Illustrations of ISM and CoDA for 24HAC data analysis

Yinxiang Wu

In this document, we illustrated and repeated the analysis we performed in our paper *Analysis of the 24-Hour Activity Cycle: An illustration examining the association with cognitive function in the Adult Changes in Thought (ACT) Study [arXiv submission pending]*, based on the hypothetical data.

The dataset contains a continuous outcome variable y, a categorical variable *sex*, and compositional variables for *sit*, *stand*, *step*, and *sleep*, each of which ranges from 0 to 1, and they sum to 1. The compositional variables were simulated from the distribution of the real data that were used in the paper. The continuous outcome was generated based on the model

$$y = -0.04z_1 - 0.06z_2 + 0.68z_3 + 0.5sex + \epsilon$$

where z_1, z_2, z_3 are a set of isometric log-ratio (ilr) coordinates of the compositional variables (see below), and $\epsilon \sim N(0, 0.25)$. For more details of the ilr-coordinates, please refer to the CoDA section below and see the section 3.2 in the paper.

$$z_1 = \sqrt{\frac{3}{4}} ln(\frac{Sit}{(Stand \times Step \times Sleep)^{\frac{1}{3}}})$$

$$z2 = \sqrt{\frac{2}{3}} ln(\frac{Stand}{(Step \times Sleep)^{\frac{1}{2}}})$$

$$z_3 = \sqrt{\frac{1}{2}} ln\frac{Step}{Sleep}$$

Descriptive table of the data

The table 1 presents descriptive statistics of the sample.

Table 1: Descriptive statistics of the sample

	Overall
n	1000
sex = Female (%)	586 (58.6)
Sit (hrs/day), mean (SD)	10.12 (2.14)
Stand (hrs/day), mean (SD)	3.87 (1.72)
Step (hrs/day), mean (SD)	1.42 (0.82)
Sleep (hrs/day), mean (SD)	8.59 (1.27)

ISM approach

Linear ISM

Consider an example where each minute of the 24-hour day is classified into one of four activities: sleeping, sitting, standing, and stepping. The ISM is formulated by including the total activity and all but one of the activity variables – the activity you will explore displacing – in the model. For example, with a continuous health outcome an ISM that leaves out the time stepping can be formulated, as below:

$$E(Y) = \beta_0 + \beta_1 Sit + \beta_2 Stand + \beta_3 Step + \beta_4 Ttoal + \gamma Sex$$

where $E(Y)$ abbreviates the conditional mean of the health outcome given the time allocation variables (Sit, Stand, Sleep, Total measured on the same unit, e.g., hours in a 24-hour day), and the covariate sex. When *Total* is exactly a constant 24 hours/day for every subject, like in this hypothetical example, only one of the intercept or the Total terms can be included in the model. For more details, you can refer to the section 3.1 in the paper.

Four linear ISMs adjusted for sex were fit to the data, with each of the four activities omitted from the model one at a time. Table 2 summarizes the coefficient estimates. This is a table similar to the Table 2 in the paper.

Table 2: Isotemporal Substitution of Activities, per 1 hr/Day Increase

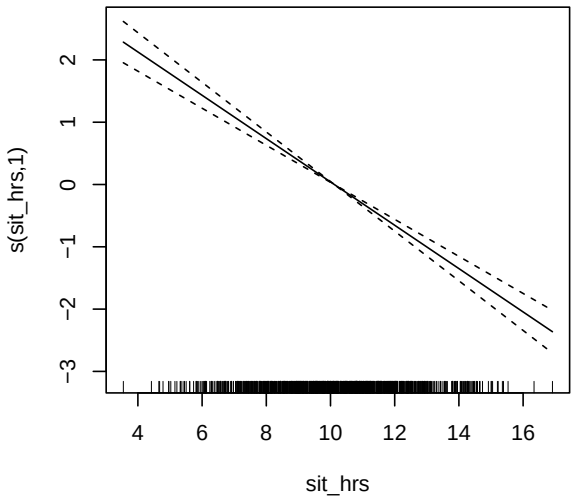
	Model with sit dropped		**Model with stand dropped**		**Model with step dropped**		**Model with sleep dropped**	
	Beta (95% C.I.)	p-value	Beta (95% C.I.)	p-value	Beta (95% C.I.)	p-value	Beta (95% C.I.)	p-value
Sit (hours)	Dropped	Dropped	0.01 [-0.02, 0.03]	0.656	-0.35 [-0.40, -0.30]	<0.001	0.05 [0.02, 0.08]	0.001
Stand (hours)	-0.01 [-0.03, 0.02]	0.656	Dropped	Dropped	-0.35 [-0.42, -0.29]	<0.001	0.04 [0.01, 0.08]	0.010
Step (hours)	0.35 [0.30, 0.40]	<0.001	0.35 [0.29, 0.42]	<0.001	Dropped	Dropped	0.40 [0.35, 0.45]	<0.001
Sleep (hours)	-0.05 [-0.08, -0.02]	0.001	-0.04 [-0.08, -0.01]	0.010	-0.40 [-0.45, -0.35]	<0.001	Dropped	Dropped

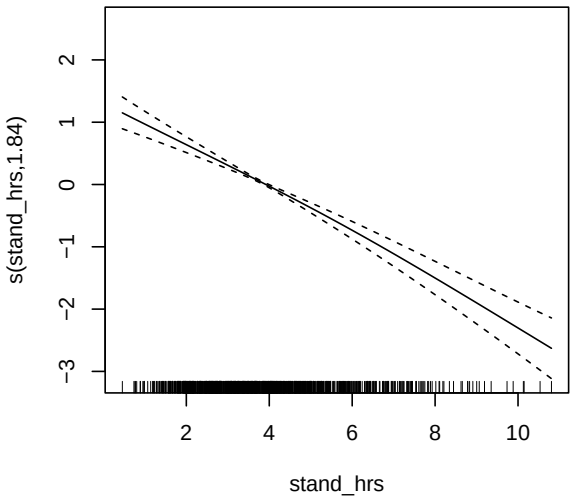
The associations of 1-hr time reallocations between any two types of activity are summarized in Table 2. For example, the ISM model with Step dropped suggested that reallocating 1 hr/day from sitting, standing, or sleeping to stepping was associated with 0.35 [0.30, 0.40], 0.35 [0.29, 0.42], and 0.40 [0.35, 0.45] units higher mean (95% CI) outcome, respectively.

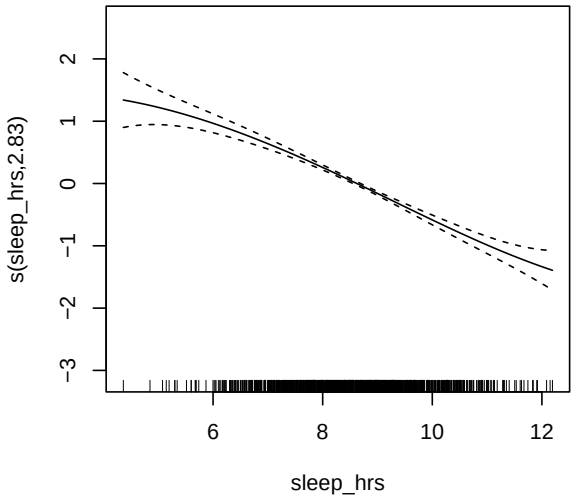
Nonlinear ISM

A more flexible ISM could be fit with each activity term modeled by a penalized spline function, while keeping the total activity as a linear term. The slope of a spline term represents the instantaneous effect of increasing a small amount of time in the activity the spline term corresponds to, while decreasing the same small amount of time in the activity that is left out from the model. The optimal trade-off between smoothness and goodness of fit can be determined by either performing cross validation or minimizing the generalized cross validation (GCV) criteria. The significance of the association for each behavior in the nonlinear ISM model can be tested via a Wald like test [Wood SN 2006]. The nonlinear ISM analysis can be done in R with package "mgcv".

For example, we fit a nonlinear ISM dropping step time adjusted for sex. The estimated smoothing terms for sit, stand, step are shown below. At the mean composition, all the smoothing terms equal to 0 because of the identifiablility constrain. We observed that increasing time in each of the behavior was associated with worse mean outcome, and in the main range of the data, the associations were approximately linear. This shows consistent results to that obtained from linear ISMs, and indicates that linear ISM would fit the data equally well and could be a better option for this hypothetical data because of its parsimony.


s(sit_hrs,1)
sit_hrs
−3
−2
−1
0
1
2
4
6
8
10
12
14
16


s(stand_hrs,1.84)
stand_hrs
−3
−2
−1
0
1
2
2
4
6
8
10

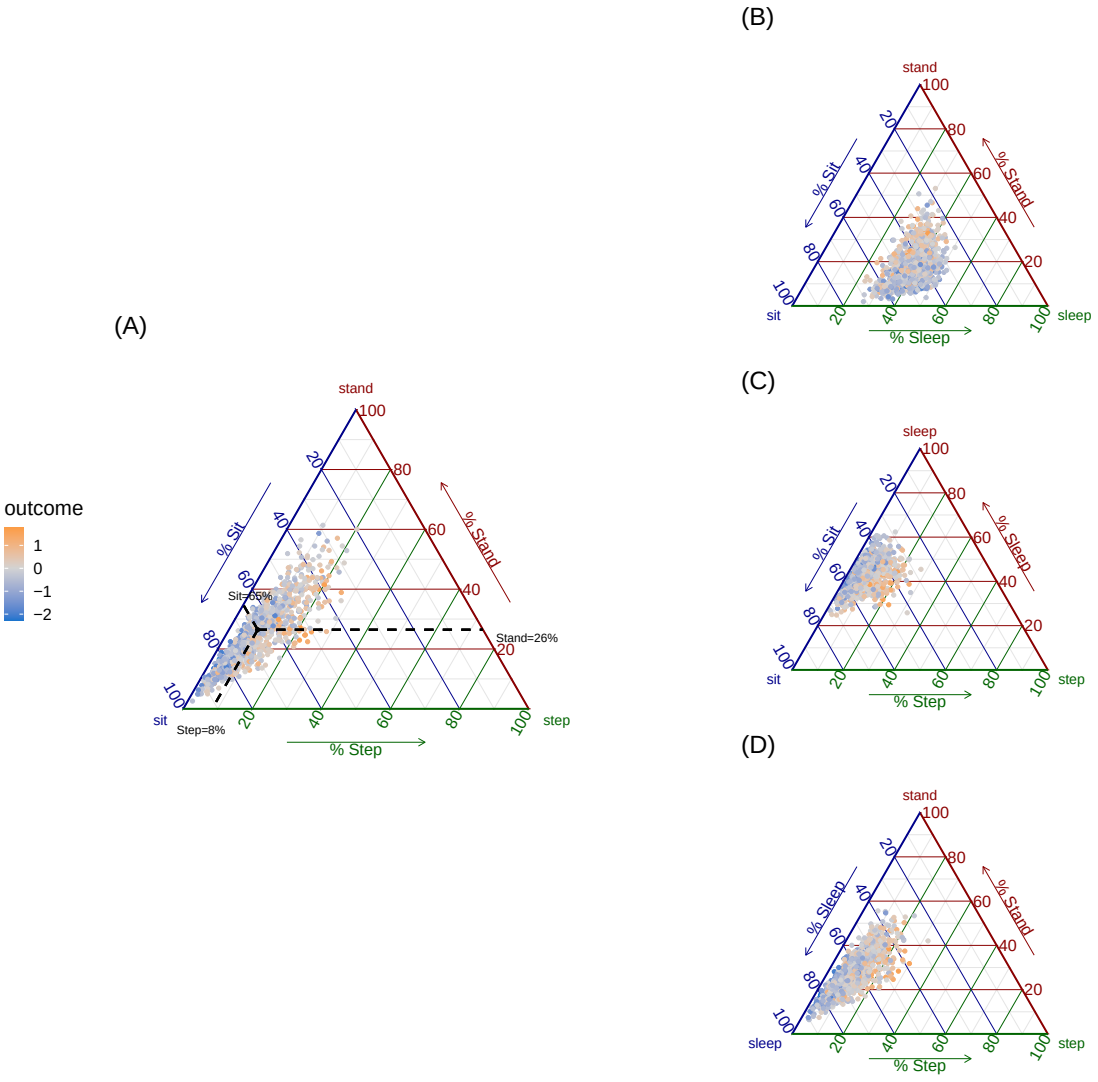

s(sleep_hrs,2.83)
sleep_hrs
−3
−2
−1
0
1
2
6
8
10
12

CoDA

Unlike ISM treating each activity behavior as an univariate variable on the original time scale, the fundamental unit of observation is the multivariate vector of the proportions or percentages of the 24 hours that are spent in each type of activity.

Visualizations and compositional descriptive statistics of 24HAC can be helpful, before fitting any models. The Figure below (similar to the Figure 1 in the paper) displays 24HAC compositions of sit, stand, step, and sleep for the fake data using ternary diagrams, a common tool to visualize composition with 3 parts. Since the 24HAC of interest here consists of four activity behaviors, we plotted four ternary diagrams (A-D below), with each graph representing a sub-composition of three activity behaviors. From the figure below, we can see how sub-compositions are distributed and possibly associated with the outcome.

Simplex plot



Comparisons of compositional means by groups

The compositional mean is a common descriptive statistics to describe central tendency of compositional data. It is defined as the vector of geometric means of each behavior, rescaled to sum to 1. Please refer to the Supplemental material A1 for more details about this definition. Since the components of a composition are inter-correlated, it is not sensible to calculate the variance of a single component. In stead, a variation matrix for the log-ratio is used to describe the interdependence between every pair of behaviors i.e. each element of that matrix is the variance of log-ratio between two components. An off-diagonal value close to 0 means the two parts are highly proportional in the observed data. Both compositional mean and variation matrix can be easily coded in R.

For inferential analysis such as hypothesis testing and regressions, CoDA relies on the isometric log-ratio (ilr) transformation, which transforms each D-part composition to a unique D-1 vector on a new coordinate system where each new coordinate is a log-ratio which falls along the real line. For example, a possible transformation is as follows:

$$z_1 = \sqrt{\frac{3}{4}} ln(\frac{Sit}{(Stand \times Step \times Sleep)^{\frac{1}{3}}})$$
$$z2 = \sqrt{\frac{2}{3}} ln(\frac{Stand}{(Step \times Sleep)^{\frac{1}{2}}})$$
$$z_3 = \sqrt{\frac{1}{2}} ln\frac{Step}{Sleep}$$

In R, we used the function *pivotCoord()* from the package *robCompositions* for this transformation.

With transformed data (z_1, z_2, and z_3) and under normality assumptions, we performed James multivariate analysis of variance with unequal variances to test the difference in the compositional mean between sex. The James test was available in the R pacage *Compositional*.

The table below (similar to the Table S2 in the paper) presents the compositional means in the overall sample and the groups defined by sex. P-value = 0.348 indicating insufficient evidence to reject the null hypothesis that the compositional means are equal between males and females.

Table 3: Compositional mean in subgroups

	N (%)	Sit	Stand	Step	Sleep	P-value
Overall	1000 (100 %)	10.3 (42.8%)	3.6 (15.2%)	1.3 (5.2%)	8.8 (36.8%)	
Sex						0.348
Male	414 (41.4 %)	10.23 (42.6 %)	3.73 (15.5 %)	1.27 (5.3 %)	8.78 (36.6 %)	
Female	586 (58.6 %)	10.3 (42.9 %)	3.58 (14.9 %)	1.25 (5.2 %)	8.87 (36.9 %)	

Note:
p-value from multivariate analysis of variance on the isometric log-transformed time use variables without assuming equal variance across subgroups.

CoDA regressions and interpretations

Next, we applied CoDA to estimate a type of time reallocation i.e. increasing time in one activity while simultaneously proportionally decreasing time in the other activities. To achieve this, it is convenient to create four sets of ilr-coordinates with each behavior in turn being singled out as the numerator in the pivot

coordinate z_1. Four linear regression models were fit with the continuous outcome, and with the resulting ilr-coordinates (z_1, z_2, z_3) as predictors. Each regression model is adjusted for sex.

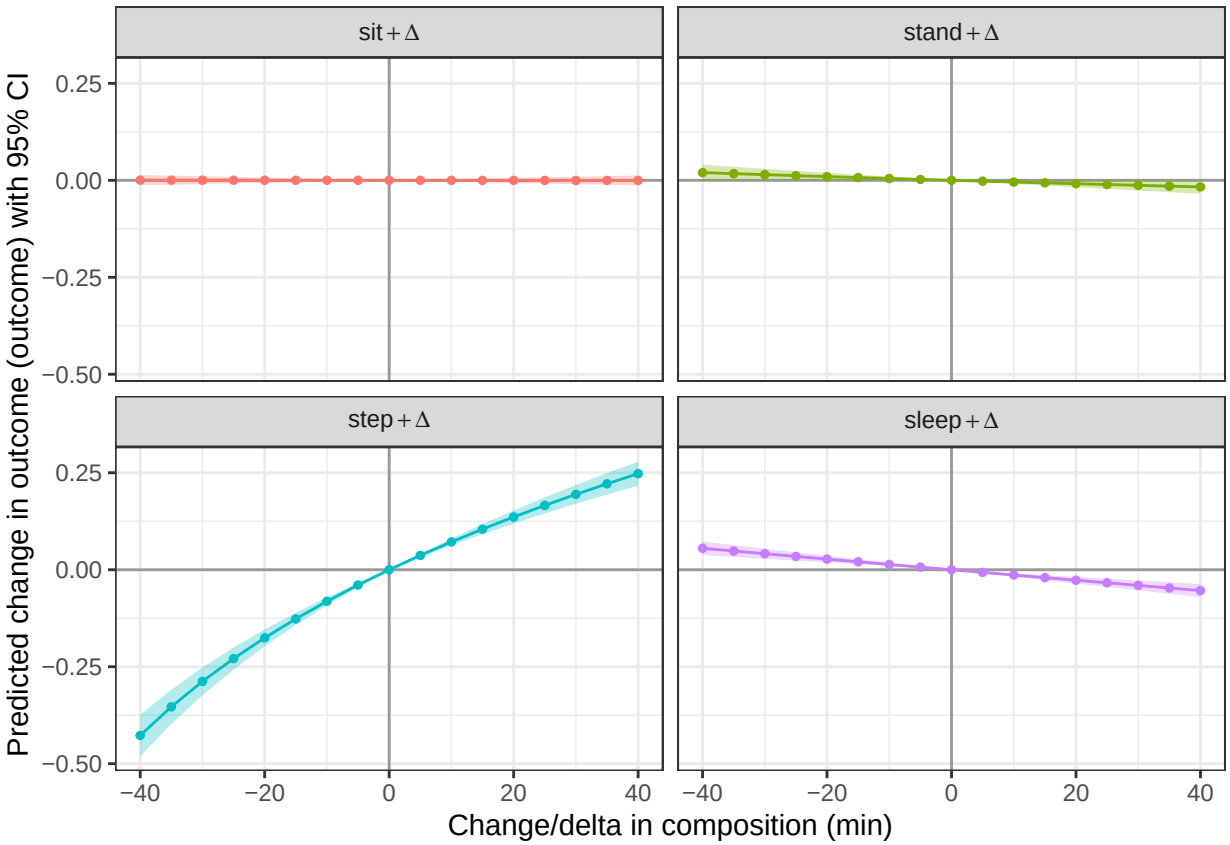
Table below (similar to the Table 3 in the paper) summarized regression coefficient estimates $\hat{\beta}_1$ for the four CoDA pivot coordinates, each of which quantifies the effect of increasing time in one behavior by a factor while simultaneously decreasing time in the other behaviors by another factor. To make meaningful interpretation of those $\hat{\beta}_1$ estimates, we need to consider a referent composition in order to inform what magnitude difference in z_1 is a meaningful difference. Suppose the compositional mean calculated over the entire sample is chosen as the referent composition, and we are interested in the effect of increasing step by a factor of $1 + r$. Then, all the other components should simultaneously be decreased by another factor $1 - s$ to maintain z_2 and z_3 constant. Some derivations show that the difference in the mean outcome for such a time reallocation equals to $\hat{\beta}_1\sqrt{\frac{3}{4}}log(\frac{1+r}{1-r})$. We created a R function *comp_contrast* that can output the difference between any two compositions in terms of ilr-coordinates (using *pivotCoord* function from the package *robCompositions*). More specifically, with a fitted CoDA regression model with z_1, z_2, z_3 (from pivotCoord) as predictors and a given referent composition say c_1, to estimate the effect of a specific time reallocation, we only need to know the composition c_2 after such time reallocation, and enter c_1 and c_2 into the *comp_contrast* function, we can obtain the difference between the two compositions in terms of z_1, z_2, and z_3. The estimated effect of such time reallocation is then a linear combination of that difference with linear coefficients as $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$. Note that CoDA regression results should be the same regardless of the form of ilr-transformations i.e. which set of ilr-coordinates is used.

Table 4: Regression of pivot coordinates against the outcome. Analysis controlled for sex. Remaining = remaining behaviors

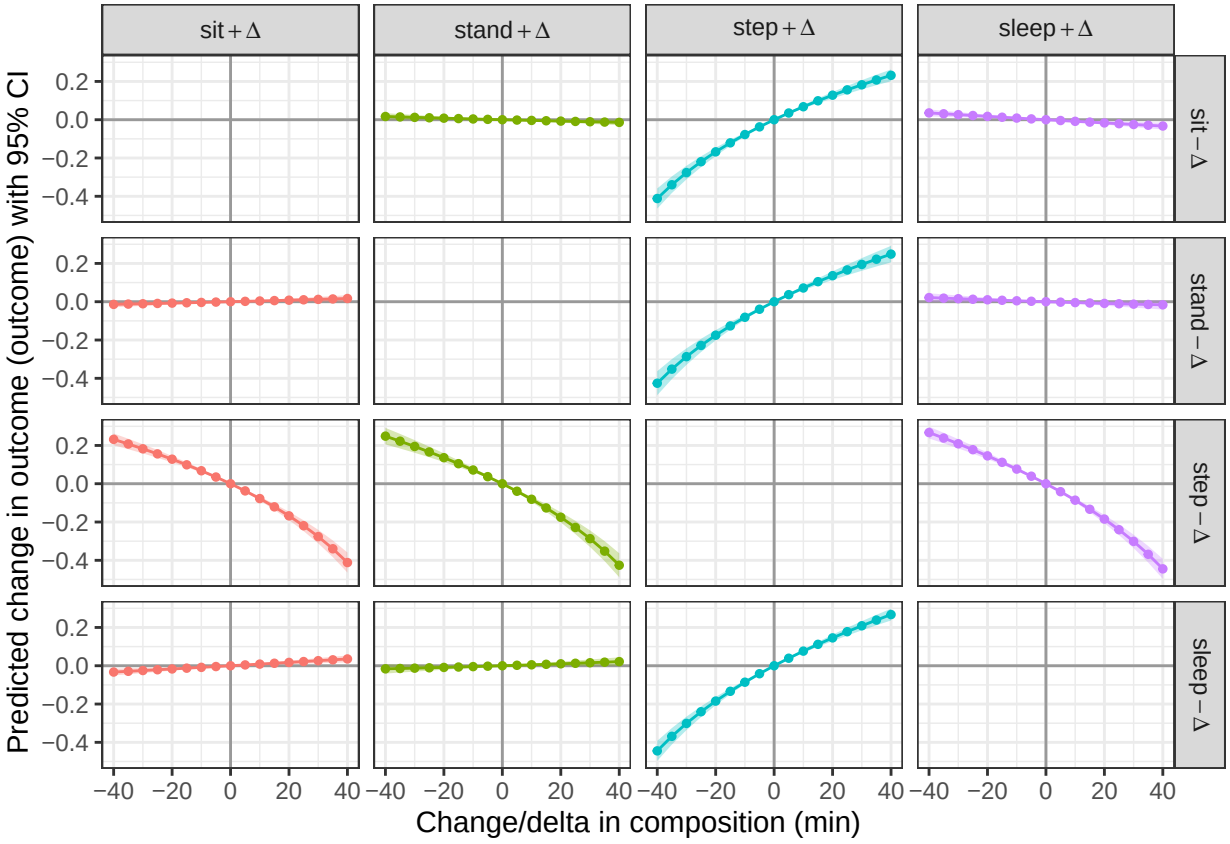
	Estimate (95% C.I.)	P-value
Sit vs Remaining	-0.00 [-0.14, 0.13]	0.949
Stand vs Remaining	-0.10 [-0.20, 0.00]	0.059
Step vs Remaining	0.63 [0.55, 0.71]	<0.001
Sleep vs Remaining	-0.53 [-0.69, -0.36]	<0.001

Based on the results in Table 4, we can see increasing time in step while proportionally decreasing time in the other activities is associated with higher mean outcome. In contrast, increasing time in sleep while proportionally decreasing time in the other activities is associated with lower mean outcome. More specifically, increasing 30 minsin step while proportionally decreasing time in the other behaviors is associated with 0.19 [0.17, 0.22] increase in the outcome. Increasing 30 mins in sleep while proportionally decreasing time in the other behaviors is associated with a decrease in mean outcome of 0.04 [0.03, 0.05].

The results can be visualized by using the package *codaredistlm* available on Github: github.com/tystan/codaredistlm. Please see the blow plots created by using the function *pred_df* and *plot_delta_comp*. The figure below is similar to Figure 2 in the paper.


sit + Δ
stand + Δ
step + Δ
sleep + Δ
Predicted change in outcome (outcome) with 95% CI
0.25
0.00
−0.25
−0.50
−40
−20
0
20
40
Change/delta in composition (min)

We can use the same functions to estimate and visualize the effect of time-reallocation between any pair of behaviors, e.g. reallocate time only between step and sit. See the plot below (similar to Figure 3 in the paper).



LPA

LPA is a more exploratory method used to identify distinct latent subgroups with respect to activity profiles based on observed 24HAC data. This analysis can be done in R using the package *tidyLPA* (https://cran.r-project.org/web/packages/tidyLPA/vignettes/Introduction_to_tidyLPA.html). Another objective of LPA is to analyze the potential correlates of latent profiles and the associations of the profiles with outcomes. However, this analysis requires specialized regression methods that account for class assignment uncertainty, which can be performed in either *Mplus* or *LatentGold*. Both are commercial software. To the best of our knowledge, those methods have not been implemented in any R package. To perform similar analysis as in our paper, please find the LatentGold syntax in the supplemental materials (A2.2) of our paper.

Here, we only performed LPA to identify latent subgroups based on the compositional variables *sit*, *stand*, *step*, and *sleep*, so-called profile indicators.

LPA assumes that the profile indicators follow a finite mixture of multivariate normal distributions with each latent subgroup having its own mean and possibly distinct variance-covariance structure. The key part in running LPA is to determine the best number of classes i.e. latent groups. Usually, a series of models with different number of latent classes are fit. The best model is selected based on the mix of several criteria e.g. fit statistics, statistical comparisons between models, smallest number of subjects assigned to a class, interpretability of the classes, etc.

A key feature of 24HAC data is the co-dependence between activity behaviors, which prevents us from applying LPA to all 24HAC variables because it will lead to a degenerate (rank-deficient) covariance matrix.

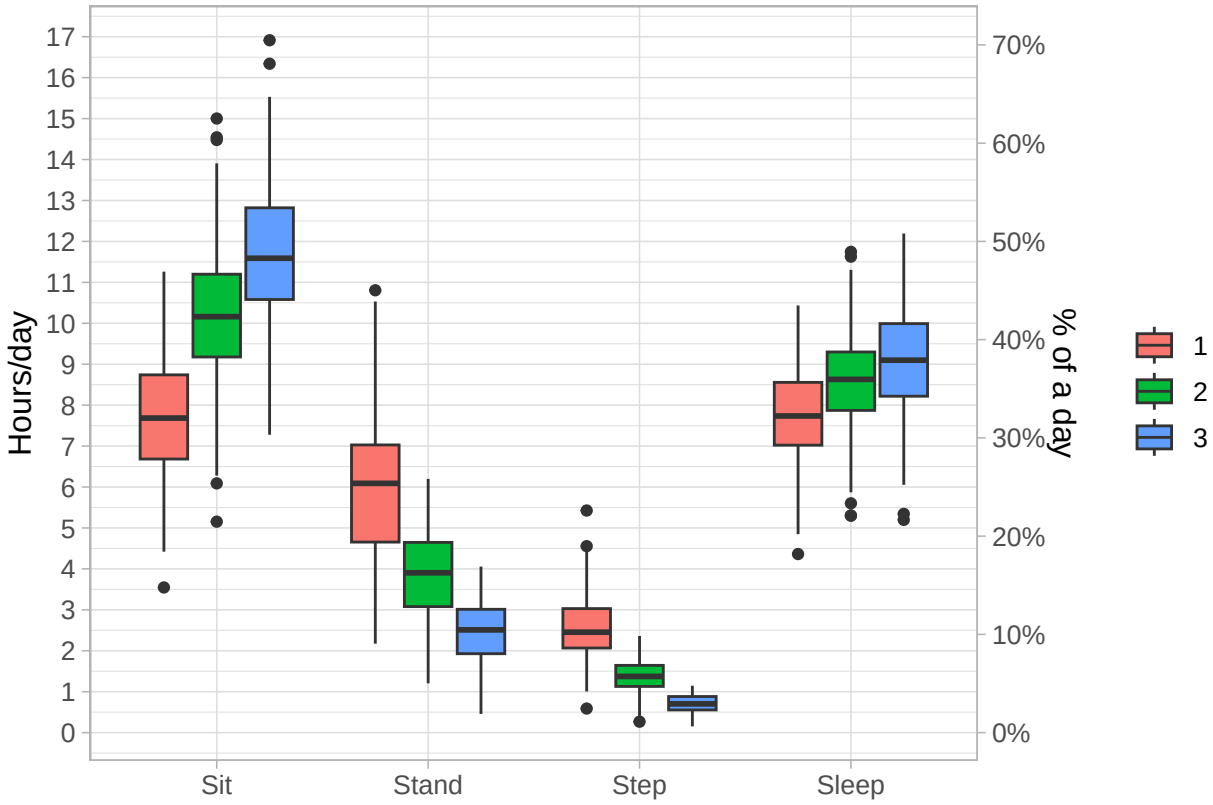
Two possible solutions to consider are (1) apply LPA to the same ilr-transformed variables used in CoDA (see CoDA section) or (2) drop one activity behavior variable from the LPA. Here, we chose to drop sleep from LPA. See our paper for more discussions.

We fit a series of LPA models with the number of classes ranging from 2 to 6 and allowed both variances and covariances of profile indicator variables to differ across latent classes (most flexible models). The Table 5 below shows the fit statistics for fitted models. AIC, BIC, CAIC, SABIC are all statistics balancing the model fit and complexity of the model. The difference between them is how they weigh the two objectives. In our case, they respectively favor 4-class model, 3-class model, 3-class model, 3-class model. Based on the bootstrap likelihood ratio test (BLRT) comparing the model with k and k-1 classes, we have strong evidence that the 3-class model is better than the 2-class model, but the 4-class model is not better than the 3-class model. Thus, 3-class model is used as the final model. For most cases, the decision process may be more complicated than here, and may need to be based on a mix of criteria and sometimes subjective.

Table 5: Fit statistics for latent profile models with 2-6 profiles

	Classes	LogLik	AIC	BIC	CAIC	SABIC	ICL	n_min	BLRT_p	Entropy
2-class model	2	5221.40	-10404.81	-10311.56	-10292.56	-10371.91	10067.89	0.35	0.01	0.64
3-class model	3	5278.90	-10499.80	-10357.48	-10328.48	-10449.58	9962.56	0.23	0.01	0.64
4-class model	4	5291.69	-10505.38	-10313.97	-10274.97	-10437.84	9852.59	0.08	0.08	0.67
5-class model	5	5296.22	-10494.44	-10253.96	-10204.96	-10409.59	9642.45	0.11	0.41	0.63
6-class model	6	5301.38	-10484.77	-10195.21	-10136.21	-10382.60	9550.89	0.06	0.84	0.65

Once we obtain the final model, we can get the probability of each subject belonging to different latent classes (a.k.a posterior probability of class membership). The figure below shows the distribution of four activities across the three classes after we assign every subject to the class with the highest probability (a.k.a modal assignment).



In most applications, after obtaining a model for latent classes, we are also interested how the latent classes are associated different covariates or health outcomes. Most studies in the literature simply assign every individual to the class with the highest posterior probability of class membership and treat that class assignments as known information to do further inferential analysis. However, this could lead to bias effects and underestimated SEs. In our paper, we discussed this issue and current approaches to dealing with that. Those approaches are currently only available in LatentGold and Mplus, but not in R, and hence not presented here.