Appendix: Identifiability Under Latent Cell-State Confounding

A1. Derivation details

A1.1 Observational slope decomposition

Given - `X = alpha Z + eps_x` - `Y = beta X + gamma Z + eps_y` with zero-mean independent noises and `Var(Z)=1`,

`Var(X) = alpha^2 + sigma_x^2`

`Cov(X,Y) = Cov(X, beta X + gamma Z + eps_y) = beta Var(X) + gamma Cov(X,Z) = beta Var(X) + gamma alpha`

Hence observational slope:

`Cov(X,Y)/Var(X) = beta + (gamma alpha)/Var(X)`

which shows confounding-induced shift away from `beta`.

A1.2 Anchor-adjusted bias formula

Let anchor `A = Z + u`, `Var(u)=sigma_u^2`, and regress `Y` on `(X, A)`. By Frisch-Waugh-Lovell, coefficient on `X` equals slope of `Y` on residualized `X_r = X - proj_A(X)`.

`proj_A(X)` coefficient is `Cov(X,A)/Var(A) = alpha/(1+sigma_u^2)`.

So `X_r = X - [alpha/(1+sigma_u^2)]A`

`Cov(Z, X_r) = alpha - alpha/(1+sigma_u^2) = alpha sigma_u^2/(1+sigma_u^2)`

`Var(X_r) = Var(X) - Cov(X,A)^2/Var(A) = (alpha^2 + sigma_x^2) - alpha^2/(1+sigma_u^2) = sigma_x^2 + alpha^2 sigma_u^2/(1+sigma_u^2)`

Bias term from residual confounding is

`Bias_anchor = gamma * Cov(Z,X_r)/Var(X_r) = gamma * alpha * sigma_u^2 / [sigma_x^2(1+sigma_u^2) + alpha^2 sigma_u^2]`.

This matches observed degradation with larger anchor noise in simulations.

A2. Output inventory

Synthetic suite outputs: - `/Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_latent_`
- `/Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_latent_cell_state_confoundi`
- `/Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_latent_cell_state_confoundi`
- `/Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_latent_cell_state_confoundi`
- `/Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_latent_cell_state_confoundi`
- `/Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_latent_cell_state_confoundi`

Tissue diagnostics outputs: - /Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_late
- /Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_latent_cell_state_confoundi
- /Users/ihorkendiukhov/biodyn-work/subproject_10_identifiability_latent_cell_state_confoundi

A3. Additional result slices

A3.1 Best and worst environment-stress cells

From synthetic_environment_stress.csv:

Best gain cell: - alpha_shift=0.0, variance ratio 1.8 - pooled RMSE 0.6280 - invariance RMSE 0.1133 - gain +0.5147

Worst gain cell: - alpha_shift=0.0, variance ratio 1.0 - pooled RMSE 0.7484 - invariance RMSE 4.4592 - gain -3.7109

Interpretation: invariance estimator can be numerically explosive when environments are not sufficiently separated.

A3.2 Tissue pass/fail summary

From tissue_edge_invariance_fits.csv: - Pass edges: 22/76. - Median R^2 pass set: 0.738. - Median R^2 fail set: 0.130.

Top pass edges by fit: 1. RELA -> CXCL8 (R^2=0.989, RMSE=0.0037) 2. GATA2 -> NR3C2 (R^2=0.981, RMSE=0.0622) 3. NFKB2 -> CXCL8 (R^2=0.952, RMSE=0.0229)

Worst-fit examples: 1. HIF1A -> XXYLT1 (R^2=0.0006) 2. STAT3 -> TYK2 (R^2=0.0009) 3. GATA3 -> RBMS1 (R^2=0.0022)

A4. Practical implementation note

For production use, the environment-based estimator should include a hard precondition check: - reject estimation when |1/VarX1 - 1/VarX2| < epsilon.

This prevents unstable divisions and aligns estimator behavior with identifiability requirements.