Matrix Linear Models for High-Throughput Chemical Genetic Screens

Jane W. Liang,* Robert J. Nichols,† and Śaunak Sen‡,1

*Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, †Department of Microbiology and Immunology, University of California, San Francisco, California 94143, and ‡Department of Preventive Medicine, University of Tennessee Health Science Center, Memphis, Tennessee 38163

ORCID IDs: 0000-0002-2302-3809 (J.W.L.); 0000-0003-4519-6361 (Ś.S.)

ABSTRACT We develop a flexible and computationally efficient approach for analyzing high-throughput chemical genetic screens. In such screens, a library of genetic mutants is phenotyped in a large number of stresses. Typically, interactions between genes and stresses are detected by grouping the mutants and stresses into categories, and performing modified *t*-tests for each combination. This approach does not have a natural extension if mutants or stresses have quantitative or nonoverlapping annotations (*e.g.*, if conditions have doses or a mutant falls into more than one category simultaneously). We develop a matrix linear model (MLM) framework that allows us to model relationships between mutants and conditions in a simple, yet flexible, multivariate framework. It encodes both categorical and continuous relationships to enhance detection of associations. We develop a fast estimation algorithm that takes advantage of the structure of MLMs. We evaluate our method's performance in simulations and in an *Escherichia coli* chemical genetic screen, comparing it with an existing univariate approach based on modified *t*-tests. We show that MLMs perform slightly better than the univariate approach when mutants and conditions are classified in nonoverlapping categories, and substantially better when conditions can be ordered in dosage categories. Therefore, it is an attractive alternative to current methods, and provides a computationally scalable framework for larger and complex chemical genetic screens. A Julia language implementation of MLMs and the code used for this paper are available at https://github.com/janewliang/GeneticScreen.jl and https://bitbucket.org/jwliang/mlm_gs_supplement, respectively.

KEYWORDS chemical genetic screens; *E. coli*; linear models; high-throughput data

HIGH-THROUGHPUT genetic screens have revolutionized biology by their ability to answer complex, large-scale scientific questions. This is made possible by advances in automation and multiplexing, the availability of large and comprehensive collections (such as mutant libraries and sequenced genomes), and advances in computational and statistical methodology. Here, we consider a high-throughput genetic screen to observe the fitness of a library of mutants in a variety of growth conditions. Potential goals of such a screen would be to analyze gene × condition interactions or to predict the effect of a new, but related, antibiotic. Matching genes with phenotypes is a particularly valuable application of high-throughput experiments in the age of rapid sequencing technology (van Opijnen and Camilli 2012). These techniques can shed light on the physiological roles of partially redundant gene functions (Typas *et al.* 2011) and the physiological pathways that are involved with responses to different environmental factors (Ivask *et al.* 2013). They can reveal relationships between unknown or seemingly unrelated genes (Oh *et al.* 2011), and provide insights into genes involved in multiple antibiotic resistance (Nichols *et al.* 2011).

It is now possible to run high-throughput genetic screens cost-effectively in bulk, but the unprecedented scale of these types of studies necessitates the development of generalizable and efficient methods for analyzing their results. Such studies are essentially multivariate problems, but most traditional methods turn them into several univariate problems for computational feasibility. Doing so fails to take advantage of

doi: https://doi.org/10.1534/genetics.119.302299
Manuscript received May 7, 2019; accepted for publication June 6, 2019; published Early Online June 26, 2019.
Supplemental material available at FigShare: https://doi.org/10.25386/genetics.8316881.
[1]Corresponding author: Department of Preventive Medicine, University of Tennessee Health Science Center, 66 N. Pauline St., Suite 656, Memphis, TN 38163. E-mail: sen@uthsc.edu

known groupings and correlations in the observations. In the case of the genetic screening example, one can group mutants by gene or gene family, group growth conditions by antibiotic class or temperature, and consider spatial correlation on plates.

We present matrix linear models (MLMs), which provide a formal statistical framework for encoding such known, but perhaps nonexplicit, underlying relationships to enhance detection of associations that might otherwise be masked. This straightforward, multivariate approach can take into account any number of continuous or categorical covariates. Existing methods can encode for different mutant strains and condition types, but are univariate; these approaches are akin to ANOVA or Student's *t*-tests. In addition to categorical groupings, MLMs have the distinct advantage of being able to explicitly model more complex and potentially noncategorical information, such as dosage-response levels, null/heterozygous/homozygous genotypes, and spatial correlation on the plates (colonies located on the edge of the plate are expected to exhibit greater growth, since they have fewer neighbors with whom to compete for resources). MLMs offer flexibility over existing methods analogous to what linear regression offers over Student's *t*-tests.

Estimation of MLMs is fast even in moderately large dimensions. Using simulations and data from an *Escherichia coli* genetic screen (Nichols *et al.* 2011), we show that our method produces results comparable to those of the univariate *S* score approach (Collins *et al.* 2006), but with considerably more efficient computation time. We also analyze the data while encoding for dosage response of growth conditions, to demonstrate the method's ability to incorporate information from continuous covariates and assess relationships more generally.

This paper is organized as follows. The *Materials and Methods* introduces a motivating *E. coli* chemical genetic screen data set, and describes the statistical model and estimation (*Statistical analysis*). The *Results* section evaluates our method using simulated and real data. We conclude with a *Discussion*.

Materials and Methods

Colony opacity was recorded for mutant strains grown at high density on agar plates with a range of conditions (Nichols *et al.* 2011; Shiver *et al.* 2016). Six "plate arrangements" of mutants were used, with 1536 colonies grown per plate. In this context, a plate arrangement refers to the choice of mutants and exposures, as well as their positioning in the 1536 wells. The 3983 mutant strains were taken from the Keio single-gene deletion library (Baba *et al.* 2006), essential gene hypomorphs [C-terminally tandem-affinity tagged (Butland *et al.* 2008) or specific alleles], and a small RNA/small protein knockout library (Hobbs *et al.* 2010). The colonies were grown in 307 conditions representing different *E. coli* stresses. More than half were antibiotic/antimicrobial treatments, but other types of conditions, such as temperature and pH, were included. Among the six plate arrangements, a total of 982,902 condition × gene interactions need to be estimated, along with main effects across interactions for the growth conditions and mutant strains. The study aimed to examine the interaction effects between the *E. coli* genes and the growth conditions. This information can be used to study potential drugs with unknown targets and the mechanism behind drug interactions, as well as identify the genes necessary to support growth in different conditions (Nichols *et al.* 2011).

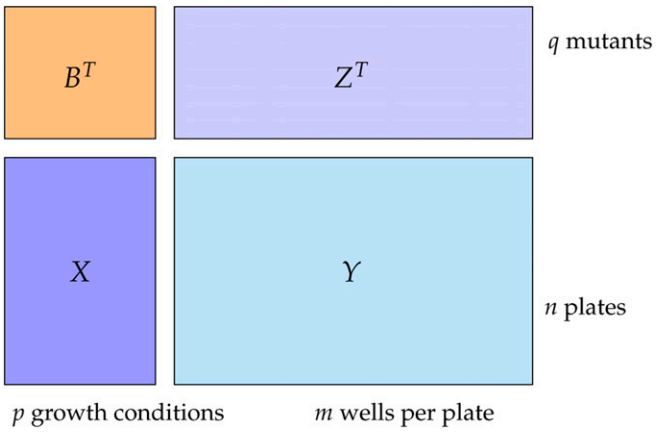

Figure 1 A visualization of the response, covariate, and coefficient matrices for a matrix linear model. The dimensions shown are for illustration only and are not necessarily to scale.

Statistical analysis

Model: Let Y be an $n \times m$ matrix of quantitative colony growth from a high-throughput genetic screen, similar to the one described in the previous section. The growth conditions are annotated by $X_{n\times p}$ along the n rows and the different mutant strains are annotated by $Z_{m\times q}$ along the m columns (Figure 1). MLMs are thus given by

$$Y = XBZ^T + E \tag{1}$$

with the main and interaction effects contained in $B_{p\times q}$ and errors in $E_{n\times m}$. The statistical form of the model is similar to that used by Xiong *et al.* (2011) for genetic analysis of function-valued phenotypes.

Let $\otimes$ denote the Kronecker product and vec be the vectorization operator that stacks columns of a matrix into a single-column vector. The vectorized equivalent of the problem is

$$\mathrm{vec}(Y^T) = (X \otimes Z) \cdot \mathrm{vec}(B^T) + \mathrm{vec}(E^T). \tag{2}$$

This resembles the familiar linear regression model of $y = \mathbf{X}\beta + \epsilon$, and so MLMs can theoretically be analyzed using existing software for least squares linear regression methods. Unfortunately, doing so is frequently computationally inefficient or even infeasible when the Kronecker product is too large for memory. We leverage the fact that all of the information in the design matrix is contained in two smaller

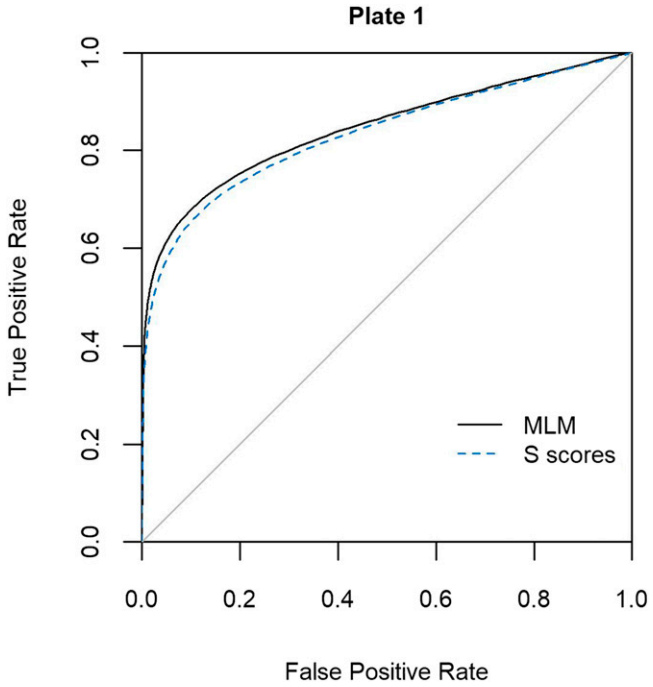

Figure 2 ROC curves comparing MLM to *S* scores applied to data simulated using a framework of first plate arrangement. Interactions corresponding to adaptive Benjamini–Hochberg adjusted *P*-values (Benjamini and Hochberg 2000; MuToss Coding Team *et al.* 2017) below a given cutoff are considered significant; these results are compared to the simulated interactions. Curves were generated by plotting the true-positive rates and false-positive rates at varying *P*-value cutoffs. The two methods perform very similarly, with MLM [AUC of 0.845 (Ekstrøm 2018)] performing slightly better than the *S* scores (AUC of 0.833). MLM, matrix linear model; ROC, receiver operating characteristic.

matrices. By doing so, we avoid the cumbersome Kronecker product and solve two smaller least squares problems. From a philosophical standpoint, vectorizing the data destroys its natural structure, reducing the interpretability of the results.

If the n plates, each exhibiting a certain growth condition, are independent with a common covariance matrix, then $\text{var}(\text{vec}(E^T)) = I_n \otimes \Sigma$. The residual covariance matrix Σ is generally unknown and must be estimated using the data.

Estimation: To obtain least squares estimates, we choose B to minimize the residual sum of squares,

$$\text{RSS}(B) = \text{vec}\left(Y^T - f(X,Z,B)^T\right)^T \text{vec}\left(Y^T - f(X,Z,B)^T\right) \tag{3}$$

where Y is the observed response matrix of colony growth and $f(X,Z,B) = XBZ^T$ is the linear model function, taking X and Z to be fixed. That is, we want to find B such that f is the best linear combination of X and Z for approximating Y.

This approach may be viewed as a generalized estimation equations approach (Xiong *et al.* 2011). The solution has a closed form,

Table 1 Area under the curve for simulations based on each of the six plate arrangements

Plate	1	2	3	4	5	6
MLM	0.845	0.843	0.853	0.852	0.847	0.853
S scores	0.833	0.838	0.852	0.850	0.840	0.852

This is computed directly from the ROC curves, as in Figure 2 (Ekstrøm 2018). Matrix linear models outperform Collins's *S* scores slightly but consistently in each of the six cases. MLM, matrix linear model.

$$\hat{B} = \left(X^T X\right)^{-1} X^T Y Z \left(Z^T Z\right)^{-1}, \tag{4}$$

which can be computed directly without the use of iterative solvers or specifying decision variables. The solution can be viewed as a combination of a least squares on the mutant strains (the Z part) and another on the conditions (the X part). The resulting estimate is asymptotically unbiased (Liang and Zeger 1986).

The variance–covariance matrix of the estimated coefficient is

$$\tau = \text{var}\left[\text{vec}\left(\hat{B}^T\right)\right] = \left(X^T X\right)^{-1} \otimes \left(\left(Z^T Z\right)^{-1} Z^T \Sigma Z \left(Z^T Z\right)^{-1}\right). \tag{5}$$

If a consistent estimate of the residual covariance matrix Σ can be obtained, so can a consistent estimate of the variance of the coefficient estimates. Note that the choice of an estimator of Σ (made separately from the least squares minimization problem) will affect the formulation of τ, but not of B (see Appendix for the derivation of the least squares estimate and variance).

The fitted values are easily computed as

$$\hat{Y} = X\hat{B}Z^T = X\left(X^T X\right)^{-1} XYZ\left(Z^T Z\right)^{-1} Z^T. \tag{6}$$

Testing: Similar to the *t*-test statistic used to assess the coefficients from the univariate linear regression model, we can define a test statistic for our method as

$$\text{vec}\left(t^T\right) = \frac{\text{vec}\left(\hat{B}^T\right)}{\sqrt{\text{diag}(\hat{\tau})}}, \tag{7}$$

where $\text{diag}(\hat{\tau})$ is the vector of the diagonal entries of $\hat{\tau}$, the estimator of the variance–covariance matrix of the estimated coefficient τ as defined in the previous subsection. Obtaining $\hat{\tau}$ requires specifying an estimate of the residual covariance matrix, which we took to be the consistent $\hat{\Sigma} = \frac{1}{n}(Y - \hat{Y})^T (Y - \hat{Y})$. Both the division and square root functions in the test statistic formula are element-wise operators.

Based on simulations and real data, we empirically observed that the test statistics approximate a skewed t distribution. Rather than attempting to specify appropriate parameters for a skewed t distribution, we used permutation tests to obtain *P*-values for our analysis. At each permutation, we randomly permuted the rows of the Y matrix, *i.e.*, the

Table 2 False-positive rate for MLMs, calculated as the proportion of *P*-values below cutoffs of 0.01, 0.05, and 0.1, based on 100 simulated data sets

Plate	1	2	3	4	5	6
0.01	0.010 (0.009, 0.011)	0.010 (0.009, 0.011)	0.010 (0.009, 0.010)	0.010 (0.010, 0.010)	0.010 (0.009, 0.011)	0.010 (0.009, 0.011)
0.05	0.050 (0.049, 0.051)	0.050 (0.049, 0.051)	0.050 (0.049, 0.051)	0.050 (0.049, 0.051)	0.050 (0.049, 0.051)	0.050 (0.049, 0.051)
0.1	0.100 (0.098, 0.102)	0.100 (0.098, 0.102)	0.100 (0.099, 0.101)	0.100 (0.099, 0.101)	0.100 (0.098, 0.102)	0.100 (0.098, 0.101)

Our method appears to preserve type I error.

independent units were taken to be the plates. A null distribution was constructed from the test statistics estimated using the permuted data. *P*-values for each coefficient were computed by comparing the observed test statistics to this null distribution.

MLMs handle only complete data, so all missing values should be appropriately dropped, smoothed, or estimated beforehand. No such considerations were needed for the *E. coli* data set, which had no missing values. Weighted least squares can be used for heteroskedastic data by weighing plates and/or colonies (see Appendix for derivation).

Data availability

We implemented our method using the Julia programming language (Bezanson *et al.* 2017); it is available as a package at https://github.com/janewliang/GeneticScreen.jl. Data preprocessing and visualization was done in R (R Core Team 2018), due to its robust ecosystem of packages for statistics and data analysis (Wickham 2007; Dowle and Srinivasan 2018). Code for data analysis and generating all figures in this paper is available at https://bitbucket.org/jwliang/mlm_gs_supplement. Supplemental material available at FigShare: https://doi.org/10.25386/genetics.8316881.

Results

We applied our method to simulated data and *E. coli* genetic screening data (Nichols *et al.* 2011). We compared the results and computation times for MLMs, and the *S* score, a popular existing method for analyzing high-throughput genetic screening data (Collins *et al.* 2006). An *S* score is essentially a Student's *t*-test statistic comparing the observations for a given mutant and condition with the observations for a given mutant over all conditions. It is given as

$$S = \frac{\mu_{\text{Exp}} - \mu_{\text{Cont}}}{\sqrt{s_{\text{Var}}/n_{\text{Exp}} + s_{\text{Var}}/n_{\text{Cont}}}} \tag{8}$$

where

$$s_{\text{Var}} = \frac{\text{var}_{\text{Exp}} \times (n_{\text{Exp}} - 1) + \text{var}_{\text{Cont}} \times (n_{\text{Cont}} - 1)}{n_{\text{Exp}} + n_{\text{Cont}} - 2}. \tag{9}$$

μ_{Exp}, var_{Exp}, and n_{Exp} are the mean, variance, and number of measurements for normalized colony growth for a mutant *and* a condition of interest. μ_{Cont}, var_{Cont}, and n_{Cont} are the median, variance, and median number of measurements for normalized colony growth for a mutant of interest *over all conditions*. Both var_{Exp} and var_{Cont} are subject to a minimum bound on the SD given by the expected SD in normalized colony growth for a mutant with similar growth phenotypes.

Unlike MLMs, which only assume that the rows of Y are independent, *S* scores assume that both the rows (plates) and columns (colonies) are independent. So the method is expected to make improvements over *S* scores, especially when this assumption is violated, such as when the columns of Y (the colonies) are spatially correlated. MLMs go beyond the *S* score ANOVA-like approach and allow for encoding more complex categorical or continuous relationships, similar to linear regression.

Simulation studies

Using the framework of the X and Z matrices from the *E. coli* data's six plate arrangements, we simulated data with 1/2 nonzero main effects and 1/4 nonzero interactions drawn from a Normal(0, 4) distribution. The errors were independent and identically distributed from the standard normal distribution. We then applied both our multivariate method and the *S* score's univariate approach to estimate ~180,000 interactions. Using permutation tests, we obtained *P*-values corresponding to each interaction for each approach. We used the adaptive Benjamini–Hochberg adjustment (Benjamini and Hochberg 2000) (Team *et al.* 2017) to account for multiple testing. In the adaptive Benjamini–Hochberg step-up procedure, one controls for the false discovery rate by first estimating the number of "true" null hypotheses (Hochberg and Benjamini 1990) and then applying the Benjamini–Hochberg procedure (Benjamini and Hochberg 1995). For the likely scenario that some of the tested hypotheses are not true, the usual Benjamini–Hochberg adjustment may be underpowered. The initial step helps keep the results from being too conservative.

To compare the results for each plate arrangement, we plotted the receiver operating characteristic (ROC) curve generated by obtaining true-positive rates (TPRs) and false-positive rates (FPRs) at varying *P*-value cutoffs for both methods (Figure 2). The cutoffs were used to determine which adaptive Benjamini–Hochberg-adjusted *P*-values corresponded to significant (nonzero) interactions, and these results were then compared to the simulated interactions to obtain the varying TPRs and FPRs. Figure 2 is the ROC plot for the first plate arrangement. The gray reference line that cuts diagonally from the lower left to the upper right is what

we would expect the curve to look like for a method that just produces random noise. Its area under the curve (AUC) is 0.5. A method that performs well will have a curve that closely aligns with the upper left corner and an AUC approaching 1 (its TPR will be high even if its FPR is low for a given cutoff). See Figure S1 for ROC plots for the remaining five plates.

The AUCs for our method and Collins's method were 0.845 and 0.833, respectively (Ekstrøm 2018). Based on both the visual and quantitative summaries, we can observe that MLMs perform as least as well as the *S* scores. However, this slight positive difference is consistent across all six plate arrangements (Table 1).

Type I error: We simulated 100 data sets based on the six plate arrangements with 1/2 nonzero main effects and 1/4 nonzero interactions, as described in the section above. To approximate a null distribution, we permuted the rows of each simulated *Y*, ran MLMs, and obtained permutation *P*-values for the interactions. For each data set, we computed the FPR as the proportion of *P*-values below cutoffs of 0.01, 0.05, and 0.1. Table 2 displays the mean proportions across all 100 simulations with 95% normal approximation C.I.s calculated based on the SD across all 100 simulations. The mean proportions are consistently at or slightly below their corresponding thresholds, suggesting that our method preserves type I error.

Dosage-response simulation: A more interesting case that illustrates the flexibility and benefits of using MLMs is to consider a genetic screen whose plate conditions have multiple dosage levels, and for whom a monotonic dose response is expected. Suppose a given condition has three dosage levels. The *S* score approach will analyze these condition × gene interactions separately for each of the dosage levels, *i.e.*, ConditionLevel1 × gene, ConditionLevel2 × gene, and ConditionLevel3 × gene. One can analyze the data analogously using MLMs by encoding each of the condition–dosage combinations as separate dummy variables.

However, it is also possible for our method to encode this information as dosage-response levels for a given condition. Instead of treating this hypothetical condition as essentially three unrelated conditions, we can encode the three dosage-response levels together as a single variable corresponding to the condition. The simplest way to do this is to assign a different "slope" to each level in a condition. If the magnitude of the effect of the hypothetical condition is expected to increase (either in a beneficial or harmful manner) with each dosage level, one could create a single-vector variable where ConditionLevel1 is encoded as 1, ConditionlLevel2 is encoded as 2, and ConditionLevel3 is encoded as 3.

To examine this scenario more closely, we used the *Z* matrix frameworks for mutant strains from each of the six plate arrangements of the *E. coli* data. For each plate, we simulated effects for an experiment with 10 different conditions, each with three dosage levels and three replicates. First, 1/4 nonzero interactions were drawn from a Normal(0, 1/4) distribution. For these 1/4 nonzero interactions, we further simulated monotonic dosage effects. We did this by first randomly selecting a direction for the condition's effect (positive or negative with equal probability). Then, for the first dosage level, we simulated an effect from an Exponential(0.5) distribution. For the second dosage level, we took the effect from the first dosage level and added it to a random effect drawn from an Exponential(0.5 $\alpha = 0.8$) distribution. For the third dosage level, we summed the second dosage level's effect with a random effect drawn from Exponential(0.5 α^2). The dosage effects for a given condition were then assigned the appropriate direction that was randomly selected in the first step. This simulation can be extended for any number of conditions with any number of monotonic dosage levels. Simulating monotonic effects is reasonable based on the study design of the experimental data. The chemicals and their concentrations were chosen to be in a range where monotonic concentration-dependent effects are observed.

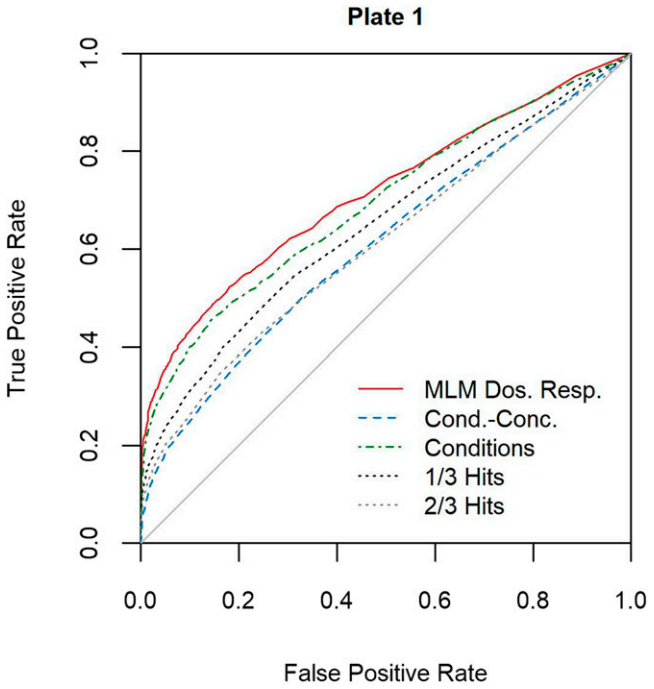

Figure 3 ROC curves for plate 1 simulations comparing dosage-response-encoded MLMs to categorically encoded *S* scores. Dosage-response-encoded MLMs outperform all other methods shown. These include *S* scores for data encoded with categorical condition–dosage combinations, for any of the three representations (Cond.-Conc, 1/3 Hits, and 2/3 Hits), as well as *S* scores for data encoded with just the conditions. Cond.-Conc.; condition concentration; Dos. Resp., dosage response; MLM, matrix linear model; ROC, receiver operating characteristic.

The ROC curves in Figure 3 illustrate the performance of the dosage-response-encoded approach compared with Collins's *S* scores, which can only encode categorical information, for the first plate arrangement. For each method, the ROC curve was generated by obtaining adaptive Benjamini–Hochberg-adjusted *P*-values (Benjamini and Hochberg 2000) (Team *et al.* 2017) from permutations and varying the cutoff

Table 3 AUCs for plate 1 ROC curves

MLM	Cond.-conc.	Conditions	1/3 hits	2/3 hits
0.714	0.612	0.694	0.650	0.612

Dosage-response-encoded matrix linear models outperform all other methods shown (Figure 3). These include *S* scores for data encoded with categorical condition–dosage combinations, for any of the three representations (Cond.-Conc, 1/3 Hits, and 2/3 Hits), as well as *S* scores for data encoded with just the conditions. Cond.-conc.; condition concentration; MLM, matrix linear model.

Table 4 Computation time in seconds to estimate condition × gene interactions, plus main effects, for each plate (a total of ~1 million interactions)

Plate	1	2	3	4	5	6	Total
MLM	0.52	0.47	0.67	0.61	0.52	0.50	3.29
S scores	22.81	20.83	30.52	29.94	23.82	24.08	151.99

Matrix linear models are considerably less computationally expensive. MLM, matrix linear model.

for determining significant interactions. These were then compared to the true simulated interaction effects to calculate the TPR and FPR.

The red solid line plots the results for MLMs with dosage-response encoding of the conditions.

The blue dashed line plots the results for Collins's *S* scores with categorical encoding of condition–dosage combinations, which is the conventional approach.

The green dot-dashed line plots the results for Collins's *S* scores with categorical encoding of only conditions, *i.e.*, all the dosage levels for a given condition are encoded as one categorical condition.

The black and gray dotted lines offer an alternate visualization of the Collins's *S* scores results, encoded for categorical condition–dosage combinations. For a given condition × gene interaction, there are three corresponding *S* scores/adjusted *P*-values (one for each condition–dosage level). When plotting the black dotted "1/3 Hits" line, a true positive is counted when at least one out of the three adjusted *P*-values for a significant simulated condition × gene interaction is below the cutoff. A false positive is counted when at least one out of the three adjusted *P*-values for a nonsignificant interaction is below the cutoff. The gray dotted "2/3 Hits" line similarly requires at least two out of the three adjusted *P*-values to be below the cutoff.

The corresponding AUCs (Ekstrøm 2018) for the plate 1 simulation are shown in Table 3.

Our proposed dosage-response MLMs approach (red solid line) outperforms Collins's *S* scores encoded with categorical condition–dosage combinations, regardless of representation (blue dashed, black dotted, and gray dotted lines). It also outperforms Collins's *S* scores when they only encode for the conditions without regard for dosage level (green dot-dashed line). These trends are consistent across simulations based on the other plates' arrangements (Figure S2 and Table S1). From an interpretation standpoint, this encoding can be useful if the investigator is interested in the overall effect of a plate condition as opposed to separately considering the different dosage levels. The more general approach to encoding covariates in MLMs leads to superior analysis in this situation.

Data analysis

We then applied the our method to each of the six plate arrangements in the *E. coli* genetic screen. The colony opacities were standardized by subtracting the median colony opacity of each plate (which has multiple mutant strains growing under a given condition) and dividing by the interquartile range.

Computational considerations: When we ran our MLM and *S* score implementations, encoded with condition–dosage combinations, on the entire data set of six plates, the latter required significantly more computation time (Table 4 and Table 5). A computer with 128 GB memory and a 3.00 GHz dual-core processor was used to obtain the times as averages of 10 runs. MLMs only take ~3.5 sec to estimate the roughly 1 million interactions, plus main effects (Table 4). In comparison, Collins's *S* scores require ~2.5 min. However, much greater computation time is needed to obtain permutation *P*-values. If the *P*-values are calculated based on 1000 permutations (parallelized over five cores), MLMs take just over 18 min (Table 5). *S* scores require over 8 hr to complete the same procedure (again, parallelized over five cores). This dramatic difference in computation time can have a considerable impact on the scope and feasibility of analyzing such data sets.

Auxotroph analysis: To assess whether our method identifies significant interactions in the expected manner, we analyzed auxotrophs. Auxotrophs are mutant strains that have lost the ability to synthesize a particular nutrient required for growth. These might include knockout strains for a certain amino acid. Since they should experience little to no colony growth under specific conditions where the required nutrient is not present, we expect negative interactions between auxotrophic mutants and minimal media growth conditions. Auxotrophs are useful as controls, since the phenotype under particular conditions for a mutant strain is typically not known.

In the original univariate analysis of the colony size data, Nichols *et al.* empirically identified 102 auxotrophs (Nichols *et al.* 2011). Likewise, a previous study of the Keio Collection auxotrophs, based on colony size, found 238 auxotrophs, 110 of which were mutants included in this data set (Joyce *et al.* 2006). Nichols *et al.* and Joyce *et al.* overlapped by 70%, despite significant experimental differences (*e.g.*, growth in liquid *vs.* solid media).

In a similar fashion, we empirically identified auxotrophs based on the MLM estimates. We did this by obtaining the quantiles of the MLM interaction scores for each mutant strain under minimal media conditions. Mutants whose 95% quantile for interaction scores with minimal media conditions fell below zero were classified as auxotrophs. Our auxotrophs had

Table 5 Computation time in minutes to estimate permutation *P*-values for the condition × gene interactions for each plate (a total of ~1 million interactions)

Plate	1	2	3	4	5	6	Total
MLM	2.82	2.78	3.46	3.28	2.82	3.00	18.16
S scores	77.62	72.59	93.37	94.22	76.31	77.32	491.42

The computational advantages of matrix linear models are even more apparent when permutation *P*-values are desired. MLM, matrix linear model.

an 83% overlap with the Nichols *et al.* auxotrophs and a 72% overlap with those of Joyce *et al.* The slightly larger intersection of auxotrophs when comparing with Nichols *et al. vs.* comparing with Joyce *et al.* is to be expected, since we are analyzing the same data set. While not all of the MLM interactions between the Nichols *et al.* and Joyce *et al.* auxotrophs and minimal media conditions were negative, the vast majority of them were. Figure 4 and Figure S3 are visualizations of the distributions of each auxotroph's interaction scores across minimal media conditions. The interaction scores are plotted as points, and the median for each auxotroph is plotted as a horizontal bar; most fall below zero.

Some of the discrepancy may be due to differences between analyzing colony opacity, as we did, and analyzing colony size, as Nichols *et al.* and Joyce *et al.* did. Kritikos *et al.* ran a study of the Keio collection using colony opacity (Kritikos *et al.* 2017). We used the publicly available Kritikos *et al. S* scores to replicate the auxotroph-identifying process for MLM *t*-test statistics. Of the 19 Nichols *et al.* auxotrophs that MLM was not able to identify, the Kritikos *et al. S* scores were unable to detect 12. This result suggests that about two-thirds of the Nichols *et al.* auxotrophs that MLM was unable to detect can be accounted for by differences in the types of measurements used to quantify growth.

Among the remaining 5 out of 19 auxotrophs, fepC is a mutant that results in the loss of ferric enterobactin uptake. There are minimal media conditions for both "high iron" and "low iron." The MLM *t*-test statistics are generally positive for interactions between fepC and both iron growth conditions; the Kritikos *et al. S* scores are positive only for high iron. Thus, fepC might be considered a borderline case in auxotroph determination depending on whether or not it exhibits growth under the low-iron condition. Additionally, there were two auxotrophic mutants that were found through analysis of the MLM *t*-test statistics, but not through analysis of the Kritikos *et al. S* scores.

Out of the 31 Joyce *et al.* auxotrophs that the MLM *t*-test statistics were unable to find, 22 were also undetectable to the Kritikos *et al. S* scores. Once again, this suggests that about two-thirds of the discrepant auxotrophs can likely be explained by differences between analyzing colony size and colony opacity (Kritikos *et al.* 2017).

ROC plots (Figure 5 and Figure S4) can assess the ability of MLMs to correctly identify auxotrophs found by Nichols *et al.* and Joyce *et al.* To get the TPRs and FPRs for Figure 5, we took the auxotrophs identified by Nichols *et al.* to be the true auxotrophs. We then obtained TPRs and FPRs by varying

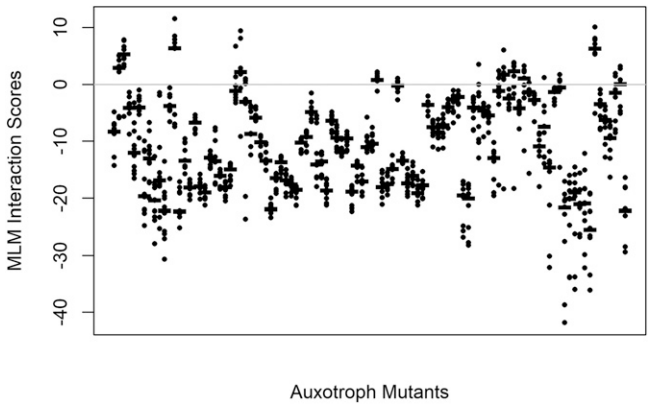

Figure 4 Distributions of MLM interaction estimates for auxotrophs identified by Nichols *et al.* (2011) over minimal media conditions. The Nichols *et al.* auxotrophs are plotted along the *x*-axis. The MLM interactions between the auxotrophs and minimal media conditions are plotted along the *y*-axis, with the horizontal bars indicating the median value. Most interactions fall below zero, indicating little growth. MLM, matrix linear model.

cutoffs for the median minimal media interaction score for the auxotrophs that we identified. Figure S4 was obtained analogously for the Joyce *et al.* auxotrophs. The AUCs were 0.884 and 0.824, respectively (Ekstrøm 2018); there is high concordance between the three sets of auxotrophs.

Dosage-response analysis

We also analyzed the data set by running MLMs with dosage-response levels encoded in the *X* matrix growth conditions. It should be noted that many of the conditions in the Nichols *et al.* data set had only one dosage level, which makes it somewhat less-than-ideal for illustrating this encoding approach. We then compared the dosage-response results with the results from applying MLMs and Collins's *S* scores on data encoded for condition–dosage combinations.

Figure 6 examines the performance of these three approaches for analyzing the first plate arrangement. The other five plate arrangements, shown in Figure S5, produced similar results. For each of the three methods, we used permutation tests to obtain adaptive Benjamini–Hochberg-adjusted *P*-values (Benjamini and Hochberg 2000) (Team *et al.* 2017). We then plotted the proportion of adjusted *P*-values below varying thresholds to generate the three curves. The dosage-response-encoded MLMs were able to detect more significant interactions (adjusted *P*-values below a given cutoff) than Collins' method at nearly every threshold. For lower cutoffs, this continuous encoding strategy also detects more significant interactions than MLMs with categorical encoding.

This simple example illustrates the potential gains in performance, and detection of significant interactions, when taking advantage of MLMs' regression-like ability to encode for complex and continuous covariates. Encoding the conditions in this manner also provides interpretable results about the effects of dosage in the interactions as well as the effects of the conditions themselves.

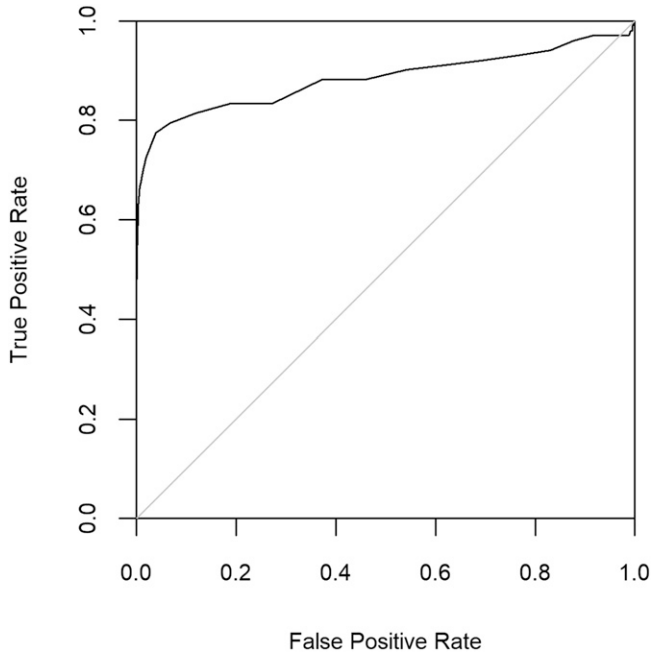

Figure 5 ROC curve for the auxotrophs we empirically identified, compared against those identified by Nichols *et al.* (2011) as the reference. True-positive rates and false-positive rates were calculated based on the median minimal media interaction score for the each of the auxotrophs we identified, at varying cutoffs. The AUC was 0.884 (Ekstrøm 2018). ROC, receiver operating characteristic.

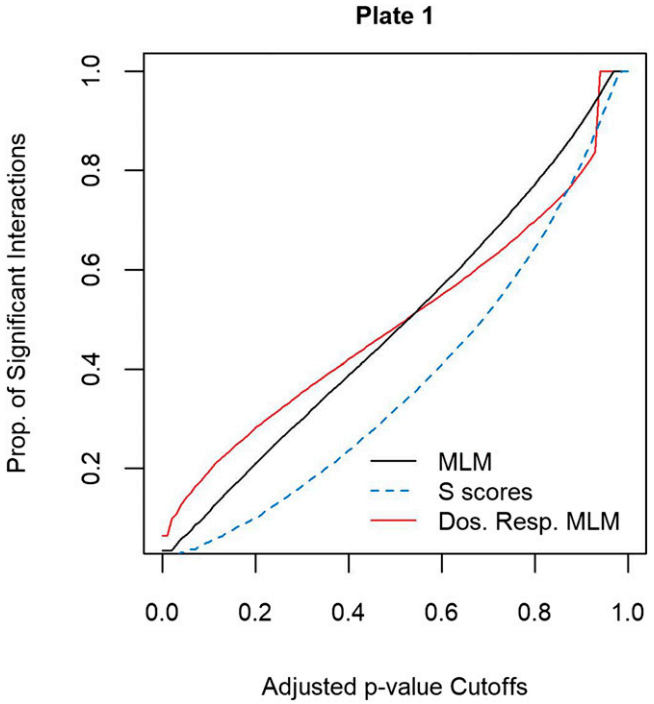

Figure 6 Comparing the proportion of significant interactions detected when encoding for dosage response and when encoding dosage–condition combinations categorically, for the first plate arrangement. The former is only possible in MLMs. The latter is shown for both MLMs and *S* scores. The curves were generated by obtaining the adaptive Benjamini–Hochberg-adjusted *P*-values from permutation tests for each method and identifying the proportion of adjusted *P*-values below varying cutoffs. Dos. Resp., dosage response; MLM, matrix linear model.

Discussion

We have presented MLMs, a simple framework for encoding relationships and groupings in high-throughput genetic screens. This approach is computationally efficient, running faster than the univariate *S* score method (Collins *et al.* 2006). The speed advantage is especially noticeable when for large data sets and when doing permutation tests. We gain this speed by harnessing both the matrix structure of high-throughput data and the strengths of least squares estimation.

MLMs can also improve detection of interaction effects that might otherwise be masked. Comparing our method to the *S* score approach in simulations and an *E. coli* genetic screen (Nichols *et al.* 2011), we show that we achieve comparable results at much less computational expense. Furthermore, unlike the *S* score, MLMs are not limited to encoding categorical groupings. In this way, the relationship between *S* scores and MLMs is analogous to that between ANOVA/Student's *t*-tests and linear regression. Analysis of simulated and *E. coli* data demonstrates that MLMs provide a more flexible and powerful approach for scenarios with noncategorical covariates, such as multidose conditions.

In this paper, we utilized a generalized estimating equation approach using least squares as the computational engine. Several extensions are possible. We may want a robust estimation procedure that downweights extreme observations (least squares is well known to be sensitive to outliers). This can be achieved by modifying the loss function from a sum of squares to a robustified version, such as Huber's loss function (Hastie *et al.* 2009). The resulting optimization problem is expected to be more complex. Another extension might be to fit penalized MLMs with an L_1 penalty. This optimization problem is also challenging, and we expect to report progress in future work. Finally, although we were motivated by high-throughput chemical genetic screens, many other high-throughput data, such as metabolomic data or cancer cell line drug screens, have a similar structure and might benefit from a similar approach.

Acknowledgments

We thank Carol Gross of the University of California, San Francisco (UCSF) for permission to use and share the chemical genetic screen data, and Anthony Shiver of Stanford University for help with accessing, processing, and interpreting the data. This work started when J.W.L. was a summer intern at UCSF, and continued when she was a scientific programmer at the University of Tennessee Health Science Center (UTHSC). We thank both UCSF and UTHSC for funding, and a supportive environment. S.S. was partly supported by National Institutes of Health grants GM-070683, GM-078338, GM-123489, ES-022841, and DA-044223.

Literature Cited

Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura *et al.*, 2006 Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Syst. Biol. 2: 2006.0008.

Benjamini, Y., and Y. Hochberg, 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. B 57: 289–300.

Benjamini, Y., and Y. Hochberg, 2000 On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Educ. Behav. Stat. 25: 60–83. https://doi.org/10.3102/10769986025001060

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah, 2017 Julia: a fresh approach to numerical computing. SIAM Rev. 59: 65–98. https://doi.org/10.1137/141000671

Butland, G., M. Babu, J. J. Díaz-Mejía, F. Bohdana, S. Phanse *et al.*, 2008 eSGA: E. coli synthetic genetic array analysis. Nat. Methods 5: 789–795. https://doi.org/10.1038/nmeth.1239

Collins, S. R., M. Schuldiner, N. J. Krogan, and J. S. Weissman, 2006 A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. Genome Biol. 7: R63. https://doi.org/10.1186/gb-2006-7-7-r63

Dowle, M., and A. Srinivasan, 2018 Data.table: extension of 'data.frame'. R package version 1.11.8. https://cran.r-project.org/web/packages/data.table/index.html.

Ekstrøm, C. T., 2018 *MESS: miscellaneous esoteric statistical scripts*. R package version 0.5.2. https://rdrr.io/cran/MESS/.

Hastie, T., R. Tibshirani, and J. Friedman, 2009 *The Elements of Statistical Learning*, Ed. 2nd. Springer-Verlag, New York. https://doi.org/10.1007/978-0-387-84858-7

Hobbs, E. C., J. L. Astarita, and G. Storz, 2010 Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in Escherichia coli: analysis of a bar-coded mutant collection. J. Bacteriol. 192: 59–67. https://doi.org/10.1128/JB.00873-09

Hochberg, Y., and Y. Benjamini, 1990 More powerful procedures for multiple significance testing. Stat. Med. 9: 811–818. https://doi.org/10.1002/sim.4780090710

Ivask, A., A. ElBadawy, C. Kaweeteerawat, D. Boren, H. Fischer *et al.*, 2013 Toxicity mechanisms in Escherichia coli vary for silver nanoparticles and differ from ionic silver. ACS Nano 8: 374–386. https://doi.org/10.1021/nn4044047

Joyce, A. R., J. L. Reed, A. White, R. Edwards, A. Osterman *et al.*, 2006 Experimental and computational assessment of conditionally essential genes in Escherichia coli. J. Bacteriol. 188: 8259–8271. https://doi.org/10.1128/JB.00740-06

Kritikos, G., M. Banzhaf, L. Herrera-Dominguez, A. Koumoutsi, M. Wartel *et al.*, 2017 A tool named Iris for versatile high-throughput phenotyping in microorganisms. Nat. Microbiol. 2: 17014. https://doi.org/10.1038/nmicrobiol.2017.14

Liang, K.-Y., and S. L. Zeger, 1986 Longitudinal data analysis using generalized linear models. Biometrika 73: 13–22. https://doi.org/10.1093/biomet/73.1.13

Nichols, R. J., S. Sen, Y. J. Choo, P. Beltrao, M. Zietek *et al.*, 2011 Phenotypic landscape of a bacterial cell. Cell 144: 143–156. https://doi.org/10.1016/j.cell.2010.11.052

Oh, E., A. H. Becker, A. Sandikci, D. Huber, R. Chaba *et al.*, 2011 Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. Cell 147: 1295–1308. https://doi.org/10.1016/j.cell.2011.10.044

R Core Team, 2018 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Shiver, A. L., H. Osadnik, G. Kritikos, B. Li, N. Krogan *et al.*, 2016 A chemical-genomic screen of neglected antibiotics reveals illicit transport of kasugamycin and blasticidin S. PLoS Genet. 12: e1006124 [corrigenda: PLoS Genet. 13: e1006902 (2017)]. https://doi.org/10.1371/journal.pgen.1006124

MuToss Coding Team (Berlin 2010), G. Blanchard, T. Dickhaus, N. Hack, F. Konietschke, *et al.*, 2017 mutoss: unified multiple testing procedures. R package version 0.1–12. https://rdrr.io/cran/mutoss/.

Typas, A., M. Banzhaf, C. A. Gross, and W. Vollmer, 2011 From the regulation of peptidoglycan synthesis to bacterial growth and morphology. Nat. Rev. Microbiol. 10: 123–136. https://doi.org/10.1038/nrmicro2677

van Opijnen, T., and A. Camilli, 2012 A fine scale phenotype–genotype virulence map of a bacterial pathogen. Genome Res. 22: 2541–2551. https://doi.org/10.1101/gr.137430.112

Wickham, H., 2007 Reshaping data with the reshape package. J. Stat. Softw. 21: 1–20. https://doi.org/10.18637/jss.v021.i12

Xiong, H., E. H. Goulding, E. J. Carlson, L. H. Tecott, C. E. McCulloch *et al.*, 2011 A flexible estimating equations approach for mapping function-valued traits. Genetics 189: 305–316. https://doi.org/10.1534/genetics.111.129221

Communicating editor: M. Beaumont

Appendix

In this appendix, we provide details on the derivation of the estimation formulas introduced in the *Statistical analysis* section. We will use some well-known results from linear regression and matrix algebra, also stated below, to help the reader follow the main argument.

Recap: Kronecker Products

For matrices A, B, C, D, and X, the following identities hold.

$$(A \otimes B)\ (C \otimes D) = AC \otimes BD$$

$$(A \otimes B)^T = A^T \otimes B^T$$

$$\mathrm{vec}\left(AXB^T\right) = (B \otimes A)vec(X)$$

Recap: Linear Regression

Consider the linear model

$$y = \mathbf{X}\beta + \boldsymbol{\epsilon},$$

where $\mathrm{var}(\boldsymbol{\epsilon}) = \Lambda$. The least squares estimate of β minimizes

$$\mathrm{RSS}(\beta) = (y - \mathbf{X}\beta)^T (y - \mathbf{X}\beta),$$

and is

$$\hat{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T y = P_{\mathbf{X}}\, y,$$

where $P_{\mathbf{X}}$ is the projection matrix corresponding to $\mathbf{X}$.

The variance of this estimate is

$$\mathrm{var}\left(\hat{\beta}\right) = \mathrm{var}(P_{\mathbf{X}} y) = P_{\mathbf{X}}\mathrm{var}(y)\, P_{\mathbf{X}}^T = P_{\mathbf{X}}\Lambda P_{\mathbf{X}}^T = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\Lambda\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}.$$

If we substitute Λ in the above formula with a consistent estimator, then we get the so-called "sandwich estimator" of the variance.

Recap: Weighted Linear Regression

If we minimize a weighted least squares criterion, given a known (positive definite) weight matrix W, we minimize

$$\mathrm{RSS}_W\ (\beta) = (y - \mathbf{X}\beta)^T W^{-1} (y - \mathbf{X}\beta),$$

and the solution is

$$\hat{\beta}_W = \left(\mathbf{X}^T W^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T W^{-1} y = P_{\mathbf{X}}^W\, y,$$

where $P_{\mathbf{X}}^W$ is the weighted projection matrix of X with weights W. $P_{\mathbf{X}}^W$ reduces to $P_{\mathbf{X}}$ when the weight matrix is the identity matrix. The formula for the variance estimate is identical to the variance estimate for the least squares estimate, replacing the projection matrix ($P_{\mathbf{X}}$) with the weighted version ($P_{\mathbf{X}}^W$).

These results can be used to derive least squares and weighted least squares formulas for the MLM coefficient estimates, and their variances.

Vectorized Version of the Model

Applying the vectorization and transpose operators on Equation 1,

$$\mathrm{vec}(Y^T) = \mathrm{vec}\left((XBZ^T)^T\right) + \mathrm{vec}(E^T) = \mathrm{vec}(ZB^TX^T) + \mathrm{vec}(E^T) = (X \otimes Z)\mathrm{vec}(B^T) + \mathrm{vec}(E^T),$$

which gives us Equation 2. Thus, we can see that this specification is in the form of the familiar linear model $y = \mathbf{X}\beta + \epsilon$ with

$$y \equiv \mathrm{vec}(Y^T)$$

$$\mathbf{X} \equiv X \otimes Z$$

$$\beta \equiv \mathrm{vec}(B^T)$$

$$\epsilon \equiv \mathrm{vec}(E^T)$$

$$\Lambda \equiv I_n \otimes \Sigma.$$

Least Squares Estimate

With the above equivalence, we can prove the least squares estimate formula using substitution. The projection matrix for the design matrix corresponding to the vectorized model is

$$\begin{aligned} P_{X\otimes Z} &= \left((X\otimes Z)^T(X\otimes Z)\right)^{-1}(X\otimes Z)^T \\ &= \left((X^T\otimes Z^T)(X\otimes Z)\right)^{-1}(X\otimes Z)^T \\ &= \left((X^TX)\otimes(Z^TZ)\right)^{-1}(X\otimes Z)^T \\ &= \left((X^TX)^{-1}\otimes\ (Z^TZ)^{-1}\right)\ (X^T\otimes Z^T) \\ &= \left((X^TX)^{-1}X^T\right)\otimes\left((Z^TZ)^{-1}Z^T\right) \\ &= P_X\otimes P_Z, \end{aligned}$$

where we use Kronecker product identities repeatedly to simplify the expression. Therefore,

$$\mathrm{vec}\left(\hat{B}^T\right) = P_{X\otimes Z}\mathrm{vec}(Y^T) = (P_X\otimes P_Z)\mathrm{vec}(Y^T) = \mathrm{vec}(P_ZY^TP_X^T).$$

Transposing, unvectorizing, and substituting the projection matrices, we obtain

$$\hat{B} = P_XYP_Z^T = (X^TX)^{-1}X^TYZ(Z^TZ)^{-1},$$

which establishes Equation 4.

Variance of Estimate

Again by substitution, we can see that

$$\begin{aligned} \mathrm{var}(\mathrm{vec}(B^T)) &= P_{X\otimes Z}(I_n\otimes\Sigma)P_{X\otimes Z}^T \\ &= (P_X\otimes P_Z)(I_n\otimes\Sigma)(P_X^T\otimes P_Z^T) \\ &= (P_XP_X^T)\otimes(P_Z\Sigma P_Z^T). \end{aligned}$$

Substituting the projection matrices and simplifying, we get Equation 5.

Weighted Estimation

The most common form of weighted estimation occurs when we standardize all the rows or columns of the data matrix. This corresponds to using weighted regression with a weight matrix that is the Kronecker product of two diagonal weight matrices for the vectorized version of the problem. More generally, if we consider weight matrices of the form $W = W_X \otimes W_Z$, where W_X and W_Z are positive definite, then it is easily seen that the formulas for the weighted version of $\hat{B}$ and its variance are the same as those for the unweighted version, replacing the projection matrices (P_X and P_Z) with their weighted counterparts ($P_X^{W_X}$ and $P_Z^{W_Z}$).