columns of B . The K columns of X are linearly independent since $X^T X$ is invertible by assumption. Thus, the space of solutions to $X^T b_l = 0$ has dimension $I - K$ by the rank-nullity theorem; in other words, b_l has $I - K$ free parameters. Thus, over all L columns of B , the total number of free parameters to be estimated, given the constraint $X^T B = 0$, is $(I - K)L$.

S3.2 RR-GBM setting

In the RR-GBM, B is parametrized as $B = Q\Lambda R^T$, where $Q \in \mathbb{R}^{I \times N}$, $\Lambda \in \mathbb{R}^{N \times N}$, $R \in \mathbb{R}^{L \times N}$, and Λ is diagonal with non-zero entries on the diagonal, and from Section S2, the assumed identifiability constraints are:

1. $Q^T Q = I$,
2. $R^T R = I$,
3. $\lambda_{11} > \lambda_{22} > \dots > \lambda_{NN} > 0$, and
4. $X^T Q \Lambda R^T = 0$.

We also assume $N < \min\{L, I\}$ and $K < I$, which are satisfied in all our applications of the model. To count the number of free parameters in B under these constraints, we first argue that given constraints 1-3, the fourth constraint ($X^T Q \Lambda R^T = 0$) is equivalent to $X^T Q = 0$. To see this, let $S = X^T Q \Lambda$ and note that the constraint becomes $S R^T = 0$. Since $R^T R = I$, the columns of R are linearly independent, and thus its null space is $\{0\}$. Hence, $S R^T = 0$ implies $S = 0$ since the left null space of R^T is same as the null space of R . Thus, since $R^T R = I$, the constraint $X^T Q \Lambda R^T = 0$ holds if and only if $X^T Q \Lambda = 0$. Furthermore, since Λ is a diagonal matrix of non-zero entries, $X^T Q \Lambda = 0$ if and only if $X^T Q = 0$. Therefore the constraints can equivalently be written:

1. $Q^T Q = I$,
2. $R^T R = I$,
3. $\lambda_{11} > \lambda_{22} > \dots > \lambda_{NN} > 0$, and

4. $X^T Q = 0$.

To count the free parameters, begin by noting that Λ has N free parameters. Next, the constraint that $R^T R = I$ is equivalent to $R \in V_N(\mathbb{R}^L)$, where $V_N(\mathbb{R}^L)$ denotes the Stiefel manifold represented by the set of all orthonormal N -frames in $\mathbb{R}^L$. Thus, R has $LN - N(N + 1)/2$ free parameters (Chakraborty and Vemuri, 2019). Finally, Q is subject to the constraints that $Q^T Q = I$ and $X^T Q = 0$. The $X^T Q = 0$ constraint forces Q to lie in the null space of X^T , which has dimension $I - K$ since X is rank K . Thus, these constraints force Q to be an orthonormal N -frame in an $I - K$ dimensional subspace, which is equivalent to $V_N(\mathbb{R}^{I-K})$, and this Stiefel manifold has dimension $(I - K)N - N(N + 1)/2$. Hence, Q has $(I - K)N - N(N + 1)/2$ free parameters.

Summing up the number of free parameters for Λ , R , and Q yields a total of

$$N + LN - \frac{N(N + 1)}{2} + (I - K)N - \frac{N(N + 1)}{2},$$

which simplifies to $N(I - K + L - N)$.

S3.3 Reduction in parameter count for RR-GBM relative to full GBM

Let p_{GBM} and p_{RRGBM} denote the number of parameters in B under the GBM and the RR-GBM, respectively. By Sections S3.1 and S3.2, the fraction of parameters under the RR-GBM relative to the GBM is

$$\frac{p_{\text{RRGBM}}}{p_{\text{GBM}}} = \frac{N(I - K + L - N)}{(I - K)L} = \frac{N}{L} + \frac{N}{I - K} - \frac{N}{L} \frac{N}{I - K}.$$

Meanwhile, the reduction in the number of parameters, represented as a fraction, is

$$1 - \frac{p_{\text{RRGBM}}}{p_{\text{GBM}}} = 1 - \frac{N}{L} - \frac{N}{I - K} + \frac{N}{L} \frac{N}{I - K} = \left(1 - \frac{N}{I - K}\right) \left(1 - \frac{N}{L}\right).$$

This expression highlights the reliance on both the ratio of N to L and also the ratio of N to $I - K$.

In particular, if $I - K \gg N$, then

$$\frac{p_{\text{RRGBM}}}{p_{\text{GBM}}} \approx \frac{N}{L} \quad \text{and} \quad 1 - \frac{p_{\text{RRGBM}}}{p_{\text{GBM}}} \approx 1 - \frac{N}{L}.$$

S4 Details of the simulation studies

In this section, we provide additional details on the simulation studies in Section 3.

S4.1 Simulation of covariates, true parameters, and data

The true parameters A_0 , C_0 , U_0 , Σ_0 , V_0 , S_0 , T_0 , and ω_0 as well as the covariate matrices X and Z are generated using the same procedure as described by section S2 in Miller and Carter (2020), where S_0 , T_0 , and ω_0 are only used in the Negative Binomial setting. The additional RR-GBM true parameters $Q_0 \in \mathbb{R}^{I \times N}$, $\Lambda_0 \in \mathbb{R}^{N \times N}$, and $R_0 \in \mathbb{R}^{L \times N}$ are simulated as follows, assuming $I > L$. We generate Q_0 and R_0 uniformly from the Stiefel manifolds $V_N(\mathbb{R}^I)$ and $V_N(\mathbb{R}^L)$, respectively. For instance, in the case of Q_0 , this is done by first generating $\tilde{Q} \in \mathbb{R}^{I \times N}$ with i.i.d. entries $\tilde{q}_{in} \sim \mathcal{N}(0, 1)$, and then computing the thin QR decomposition of $\tilde{Q}$ to obtain $\tilde{Q} = Q_0 Q_R$ where $Q_0 \in \mathbb{R}^{I \times N}$ and $Q_R \in \mathbb{R}^{N \times N}$. It can be shown that this Q_0 is uniformly sampled from the Stiefel manifold $V_N(\mathbb{R}^I)$ (Tropp, 2012). The diagonal entries of Λ_0 are chosen to be evenly spaced from $\sqrt{I} + \sqrt{L}$ to $2(\sqrt{I} + \sqrt{L})$, following the approach of Miller and Carter (2020), which is based on the distribution of singular values in random matrices (Marčenko and Pastur, 1967). To make the B_0 matrix have approximate rank N (but exact rank L), we multiply the last $L - N$ diagonal entries by a small scaling factor equal to 0.01 by default. (To make B_0 exactly reduced-rank, we use a multiplier of 0 here.) We then compute $\tilde{B} = Q_0 \Lambda_0 R_0^T$ and define the true B matrix to be $B_0 = \tilde{B} - X(X^+ \tilde{B})$, where X^+ is the Moore–Penrose pseudoinverse of X . This last projection step ensures that B_0 satisfies the constraint that $X^T B_0 = 0$, while preserving the reduced-rank or approximate reduced-rank property. We then apply the SVD to B_0 once more to get the final true values of Q_0 , Λ_0 , and R_0 .

Finally, the data matrix Y is then generated in the same way as described in section S2 in Miller and Carter (2020), using the log link function and either the Poisson or Negative Binomial distribution for sampling as indicated.

S4.2 Structure in the sample covariate effects matrix

Given a set of sample covariates Z , we can simulate any desired latent sample covariate effect structure through the controlled generation of R_0 , enabling the construction of an R_0 with any

desired column space. The different sample covariate effect structures were all simulated with the same underlying hyperparameters: $I = 5000$, $J = 2000$, $K = 1$, $L = 10$, and $M = 1$. The R_0 matrix was simulated to meet the identifiability constraints with the desired structure, denoted by $\tilde{R}$, a matrix of dimension $L \times P$ where P is the desired number of factors with $P \leq N$. It is assumed that the columns of $\tilde{R}$ are orthogonal, as if they were not the final R_0 would not lie in the Stiefel manifold. An example of $\tilde{R}$ could be a binary matrix encoding a desired sparsity pattern such as that found in Figure 5. The sampling process can be described as uniform sampling from the Stiefel manifold with the respective dimensions subject to the constraint that the column normalized $\tilde{R}$ is a submatrix of the final R_0 . The steps of the sampling procedure are as follows:

1. Set factor $\tilde{R} \in \mathbb{R}^{L \times P}$ to have the desired latent sample covariate structure.
2. $R_{*j} = \tilde{R}_{*j} / \|\tilde{R}_{*j}\|$ (column-wise normalization)
3. Simulate $G \in \mathbb{R}^{L \times (N-P)}$ where $G_{ij} \sim \mathcal{N}(0, 1)$.
4. $H = (\mathbf{I} - RR^T)G$
5. Take QR decomposition of H to obtain H_Q where $H = H_Q H_R$.
6. Create the simulated true R_0 matrix via the following construction $R_0 = \begin{bmatrix} R & H_Q \end{bmatrix}$ (or placing R in the desired column locations).

The resulting R_0 matrix has the constructed effect on the sample covariate effects as represented by $\tilde{R}$ since the remaining H_Q columns are the result of projecting Gaussian noise onto the orthogonal complement of the space spanned by $\tilde{R}$.

For the situation when $\tilde{r} = \tilde{R}$ is a vector, the same procedure holds and generates the desired R_0 .

S4.2.1 Structure in sample covariate effect matrices used

Figure 6(a) was generated using the following $\tilde{R}$:

$$\tilde{R} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \\ 1 & -1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & -1 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Figure 6(b) was generated using the following $\tilde{R}$; here, column 3 was obtained by applying Gram-Schmidt orthogonalization to the matrix $\tilde{R}_0$ shown on the left:

$$\tilde{R}_0 = \begin{bmatrix} 0.8 & 0 & 0 \\ -0.8 & 1 & 0 \\ 0.8 & 1 & 0 \\ -0.5 & 0 & 0 \\ 0.8 & 0 & 1 \\ 0.6 & 0 & 0 \\ -0.5 & 0 & 0 \\ 0.8 & 0 & 0 \\ 0.9 & 0 & 0 \\ -0.6 & 0 & 0 \end{bmatrix}, \quad \tilde{R} = \begin{bmatrix} 0.8 & 0 & -0.13 \\ -0.8 & 1 & 0.13 \\ 0.8 & 1 & -0.13 \\ -0.5 & 0 & 0.08 \\ 0.8 & 0 & 0.94 \\ 0.6 & 0 & -0.10 \\ -0.5 & 0 & 0.08 \\ 0.8 & 0 & -0.13 \\ 0.9 & 0 & -0.15 \\ -0.6 & 0 & 0.10 \end{bmatrix}.$$

S5 Supplemental figures

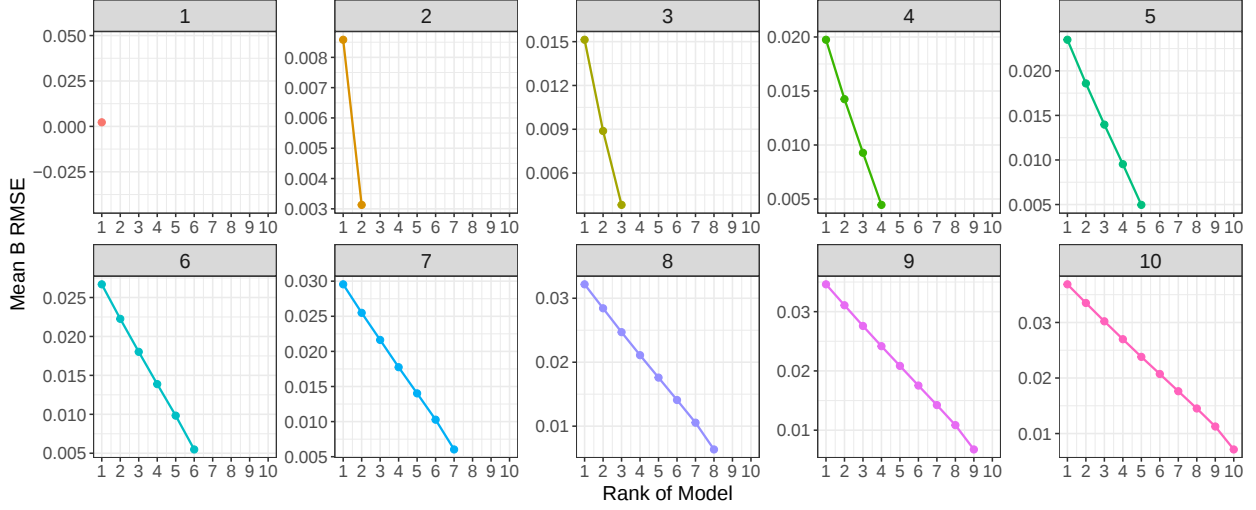


Figure S1: **RR-GBM estimation accuracy as a function of true rank and model rank.** Panels are labeled according to N_0 , the true rank of B used to simulate data. The x-axis is N , the rank of B used to fit the data. Each point is the average RMSE over 100 simulations, where for each simulation we compute the RMSE between the estimate $\hat{B} = \hat{Q}\hat{\Lambda}\hat{R}^T$ and the true matrix B_0 . Negative Binomial data was generated with 1 latent factor ($M = 1$). Each plot only goes up to $N = N_0$ since models with $N > N_0$ are degenerate and the estimation procedure does not converge.

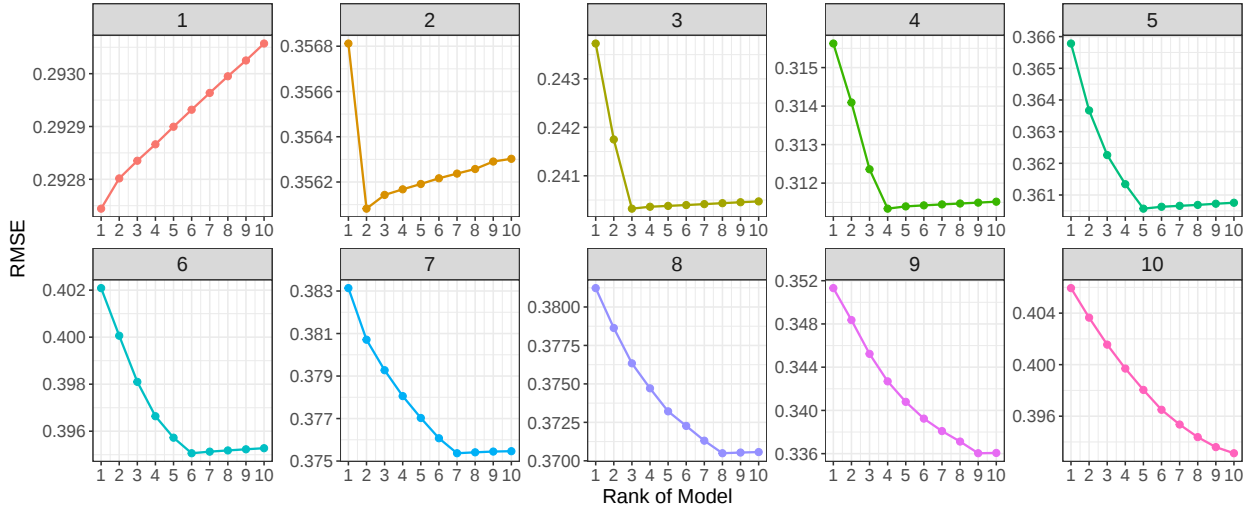


Figure S2: **Empirical selection of the rank in RR-GBM on a single replicate dataset.** Panels are labeled according to the true approximate rank of B , where the corresponding $L - N_0$ singular values are 0.01 times their randomly generated value to create a B with approximate rank N_0 . Each dot indicates the RMSE between the true values and predicted values from one simulation run. Poisson data was generated and fit using a model with one latent factor ($M = 1$).

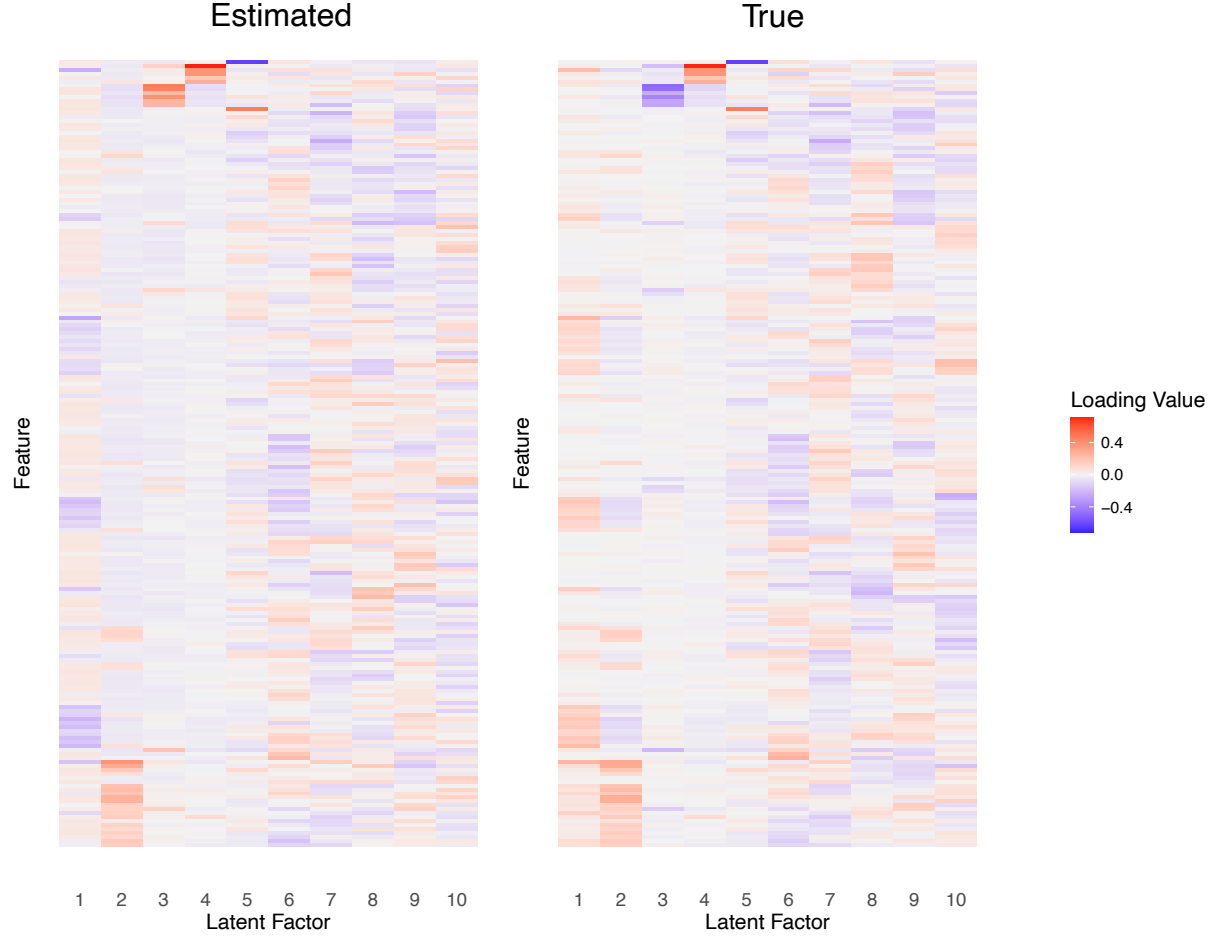


Figure S3: **Visualizing the effect of sample covariates on feature level latent factors in RR-GBM.** The estimate $\hat{Q}$ and true matrix Q_0 are shown. Data were simulated with 20,000 features, 200 samples, 1 latent factor, and 20 Bernoulli sample covariates with a Negative Binomial outcome distribution. The feature level coefficients for the five sample covariates 18, 11, 16, 3, 19, were altered at random with $\{1\%, 5\%, 10\%, 30\%, 50\%\}$ of feature coefficients being doubled, respectively. The true B_0 matrix was simulated as described in Section S4, before the selected coefficients were altered. The rows are hierarchically clustered in the true Q_0 and the respective ordering is applied to the estimated $\hat{Q}$ matrix. All 20,000 features are plotted, with runs of features with similar loadings being blocked together by the limited resolution of the figure.