Fiona Steele, Paul Clarke and Jouni Kuha

Modeling within-household associations in household panel studies

Article (Accepted version) (Refereed)

Submitted to the Annals of Applied Statistics

MODELING WITHIN-HOUSEHOLD ASSOCIATIONS IN HOUSEHOLD PANEL STUDIES

By Fiona Steele[*], Paul Clarke[†] and Jouni Kuha[*]

London School of Economics & Political Science[*] *and University of Essex*[†]

Household panel data provide valuable information about the extent of similarity in coresidents' attitudes and behaviours. However, existing analysis approaches do not allow for the complex association structures that arise due to changes in household composition over time. We propose a flexible marginal modeling approach where the changing correlation structure between individuals is modeled directly and the parameters estimated using second-order generalized estimating equations (GEE2). A key component of our correlation model specification is the 'superhousehold', a form of social network in which pairs of observations from different individuals are connected (directly or indirectly) by coresidence. These superhouseholds partition observations into clusters with nonstandard and highly variable correlation structures. We thus conduct a simulation study to evaluate the accuracy and stability of GEE2 for these models. Our approach is then applied in an analysis of individuals' attitudes towards gender roles using British Household Panel Survey data. We find strong evidence of between-individual correlation before, during and after coresidence, with large differences among spouses, parent-child, other family, and unrelated pairs. Our results suggest that these dependencies are due to a combination of non-random sorting and causal effects of coresidence.

1. Introduction. In the social sciences, there is considerable interest in studying dependencies in the attitudes and behaviors of members of the same household. Previous research on couples suggests that such dependencies can be mainly explained by homogamy or a causal effect of coresidence (Brynin, Longhi and Martínez Pérez, 2008; Butterworth and Rodgers, 2006; Davillas and Pudney, 2017). Homogamy is a form of assortative mating wherein individuals select partners with similar social, cultural and demographic characteristics (e.g. Blackwell and Lichter, 2004; Kalmijn, 1998), and is a special case of homophily which refers to the tendency for people to form social connections with people like themselves (McPherson, Smith-Lovin and Cook, 2001). In contrast, a causal effect of coresidence arises when

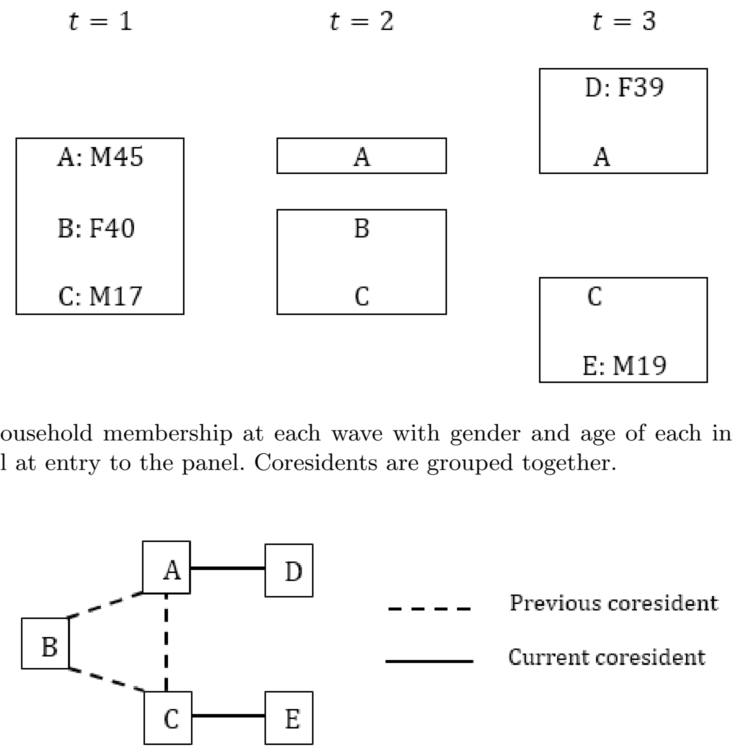
MSC 2010 subject classifications: Primary 60K35, 60K35; secondary 60K35

Keywords and phrases: household effects, household correlation, longitudinal households, homophily, multiple membership multilevel model, marginal model, generalised estimating equations

the (possibly reciprocal) influence of one coresident partner on another, and shared experiences and influences of family, friends and lifestyle factors, causes their attitudes and behaviors to converge over time. There is evidence of couple concordance in social and political attitudes (Brynin, Longhi and Martínez Pérez, 2008) and health indicators (Davillas and Pudney, 2017). Moving beyond couples to include other coresidents, within-household correlations have been found across a range of individual outcomes such as voting in political elections (Johnston et al., 2005), happiness and well-being (Ballas and Tranmer, 2012), and self-rated health (Chandola et al., 2003; Sacker, Wiggins and Bartley, 2006).

In this paper, we examine individuals' views about the relative contributions made by men and women to household income and looking after home and family, using longitudinal data from the British Household Panel Survey (BHPS). Previous studies of attitudinal change have focused on either the effects of individual and household characteristics on these attitudes (Berridge, Penn and Ganjali, 2009; Sweeting et al., 2014), or the concordance of attitudes within couples (Brynin, Longhi and Martínez Pérez, 2008). We extend this work to explore whether the association between an individual's attitudes and those of other individuals changes before, during and after they were members of the same household, and test whether the similarity found between married and cohabiting partners found elsewhere (Brynin, Longhi and Martínez Pérez, 2008) extends to parents and their children, other family pairs, and unrelated sharers.

The main methodological challenges are estimating covariate effects on individual outcomes and modeling the complex association structures for these outcomes. This complexity arises because of the changes in household composition over time following, for example, union formation and dissolution, and children leaving or returning to the parental home. Such changes are commonly reflected in the design of household panel studies, which follow the original sample members and their new coresidents. This causes problems because household clusters are defined entirely in terms of their members (usually as groups of people sharing living accommodation or one meal a day). In contrast, clusters such as schools and areas are uniquely identified entities which remain fixed no matter what membership changes occur. Hence, while identifiers for fixed entities like schools and areas are naturally time-invariant, it is unclear how to define "longitudinal households" and attempts to do so have been described as "futile" (Duncan and Hill, 1985). An alternative view of households, which we adopt in this paper, is as "evolving social networks" (Murphy, 1996). Figure 1 illustrates the formation of such a network over three waves. At the first wave, there is one household consisting of a couple (A,B) and their son C (Figure 1a). The couple has split by the second wave, with the man A forming a new single-person household and the woman B remaining with her son C. By the third wave, A has formed a new partnership with D, C has left home to live with friend E, and B's household does not respond. The network at wave 3, containing all five individuals, is shown in the graph of Figure 1b. Other examples of clusters that could also be viewed as evolving networks are peer groups, defined as children taught in the same class or living in the same neighbourhood, and friendship networks.



(a) Household membership at each wave with gender and age of each individual at entry to the panel. Coresidents are grouped together.

(b) Network members with coresidence status at $t = 3$.

Fig 1: Illustration of the evolution of a household network over 3 waves.

The difficulty in defining longitudinal households is reflected in the methods commonly used for panel data analysis. The standard approach to the analysis of an individual-level outcome is simply to ignore household effects, and account for changes in coresidents through the inclusion of covariates which index these changes. Those studies which have considered household effects have focused on outcomes from one wave (Chandola et al., 2003; Ballas and Tranmer, 2012; Johnston et al., 2005), or restricted analysis to households (usually couples) that have remained together for the entire observation period (e.g. Keizer and Schenk, 2012). The first of these approaches does not fully exploit the available panel data, while the second leads to highly selective analysis samples in long panels.

To date, the only approach explicitly allowing changes in household composition is the multiple membership random effects model (Goldstein et al., 2000). We argue that this approach is too restrictive because it constrains the association structure among coresidents in an unrealistic way and, more generally, that random effects models are less suitable when clusters are defined entirely by their members. Instead, we propose a more flexible marginal modeling approach that allows us to directly model and estimate the association structure between coresidents. By taking the individual as the unit of analysis, and incorporating household-composition changes directly into the association model, it is unnecessary to define longitudinal households, or restrict analyses to fixed-membership households. Our approach thus reflects the view that a household panel study is "a study of individuals in their changing household contexts" (Buck and McFall, 2012, p.7).

2. Panel models.

We now introduce notation and set out a general panel model for the mean outcome and the between-outcome covariance structure for panel data on individuals and their coresidents. Random effects and marginal formulations of this model are respectively described in Sections 3 and 4.

Let Y_{ti} be the outcome at wave t $(t = 1, \ldots, T)$, and $\mathbf{Y}_i = (Y_{1i}, \ldots, Y_{Ti})'$ the vector of all outcomes, for individual i $(i = 1, \ldots, n)$. We make the usual simplifying assumption that all individuals are interviewed at the same point in calendar time, and that every between-wave interval is of equal length. The outcomes are taken to follow the marginal model

$$Y_{ti} = \mu_t(\mathbf{x}_{ti}) + r_{ti} \tag{1}$$

where $\mathbf{x}_{ti}$ is a vector of explanatory variables, $\mu_t(\mathbf{x}_{ti}) = \mathrm{E}(Y_{ti}|\mathbf{x}_{ti})$ is the mean outcome, and r_{ti} is the zero-mean model residual. For the application in this paper, we take the mean outcome to follow the linear model

$$\mu_t(\mathbf{x}_{ti}) = \mathbf{x}'_{ti}\boldsymbol{\beta} \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients. We are equally interested in the between-outcome covariances

$$\mathrm{cov}(Y_{ti}, Y_{t'i'}|\mathbf{x}_{ti}, \mathbf{x}_{t'i'}) = \sigma_{ti,t'i'}(\mathbf{x}_{ti}, \mathbf{x}_{t'i'}), \tag{3}$$

including the pairs where $i = i'$ or $t = t'$. Note that μ_t and $\sigma_{ti,t'i'}$ may involve different components of $\mathbf{x}_{ti}$ because each model has separate covariates of direct substantive interest.

The way in which household effects are accounted for in this model depends on the modeling approach used. For the random effects models which we discuss in Section 3, household enters through the decomposition of residual r_{ti} into distinct components which include one for household. Conversely, for the marginal models we develop in Section 4, household enters through parameterizing $\sigma_{ti,t'i'}$ to reflect whether and how individuals $i \neq i'$ are connected by coresidence.

Whichever modeling approach is used, we consider nonzero covariances to be plausible in two situations: for variances and (auto)covariances on the same individual (that is, when $i = i'$); and for different individuals connected through having been residents in the same household(s). We elaborate on this below, but first we introduce the coresidence-status indicator for i and i' at wave t:

$$c_t(i, i') = \begin{cases} 1 & \text{if } i \text{ and } i' \text{ are coresident at wave } t \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $c_t(i, i') = 1$ if $i' = i$. The members of individual i's household at wave t can thus be denoted by the set $M_{ti} = \{i' : c_t(i, i') = 1\}$, where $M_{ti} = M_{ti'}$ for all coresident pairs (i, i'). The households are labelled $h = 1, \ldots, H$ where $h_{ti} \in \{1, \ldots, H\}$ is the label for M_{ti} such that $h_{ti} = h_{ti'}$ for all coresidents.

3. Random effects models.

We first review and critique random effects specifications of (1)–(3). We begin with hierarchical models that incorporate household effects for the situation where the composition of each individual's household remains fixed for the duration of the panel. This is followed by a description of non-hierarchical multiple membership models that allow for changes in household membership over time.

3.1. *Three-level models for fixed household membership.*

Consider first the special case where each household remains fixed across waves, so that $h_{ti} = h_i$ for all t, i. Conventional random effects models can then be used to fit (1)–(3) and its nonlinear generalisations (Goldstein, 2010). In particular, the classical three-level hierarchical model takes observations Y_{ti} to be nested within individuals, and individuals within the unchanging households. It decomposes the residual term in (1) as $r_{ti} = e_{ti} + a_i + u_{h_i}$ where e_{ti} is an outcome-specific residual, a_i an individual-level random effect and u_{h_i} the random effect for household h_i, and all of these terms are taken to have zero mean and to be homoscedastic and mutually uncorrelated. Letting $\sigma_u^2 = \text{var}(u_{h_i})$ and $\sigma_a^2 = \text{var}(a_i)$, the conditional covariances are then

$$\sigma_{ti,t'i'}(\mathbf{x}_{ti}, \mathbf{x}_{t'i'}) = \mathrm{I}(i = i')\, \sigma_a^2 + I(h_i = h_{i'})\, \sigma_u^2. \tag{5}$$

This implies that there are two kinds of nonzero conditional dependencies of the outcomes: the within-individual autocovariances $\sigma_{ti,t'i}(\mathbf{x}_{ti}, \mathbf{x}_{t'i}) = \sigma_a^2 + \sigma_u^2$ for all $t \neq t'$, and the within-household covariances $\sigma_{ti,t'i'}(\mathbf{x}_{ti}, \mathbf{x}_{t'i'}) = \sigma_u^2$ between all individuals $i \neq i'$ in the same household.

Three-level random effects models have been proposed for the analysis of repeated-measures data on individuals in couples or families (Atkins, 2005). An alternative but closely related approach suitable for couples and other family dyads is a bivariate two-level model (Raudenbush, Brennan and Barnett, 1995). These approaches have been applied widely in couple research, with analyses based on household panel data restricted to individuals who remain with the same partner throughout the observation period (e.g. Keizer and Schenk, 2012). A three-level model was also used in a study of household effects that included all adult respondents, rather than only couples, but the treatment of households which change over time was not discussed (Milner et al., 2014).

Another type of three-level model is the dynamic group model which includes time-varying group-level random effects (Bauer et al., 2013). While changing group membership is potentially a reason to allow time-varying random effects, the application of these models is limited to groups defined by entities that remain fixed over time even if group membership changes. Moreover, dynamic group models were explicitly formulated to answer research questions about temporal patterns in group effects over time - for example, the stability of school effects in Leckie and Goldstein (2009) - whereas our focus is specifically on questions concerning correlations between group members.

3.2. *Multiple membership models for time-varying household membership.* We now return to the general situation where the composition of an individual's household may change over time. The only approach which has up to now been available for this case is multiple membership random effects modeling as proposed by Goldstein et al. (2000). These models are again based on decomposing the residual r_{ti} in (1) as $r_{ti} = e_{ti} + a_i + u_{ti}$, where u_{ti} is now a random effect for the time-varying household. It is specified as

$$u_{ti} = \sum_{h=1}^{H} w_{h(ti)} u_h^* \tag{6}$$

where the u_h^* are identically distributed zero-mean random variables for all the distinct households, taken to have $\mathrm{var}(u_h^*) = \sigma_u^2$ and to be independent of each other and of the explanatory variables $\mathbf{x}$. The weights $w_{h(ti)}$ are specified by the analyst and are nonzero if i was a member of h for at least one wave, and zero otherwise. In other words, the household effect u_{ti} for an individual i at wave t is formed as a weighted sum of effects contributed by the individual's different households over time. If the variances of e_{ti} and a_i are constant, the variance of r_{ti} is also constant if $\sum_h w_{h(ti)}^2 = 1$ (Goldstein et al. (2000) assumed, instead, that $\sum_h w_{h(ti)} = 1$).

In a multiple membership model, the conditional dependencies of observations Y_{ti} are implied by the choice of the weights $w_{h(ti)}$. For instance, suppose that the household effect u_{ti} is defined as an equally weighted sum of the effects of the d_{ti} distinct households that individual i has belonged to in the observed waves up to t, so that $w_{h(ti)} = 1/\sqrt{d_{ti}}$ for these households and $w_{h(ti)} = 0$ otherwise. Then, for example, the conditional correlation between Y_{ti} and $Y_{ti'}$ is proportional to $s_{t(ii')}/\sqrt{d_{ti}d_{ti'}}$, where $s_{t(ii')}$ is the number of distinct households that individuals i and i' have shared up to wave t.

Such implied correlation structures are not always substantively satisfactory. In the case introduced above, for example, the correlation depends on how many households the individuals have shared, but not *when* this sharing took place. It also depends on the total numbers of households that have so far been observed for the individuals. Since this can be no larger than the current wave t, for a given $s_{t(ii')}$ the correlation is often higher for early waves of the study than for later ones. Each of these features could be changed by modifying the specification of the weights, but any such choice would introduce problems of its own. We are not aware of any way of defining a multiple membership household effect (6) which would not imply counterintuitive patterns of association in some situations.

Even if a multiple membership model always gave coherent associations, it would still be poorly suited to our goals. This is because, as observed by Prentice and Zhao (1991, p. 827), "a given random effects model and distributional assumption implies a corresponding covariance structure on the response vector. This structure may involve a parameterization that is not sufficiently flexible or interpretable, especially if the covariances are of substantive interest". In our application the changing association structure between individuals *is* of interest, and we want to model it directly and to examine specific hypotheses about it. Instead of random effects models, this goal is better achieved by using marginal modeling. In the rest of this article we focus on this approach for data with time-varying household membership.

4. A marginal modeling approach.

We now set out a marginal modeling approach for household panel data. The joint model comprises distinct marginal models for the mean of each outcome, the variance of these outcomes, and the pairwise correlations between the outcomes of different individuals (or of the same individual at different waves). A key component of the correlation model specification is what we refer to as a superhousehold. This is an artificial group constructed to contain individuals whose outcomes are potentially correlated because they have experienced shared influences from the same (cross-sectional) households over time; conversely, there is no correlation between individuals in different superhouseholds. In this way, superhouseholds impose a loose cluster structure on the correlation matrix but, in contrast to standard marginal models for panel data, the within-cluster structure can vary between superhouseholds.

4.1. *Specification of the marginal models.*

Suppose there are n_k individuals and m_k person-wave observations in cluster k $(k = 1, \ldots, K)$ (noting that $\sum_k n_k = n$), and let Y_{tik} be the response at wave t for person i in cluster k. (The definition of the superhouseholds used to determine these clusters is deferred until Section 4.3.) Define the conditional expectation, scale and pairwise correlation as $\mu_{tik} = \mathrm{E}(Y_{tik}|\mathbf{x}_{1,tik})$, $\phi_{tik} = \mathrm{var}(Y_{tik}|\mathbf{x}_{2,tik})/v_{tik}$ and $\rho_{tik,t'i'k} = \mathrm{cor}(Y_{tik}, Y_{t'i'k}|\mathbf{x}_{3,tik,t'i'k})$ where v_{tik} is the variance function. The covariate vectors $\mathbf{x}_{1,tik}$ and $\mathbf{x}_{2,tik}$ may contain a mix of time-varying and individual-specific characteristics, while $\mathbf{x}_{3,tik,t'i'k}$ may contain variables that characterise the pair of person-wave observations $(ti, t'i')$, for example the coresidence status of individuals i and i' at waves t and t'.

Collating for cluster k, we let $\mathbf{Y}_k = (Y_{11k}, ..., Y_{Tn_kk})$ be the $m_k \times 1$ response vector for cluster k, and $\boldsymbol{\mu}_k$, $\boldsymbol{\phi}_k$ and $\boldsymbol{\rho}_k$ be the corresponding $m_k \times 1$ mean and scale vectors and $m_k(m_k - 1)/2 \times 1$ correlation vector, respectively. We further let $\mathbf{X}_{1k}$, $\mathbf{X}_{2k}$ and $\mathbf{X}_{3k}$ be the covariate matrices for the mean, variance and correlation functions respectively. Following Yan and Fine (2004), we specify generalized linear models for the marginal conditional expectation, scale and correlation of $\mathbf{Y}_k$ as

$$g_1(\boldsymbol{\mu}_k) = \mathbf{X}_{1k}\boldsymbol{\beta} \tag{7}$$

$$g_2(\boldsymbol{\phi}_k) = \mathbf{X}_{2k}\boldsymbol{\gamma} \tag{8}$$

$$g_3(\boldsymbol{\rho}_k) = \mathbf{X}_{3k}\boldsymbol{\alpha} \tag{9}$$

where $g_1(\cdot)$, $g_2(\cdot)$ and $g_3(\cdot)$ are link functions and $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are parameter vectors. To ensure positive variance estimates and correlation estimates in the interval $(-1, 1)$, common choices for $g_2(\cdot)$ and $g_3(\cdot)$ are, respectively, the exponential and hyperbolic tangent functions. However, when $\mathbf{X}_{2k}$ and $\mathbf{X}_{3k}$

consist only of indicator variables, identity links may be adequate, and lead to more interpretable parameters. In our application (see Section 7), the parameters of (7) and (9) are of primary substantive interest and therefore $\mathbf{X}_{2k}$ is specified as a constant vector; $\mathbf{X}_{3k}$ will contain a set of characteristics for each pair of observations $(Y_{tik}, Y_{t'i'k})$ in cluster k, including indicators that distinguish between observations on the same person $(t < t', i = i')$ or on two coresidents $(t \leq t', i \neq i')$; for coresidents, indicators are also defined to denote their coresidence status (future, current or past) at t and t'.

4.2. *Estimation.* We use an extension of the generalized estimating equations (GEE) approach of Liang and Zeger (1986) for simultaneous estimation of the marginal mean and association structure of a multivariate response (Prentice and Zhao, 1991; Liang, Zeger and Qaqish, 1992). This approach, commonly referred to as second-order GEE (GEE2), is appropriate in situations where the association structure is of primary substantive interest. The advantage of standard first-order GEE over GEE2 is that it produces estimates of the mean parameters which are robust to incorrectly specified covariance models, but this is only an advantage if the parameters of the covariance matrix are not of substantive interest.

We adopt the approach of Yan and Fine (2004) by modeling the association structure with separate estimating equations for the scale and correlation. One advantage of modeling correlations rather than covariances is that correlation parameters have a more natural interpretation. The system of three estimating equations for the mean, scale and correlation parameters is

$$u(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \begin{pmatrix} D_{1k} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & D_{2k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & D_{3k} \end{pmatrix}' \begin{pmatrix} V_{1k} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & V_{2k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & V_{3k} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{Y}_k - \boldsymbol{\mu}_k \\ \mathbf{s}_k - \boldsymbol{\phi}_k \\ \mathbf{r}_k - \boldsymbol{\rho}_k \end{pmatrix}$$

where $\mathbf{s}_k$ and $\mathbf{r}_k$ are the vectors of empirical variances and pairwise correlations, D_{1k}, D_{2k} and D_{3k} are matrices of first derivatives of $\boldsymbol{\mu}_k$, $\boldsymbol{\phi}_k$ and $\boldsymbol{\rho}_k$ with respect to parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, and V_{1k}, V_{2k} and V_{3k} are the conditional working covariance matrices of $\mathbf{Y}_k$, $\mathbf{s}_k$ and $\mathbf{r}_k$. The equations may be solved using a modified Fisher scoring algorithm which has been implemented in the R package `geepack` (Højsgaard, Halekoh and Yan, 2006). Standard errors may be obtained using jacknife variance estimators, but less computationally intensive robust sandwich variance estimators were found to perform well in the simulation study and application that follow.

4.3. *Superhousehold definition.* We now formally define the superhousehold clusters for the joint marginal model. As discussed in Section 1, individuals changing household over time can be thought of as an "evolving network" (Murphy, 1996). In this context, the social network evolves as individuals move from their original households at entry to the panel, and we can define a superhousehold by grouping together individuals who are connected by having ever been coresident. Figure 1 made clear that connections between individuals can be represented by a network graph in which the coresidence of two individuals is indicated by an edge between them. In general, a superhousehold is a group of individuals linked by pathways of edges in the network graph; if no such pathway can be found between a pair of individuals then they must necessarily be in different superhouseholds. A pathway will thus exist if the pair were coresident at any wave (a direct connection), or if they were never coresidents but one them was coresident with a third person who was ever coresident with the other member of the pair (an indirect connection). In Figure 1, for example, B and D have never lived together, but are indirectly connected through their coresidence with A at different waves. The cluster at t contains all observations contributed by this set of individuals for waves $1, \dots, t$. The clusters in the model specified by (7)–(9) correspond to the superhouseholds at the final observed wave T.

Using this formulation, we can focus on parameterizing the correlation between individuals in superhouseholds, in line with how each pair is connected. The construction of a network graph and the identification of superhousehold clusters only requires that we are able to identify the (cross-sectional) coresidence status of individuals at each wave. In contrast to the three-level modeling approach, we do not need to choose between unsatisfactory definitions of a longitudinal household.

More generally, denote by $\mathcal{N} = (\mathcal{V}, \mathcal{E})$ the undirected network graph at wave T, where $\mathcal{V} = \{1, ..., n\}$ is the set of vertices/individuals, $\mathcal{E} = \{C_T(i, i') : \text{for all } i \neq i' \in \mathcal{V}\}$ is the set of edges between them, and

$$C_t(i, i') = \begin{cases} 1 & \text{if } \sum_{t'=1}^{t} c_{t'}(i, i') > 0, \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

for $t = 1, ..., T$ is the superhousehold coresidence indicator at wave t. Using this notation, the members of individual i's superhousehold are $S_i = \{i' : \text{path}_{\mathcal{N}}(i, i') = 1\}$, where $\text{path}_{\mathcal{N}}(i, i') = 1$ if $\mathcal{N}$ contains a pathway between i and i' or else zero, and $\text{path}_{\mathcal{N}}(i, i) = 1$. If the superhousehold clusters are indexed by k, then the index of superhousehold S_i can be denoted by k_i such that $k_i = k_{i'}$ if $S_i = S_{i'}$ and $k_i \neq k_{i'}$ if $S_i \neq S_{i'}$.

An important point to note is that the edges in $\mathcal{N}$ represent the presence of *conditional* associations between pairs of individuals given the rest, and not pairwise-marginal associations. This is not problematic for normally distributed residuals because although conditional covariances are nonzero for pairs connected by pathways in $\mathcal{N}$, and zero otherwise (e.g. Jones and West, 2005), this implies that all marginal correlations between pairs of individuals within the same superhousehold are nonzero.

4.4. *Positive definite correlation matrices.* As noted above, we can ensure that the GEE2 estimator of $\boldsymbol{\alpha}$ yields estimates of the pairwise correlations in the $(-1, 1)$ range by using a hyperbolic tangent link function. Nevertheless, the GEE2 estimator does not constrain the fitted correlation matrix under model (9) to be positive definite. While other approaches are available which could potentially do this, we argue that these are unsuitable for the present application, in which household transitions lead to superhouseholds with distinct and unpatterned correlation structures.

Within the GEE framework, quasi least squares regression can be used to ensure the fitted correlation matrix is feasible in the sense of being positive definite (Chaganty, 1997). However, the analyst must derive bounds for the correlations based on the structure imposed on the correlation matrix. Bounds have been derived for a range of longitudinal and nested structures (Shults and Hilbe, 2014), but in our case the feasible parameter space cannot easily be calculated as superhouseholds do not have patterned correlation structures. Most other approaches are based on constrained maximum likelihood estimation of joint mean-covariance models (e.g. Jennrich and Schluchter, 1986; Pourahmadi, 1999), but to date the implementation of these methods has been confined to patterns where the form of the within-cluster covariance matrix is the same across clusters. More general approaches present substantial computational challenges (Chiu, Leonard and Tsui, 1996), or are suitable for observations with a spatial or temporal structure (Gneiting, 2002).

We thus propose to use GEE2, despite its not being able to guarantee positive-definite correlation matrices, because our substantive interest lies in obtaining accurate estimates of the population correlation parameters rather than the correlation matrix for each superhousehold. If our correlation model were correctly specified (or at least not badly mis-specified) then the impact of failing to constrain our estimates to produce positive-definite correlation matrices would be an estimator with increased bias and lower precision, but we could interpret our correlation-parameter estimates in the same way even if one or more of the estimated superhousehold correlation matrices were not positive definite. Conversely, if the correlation model were badly mis-specified, Crowder (1995) showed that a mis-specified working correlation matrix could result in non-positive-definite correlation matrices and inconsistent joint estimators for which the usual asymptotic results do not hold. However, as Crowder (p.410) suggests, "in practice, statistical judgement would normally be employed in an attempt to avoid such hidden pitfalls."

We investigate this issue empirically in the simulation study by considering the frequency with which non-positive definite estimates of the superhousehold correlation matrices occur and, in the application to gender role attitudes, by checking whether the estimated superhousehold correlation matrices are positive definite or not (in the event, every correlation matrix was estimated to be positive definite). Further discussion of positive-definite correlation matrices can be found in supplementary materials.

4.5. *Missing data.* In the presence of missing data, our estimates use the data from all respondent-wave observations where $(Y_{ti}, \mathbf{x}_{ti})$ are completely observed. GEE estimators based on these data are generally consistent only if the data are missing completely at random, that is, nonresponse is independent of any variable (covariate or outcome) in the model. Inferences from missing at random data, where nonresponse depends only on the values of the observed variables, can be obtained by incorporating a model for the nonresponse probability into the estimating equations (Robins, Rotnitzky and Zhao, 1995).

5. British Household Panel Survey. Our data are from the British Household Panel Survey (BHPS), which began in 1991 with 10,300 adult (aged 16 or over) residents in 5,500 households (ISER, 2009). These original sample members (OSMs) are followed up and interviewed annually. People who form households with OSMs after 1991 are referred to as temporary sample members, unless they have children with OSMs in which case they become permanent sample members (PSMs); children of OSMs also become PSMs after turning 16. Like OSMs, PSMs are then followed regardless of whether they remain coresident with an OSM. Tracking of OSMs and PSMs and their households allows identification of correlations between future, current and previous coresidents. We use data from waves 1-17 between 1991 and 2008.

5.1. *Household structures.* A major motivation for our modeling approach is that changes in household composition over time lead to complex non-hierarchical structures where person-wave observations are not nested in households. In BHPS, 12.7% of adult members of a sample household for two consecutive waves experience a change in their adult coresidents between waves t and $t+1$ ($t = 1, \ldots, 16$; $n = 204,367$ person-waves), where an adult is defined as a potential BHPS respondent. The proportion experiencing a change increases to 31.8% over a 5-year interval ($n = 106,795$), 44.0% for a 10-year interval ($n = 41,708$) and 50.5% for a 15-year interval ($n = 9,779$). Thus there is appreciable churn in household membership during the period of observation. Further analysis of the types of event that lead to household change is given in supplementary material (Table S1).

The superhouseholds we use to define clusters are created using cross-wave individual and cross-sectional household identifiers. These identifiers allow us to infer the coresidence of any pair of respondents at waves t and t'. In practice, the construction of superhouseholds is challenging because household panel studies have complex designs. In particular, any algorithm must account for new entrants (from single individuals to entire households) at each wave, individuals who rejoin previous coresidents (e.g. children returning to the parental home), and wave non-response of households and individuals within households. Further details are provided in the supplementary material.

5.2. *Response variable and covariates for the mean and correlation functions.* The response variable is an index of attitudes towards gender roles obtained from a principal components analysis of six ordinal items. The items measure strength of agreement with statements such as “family life suffers when the woman has a full time job” and “both the husband and wife should contribute to the household income.” The response is the standardized score for the first principal component, with high values corresponding to more egalitarian attitudes. These questions were asked of the adult respondents every two years so that each individual is observed for a maximum of nine waves (although the mean is 3.9 waves due to a combination of late entry into the study, wave nonresponse and attrition). The analysis sample contains 27,033 adult individuals who contribute 106,060 person-wave observations. There are 11,460 superhousehold clusters at wave 17, ranging in size from 1 to 100 person-wave observations (mean=9.4, SD=9.0).

The marginal model for the mean attitude includes the following covariates: age in years (centred at 45), gender, highest academic qualifications (none, below university level, university degree), marital status (married/civil partnership, cohabiting, widowed, separated/divorced, never married), housing tenure (own outright, own with mortgage, social housing, private rental), and survey year. The choice of covariates was informed by previous studies of gender role attitudes using BHPS data (e.g. Berridge, Penn and Ganjali, 2009; Sweeting et al., 2014).

However, the correlation structure is the focus of our application. We wish to estimate the correlation between different individuals' attitudes while they are living together ('current' coresidents) and test whether this correlation persists after coresidence ends ('past' coresidents). Specific research questions, and details of the covariates $\mathbf{X}_{3k}$ from (9) are given in Section 7. This design matrix contains 891,951 pairs of person-wave observations across the superhouseholds. A small number are of size 1 (single-person households observed for only one wave) and so do not contribute to the estimation of the correlation structure. The number of pairs per superhousehold ranges from 1 to 4950 (=0.5×100×99 as 100 is the maximum cluster size).

6. Simulation study.

6.1. *Design.* As discussed earlier, there may be considerable between-superhousehold variation in the correlation structure $\boldsymbol{\rho}_k$. This can lead to covariance matrices with irregular structures that can be problematic to estimate (Dempster, 1972). We hence carry out a simulation study to assess the performance of the GEE estimator across a range of complex correlation structures.

We generate BHPS-like data with complex dependence structures using a two-stage approach. The first stage involves generating superhouseholds by sampling the actual changes in household membership observed in the BHPS. This is done by listing the superhouseholds constructed in Section 4.3 and selecting superhouseholds from this list. Once a superhousehold is selected, the time-varying household structure of its members is fixed for the second stage of the simulation. The second stage involves generating realisations of the outcome variable using a data generating model (DGM) which respects the within-superhousehold correlation structure.

For balanced designs, one superhousehold is selected and the clusters formed by generating M realisations from the DGM described below. For unbalanced designs, M superhouseholds are sampled with replacement from the 11,460 BHPS superhouseholds in the analysis sample, and a realisation generated for each under the DGM. The sampling of superhouseholds was repeated to generate each replicate of the simulation. The DGM itself is a simplified version of the models we fit in Section 7. For each correlation structure, it has a single parameter α_1 for within-individual correlations and three between-individual correlation parameters for individuals i and i' at waves t and t' ($t \leq t'$): α_2 if coresident at both t and t', α_3 if coresident at t but not t', and α_4 if past coresidents at both t and t'. As was discussed in Section 4.3, it is unrealistic to generate zero correlations between individuals in the same superhousehold, so we also specify α_5 to be the correlation for future or never coresidents (set to 0.15 in the DGM).

The design matrix for the correlation model was formed from the observed design matrices for the selected superhousehold(s). The DGM for the mean model includes an intercept, a dummy for female, and a linear effect of age (centred at 45 years), with associated parameters $(\beta_0, \beta_1, \beta_2)$. The mean structure was generated using the observed values of these covariates for the person-wave observations in the selected BHPS superhouseholds. The DGM for the residual variance or scale function contains only an intercept term (γ). The identity link was used for all three submodels. All parameter values are based on the estimates obtained from fitting the true model to the full BHPS sample with standardized gender-role attitudes as the response.

The results for each scenario described below are based on 500 simulated datasets. A series of models with different correlation structures was fitted to each simulated data set, ranging from M1 (within-individual autocorrelation only) through to the correct model (M4). Each fitted model includes an additional 'other' correlation parameter $\bar{\alpha}$ (which equals α_5 under M4) for the complement of the other indicators in the model (that is, 1 minus the sum of the other indicators). This additional parameter is not of substantive interest, but avoids imposing any zero constraints on the within-cluster correlations. The mean and scale functions were correctly specified in all fitted models.

6.2. *Results.* We considered a number of balanced designs each based on $M = 5000$ copies of one selected BHPS superhousehold (or $M = 1000$ copies for larger superhouseholds). The simplest superhousehold contained a couple observed together at all waves with no other adult coresidents, leading to a three-level hierarchical structure. More complex and irregular superhousehold structures arising from multiple changes in household membership over the observation period were also considered. For example, supplementary Table S2 shows the results for a superhousehold containing 25 person-wave observations from five individuals. For this and every other balanced design considered, the convergence rate was 100%, the implied fitted correlation matrix was always positive definite, estimates and standard errors were unbiased, and the confidence interval coverage probabilities were close to the nominal 95% level.

We now turn to the unbalanced case. Table 1 shows the results for replicates where the fitted model converged (determined by the difference between successive iterations being less than 0.001 for every parameter). The estimator and standard error are almost unbiased with good confidence interval coverage. However, nonconvergence is possible even if the correct model (M4) is specified (increasing the maximum number of iterations from

25 to 50 does not improve the convergence rate). Model M1 always converges because of its simple correlation structure. When convergence was not achieved, there was a small bias for the parameter estimates, but a large positive bias for the standard errors (see supplementary Table S3). As discussed in Section 4.3, imposing zero constraints on correlations is undesirable and doing so will lead to sparse superhousehold correlation matrices. For the unbalanced design, the estimator shows a very high chance of nonconvergence when the 'other' parameter ($\bar{\alpha}$) is excluded from fitted models M2–M4. Further discussion of the simulation results can be found in the supplementary materials.

Table 1

Simulation results for an unbalanced design with $r = 500$ replicates of $M = 5000$ superhouseholds selected with replacement from the BHPS data. Results are shown for the r_C replicates for which convergence was achieved.

	Mean function			Scale	Correlation function			
	β_0	β_1	β_2	γ	α_1	α_2	α_3	α_4
True	-0.1	0.25	-0.015	0.9	0.6	0.3	0.2	0.2
M2 ($r_C = 323$)								
Mean	-0.100	0.249	-0.015	0.899	0.599	0.296	–	–
Mean SE	0.016	0.017	0.001	0.011	0.009	0.014	–	–
SD	0.013	0.014	0.000	0.011	0.009	0.015	–	–
95% coverage	0.960	0.957	0.954	0.947	0.935	0.913	–	–
M3 ($r_C = 465$)								
Mean	-0.099	0.249	-0.015	0.900	0.600	0.299	0.202	–
Mean SE	0.012	0.013	0.000	0.011	0.009	0.014	0.019	–
SD	0.012	0.013	0.000	0.011	0.010	0.015	0.020	–
95% coverage	0.951	0.951	0.957	0.955	0.935	0.914	0.938	–
M4 ($r_C = 390$)								
Mean	-0.100	0.249	-0.015	0.899	0.599	0.298	0.202	0.200
Mean SE	0.012	0.013	0.000	0.011	0.009	0.014	0.019	0.027
SD	0.013	0.014	0.000	0.011	0.009	0.015	0.020	0.029
95% coverage	0.949	0.959	0.949	0.954	0.941	0.921	0.926	0.921

As noted in Section 4.4, the fitted correlation matrix implied by the GEE estimates of $\boldsymbol{\alpha}$ may not be positive definite. To explore how often this might occur in practice, we computed the number of superhouseholds with a nonpositive definite correlation matrix for M1–M4 for the first 50 replicates of the unbalanced design with 5000 superhouseholds. (For replicates where a model did not converge, the correlation matrices after 25 iterations were used.) The fitted correlation matrices are always positive definite for M1. For the other models, the final estimates of $\boldsymbol{\alpha}$ sometimes imply a non-positive definite correlation matrix, but in each case this affected only a small number of superhouseholds. The proportion of superhouseholds with a non-positive definite matrix for replicates where convergence was achieved ranges from 0% to 4.3% for M2, 0% to 2.3% for M3 and 0% to 1.5% for M4, and there is no discernible effect on the performance of the GEE estimator for any parameter (Table 1). When a model did not converge, the maximum proportions increase to 7.3%, 3.2% and 5.5% for M2–M4, so there is a weak association between non-convergence and non-positive definite correlation matrices and both problems can arise even when the model is correctly specified (M4).

7. Application: association structure of gender role attitudes in Britain.

7.1. *Research questions.* The application is a longitudinal analysis of gender role attitudes (GRA) using data from BHPS, with a focus on the association structure of coresidents' attitudes. Previous research on GRA has used longitudinal models that either ignore household effects (e.g. Berridge, Penn and Ganjali, 2009) or studied only the cross-sectional similarity between spouses (e.g. Brynin, Longhi and Martínez Pérez, 2008). We allow for temporal changes in a person's coresidents, and exploit the panel design to investigate questions about the nature of between-individual correlation before, during and after coresidence. We also extend earlier analyses of couple concordance by including all adults in a household and testing for differences in the between-individual correlation among family and unrelated dyads. The analysis considers the following specific research questions:

(i) What is the extent of the correlation between coresidents at a given wave t and across waves $t \neq t'$? Cross-wave correlations during coresidence may be explained by the presence of shared unmeasured time-invariant influences on attitudes, or the causal effect of one individual's attitudes on another's that persists over time.

(ii) Does between-individual correlation in GRA continue after coresidence has ended? A decay in correlation with duration since the end of coresidence would be expected if similarity in GRA is largely due to reciprocal influences during coresidence.

(iii) How does the correlation in GRA differ for couples, other family dyads and unrelated sharers? Parent-child correlation is most likely explained by an influence of the parent on the child, while for unrelated dyads homophily (non-random sorting) and reciprocal influences may both play a role.

(iv) How do between-individual correlations change after accounting for individual and household covariates in the model for mean attitudes?

7.2. *Specification of the within-cluster correlation structure.* We consider models of the form (7)–(9) with identity links for the mean, variance and correlation functions. The model for the mean includes the individual and household characteristics described in Section 5.2. Although the general model of (8) allows the scale $\boldsymbol{\phi}_k$ to depend on covariates, we assume a constant residual variance. In this section we set out the specification of the correlation structure, and the indicators that form the design matrix $\mathbf{X}_{3k}$ in (9), to investigate questions (i) to (iii) above.

The within-person autocorrelations are assumed to have a Toeplitz structure, starting with a separate parameter for each lag $t' - t$ (for $t < t'$), measured in two-year intervals:

$$\rho_{tik,t'ik} = \alpha^W_{t'-t}, \qquad t' - t = 1, 2 \ldots, 8.$$

Next, we specify the between-person correlations for current, past and future coresidents. For individuals i and i' in the same superhousehold k, we allow the correlation between their responses at waves t and t' to depend on their coresidence status at each wave. For $t \leq t'$, we can distinguish the four situations described below, which are illustrated in Figure 2 for individuals (A, C, D) from Figure 1.

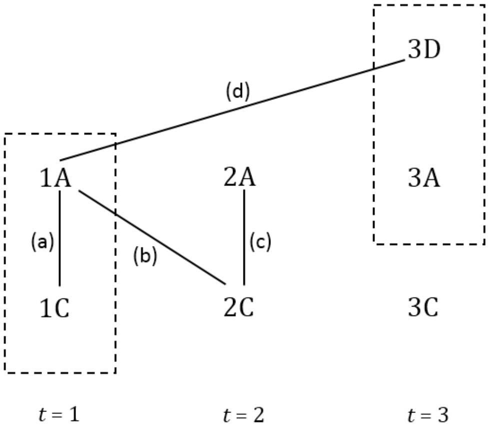


Fig 2: Classification of between-individual correlations at waves t and t' $(t \leq t')$ by coresidence status for individuals A, C and D of Figure 1. Dashed borders indicate coresidence at that wave, and paths represent one correlation of each of the following types: (a) coresident at t and t', (b) coresident at t but not t', (c) past coresidents, and (d) future coresidents at t.

(a) **Coresident at t and t'.** Assuming a Toeplitz structure gives

$$\rho_{tik,t'i'k} = \alpha^{B}_{1,t'-t}, \qquad t'-t = 0, 1 \ldots, 8.$$

For the example in Figure 1, the set of person-wave observations $(ti, t'i')$ with this property is $\{(1A, 1B), (1A, 1C), (1B, 1C), (2B, 2C), (3A, 3D), (3C, 3E)\}$. We investigate question (i) by testing whether $\alpha^{B}_{1,0} = 0$ and considering the change in the correlation with $t' - t$.

For pairs where at least one of t and t' is after or before the period of coresidence, additive adjustments are made to $\alpha^{B}_{1,t'-t}$ as described below. This parameterisation differs from that used in the simulation study where the correlations according to coresidence status at t and t' were estimated directly. A more general parameterisation is necessary when the linear predictor for the correlation is extended to allow correlations to vary across more than one dimension.

(b) **Coresident at t but no longer at t'.**

$$\rho_{tik,t'i'k} = \alpha^{B}_{1,t'-t} + \alpha^{B}_{2}, \qquad t'-t = 0, 1 \ldots, 8,$$

where we expect $\alpha^{B}_{2} < 0$ if the correlation is lower when i and i' are coresident at only t relative to (a) when coresident at both t and t'. In Figure 1, pairs with this property are $\{(1A, 2B), (1A, 2C), (1A, 3C), (1C, 2A), (1C, 3A)\}$.

(c) **Past coresidents: last coresident at $s < t \leq t'$.**

$$\rho_{tik,t'i'k} = \alpha^{B}_{1,t'-t} + \alpha^{B}_{3,t-s}, \qquad t-s = 1, 2, \ldots,$$

where we expect $\alpha^{B}_{3,t-s} < 0$ for all $t - s$ and $|\alpha^{B}_{3,1}| < |\alpha^{B}_{3,2}| < \ldots$, i.e. the correlation is reduced when i and i' are no longer coresident relative to when they were coresident, and the correlation decreases as $t-s$ increases. This applies to pairs $\{(2A, 2B), (2A, 2C), (2A, 3C), (2C, 3A)\}$. Investigation of question (ii) involves tests of $\alpha^{B}_{2} = 0$ and exploring the change in $\alpha^{B}_{3,t-s}$ with $t - s$.

(d) **Future coresidents at t, current or past coresidents at t'.** To allow for the possibility that individuals with similar attitudes may select into coresidence, the correlation function for individuals who live together after t is

$$\rho_{tik,t'i'k} = \alpha^{B}_{1,t'-t} + \alpha^{B}_{4}, \qquad t'-t = 0, 1 \ldots, 8.$$

> If individuals influence each others attitudes during coresidence, we expect $\alpha_4^B < 0$ while $\alpha_{1,t'-t}^B + \alpha_4^B > 0$ would be consistent with homophily. For the example in Figure 1, at $t = 1, 2$ individuals A and C are future partners of D and E who enter at $t = 3$, and pairs contributing to the estimation of α_4^B are $\{(1A, 3D), (2A, 3D), (1C, 3E), (2C, 3E)\}$. Due to the design of BHPS, it is extremely rare for future coresidents to both be observed prior to living together, and thus for both t and t' to be before coresidence.

The elements of $\mathbf{X}_{3k}$ which define the above correlation structure can be expressed respectively in terms of the household and superhousehold coresidence indicators $c_t(i, i')$ from (4) and $C_t(i, i')$ from (10) (see Table 2).

TABLE 2
Within-superhousehold parameters and their corresponding indicator variables for residual correlations between person-wave observations for individuals i and i' at occasions $t \leq t'$.

$\boldsymbol{\alpha}$	Covariates in $\mathbf{x}_{3,tik,t'i'k}$	Description
α_{1l}^W	$\mathrm{I}(i = i')\,\mathrm{I}(t' - t = l)$ $l = 1, 2, \ldots$	Within person
α_{1l}^B	$\mathrm{I}(i \neq i') c_t(i, i') c_{t'}(i, i')\,\mathrm{I}(t' - t = l)$ $l = 0, 1, \ldots$	Coresident at both t and t'
α_2^B	$\mathrm{I}(i \neq i')\, c_t(i, i')\, \{1 - c_{t'}(i, i')\}$	Coresident at t but not t'
α_{3l}^B	$\mathrm{I}(i \neq i')\, \{1 - c_t(i, i')\}\, \{1 - c_{t'}(i, i')\}$ $\times\, C_s(i, i')\,\mathrm{I}(t - s = l)$ $l = 1, 2, \ldots$	Not coresident at t or t' but were coresident at $s < t$
α_4^B	$\mathrm{I}(i \neq i')\, \{1 - c_t(i, i')\}\, C_{t'}(i, i')$	Not coresident at t, only at or after t'

In addition to the indicators of coresidence status, we examine question (iii) by defining indicators for the relationship between individuals i and i' (couple, parent-child, other family, or unrelated) with coefficient vector $\boldsymbol{\alpha}_5^B$. Table 3 shows the distribution of pairs of person-wave observations according to the timing of t and t' with respect to the period of coresidence and, among pairs contributed by ever coresidents, the distribution by their relationship.

TABLE 3

Classification of person-wave pairs within superhouseholds by coresidence status and relationship type for occasions $t \leq t'$.

	Number of pairs	Percent
Coresidence status at occasions t and t' ($n = 891,951$)		
Same person ($t \neq t'$)	256,505	28.8
Never coresident during observation period	139,712	15.7
Future coresidents at t; current or past at t'	28,302	3.2
Current coresidents at both t and t'	272,229	30.5
Current coresidents at one wave, previous at other	102,683	11.5
Previous coresidents at both t and t'	95,520	10.4
Relationship type among coresidents ($n = 495,734$)		
Partners	228,462	46.1
Parent-child	192,852	38.9
Other family	48,963	9.9
Unrelated	25,457	5.1

7.3. *Model selection.* As research questions (i)-(iii) in Section 7.1 are concerned with the unconditional correlations in GRA among coresidents, our modeling strategy was to first build the correlation structure with only an intercept term in the mean function $\boldsymbol{\mu}_k$ before adding covariates to investigate question (iv).

The correlation model was built gradually, introducing and testing the parameters described in Section 7.2. For all fitted models convergence was achieved and the implied correlation matrix was positive definite for every superhousehold. The initial correlation structure $\boldsymbol{\rho}_k$ allowed for within-person autocorrelation and between-person correlation for any pair of individuals who lived together during the observation period. A simplified Toeplitz structure was fitted to allow $\rho_{tik,t'ik}$ to depend on lag $t' - t$, with equal correlation assumed for grouped lags 1, 2-3, 4-6 and 7-8 (measured in 2-year intervals). A Wald test of the equality of the four within-person correlation parameters indicated strong evidence of autocorrelation ($X^2 = 306.9$, df $= 3$).

For a pair of responses Y_{tik} and $Y_{t'i'k}$ for individuals i and i' who were ever observed as coresidents, a simplification of the correlation structure defined by (a)-(d) of Section 7.2 was fitted to examine how the between-individual correlation depended on whether t and t' were before, during or after coresidence, with a separate parameter for the situation where t was during and t' after coresidence. Compared to when t and t' are both during coresidence, the correlation is significantly lower when one or both of t and t' is after ($X^2 = 64.8$, df $= 1$) or before ($X^2 = 22.2$, df $= 1$) coresidence.

Three generalisations to this basic correlation structure were then considered in turn. First, to investigate question (i), $\rho_{tik,t'i'k}$ was permitted to depend on $t'-t$ for individuals i and i' who lived together at any time during the observation period, assuming a Toeplitz structure with separate correlation parameters for grouped lags 0, 1-2, 3-5 and 6-8. (The fitted parameters are $\alpha^B_{1,t'-t}$ in (a) of Section 7.2.) There was strong evidence of coresident autocorrelation ($X^2 = 32.5$, df $= 3$). The second extension, to explore question (ii), was to allow the correlation among past coresidents to depend on the time since they last lived together s ($\alpha^B_{3,t-s}$ in (c)). However, there was little indication of this form of time dependency ($X^2 = 1.8$, df $= 2$, p=0.407), possibly because of measurement error in the duration $t - s$ due to gaps in coresidence histories resulting from household nonresponse. The model was therefore simplified to include a single parameter to differentiate past and current coresidents. Finally, we investigated question (iii) by allowing $\rho_{tik,t'i'k}$ to depend on whether i and i' were a couple, parent and child, other family relations, or unrelated (parameters $\boldsymbol{\alpha}^B_5$). There was strong evidence that the between-coresident correlation varies according to relationship type ($X^2 = 142.0$, df $= 3$). There was no evidence that the effect of relationship type depends on the timing of t and t' relative to the period of coresidence.

7.4. *Results.* The estimates for the selected correlation structure are shown in Table 4. The model also includes a parameter for the marginal correlation between responses for individuals in a superhousehold who never lived together, but who were nevertheless linked through their respective coresidents. This parameter was included primarily to aid convergence and is therefore not shown in Table 4, but its estimate is small and positive (0.123, SE=0.026). In our parameterisation $\alpha^B_{1,0}$ is the coefficient of an indicator of ever coresidence over the whole observation period and represents the intercept in the correlation function for coresidents. All other coefficients are interpreted as contrasts with the reference case of a couple at $t = t'$.

In line with questions (i)-(iii), we begin with an interpretation of the unconditional correlations from the model with only an intercept in the mean function. As expected, the within-person correlation $\rho_{tik,t'ik}$ decreases with $t' - t$. The coresident correlation was modeled as a linear additive function of indicator variables for grouped $t'-t$, relationship type, and past or future coresidence.

In answer to question (i), there is substantial contemporaneous correlation between the attitudes of coresidents (estimated as 0.384 for couples). Although the negative estimates of $(\alpha^B_{1,1}, \alpha^B_{1,3}, \alpha^B_{1,6})$ imply that the between coresident correlation declines with increasing $t' - t$, the decrease is small so there is strong evidence of cross-wave correlation. Among couples, for ex-

Table 4

Analysis of gender role attitudes: estimates of correlation parameters for observations at occasions t and t' ($t \leq t'$) within superhouseholds before and after including covariates in the mean function.

	Mean: intercept		Mean: covariates	
Correlation parameter ($\boldsymbol{\alpha}$)	Est.	(SE)	Est.	(SE)
Within-person by lag* $t' - t$				
1 (α_1^W)	0.693	(0.005)	0.667	(0.005)
2-3 (α_2^W)	0.642	(0.006)	0.616	(0.007)
4-6 (α_4^W)	0.578	(0.010)	0.557	(0.010)
7-8 (α_7^W)	0.512	(0.014)	0.490	(0.015)
Between coresident				
Intercept ($\alpha_{1,0}^B$)†	0.384	(0.009)	0.368	(0.010)
Lag* $t' - t$ (ref=0)				
1-2 ($\alpha_{1,1}^B$)	-0.020	(0.004)	-0.017	(0.004)
3-5 ($\alpha_{1,3}^B$)	-0.042	(0.007)	-0.030	(0.007)
6-8 ($\alpha_{1,6}^B$)	-0.064	(0.011)	-0.044	(0.011)
Past: coresident at t but not t' (α_2^B)	-0.025	(0.014)	-0.050	(0.013)
Past: last coresident before t (α_3^B)	-0.062	(0.018)	-0.072	(0.017)
Future: coresident only after t (α_4^B)	-0.088	(0.017)	-0.120	(0.018)
Relationship type (ref=couple)				
Parent-child ($\alpha_{5,2}^B$)	-0.184	(0.016)	-0.107	(0.014)
Other family ($\alpha_{5,3}^B$)	-0.096	(0.022)	-0.169	(0.020)
Unrelated ($\alpha_{5,4}^B$)	-0.011	(0.046)	-0.163	(0.035)

Standard errors in parentheses; *Lags $t' - t$ are in 2-year intervals because Y_{ti} is measured every two years;†Intercept parameter is the correlation for observations on a couple at $t = t'$.

ample, the correlation between observations that are the maximum 16 years apart is estimated as 0.384-0.064=0.320. The persistence of the correlation suggests that there are time-invariant characteristics, such as similar social backgrounds and values, that affect the attitudes of both coresidents across time. Another possible source of the cross-time correlation is a 'memory' effect whereby one individual's attitude at t continues to influence the other's at t'.

Turning to question (ii) we find that, compared to current coresidents, the between-individual correlation is lower when at least one time point in the pair is after the end of coresidence. However, as noted earlier, it does not depend on the time since last coresidence. There is also a small reduction in the correlation when t is prior to coresidence, but it remains non-negligible which provides some support for homophily.

The magnitude of the correlation between coresidents depends strongly on their relationship (question (iii)). The estimates of $(\alpha_{5,2}^B, \alpha_{5,3}^B, \alpha_{5,4}^B)$ indicate that the correlation is highest for spouses and lowest for parent-child pairs. Table 4 also shows estimates of the residual correlation parameters after adjusting for the effects of various individual and household characteristics on mean attitudes (question (iv)). The most notable change is in the correlations by relationship type. For coresidents at $t = t'$, the parent-child correlation increases from 0.20 to 0.26, while the correlations between other family members and unrelated household members decrease from 0.29 to 0.20 and from 0.37 to 0.20 respectively. Further investigation reveals that these changes in correlations are explained by the inclusion of individual age in the mean function. Parent-child pairs are the most heterogeneous in age, and so failure to control for age has a masking effect on the correlation. In contrast, the correlation between the attitudes of individuals in other family member (mainly sibling) and unrelated pairs is partly explained by their similarity in age. Covariates also account for part of the correlation between past and future coresidents.

Table 5

Analysis of gender role attitudes: covariate effects on standardized attitudes. Higher values of the response indicate more egalitarian attitudes.

Variable	Est.	(SE)	Percent/ mean (SD)*
Female	0.226	(0.008)	54.5%
Age in years (centred at 45)	-0.015	(0.0004)	1.07 (17.82)
Marital status (ref=married)			
Cohabiting	0.083	(0.012)	10.8%
Widowed	0.082	(0.016)	7.6%
Separated/divorced	0.064	(0.014)	7.4%
Never married	0.120	(0.013)	17.9%
Highest academic qualification (ref=none)			
Below degree level	0.054	(0.012)	54.5%
Degree	0.144	(0.017)	12.1%
Housing tenure (ref=owned outright)			
Owned with mortgage	0.044	(0.009)	46.4%
Social rent	-0.007	(0.013)	18.3%
Private rent	0.035	(0.013)	9.0%
Survey year (in 2-year intervals)	0.013	(0.0007)	–
Constant	-0.342	(0.015)	–

Standard errors in parentheses; *Descriptive statistics show the distribution of each covariate across $n = 106,060$ person-waves: percentages for categorical variables and mean (SD) for continuous variable age.

Parameter estimates for the mean function are given in Table 5. The following individual characteristics are associated with more egalitarian gender role attitudes: female gender, younger age, marital status other than marriage (or same-sex civil partnership) and higher levels of education. There is also evidence that attitudes have become less traditional over time. Housing tenure is included, together with education, as a proxy for socioeconomic position. Individuals living in a house that is owned outright have on average more traditional attitudes than homeowners with a mortgage or renters.

8. Discussion. Household panel surveys offer the potential to learn about the nature of the associations among the outcomes of people who share a household. Longitudinal data on individuals and their households permit separation of within-individual and between-individual within-household variability, and investigation of between-individual correlations across time and after coresidence ends. However, previous research has been unable to investigate these questions because of challenges arising from changes in household composition. While multiple membership multilevel models appeared a promising way forward, we have demonstrated these impose strong and unrealistic assumptions on the between-individual association structure. We instead proposed a flexible marginal modeling approach where the correlation between a pair of person-wave observations is modeled directly as a function of characteristics of the pair. In our analysis of gender role attitudes, we considered the effects of coresidence status at each wave and the relationship between the individuals. Examples of other possible covariates include the age of one member of the pair and their age difference, their gender composition, and their religion.

Household panel surveys provide only limited information for disentangling homophily and causal effects of coresidence as explanations for between-coresident associations. To explore homophily requires data on individual outcomes before they become coresidents but, in common with other household panels, the design of BHPS does not allow us to observe pairs of individuals prior to living together, so our estimate of the correlation between ‘future’ coresidents is based on pairs where one observation is before and the other during the period of coresidence. Data on the duration of coresidence are required to investigate causal effects of coresidence, where a pattern of increasing correlation with duration would suggest a (reciprocal) influence of one individual on the other. Coresidence histories in BHPS are left-truncated, although for couples it is possible to infer duration of coresidence from retrospective union histories.

The proposed method can be applied to any household panel survey, many of which have a similar design to BHPS with individuals and their coresidents tracked over time. Longitudinal individual and household data are also available from some national population registers. More generally, clusters with a network structure arise in cross-sectional and longitudinal studies of peer group effects, and marginal models offer a flexible way of studying the dependency in behavioural and educational outcomes for members of the same network. Examples of potential applications include longitudinal analyses of risk-taking among friendship groups (Pearson and West, 2003) and happiness in family, friend and coworker networks (Fowler and Christakis, 2008). Beyond the social sciences, examples of evolving networks can be found in studies of animal populations. In veterinary epidemiology, for example, movement of cattle between herds leads to a network structure where animals who have shared contact are members of the same network and may have correlated disease risks.

A potential disadvantage of GEE2 is that it has a higher chance of nonconvergence than GEE1 when estimating models with the same mean structure (Hardin and Hilbe, 2012, p.152; Ziegler, Kastner and Blettner, 1998). In our simulation study, we found that GEE2 performed well for balanced designs, but nonconvergence was an issue when superhousehold clusters were highly unbalanced. The convergence rate and behaviour of the estimator was much improved by fitting an additional correlation parameter to capture all between-individual correlations that are not of direct scientific interest, rather than constraining these to be zero. Following this strategy, no convergence problems were encountered for the models considered in our application. Alternative approaches to estimating marginal models where the association structure is of interest are pairwise likelihood (PL) (Kuk and Nott, 2000) or a hybrid of GEE1 for the mean parameters and PL for the association parameters (Kuk, 2007). The GEE1-PL hybrid is highly flexible, avoids inversion of large cluster-specific covariance matrices, and yields robust estimates of the mean parameters when the association structure is misspecified.

REFERENCES

Atkins, D. C. (2005). Using multilevel models to analyze couple and family treatment data: basic and advanced issues. *Journal of Family Psychology* **19** 98-110.

Ballas, D. and Tranmer, M. (2012). Happy people or happy places? A multilevel modeling approach to the analysis of happiness and well-being. *International Regional Science Review* **35** 70-102.

Bauer, D. J., Gottfredson, N. C., Dean, D. and Zucker, R. A. (2013). Analyzing repeated measures data on individuals nested within groups: accounting for dynamic group effects. *Psychological Methods* **18** 1-14.

Berridge, D., Penn, R. and Ganjali, M. (2009). Changing attitudes to gender roles: a longitudinal analysis of ordinal response data from the British Household Panel Study. *International Sociology* **24** 346-367.

Blackwell, D. L. and Lichter, D. T. (2004). Homogamy among dating, cohabiting and married couples. *The Sociological Quarterly* **45** 719-737.

Brynin, M., Longhi, S. and Martínez Pérez, Á. (2008). *The social significance of homogamy* In *Changing Relationships* 5, 73-90. Routledge, New York.

Buck, N. and McFall, S. (2012). Understanding Society: design overview. *Longitudinal and Life Course Studies* **3** 5-17.

Butterworth, P. and Rodgers, B. (2006). Concordance in the mental health of spouses: analysis of a large national household panel survey. *Psychological Medicine* **36** 685-697.

Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations. *Journal of Statistical Planning and Inference* **63** 39-54.

Chandola, T., Bartley, M., Wiggins, R. and Schofield, P. (2003). Social inequalities in health by individual and household measures of social position in a cohort of healthy people. *Journal of Epidemiology and Community Health* **57** 56-62.

Chiu, T. Y. M., Leonard, T. and Tsui, K. W. (1996). The matrix-logarithm covariance model. *Journal of the American Statistical Association* **91** 198-210.

Crowder, M. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82** 407-410.

Davillas, A. and Pudney, S. (2017). Concordance of health states in couples: analysis of self-reported, nurse administered and blood-based biomarker data in the UK Understanding Society panel. *Journal of Health Economics* **56** 87-102.

Dempster, A. P. (1972). Covariance selection. *Biometrics* **28** 157-175.

Duncan, G. and Hill, M. (1985). Conceptions of longitudinal households: fertile or futile? *Journal of Economic and Social Measurement* **13** 361-375.

Fowler, J. H. and Christakis, N. A. (2008). Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the Framingham Heart Study. *British Medical Journal* **337** a2338.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space-time data. *Journal of the American Statistical Association* **97** 590-600.

Goldstein, H. (2010). *Multilevel Statistical Models*, 4th ed. Wiley, London.

Goldstein, H., Rasbash, J., Browne, W. J., Woodhouse, G. and Poulain, M. (2000). Multilevel models in the study of dynamic household structures. *European Journal of Population* **16** 373-387.

Hardin, J. W. and Hilbe, J. M. (2012). *Generalized Estimating Equations*, 2nd ed. CRC Press, Boca Raton, FL.

Højsgaard, S., Halekoh, U. and Yan, J. (2006). The R package `geepack` for generalized estimating equations. *Journal of Statistical Software* **15** 1-11.

ISER (2009). *British Household Panel Survey: Waves 1-17, 1991-2008*, 6th ed. University of Essex, Institute for Social and Economic Research [original data producer(s)], Colchester, Essex: UK Data Archive [distributor]. SN: 5151.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42** 805-820.

Johnston, R., Jones, K., Propper, C., Sarker, R., Burgess, S. and Bolster, A. (2005). A missing level in the analyses of British voting behaviour: the household as context as shown by analyses of a 1992–1997 longitudinal survey. *Electoral Studies* **24** 201-225.

Jones, B. and West, M. (2005). Covariance decomposition in undirected Gaussian graphical models. *Biometrika* **92** 779-786.

Kalmijn, M. (1998). Intermarriage and homogamy: causes, patterns, trends. *Annual Review of Sociology* **24** 395-421.

Keizer, R. and Schenk, N. (2012). Becoming a parent and relationship satisfaction: a longitudinal dyadic perspective. *Journal of Marriage and Family* **74** 759-773.

Kuk, A. Y. C. (2007). A hybrid pairwise likelihood method. *Biometrika* **94** 939-952.

Kuk, A. Y. C. and Nott, D. J. (2000). A pairwise likelihood approach to analysing correlated binary data. *Statistics and Probability Letters* **47** 329-335.

Leckie, G. and Goldstein, H. (2009). The limitations of using school league tables to inform school choice. *Journal of the Royal Statistical Society, Series A* **172** 835-851.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13-22.

Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analyses for categorical data (with discussion). *Journal of the Royal Statistical Society B* **54** 3-20.

McPherson, M., Smith-Lovin, L. and Cook, J. M. (2001). Birds of a feather: homophily in social networks. *Annual Review of Sociology* **27** 415-444.

Milner, A., Spittal, M. J., Page, A. and LaMontagne, A. D. (2014). The effect of leaving employment on mental health: testing 'adaptation' versus 'sensitisation' in a cohort of working-age Australians. *Occupational and Environmental Medicine* **71** 167-174.

Murphy, M. J. (1996). The dynamic household as a logical concept and its use in demography. *European Journal of Population* **12** 363-381.

Pearson, M. and West, P. (2003). Drifting smoke rings: social network analysis and Markov processes in a longitudinal study of friendship groups and risk-taking. *Connections* **25** 59-76.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation. *Biometrika* **86** 677-690.

Prentice, R. L. and Zhao, L. P. (1991). Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses. *Biometrics* **47** 825-839.

Raudenbush, S. W., Brennan, R. T. and Barnett, R. C. (1995). A multivariate hierarchical model for studying psychological change within married couples. *Journal of Family Psychology* **9** 161-174.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association* **90** 106-121.

Sacker, A., Wiggins, R. and Bartley, M. (2006). Time and place: putting individual health into context. A multilevel analysis of the British Household Panel Survey, 1991-2001. *Health and Place* **12** 279-290.

Shults, J. and Hilbe, J. (2014). *Quasi-Least Squares Regression*. Chapman and Hall/CRC Press, Boca Raton, FL.

Sweeting, H., Bhaskar, A., Benzeval, M., Popham, F. and Hunt, K. (2014). Changing gender roles and attitudes and their implications for well-being around the new millennium. *Social Psychiatry and Psychiatric Epidemiology* **49** 791-809.

Yan, J. and Fine, J. (2004). Estimating equations for association structures. *Statistics in Medicine* **23** 859-874.

Ziegler, A., Kastner, C. and Blettner, M. (1998). The generalised estimating equations: an annotated bibliography. *Biometrical Journal* **40** 115-139.

Department of Statistics
London School of Economics & Political Science
Houghton Street
London WC2A 2AE, UK
E-mail: f.a.steele@lse.ac.uk
j.kuha@lse.ac.uk

Institute for Social & Economic Research
University of Essex
Wivenhoe Park, Colchester
Essex CO4 3SQ, UK
E-mail: pclarke@essex.ac.uk