

Supplement to Scale Reliant Mixed Effects Modeling of Sequence Count Data with ALDEx3

Kyle C. McGovern¹ and Justin D. Silverman^{1,2,3,4,*}

¹*Program in Bioinformatics and Genomics, Pennsylvania State University, University Park, PA, USA*

²*College of Information Sciences and Technology, Pennsylvania State University, University Park, PA, USA*

³*Department of Statistics, Pennsylvania State University, University Park, PA, USA*

⁴*Department of Medicine, Pennsylvania State University, Hershey, PA, USA*

*Corresponding Author: JustinSilverman@psu.edu

1 Designing Scale Models for SR-MEM

Defining a scale model for $W^\perp = (W_1^\perp, \dots, W_N^\perp)$ may seem daunting, as it requires specifying a joint distribution over N unobserved scale parameters. However the field of SRI has developed practical strategies that can simplify this task considerably.

For example, suppose external measurements of scale, denoted q_n , are available—such as from flow cytometry or qPCR. Further, assume that replication experiments have been conducted to estimate the measurement error, and that the variance of the log-transformed measurements is known, i.e., $\text{Var}(\log_2 q_n) = \sigma^2$. In this setting, it is often reasonable to assume conditional independence and define the scale model as $\log_2 W_n^\perp \sim \mathcal{N}(\log_2 q_n, \sigma^2)$. This strategy has been successfully employed in multiple SRI studies [1, 2, 3].

When external measurements of scale are unavailable and scale models must instead be specified using prior knowledge or biologically plausible assumptions, we generally recommend two strategies.

First, scale models can be constructed to generalize a desired normalization procedure. As discussed in prior work [1, 3, 4], many common normalizations can be expressed as

transformations of the form:

$$W_{\cdot n} = \frac{W_{\cdot n}^{\parallel}}{\phi(W_{\cdot n}^{\parallel})}, \quad (1)$$

for some scalar-valued function ϕ . For example, Total Sum Scaling (TSS) normalization corresponds to the case where $\phi(W_{\cdot n}^{\parallel}) = 1$, while Centered Log-Ratio (CLR) normalization is obtained when

$$\phi(W_{\cdot n}^{\parallel}) = \left(\prod_{d=1}^D W_{dn}^{\parallel} \right)^{1/D},$$

i.e., the geometric mean of the components of $W_{\cdot n}^{\parallel}$. Normalizations based on housekeeping genes also fit within this framework. Let $H \subset \{1, \dots, D\}$ denote a set of taxa (or genes) believed to be invariant with respect to the covariates in X . In this case, a natural choice of scaling function is

$$\phi(W_{\cdot n}^{\parallel}) = \left(\prod_{d \in H} W_{dn}^{\parallel} \right)^{1/|H|},$$

which assumes that the geometric mean abundance of the housekeeping set remains stable across experimental conditions. Regardless of the specific normalization used, Eq. (1) is unlikely to hold exactly. To account for possible deviations, we recommend generalizing the normalization by specifying a scale model of the form

$$\log_2 W_n^{\perp} \sim \mathcal{N}(-\log_2 \phi(W_{\cdot n}^{\parallel}), \sigma^2),$$

where σ^2 is a user-defined variance parameter representing the degree of uncertainty in the normalization. This formulation encodes prior belief in the normalization procedure while allowing for plausible departures from it. Additional examples and discussion of this approach can be found in prior work [1, 2, 3].

Second, scale models can be constructed using a reparameterization trick that simplifies the problem. Rather than directly specifying a distribution over the N -dimensional vector $W^{\perp} = (W_1^{\perp}, \dots, W_N^{\perp})$, we instead focus on how W^{\perp} contributes to estimation of the fixed effects θ_d . Specifically, the standard estimator for fixed effects in mixed-effects models (e.g., the maximum likelihood estimator) can be expressed as a linear transformation: $\theta_d = A \log W_d$, where $A = (X^{\top} V^{-1} X)^{-1} X^{\top} V^{-1}$ and $V = ZGZ^{\top} + R_{\rho}$. By linearity, this

estimator decomposes as

$$\begin{aligned}\theta_d &= A \log_2 W_d \\ &= \underbrace{A \log_2 W_d^{\parallel}}_{\theta_d^{\parallel}} + \underbrace{A \log_2 W_d^{\perp}}_{\theta^{\perp}},\end{aligned}\tag{2}$$

where θ_d^{\parallel} and θ^{\perp} are each P -dimensional vectors representing the compositional and scale components of the fixed effects, respectively. That is, θ_d^{\parallel} and θ^{\perp} describe how $\log_2 W_{dn}^{\parallel}$ and $\log_2 W_n^{\perp}$, respectively, vary with the covariates in X .

Instead of modeling W^{\perp} directly, we can instead model θ^{\perp} which only P -dimensional. To guide model specification, we recommend thinking of θ^{\perp} as the coefficients in a regression model of the form:

$$\log_2 W_n^{\perp} = \sum_{p=1}^P X_{np} \times \theta_p^{\perp}\tag{3}$$

where each element θ_p^{\perp} represents the effect of covariate X_{np} on the log scale. Once a distribution $\theta^{\perp} \sim P$ is specified, realizations of W^{\perp} can be generated by first sampling θ^{\perp} from P , then applying Eq. (3) to compute $\log_2 W_n^{\perp}$ for all n . Note Eq. (3) is the same as Eq. (5) in *Methods* in the main text.

Example. Consider differential abundance analysis which can be framed as a linear model with covariates $X_n = (1, c_n)$, where $c_n \in \{0, 1\}$ denotes treatment condition (e.g., placebo versus drug). In this case, Eq. (3) becomes:

$$\log_2 W_n^{\perp} = \theta_1^{\perp} + \theta_2^{\perp} c_n,$$

where θ_2^{\perp} is of primary scientific interest—it quantifies the difference in scale (e.g., total microbial load) between treatment groups. In a motivating example from [2], the authors studied changes in bacterial load as an inoculated microbial community expanded to occupy a larger volume. Based on the ratio of inoculation volume to final vessel volume, they specified the prior $\theta_2^{\perp} \sim \mathcal{N}(\log_2(\frac{100}{400} \text{ mL}), 0.5^2)$, where the variance accounts for possible

deviations due to unmeasured differences in carrying capacity across environments. Since the intercept θ_1^\perp is not of scientific interest, it can be fixed to zero without loss of generality. To sample from the corresponding scale model: draw $\theta_2^\perp \sim \mathcal{N}(\log_2 0.25, 0.5^2)$, and compute $\log_2 W_n^\perp = \theta_2^\perp c_n$, for each $n = 1, \dots, N$. This procedure yields realizations of W^\perp consistent with prior knowledge, and allows SR-MEM to propagate this uncertainty through to inference on θ_d .

References

- [1] McGovern KC, Silverman JD. Replacing normalizations with interval assumptions enhances differential expression and differential abundance analyses. *BMC Bioinformatics*. 2025 Jul;26(1).
- [2] Nixon MP, Letourneau J, David LA, Lazar NA, Mukherjee S, Silverman JD. Scale Reliant Inference. Preprint at <https://arxiv.org/abs/2201.03616> (2023).
- [3] Nixon MP, Gloor GB, Silverman JD. Incorporating scale uncertainty in microbiome and gene expression analysis as an extension of normalization. *Genome Biol*. 2025 May;26(1).
- [4] McGovern KC, Nixon MP, Silverman JD. Addressing erroneous scale assumptions in microbe and gene set enrichment analysis. *PLoS Comput Biol*. 2023 11;19(11):1-16.