Supplementary Note 1: Mathematical Principles of Dynamic Proportionality

DyProp Development Team

December 10, 2025

Abstract

This supplementary note provides the rigorous mathematical derivations underpinning the Dynamic Proportionality (`dyprop`) framework. We define the geometric constraints of the simplex, the zero-handling imputation protocols, and the CoDa-safe robustness transformations. Furthermore, we derive the vectorized kernel-weighted covariance algebra for continuous trajectory analysis and the Singular Perturbation solutions for mechanistic boundary function modeling via Vectorized Basis Projection. Finally, we map these metrics to Dynamical Systems Theory, reinterpreting proportionality instability as a loss of regulatory stiffness on a potential landscape.

1 The Geometry of Compositional Data

1.1 The Simplex and the Closure Problem

Genomic count data (e.g., scRNA-seq, scDNA-seq, metagenomics) are inherently compositional. The observed count vector $\mathbf{x}_i$ for a cell i is strictly constrained by the arbitrary sequencing depth (library size). Consequently, the data does not reside in Euclidean space ($\mathbb{R}^p$), but in the **Simplex** (S^p).

$$S^p = \left\{ \mathbf{x} \in \mathbb{R}^p \mid x_j > 0, \sum_{j=1}^{p} x_j = \kappa \right\} \tag{1}$$

Where:

- $\mathbf{x} = [x_1, \ldots, x_p]$ is the vector of observed counts for p features.
- κ is the library size constraint (sum constraint), which varies per cell.

As proven by **(author?)** [1], applying standard statistical operations (such as Pearson correlation or Euclidean distance) directly to S^p yields spurious negative associations, known as the "Closure Bias". To perform valid statistical analysis, we must project the data from S^p to $\mathbb{R}^p$ using an isomorphic transformation.

1.2 The Centered Log-Ratio (CLR) Transformation

We utilize the Centered Log-Ratio (CLR) transformation to project the composition into Euclidean space while preserving the geometric relationships between features.

$$\mathrm{clr}(\mathbf{x}) = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ln \frac{x_2}{g(\mathbf{x})}, \ldots, \ln \frac{x_p}{g(\mathbf{x})} \right] \tag{2}$$

Where:

- $g(\mathbf{x}) = \left(\prod_{j=1}^{p} x_j\right)^{1/p}$ is the geometric mean of the composition vector.

Theorem 1 (Log-Ratio Invariance): A crucial property for our framework is that the ratio between any two features j and k in CLR space is identical to the log-ratio of their raw abundances, as the denominator $g(\mathbf{x})$ cancels out.

$$\mathrm{clr}(x_j) - \mathrm{clr}(x_k) = \ln\left(\frac{x_j}{g(\mathbf{x})}\right) - \ln\left(\frac{x_k}{g(\mathbf{x})}\right) = \ln\left(\frac{x_j}{x_k}\right) \tag{3}$$

This identity allows us to compute stoichiometric dynamics efficiently using vector operations on the transformed matrix.

2 Zero Handling and Robustness Protocols

The logarithmic nature of Eq. (2) requires strictly positive data ($x_j > 0$). However, single-cell data contains zeros arising from distinct processes, as well as extreme outliers (e.g., PCR artifacts) that must be managed without violating the geometry of the simplex. We distinguish these according to the framework established by **(author?)** [2] and **(author?)** [3].

2.1 Technical Dropout (Sporadic Zeros)

Zeros resulting from low capture efficiency are modeled as missing data. We apply **k-Nearest Neighbor (kNN) Pooling** in the reduced Principal Component space to recover these values.

$$\tilde{x}_{ij} = \frac{1}{K} \sum_{m \in \mathcal{N}_K(i)} x_{mj} \tag{4}$$

Where:

- $\mathcal{N}_K(i)$ is the set of K nearest neighbors to cell i (default $K = 5$).

This step smooths technical noise but preserves biological silencing if the entire neighborhood is consistently silent.

2.2 Biological Silencing (Rounded Zeros)

Zeros representing true biological absence are treated as **Rounded Zeros**—values that exist below the Limit of Detection (δ). We impute these using **Geometric Bayesian Multiplicative Replacement (GBM)** to preserve the covariance structure [2].

$$x_{ij}^* = \begin{cases} \mathbb{E}[x_{ij} \mid x_{ij} < \delta] & \text{if } x_{ij} = 0 \\ x_{ij} \cdot \left(1 - \frac{\sum_{z \in \mathcal{Z}} x_{iz}^*}{\kappa}\right) & \text{if } x_{ij} > 0 \end{cases} \tag{5}$$

2.3 Ratio-First Robustness (CoDa-Safe Winsorization)

Standard outlier removal (e.g., clipping the counts matrix) violates Subcompositional Coherence by altering the geometric mean $g(\mathbf{x})$ non-linearly. To maintain rigor, we apply robustness protocols only *after* the ratio transformation.

Let $\mathbf{Y} \in \mathbb{R}^{n \times p}$ be the CLR-transformed data matrix across n cells. We define the trajectory vector $\mathbf{y}_{jk} \in \mathbb{R}^n$ representing the log-ratio evolution between features j and k across pseudotime:

$$\mathbf{y}_{jk} = \mathrm{clr}(\mathbf{x})_j - \mathrm{clr}(\mathbf{x})_k \tag{6}$$

We apply Winsorization (clipping top/bottom 1% quantiles) directly to the vector $\mathbf{y}_{jk}$. Since $\mathbf{y}_{jk}$ represents coordinates in Euclidean space, this operation is mathematically equivalent to robust regression and preserves the compositional integrity of the input features.

3 Dynamic Proportionality: The Kernel Framework

Standard compositional methods (e.g., `propr`, **(author?)** [4]) assume static relationships. To model dynamics, we introduce a continuous pseudotime coordinate $t \in [0, 1]$ and derive the **Instantaneous Weighted Covariance**.

3.1 Kernel Weighting

We define a Gaussian weight vector $\mathbf{w}(t_0)$ for a specific evaluation timepoint t_0:

$$w_i(t_0) = \exp\left(-\frac{(t_i - t_0)^2}{2h^2}\right) \tag{7}$$

3.2 Dynamic Instability (Φ)

We define the dynamic extension of the metrics proposed by **(author?)** [5]. $\Phi(t)$ measures the magnitude of proportionality deviation relative to the variance of the reference feature j. A spike in $\Phi(t)$ indicates a **Loss of Restoring Force** (Decoupling).

$$\Phi_{jk}(t) = \frac{\text{var}_w(\log(x_j/x_k))_t}{\text{var}_w(\text{clr}(x_j))_t} \tag{8}$$

3.3 Dynamic Coupling (ρ)

To distinguish between coherent transitions and chaotic loss of regulation, we extend the metric proposed by **(author?)** [6]. We define Dynamic Coupling $\rho(t)$, analogous to a correlation coefficient on the simplex:

$$\rho_{jk}(t) = 1 - \frac{\text{var}_w(\log(x_j/x_k))_t}{\text{var}_w(\text{clr}(x_j))_t + \text{var}_w(\text{clr}(x_k))_t} \tag{9}$$

Interpretation:

- $\rho(t) \approx 1$: Tight stoichiometric coupling (Homeostasis or Switch).
- $\rho(t) \to 0$: Complete regulatory decoupling (Network Melting).

4 Mechanistic Modeling: Vectorized Basis Projection

For gene pairs exhibiting significant $\Phi(t)$ dynamics, we model the transition mechanism using **Singular Perturbation Theory** [7]. Rather than iterative non-linear regression, we employ a high-performance **Vectorized Basis Projection**.

4.1 The Boundary Function

We define the log-ratio $y(t)$ as a state variable transitioning between equilibria θ_A and θ_B via first-order kinetics:

$$\frac{dy}{dt} = \frac{1}{\epsilon\Delta\theta}(y - \theta_A)(\theta_B - y) \tag{10}$$

The explicit solution is the Generalized Logistic Function:

$$\mu(t) = \theta_A + \frac{\theta_B - \theta_A}{1 + \exp\left(-\frac{t-\tau}{\epsilon}\right)} \tag{11}$$

where τ is the Tipping Point and ϵ is the Inverse Sharpness.

4.2 Vectorized Convolution and the Nyquist Criterion

We generate a dictionary matrix $\mathcal{M}$ of archetypal boundary functions spanning the topological parameter space. To guarantee signal capture, the grid resolution $\Delta\tau_{grid}$ is coupled to the detection limit via the **Nyquist Sampling Criterion**:

$$\Delta\tau_{grid} \leq \epsilon_{min} \tag{12}$$

The similarity between the observed log-ratio trajectory vector $\mathbf{y}$ and the basis dictionary $\mathcal{M}$ is computed via the dot product:

$$\mathbf{r} = \mathbf{y} \cdot \mathcal{M}^T \tag{13}$$

The index i maximizing $\mathbf{r}$ identifies the discrete "Best Fit Archetype".

4.3 Quadratic Sub-Grid Interpolation

To recover continuous parameter estimates from the discrete grid, we apply Parabolic Interpolation. Given the discrete maximum score R_i at grid point τ_i, and its neighbors R_{i-1} and R_{i+1}, the continuous peak $\hat{\tau}$ is derived as:

$$\hat{\tau} = \tau_i + \frac{\Delta\tau}{2} \cdot \frac{R_{i-1} - R_{i+1}}{R_{i-1} - 2R_i + R_{i+1}} \tag{14}$$

This yields infinite precision for the Tipping Point without the computational cost of iterative optimization.

5 The Physics of Regulation: Landscape Dynamics

5.1 Landscape Curvature and Stiffness

We reframe the interpretation of Dynamic Proportionality to **Dynamical Systems Theory**. We posit that the gene regulatory network (GRN) moves on a quasi-potential landscape $U(\mathbf{y})$. In a stable state (homeostasis), the system resides in a deep basin of attraction. Approximating the local potential as a harmonic oscillator:

$$U(y) \approx \frac{1}{2}k(y - \theta)^2 \tag{15}$$

The variance of the log-ratio is inversely proportional to the stiffness k of the regulatory restoring force:

$$\text{Var}(y) = \frac{D}{k} \tag{16}$$

where D represents the magnitude of stochastic noise (e.g., transcriptional bursting).

5.2 Dynamic Instability as a Metric of Phase Transition

Consequently, our metric $\Phi(t)$ serves as a proxy for the **Inverse Regulatory Stiffness** (k^{-1}).

- **Homeostasis ($\Phi \approx 0$):** High stiffness ($k \gg D$). Strong repression minimizes transcriptional noise.

- **Transition ($\Phi \gg 0$):** Loss of stiffness ($k \to 0$). The system traverses a "saddle point" (flattened landscape) to facilitate phenotypic rewiring [8].

Thus, DyProp identifies network events not by expression level, but by the collapse of regulatory constraints.

References

[1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall.

[2] Martín-Fernández, J.A., et al. (2015). Multiplicative replacement for general zero values. *Biometrika*, 102(2), 255-269.

[3] Palarea-Albaladejo, J., & Martín-Fernández, J.A. (2015). zCompositions: R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143, 85-96.

[4] Quinn, T.P., et al. (2017). propr: An R-package for identifying proportionally abundant features using compositional data analysis. *Scientific Reports*, 7, 16252.

[5] Lovell, D., et al. (2015). Proportionality: A valid alternative to correlation for relative data. *PLoS Computational Biology*, 11(3), e1004075.

[6] Erb, I., & Quinn, T.P. (2016). Differential Proportionality. *Analysis of Large and Complex Data*, 405-414.

[7] Verhulst, P.F. (1838). Notice sur la loi que la population suit dans son accroissement. *Correspondance Mathématique et Physique*, 10, 113-121.

[8] Scheffer, M., et al. (2009). Early-warning signals for critical transitions. *Nature*, 461(7260), 53-59.