

Response to Clark and Kristensen (2026)*

M. Dolores Collado, Ignacio Ortuño-Ortín, and Jan Stuhler

April 30, 2026

We thank Clark and Kristensen for their interest in our work and for a stimulating comment. We find their paper interesting and believe it raises important questions about the ability of genetic models to account for kinship correlations. In what follows, we first briefly recap the relevant findings from our paper, then discuss what we see as the valuable contributions of Clark and Kristensen’s analysis. We then raise two concerns: first, the theoretical predictions of their model appear to be incorrectly derived; second, the empirical fit of the model is substantially worse than their results suggest, once the structural restrictions of the model are imposed.

Background: Collado, Ortuño-Ortín, and Stuhler (2023)

In Collado, Ortuño-Ortín and Stuhler (2023), we estimated the strength of intergenerational persistence and assortative mating using an extended set of familial correlations. Our main objective was to quantify how strongly socioeconomic advantages are transmitted across generations, partially motivated by a wave of recent studies documenting higher persistence than captured by conventional parent–child correlations—including Clark’s (2014) influential work on the slow regression to the mean of surname-level status. Using Swedish register data, we computed correlations for 141 distinct kinship types, including not only close relatives but also in-laws up to five degrees of separation. This unusually large number of empirical moments allowed us to go beyond simple two-generation comparisons: we found that intergenerational persistence is substantially stronger than conventional parent–child correlations suggest, and we could decompose the transmission process into its assortative and intergenerational components, study the respective roles of observable and latent advantages, and conduct extensive out-of-sample testing.

*This note responds to Gregory Clark and Martin Hørlyk Kristensen, “The Mechanisms of Intergenerational Status Transmission in Modern Scandinavia,” EHES Working Paper No. 299, March 2026 (https://ehes.org/wp/EHES_299.pdf), which examines whether a genetic model can account for the kinship correlations reported in Collado, Ortuño-Ortín and Stuhler (2023). It was written in April 2026 and is posted here as written, apart from this note and the list of references. Page, table, equation and figure numbers refer to the working paper. A condensed version of the analysis appears as chapter 5 of Gregory Clark’s forthcoming book *To Have and Have Not: The Determinants of Social Status in England, 1600–2026*; the estimates and the figure discussed below correspond to Table 5.4 and Figure 5.6 there. Contact: jan.stuhler@uc3m.es.

A more difficult question is what mechanism underlies this strong persistence. As we emphasized in our paper, both the strength and the weakness of our approach is that it remains largely agnostic about specific causal channels. Our latent factor z captures all latent parental advantages—whether genetic, cultural, or social in nature—and we intentionally avoided taking a strong stance on the underlying mechanism. However, in extensions of our analysis, we studied the potential role of geography and genes. Among other results, we showed that standard models of additive genetic transmission—in particular with phenotypic assortment—cannot fit the kinship correlations well (Figure 6 in Collado et al., 2023). This does not mean that genetic models can generally not fit the data, but that such a model would need to differ in some ways from the standard Fisherian framework. Clark and Kristensen’s comment focuses on variants of this model.

Clark and Kristensen’s contribution

Clark and Kristensen (2026) revisit our finding that a simple genetic model with phenotypic assortative mating cannot fit the data, and argue that with a simple augmentation, it can. We think this is a useful discussion. There is a large and growing body of work estimating genetic effects using direct data—from classical twin studies to more recent approaches using genome-wide association studies and polygenic scores. In contrast, much less attention has been paid to the question of whether a genetic model can fit wider kinship correlations in a *statistical* sense. This depends not only on the heritability of a trait, but also on the structure of assortative mating. Ultimately, a satisfactory account of familial resemblance requires a model that is consistent both with targeted causal evidence from, e.g., twin and adoption studies, and with the broader statistical pattern across many kinship types, as documented in our paper and now also in Clark and Kristensen’s analysis.

Clark and Kristensen’s key innovation is to allow for a distinction between the phenotype as observed by the researcher and the phenotype on which spouses match. In their genetic model, the observed outcome y is a noisy measure of an underlying latent phenotype w , with $y = \sqrt{\theta}w + u$. The parameter $\theta < 1$ implies that assortative mating may be much stronger in w than what is visible in y —the observed spousal correlation θr can be far below the latent sorting parameter r . This is a useful extension, and it is also consistent with the structure of our own baseline model, in which we allowed for sorting in both observable and latent characteristics and found that latent sorting is substantially stronger than what is visible in observable outcomes alone. It is furthermore commendable that Clark and Kristensen fit the model not only to the Swedish correlations from our paper, but also to a newly constructed set of kinship correlations from Danish register data—which, interestingly, turn out to be very similar. There is thus little disagreement about the empirical inputs, that is, about which kinship patterns a

successful model would need to capture.

Concern 1: The theoretical derivations

A key issue, however, is that the theoretical predictions of the model appear to be incorrectly derived. In the Appendix below, we provide a complete derivation of the kinship correlations implied by Clark and Kristensen’s model. The correct general formula is:

$$\rho = \theta r^k h^{2s} \left(\frac{1+m}{2} \right)^n \left(\frac{1+r}{1+m} h^2 \right)^{I(\text{parents})} \quad (1)$$

where k is the number of spousal links, s is the number of sibling links, n is the total number of consanguine links (both sibling and parent–child), and $I(\text{parents})$ is a dummy equal to one when all consanguine links are parent–child (i.e., $s = 0$ and $n \geq 1$).

The derivation proceeds by successive linear projections along the kinship chain. We show in the Appendix that each spousal link contributes a factor of r , each parent–child link a factor of $(1+m)/2$, and each sibling link a factor of $h^2(1+m)/2$. When all consanguine links are parent–child links (no sibling links present), there is an additional one-time correction factor of $(1+r)h^2/(1+m)$.

Clark and Kristensen’s Table 1 formulas, which we refer to as their “Baseline” specification, correspond instead to:

$$\rho_{\text{CK}} = \theta h^2 r^k \left(\frac{1+m}{2} \right)^n \left(\frac{1+r}{1+m} \right)^{I(\text{parents})} \quad (2)$$

This formula always carries a single factor of h^2 , regardless of how many sibling links appear in the kinship chain. In contrast, our derivation shows that the correct formula contains h^{2s} —one factor of h^2 per sibling link. The two formulas therefore agree for many common kinship types (spouses, siblings, parent–child, sibling-in-law, direct aunt/uncle, etc.), but disagree for kinship types with two or more sibling links. For example, for the “sibling of sibling-in-law” (path F-MF-M, involving two sibling links), Clark and Kristensen’s formula predicts $\theta h^2 r \left(\frac{1+m}{2} \right)^2$, while the correct prediction is $\theta h^4 r \left(\frac{1+m}{2} \right)^2$ —a difference by a factor of h^2 . In addition, Clark and Kristensen’s formulas appear to contain extra factors of r for certain uncle-type and cousin-type kinship categories (e.g., an extra r for indirect uncle/aunt correlations such as M←M-FM-F, and an extra r^2 for the corresponding cousin type M←M-FM-F→M). These additional discrepancies appear to arise from a misspecification of the maternal link when the relative is connected through the paternal side.

The main discrepancy can be traced to the treatment of sibling links. In the correct derivation, each link in the kinship chain contributes a multiplicative factor to the overall correlation: a spousal link contributes r , a parent–child link contributes $(1+m)/2$, and

a sibling link contributes $h^2(1 + m)/2$. Sibling and parent–child links share the factor $(1+m)/2$ —reflecting the boost from genetic assortative mating—but sibling links carry an additional factor of h^2 , because siblings’ phenotypes are related only through their shared genotype (see the Appendix for the detailed derivation). Clark and Kristensen appear to treat both sibling and parent–child links as contributing the same factor $(1 + m)/2$, without the additional h^2 for sibling links. This yields correct predictions when there is at most one sibling link, but underpredicts the attenuation for kinship types involving two or more sibling links.

It is difficult to verify exactly where the error originates, as Clark and Kristensen’s paper does not contain the derivations for the formulas in their Table 1. Clark and Kristensen themselves leave the door open to the possibility that their baseline formulas may not be exactly correct, as they consider two alternative “coding schemes” in Appendix B. The first, which they call “Two-marriage coding,” distinguishes between terminal and internal marriages. The second, which they call CIS coding, uses a different correlation structure for internal spousal links. However, neither alternative maps exactly into the correct model. The “CIS coding” column corrects the h^{2s} issue for most kinship types with two or more sibling links, but it drops the parent–child correction factor $(1 + r)/(1 + m)$, yielding wrong predictions for direct-line relationships (parent–child, grandparent–grandchild, etc.). Moreover, the CIS coding retains the extra factors of r noted above for certain indirect uncle and cousin types. We think the derivations underlying their formulas should be explicitly shown, so that the source and extent of the discrepancy can be resolved.

Concern 2: Empirical fit

Moreover, Clark and Kristensen’s augmented genetic model does not appear to fit the data well once the restrictions implied by the model are properly imposed. Clark and Kristensen’s main empirical specification is a log-linear regression (their equation 2):

$$\ln(\rho_{kln}) = \alpha + \beta_1 k + \beta_2 n + \beta_3 dphen$$

where k counts marital links, n counts genetic steps, and $dphen$ is a dummy for parent–child relations. Note that the correctly derived formula (1) implies a log-linear specification with an *additional* regressor, the number of sibling links s :

$$\ln(\rho) = \alpha + \beta_1 k + \beta_2 n + \beta_3 dphen + \beta_4 s$$

where $\beta_4 = \ln(h^2)$. Clark and Kristensen’s specification omits this term, which is a direct consequence of their baseline formulas carrying a single h^2 regardless of the number of sibling links. Including s as an additional regressor would, if anything, improve the fit of

the unrestricted regression further, since it adds a degree of freedom.

However, the more fundamental issue is that the unrestricted log-linear regression is not a valid test of the structural model. In the correct five-regressor specification, α determines θ , β_1 determines r , and β_4 determines h^2 , which leaves two restrictions: β_2 and β_3 must both be specific functions of h^2 and r (already pinned down by β_1 and β_4). The unrestricted OLS does not impose these constraints. Indeed, when the s term is included, the unrestricted estimate of β_4 implies $h^2 = \exp(0.035) = 1.036$ —a value exceeding one that is inconsistent with the structural model. When those restrictions are left free, the OLS regression has enough flexibility to approximate the data relatively well even if the underlying structural model is misspecified. Even so, the unrestricted OLS fit is substantially poorer than the fit of our baseline model, in particular for sibling and parent–child correlations (compare Clark and Kristensen’s Figure 5 with Figure 2 in Collado et al., 2023).

To properly assess the model’s fit, we estimate it by nonlinear least squares (NLS)—choosing the three structural parameters (θ, h^2, r) to minimize the sum of squared residuals across all kinship correlations, with each predicted correlation computed directly from the structural formula (1). We also estimate the model by restricted OLS—that is, fitting the log-linear specification while imposing the cross-equation constraints implied by three structural parameters. As Figures 1 and 2 show, both approaches yield similar results: the model systematically underpredicts correlations for close relatives—particularly siblings—and overpredicts for some intermediate kinship types. Clark and Kristensen estimate both OLS and NLS specifications in their Tables 3 and A4, but they do not show a corresponding figure of the NLS fit, which might make the limitations of the model’s fit more visible.

From Figures 1 and 2, three observations are worth noting. First, the introduction of the θ parameter does improve the fit considerably relative to the standard genetic model without this additional parameter (i.e., $\theta = 1$) that we considered in our original paper (Figure 6 in Collado et al., 2023), and we agree with Clark and Kristensen that this is a plausible and potentially important augmentation. Second, the fit is nonetheless far from satisfactory, particularly for siblings. Clark and Kristensen themselves note (p. 14) that siblings are the one kinship type where “the actual correlation exceeds the model prediction,” suggesting a family effect beyond what the genetic model can capture. Third, the model requires a very high heritability to fit the overall pattern: our restricted OLS and NLS estimates yield $h^2 \approx 0.75$ and $h^2 \approx 0.72$, respectively. These values substantially exceed the heritability of educational attainment estimated in most twin studies—typically around 0.40–0.45 (Branigan et al., 2013; Silventoinen et al., 2020)—and are even further above the variance explained by polygenic scores from genome-wide association studies, which typically account for only 10–15% of the variance in educational attainment.

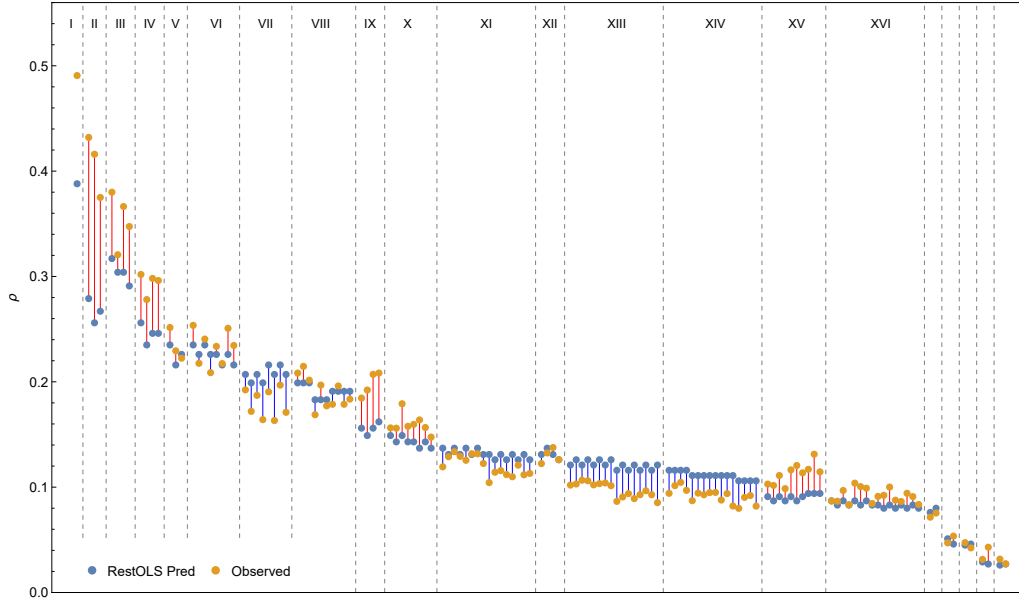


Figure 1: Clark and Kristensen’s augmented genetic model estimated by restricted OLS, using the correctly derived formula (1): predicted versus observed kinship correlations in years of education (Sweden).

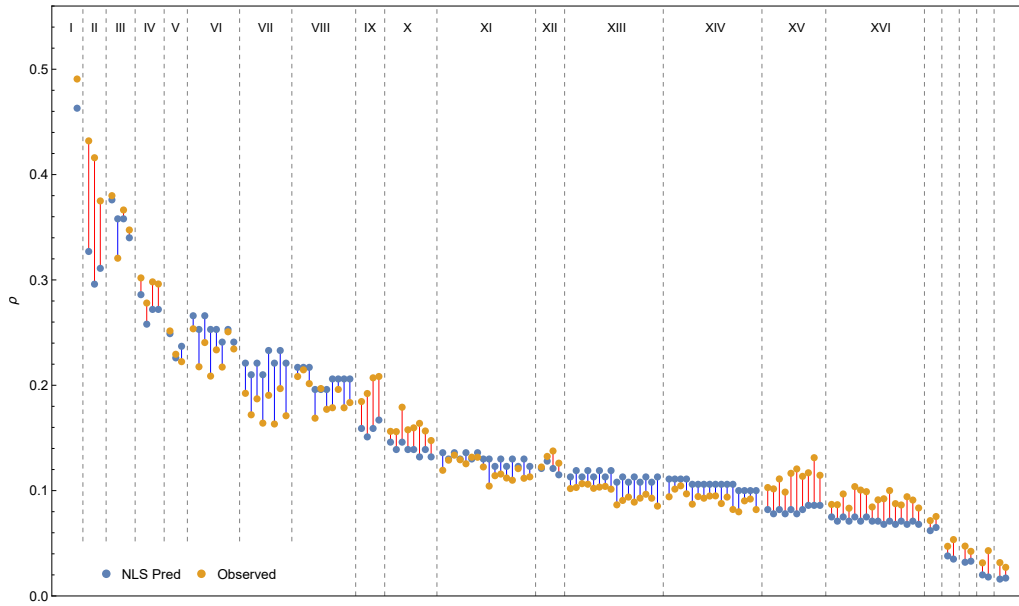


Figure 2: Clark and Kristensen’s augmented genetic model estimated by nonlinear least squares (NLS), using the correctly derived formula (1): predicted versus observed kinship correlations in years of education (Sweden).

More broadly, there is a fundamental difficulty in drawing causal conclusions from familial correlations alone. As Cavalli-Sforza and Feldman (1981) demonstrated, purely genetic and purely cultural vertical transmission models can produce identical patterns of familial correlations: “direct phenotypic transmission from parents to children could produce the appearance of high heritability even though the transmission was not genetic.”

This observational equivalence means that a good fit of a genetic model to kinship correlations could not, by itself, establish that the underlying transmission is genetic rather than cultural or social in nature.

Conclusion

To summarize, we think Clark and Kristensen pursue a valuable objective in exploring whether variants of the genetic model can account for the rich set of kinship correlations documented in our data. The introduction of the θ parameter—allowing for a wedge between observed and latent phenotypic sorting—is a natural and potentially important extension. However, two issues need to be addressed. First, the theoretical derivations underlying their predictions should be clarified and corrected. As we show in the Appendix, the correct formula differs from theirs for all kinship types involving two or more sibling links, and none of the alternative coding schemes considered in their Appendix B exactly corresponds to the correct model. Second, the fit of the model needs to be assessed using restricted estimation that properly imposes the structural constraints of the three-parameter model, rather than relying on the unrestricted log-linear OLS. When those restrictions are imposed, the augmented genetic model does not fit the Swedish kinship correlations nearly as well as the unrestricted OLS results suggest—particularly for siblings. Finally, even if the model did fit well, the potential observational equivalence between genetic and cultural vertical transmission models means that such a fit would not, by itself, establish that the underlying mechanism is genetic. We hope that this exchange will help clarify whether a properly specified genetic model can fit an extended set of kinship correlations.

References

- Branigan, Amelia R., Kenneth J. McCallum, and Jeremy Freese. 2013. “Variation in the Heritability of Educational Attainment: An International Meta-Analysis.” *Social Forces* 92(1): 109–140. <https://doi.org/10.1093/sf/sot076>
- Cavalli-Sforza, Luigi Luca, and Marcus W. Feldman. 1981. *Cultural Transmission and Evolution: A Quantitative Approach*. Princeton, NJ: Princeton University Press.
- Clark, Gregory. 2014. *The Son Also Rises: Surnames and the History of Social Mobility*. Princeton, NJ: Princeton University Press.
- Clark, Gregory, and Martin Hørlyk Kristensen. 2026. “The Mechanisms of Intergenerational Status Transmission in Modern Scandinavia.” EHES Working Paper No. 299, European Historical Economics Society. https://ehes.org/wp/EHES_299.pdf
- Collado, M. Dolores, Ignacio Ortuño-Ortín, and Jan Stuhler. 2023. “Estimating Intergen-

erational and Assortative Processes in Extended Family Data.” *Review of Economic Studies* 90(3): 1195–1227. <https://doi.org/10.1093/restud/rdac060>

Silventoinen, Karri, Aline Jelenkovic, Reijo Sund, et al. 2020. “Genetic and Environmental Variation in Educational Attainment: An Individual-Based Analysis of 28 Twin Cohorts.” *Scientific Reports* 10: 12681. <https://doi.org/10.1038/s41598-020-69526-6>

Appendix: Derivation of Kinship Correlations in Clark and Kristensen's Model

There is assortative mating in an unobserved phenotype that we denote by w and the spouses correlation in w is r . The observed phenotype y is a proxy for w

$$y_t = \sqrt{\theta}w_t + u_t$$

and therefore the spouses correlation in y is θr .

Transmission from parent to children is genetic

$$z_t = \frac{z_{t-1}^m + z_{t-1}^f}{2} + v_t$$

and the unobserved phenotype w is a function of genes

$$w_t = hz_t + \varepsilon_t$$

where $\text{Cov}(z_t, \varepsilon_t) = 0$ and $\text{Cov}(u_t, \varepsilon_t) = 0$. The variances of y_t , w_t and z_t have been normalized to 1, therefore, $h = \text{Cov}(w_t, z_t)$.

The linear projection of z_t on w_t is

$$z_t = hw_t + \omega_t$$

and the spouses correlation in genes is $m = h^2 r$.

Then, if a and a' are spouses, the cross correlations are

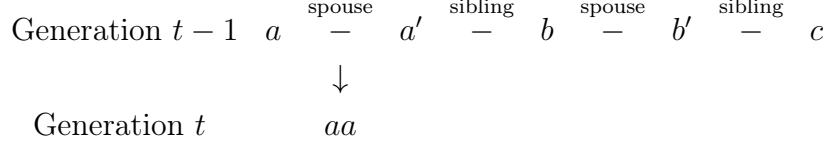
$$\begin{aligned} \text{Cov}(w_t^a, y_t^{a'}) &= \text{Cov}(w_t^a, \sqrt{\theta}w_t^{a'}) = \sqrt{\theta}r \\ \text{Cov}(w_t^a, z_t^{a'}) &= \text{Cov}(w_t^a, hw_t^{a'}) = hr \\ \text{Cov}(y_t^a, z_t^{a'}) &= \text{Cov}(\sqrt{\theta}w_t^a, hw_t^{a'}) = \sqrt{\theta}hr \end{aligned}$$

and the linear projections on the spouse w are

$$\begin{aligned} w_t^{a'} &= rw_t^a + \eta_t^w \\ y_t^{a'} &= r\sqrt{\theta}w_t^a + \eta_t^y \\ z_t^{a'} &= rhw_t^a + \eta_t^z \end{aligned}$$

where η_t^w , η_t^y and η_t^z are uncorrelated to w_t^a , y_t^a and z_t^a because the assortative mating is in w .

We next compute the correlation between relatives



Spouses $a-a'$

$$\text{Cov}(y_{t-1}^a, y_{t-1}^{a'}) = \text{Cov}\left(\sqrt{\theta}w_{t-1}^a, \sqrt{\theta}w_{t-1}^{a'}\right) = \theta r$$

Parent-child $aa-a'$

$$\begin{aligned}
\text{Cov}(y_t^{aa}, y_{t-1}^{a'}) &= \text{Cov}\left(\sqrt{\theta}hz_t^{aa}, \sqrt{\theta}w_{t-1}^{a'}\right) = \theta h \text{Cov}\left(\frac{z_{t-1}^a + z_{t-1}^{a'}}{2}, w_{t-1}^{a'}\right) \\
&= \theta h \frac{hr + h}{2} = \theta h^2 \frac{1+r}{2}
\end{aligned}$$

Siblings $a'-b$

$$\begin{aligned}
\text{Cov}(z_{t-1}^{a'}, z_{t-1}^b) &= \text{Cov}\left(\frac{z_{t-2}^m + z_{t-2}^f}{2}, \frac{z_{t-2}^m + z_{t-2}^f}{2}\right) = \frac{1+rh^2}{2} = \frac{1+m}{2} \\
\text{Cov}(z_{t-1}^{a'}, w_{t-1}^b) &= \text{Cov}(z_{t-1}^{a'}, h z_{t-1}^b) = h \text{Cov}(z_{t-1}^{a'}, z_{t-1}^b) = h \frac{1+m}{2} \\
\text{Cov}(z_{t-1}^{a'}, y_{t-1}^b) &= \text{Cov}(z_{t-1}^{a'}, \sqrt{\theta}h z_{t-1}^b) = \sqrt{\theta}h \text{Cov}(z_{t-1}^{a'}, z_{t-1}^b) = \sqrt{\theta}h \frac{1+m}{2} \\
\text{Cov}(w_{t-1}^{a'}, w_{t-1}^b) &= \text{Cov}(h z_{t-1}^{a'}, h z_{t-1}^b) = h^2 \text{Cov}(z_{t-1}^{a'}, z_{t-1}^b) = h^2 \frac{1+m}{2} \\
\text{Cov}(w_{t-1}^{a'}, y_{t-1}^b) &= \text{Cov}(w_{t-1}^{a'}, \sqrt{\theta}h z_{t-1}^b) = \sqrt{\theta}h^2 \frac{1+m}{2} \\
\text{Cov}(y_{t-1}^{a'}, y_{t-1}^b) &= \text{Cov}(\sqrt{\theta}w_{t-1}^{a'}, y_{t-1}^b) = \theta h^2 \frac{1+m}{2}
\end{aligned}$$

Siblings-in-law $a'-b'$

We have to project $y_{t-1}^{b'}$ on the next person in the chain, b . Given that b and a' are siblings, the correlation between them happens only through z , therefore, we have to project $y_{t-1}^{b'}$ on his/her spouse z_{t-1}^b .

$$y_{t-1}^{b'} = \text{Cov}\left(z_{t-1}^b, y_{t-1}^{b'}\right) z_{t-1}^b + \eta_t^{yb'zb} = \sqrt{\theta}hr z_{t-1}^b + \eta_t^{yb'zb}$$

where the error term is uncorrelated to z_{t-1}^b and therefore to his/her sibling $z_{t-1}^{a'}$ and $y_{t-1}^{a'}$. Then

$$\text{Cov}(y_{t-1}^{a'}, y_{t-1}^{b'}) = \text{Cov}\left(y_{t-1}^{a'}, r\sqrt{\theta}h z_{t-1}^b\right) = r \text{Cov}\left(y_{t-1}^{a'}, y_{t-1}^b\right)$$

Sibling of sibling-in-law $a'-c$

We have to project y_{t-1}^c on the next person in the chain, b' . Given that b' and b are spouses, the correlation between them happens only through w , therefore, we have to project y_{t-1}^c on his/her sibling $w_{t-1}^{b'}$.

$$y_{t-1}^c = \text{Cov}\left(w_{t-1}^{b'}, y_{t-1}^c\right) w_{t-1}^{b'} + \eta_t^{ycwb'} = \sqrt{\theta} h^2 \frac{1+m}{2} w_{t-1}^{b'} + \eta_t^{ycwb'}$$

where the error term is uncorrelated to $w_{t-1}^{b'}$ and therefore to his/her spouse b and to $y_{t-1}^{a'}$. Then

$$\begin{aligned} \text{Cov}(y_{t-1}^{a'}, y_{t-1}^c) &= \text{Cov}\left(y_{t-1}^{a'}, \sqrt{\theta} h^2 \frac{1+m}{2} w_{t-1}^{b'}\right) \\ &= \text{Cov}\left(y_{t-1}^{a'}, h^2 \frac{1+m}{2} y_{t-1}^{b'}\right) = h^2 \frac{1+m}{2} \text{Cov}\left(y_{t-1}^{a'}, y_{t-1}^{b'}\right) \end{aligned}$$

Uncles $aa-b$

We have to project y_t^{aa} on the next person in the chain, a' . Given that a' and b are siblings, the correlation between them happens only through z , therefore, we have to project y_t^{aa} on his/her father/mother $z_{t-1}^{a'}$.

$$y_t^{aa} = \text{Cov}(y_t^{aa}, z_{t-1}^{a'}) z_{t-1}^{a'} + \eta_t^{aa a'}$$

where the error term is uncorrelated to $z_{t-1}^{a'}$, and therefore to his/her sibling y_{t-1}^b .

We then compute

$$\begin{aligned} \text{Cov}(y_t^{aa}, z_{t-1}^{a'}) &= \text{Cov}\left(\sqrt{\theta} h z_t^{aa}, z_{t-1}^{a'}\right) = \sqrt{\theta} h \text{Cov}\left(\frac{z_{t-1}^a + z_{t-1}^{a'}}{2}, z_{t-1}^{a'}\right) \\ &= \sqrt{\theta} h \frac{\text{Cov}(z_{t-1}^a, z_{t-1}^{a'}) + \text{Cov}(z_{t-1}^{a'}, z_{t-1}^{a'})}{2} = \sqrt{\theta} h \frac{1+m}{2} \end{aligned}$$

and

$$\text{Cov}(y_t^{aa}, y_{t-1}^b) = \text{Cov}\left(\sqrt{\theta} h \frac{1+m}{2} z_{t-1}^{a'}, \sqrt{\theta} h z_{t-1}^b\right) = \frac{1+m}{2} \text{Cov}\left(y_{t-1}^{a'}, y_{t-1}^b\right)$$

General formula

We have shown that any correlation can be obtained from the three basic correlations (spouses, parents, and siblings) by multiplying by r for any spouse link added to the chain, by $\frac{1+m}{2}$ for any parent-child link, and by $h^2 \frac{1+m}{2}$ for any sibling link. The general

formula to compute all kinship correlations is therefore

$$\rho = \theta r^k h^{2s} \left(\frac{1+m}{2} \right)^n \left(\frac{1+r}{1+m} h^2 \right)^{I(\text{parents})}$$

where k is the number of spousal links, s is the number of sibling links, n is the total number of consanguine links (siblings and parent–child), and $I(\text{parents})$ is a dummy that takes value 1 only for direct parent–child relationships (i.e., $s = 0$ and $n \geq 1$).