Decompositions

Sophie Woodward

2023-06-08

Contents

Key Findings (so far)

- When the data-generating mechanism is based on the nested decomposition, the outcome model is linear, and confounding dissipates locally (within 5×5 grids), both the nested and spectral decompositions recover unbiased estimates of the treatment effect at small spatial scales.

- When the data-generating mechanism is based on the spectral decomposition, the outcome model is linear, and confounding dissipates locally (at high spectral frequencies), the spectral decomposition recovers unbiased estimates of the treatment effect at small spatial scales. The nested decomposition does not, in general. (Need to look at this more)

- When the data-generating mechanism is based on the nested decomposition, the outcome model is linear, and confounding dissipates globally (confounding within 5×5 grids but not across), the nested decomposition recovers unbiased estimates of the treatment effect at large spatial scales, but the spectral does not. (Need to look at this more and check this is the case)

- Neither decomposition can recover unbiased estimates at any scale when there is a quadratic term of exposure X is included in the outcome model.

Note: in the following plots, I mark the x axis by spatial scale. If I plot results from the nested decomposition, then there are only two points: spatial scale equal to 1 is the so-called county level (1×1 grid) and spatial scale equal to 2 is the so-called state level (5×5 grid). If I plot results from the spectral decomposition, then the spatial scale indexes the ordered eigenvalues of the graph Laplacian. So the spatial scales between spectral and nested plots should not be directly compared.

Simulation 1: Nested DGM

Denote X as exposure X, Z as unmeasured confounder, and Y as outcome. We simulate the following scenario 100 times. Across $n_1 = 25$ 5×5 grids, which we refer to as states,

$$X_{1,s} \sim \text{Exp}(1)$$
$$Z_{1,s} \sim 0.9X_2 + \sqrt{1 - 0.9^2}\text{Exp}(1)$$

for $s = 1, \ldots, n_1 = 25$. Across $n_2 = 625$ 1×1 grids, which we refer to as counties, let

$$X_{2,i} \sim \text{Exp}(1)$$
$$Z_{2,i} \sim 0.001X_2 + \sqrt{1 - 0.001^2}\text{Exp}(1)$$

and $X_i = X_{1,s(i)} + X_{2,i}$, $Z_i = Z_{1,s(i)} + Z_{2,i}$ for $i = 1, \ldots, n_2 = 625$. By construction, X, Z are nearly uncorrelated within the states of 5×5, but correlated across states. (Note that this DGM exactly follows the nested decomposition.) We let $Y_i = 2X_i - Z_i + \epsilon$ where $\epsilon_i \sim \mathcal{N}(0, 1)$ independently across i.

For each of the 100 scenarios, we decompose X, Z, Y at different spatial scales using 1) nested decomposition and 2) spectral decomposition. At each spatial scale ω, we obtain an estimate $\hat{\beta}(\omega)$ of $\beta = 2$ from a linear regression of $Y(\omega)$ on $X(\omega)$. Hypothesis: $\hat{\beta}(\omega)$ is unbiased for low ω (finer spatial scales) since by construction confounding dissipates locally.
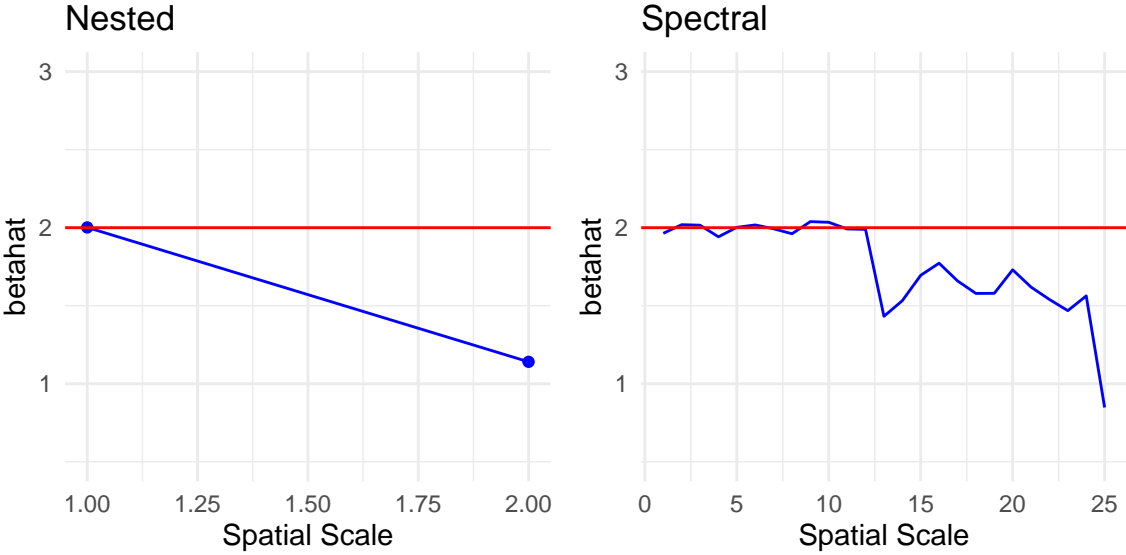
The true β is plotted in red (need to add legend). Looking at the nested plot, it's reassuring to see that the bias of the coefficient is 0 at low spatial scales when the DGM is indeed nested. Looking at the spectral plot, we see that the spectral decomposition also results in a bias of 0 for low spatial scales. The curve here takes on an almost-stepwise shape.

Alternatively, suppose Z doesn't vary at all within states. That is, for fixed 5×5 grid s', $Z_i = c \in \mathbb{R}$ $\forall i$ such that $s(i) = s'$. I accomplish this by letting $Z_{2,i} = 0$ $\forall i$ in the construction above, so $Z_i = Z_{1,s(i)}$.
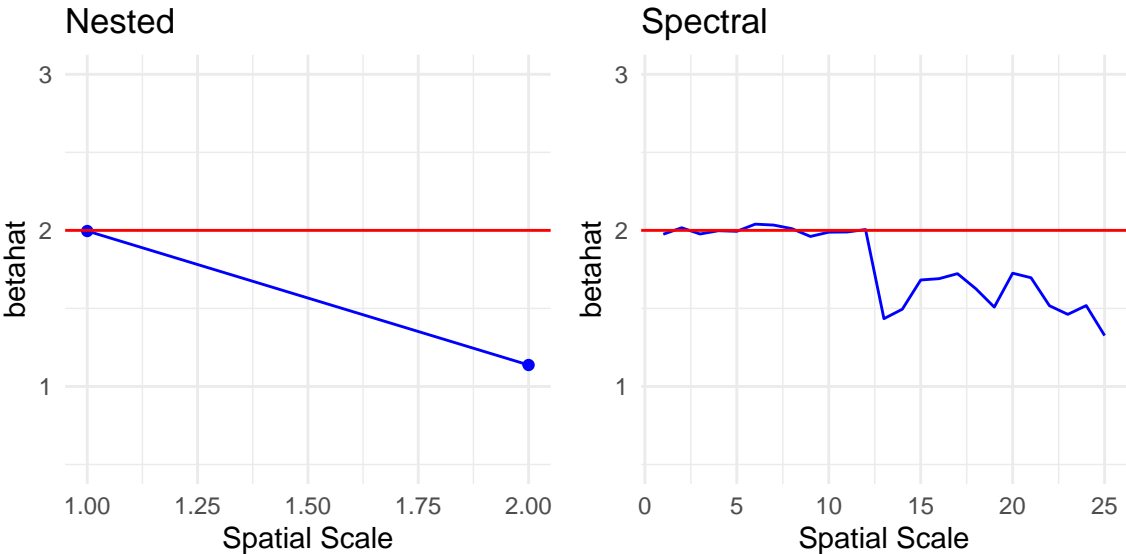
Simulation 2: Spectral DGM

We repeat simulation 1 but now the data-generating model originates from the spectral decomposition rather than the nested. In particular, we use the graph Fourier transform to project X and Z into the spectral domain. In the spectral domain,

$$X_i^* \sim \text{Exp}(1)$$

$$Z_i^* \sim \rho_i X_2 + \sqrt{1 - \rho_i^2}\text{Exp}(1)$$

for $i = 1, \ldots, 625$ where $\rho_i = 0, 1/624, 2/624, \ldots, 1$. So the covariance between X and Z goes to 0 for smaller i, which corresponds to larger eigenvalues ω of the graph Laplacian, or smaller spatial scales.

Again, for each of the 100 scenarios we decompose X, Z, Y at different spatial scales using 1) nested decomposition and 2) spectral decomposition. At each spatial scale ω, we obtain an estimate $\hat{\beta}(\omega)$ of $\beta = 2$ from a linear regression of $Y(\omega)$ on $X(\omega)$. Hypothesis: $\hat{\beta}(\omega)$ is unbiased for low ω (finer spatial scales) since by construction confounding dissipates locally.
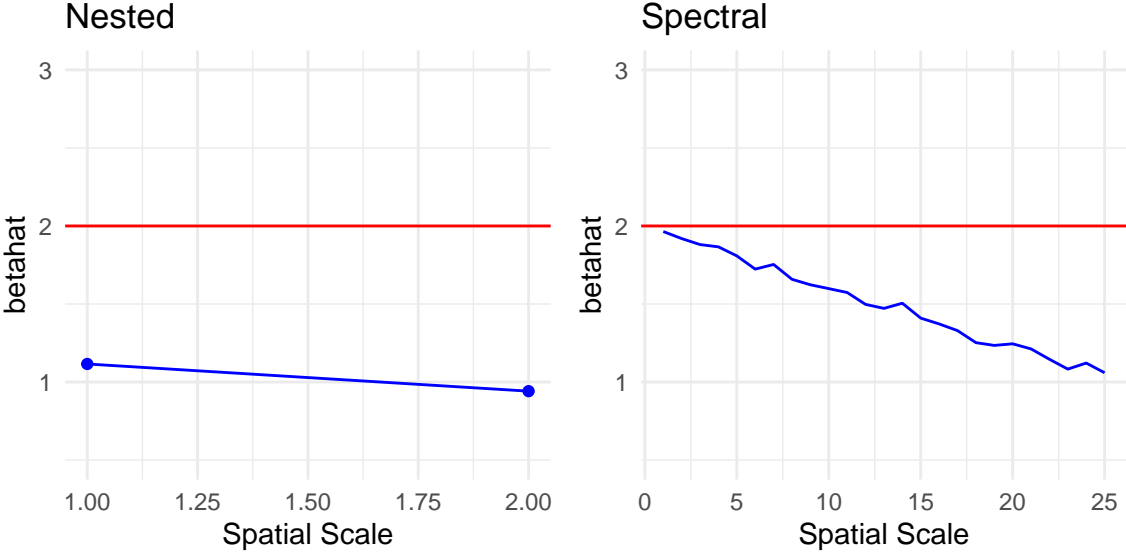

Looking at the spectral plot, it's reassuring to see that the bias of the coefficient is 0 at low spatial scales when the DGM is indeed spectral. The estimates obtained from the nested decomposition are biased at both of the two grid levels. At first I thought that made sense: by construction the covariance is continuously decreasing with spatial scale; within a 5×5 grid we will observe bias. Let's try $\rho_i = 0 \ \forall i$.
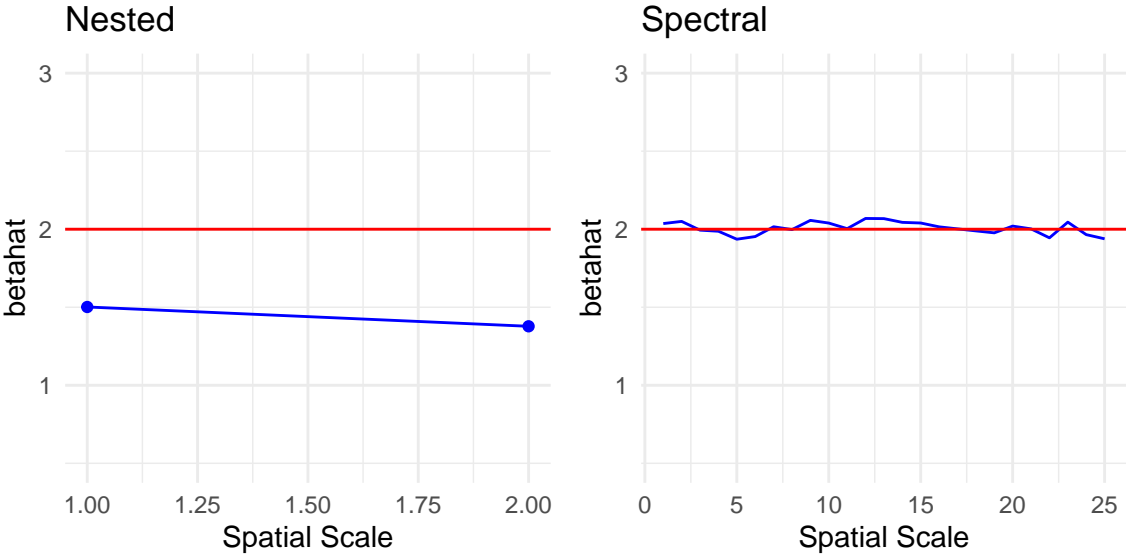
I am puzzled by the result. I thought making the spectral coherence equal to 0 for all eigenvalues would result in the nested decomposition also resulting in unbiased estimates. However, this isn't the case: The nested decomposition produces estimates that are negatively biased by 0.5. What's going on? Let's look at covariance between X and Z by spatial scale (next section).

Simulation 3: Covariance by Spatial Scale

Let's plot the correlation between X and Z by spatial scale for both decompositions when the DGM is spectral and $\text{Cov}(X^*, Z^*) = 0$ for all spatial frequencies:
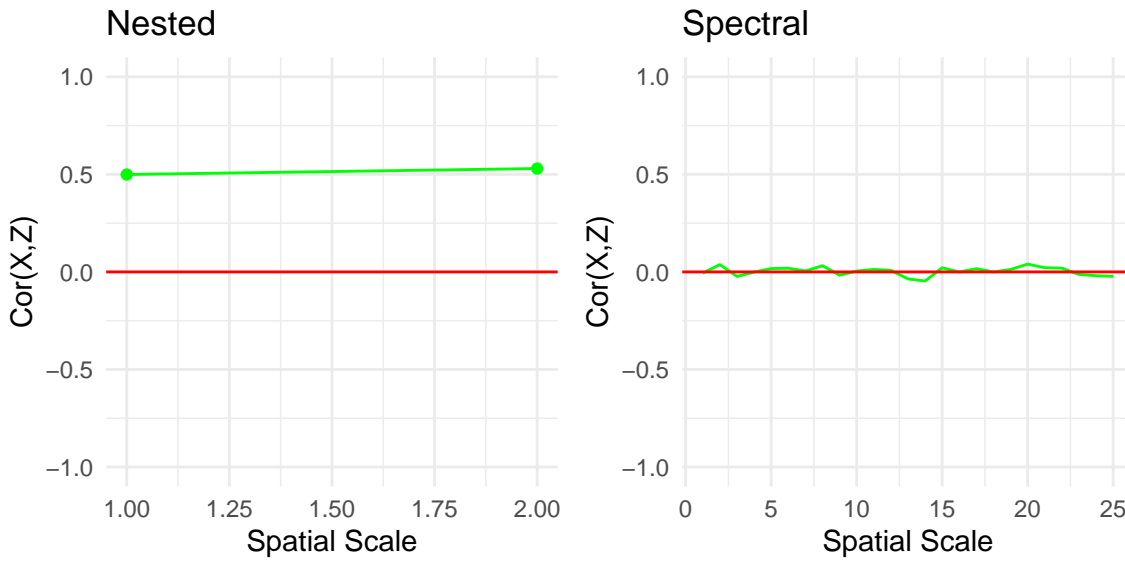
Weirdly, in the spatial domain X and Z are still correlated at both grid levels.

Let's repeat this with the nested DGM. Below I plot the correlation between X and Z by spatial scale for both decompositions when the DGM is nested and the covariance between X and Z equals 0 for both grid levels:
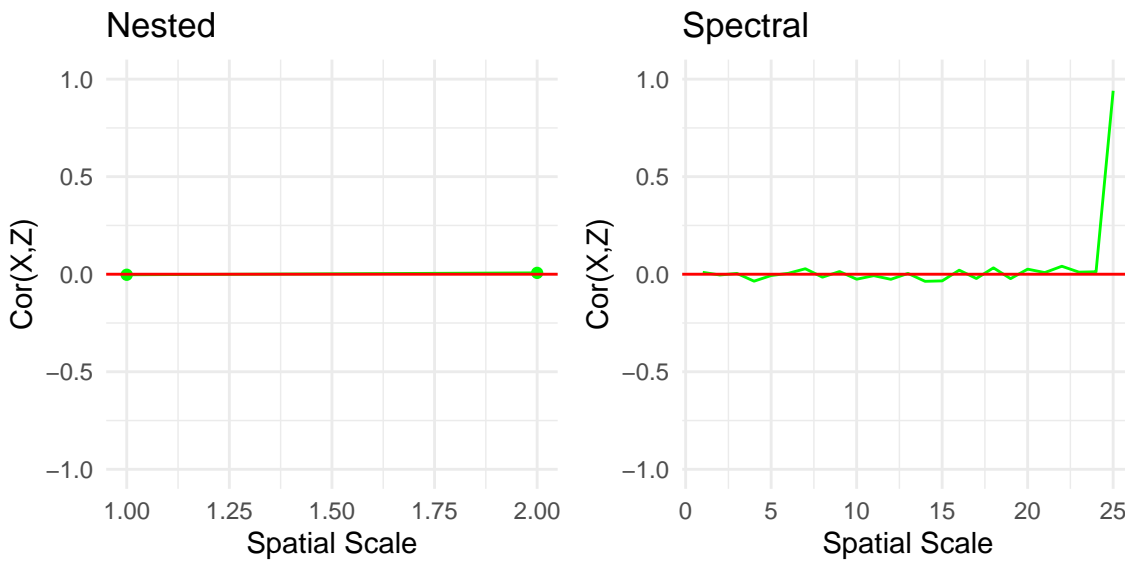
Both decompositions recover zero correlation over most spatial scales. Except the correlation for large scales of the spectral decomposition blows up...

So why is the correlation between X and Z at different levels of the nested decomposition nonzero when enforcing zero correlation at all spatial frequencies of the spectral decomposition?

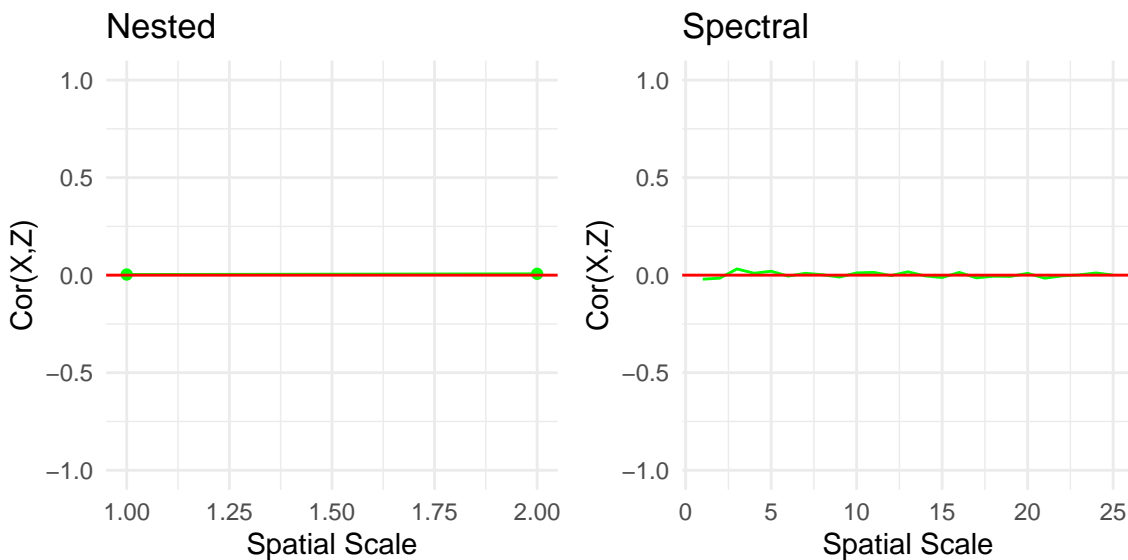Suppose Z doesn't vary at all until some spatial frequency.
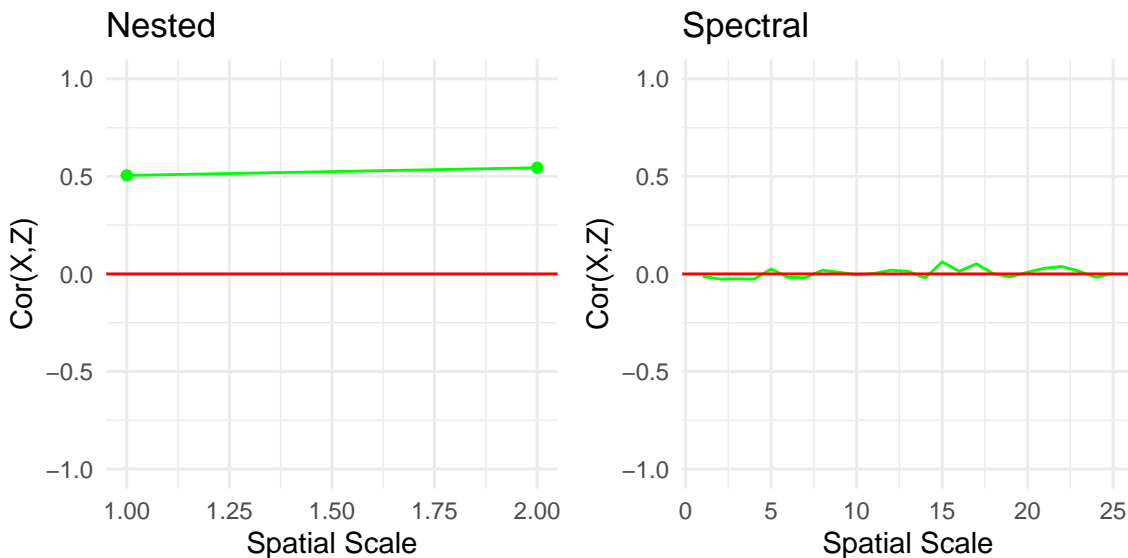
Still getting the same result.

What if we use normal distributions for X, Z instead of Expo? In the spectral domain,

$$\begin{pmatrix} X_i^* \\ Z_i^* \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_i \\ \rho_i & 1 \end{pmatrix} \right)$$

for $i = 1, \ldots, 625$ where $\rho_i = 0 \; \forall i$.

Ok, there we go - normality recovers what we would expect. So I need to play around with my DGM to figure out why Expo is producing weird results. Or, rather why normality is special (could it have to do with the fact the Fourier transform of a Gaussian is Gaussian and other nice properties of the normal? Or is this an error in the way I set up my DGM?) (TO DO)

Simulation 4: Local Confounding

Let's go back to using the Exponential DGM. We repeat simulations 1 and 2 but now confounding dissipates at larger scales rather than smaller ones.

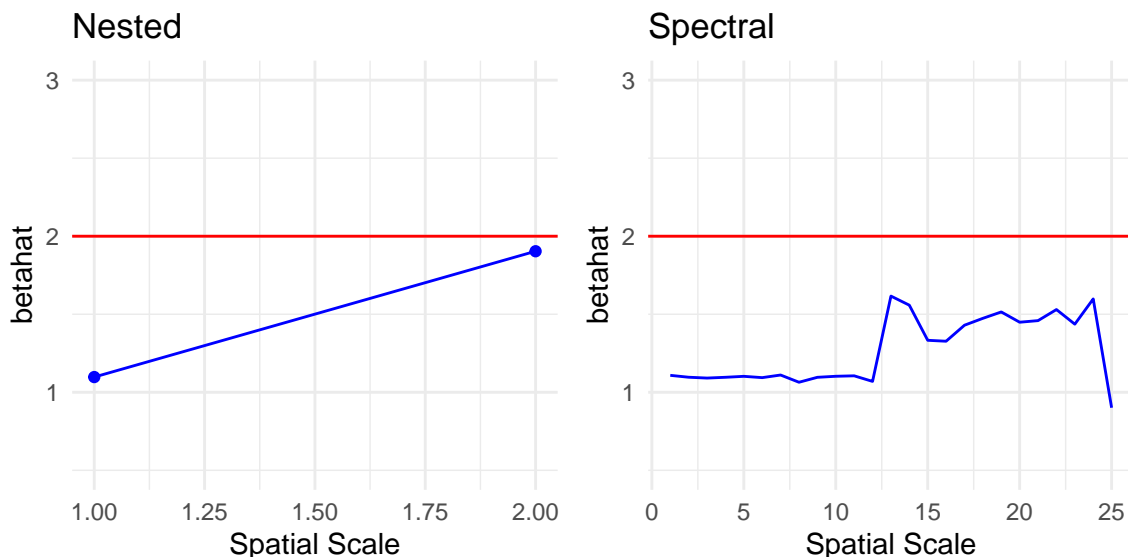For the nested DGM, this means

$$X_{1,s} \sim \text{Exp}(1)$$
$$Z_{1,s} \sim 0.001X_2 + \sqrt{1 - 0.001^2}\text{Exp}(1)$$

for each 5×5 state, $s = 1, \ldots, n_1 = 25$. Across $n_2 = 625$ counties of 1×1 grids, let

$$X_{2,i} \sim \text{Exp}(1)$$
$$Z_{2,i} \sim 0.9X_2 + \sqrt{1 - 0.9^2}\text{Exp}(1)$$

and $X_i = X_{1,s(i)} + X_{2,i}, Z_i = Z_{1,s(i)} + Z_{2,i}$ for $i = 1, \ldots, n_2 = 625$. By construction, X, Z are nearly uncorrelated across the states of 5×5, but correlated within states.

For the spectral DGM,

$$X_i^* \sim \text{Exp}(1)$$
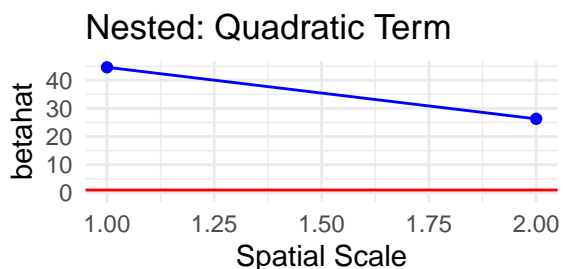$$Z_i^* \sim \rho_i X_2 + \sqrt{1 - \rho_i^2}\text{Exp}(1)$$

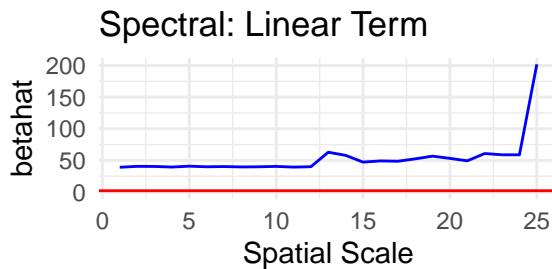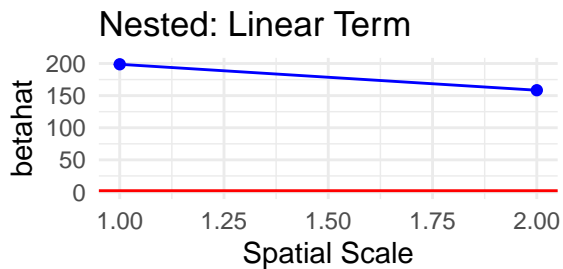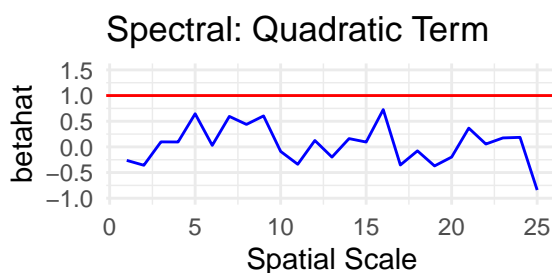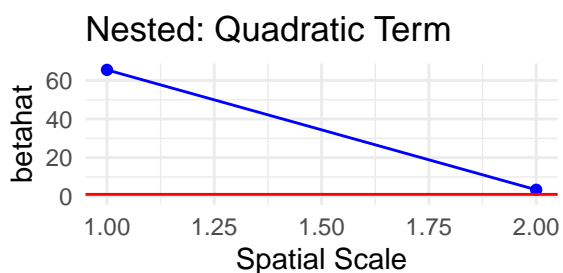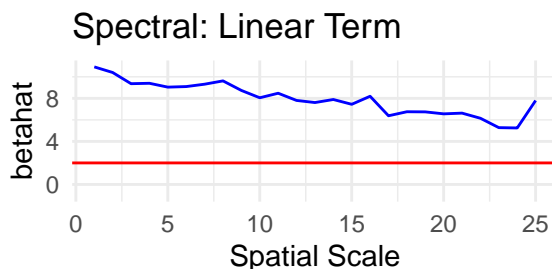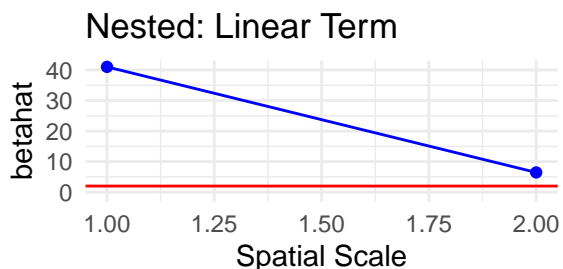for $i = 1, \ldots, 625$ where $\rho_i = 1, 623/624, 622/624, \ldots, 0$.

6
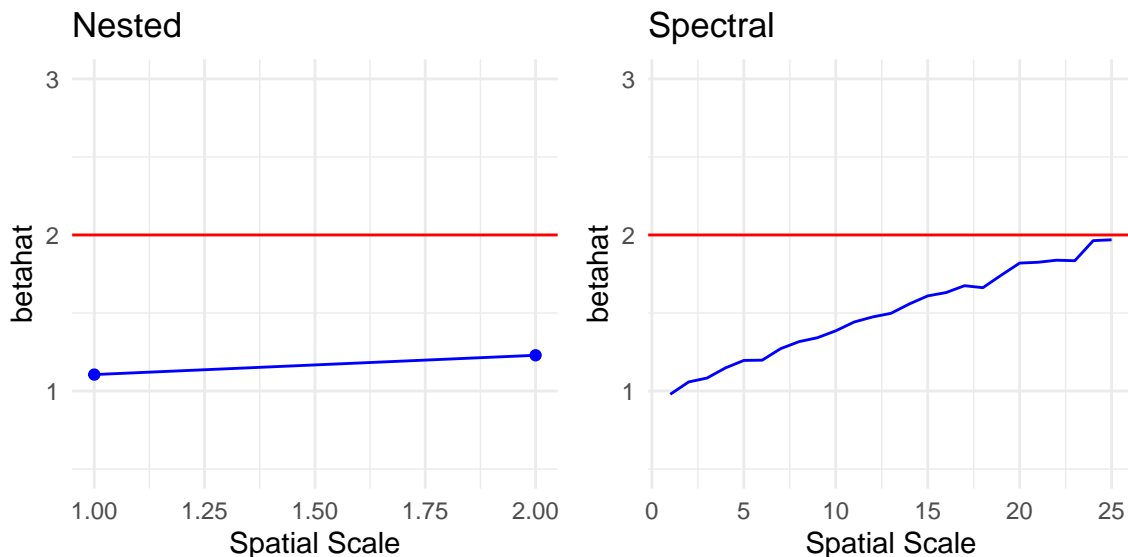
Simulation 5: Nonlinear Outcome Model

We repeat simulation 1 and 2 but the outcome model now includes a quadratic term of X. In particular, $Y_i = 2X_i + X_i^2 - Z_i + \epsilon$. The scale-specific analyses attempt to estimate both coefficients of X. Neither method does a good job at recovering either of the coefficients.
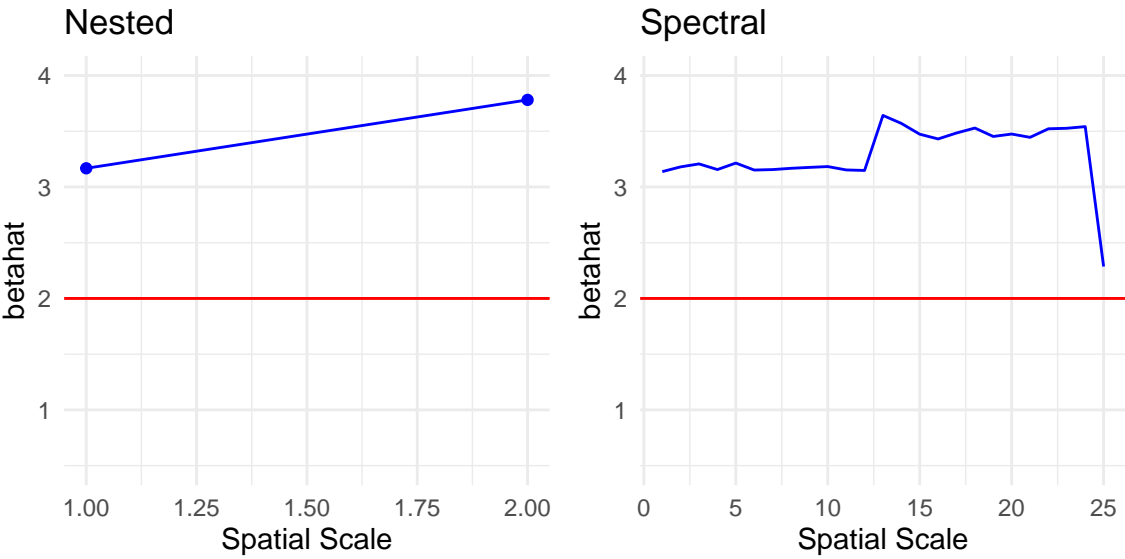
Nested DGM
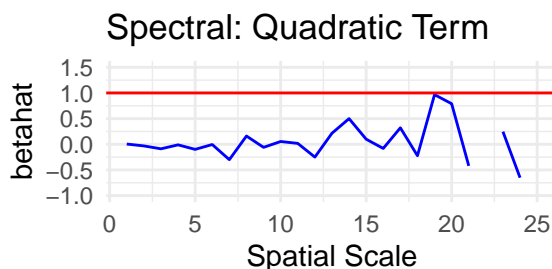
Spectral DGM

Simulation 6: Outcome Model with Interaction

We repeat simulation 1 and 2 but the outcome model now includes an interaction term between X and Z. In particular, let $Y_i = 2X_i - Z_i + 0.5X_iZ_i + \epsilon$.

Nested DGM

Spectral DGM