

# Cultural distances: controlling for history

## Introduction

We compare cultural distances between societies with linguistic distances between societies, controlling for shared history in two ways.

The first test uses mixed effects modelling. The pairing of the language family of each language (according to Glottolog) is used as a random effect. That means that the model can capture the likelihood that two languages from the Indo-European language family will be more similar to each other than two languages from different language families. The same is done with geographic area according to Autotyp.

The second test controls for history using distances from a phylogenetic tree. The tree comes from Bouckaert et al. (2012). Patristic distances between languages are used as a measure of historical distance between societies in a Mantel test. Note that the Mantel test assumes a strict distance metric, which is not necessarily the case with this data, but there are few other ways to deal with continuous pairwise distances.

## Load libraries

```
library(ape)
library(ecodist)
library(lme4)
library(sjPlot)
library(ggplot2)
library(igraph)
library(lattice)
```

## All domains

### Load data

Read the cultural distances:

```
cult = read.csv("../results/EA_distances/CulturalDistances_Long.csv", stringsAsFactors = F)
names(cult) = c("l1", "l2", "cult.dist")
cultLangs = unique(c(cult$Var1, cult$Var2))
```

Add language family:

```
l = read.csv("../data/FAIR_langauges_glottolog_xdid.csv", stringsAsFactors = F)
g = read.csv("../data/glottolog-languoid.csv/languoid.csv", stringsAsFactors = F)
l$family = g[match(l$glotto, g$id),]$family_pk
l$family = g[match(l$family, g$pk),]$name
```

Read the semantic distances

```
ling = read.csv("../data/FAIR/semantic_distances_FAIR.csv", stringsAsFactors = F)
```

Combine the linguistic and cultural distances

```

cult$l1.iso2 = 1[match(cult$l1, l$Language2),]$iso2
cult$l2.iso2 = 1[match(cult$l2, l$Language2),]$iso2

fairisos = unique(c(ling$l1, ling$l2))
cultisos = unique(c(cult$l1.iso2, cult$l2.iso2))

cult = cult[(cult$l1.iso2 %in% fairisos) & (cult$l2.iso2 %in% fairisos),]
ling = ling[(ling$l1 %in% cultisos) & (ling$l2 %in% cultisos),]

matches = sapply(1:nrow(ling), function(i){
  which(cult$l1.iso2==ling$l1[i] & cult$l2.iso2==ling$l2[i])
})

ling$cult.dist = cult[matches,]$cult.dist
ling$cult.dist.center = scale(ling$cult.dist)
cdc.s = attr(ling$cult.dist.center, "scaled:scale")
cdc.c = attr(ling$cult.dist.center, "scaled:center")
ling$cult.dist.center = as.numeric(ling$cult.dist.center)

ling$family1 = 1[match(ling$l1, l$iso2),]$family
ling$family2 = 1[match(ling$l2, l$iso2),]$family
ling$area1 = 1[match(ling$l1, l$iso2),]$autotyp.area
ling$area2 = 1[match(ling$l2, l$iso2),]$autotyp.area

fgroup = cbind(ling$family1, ling$family2)
fgroup = apply(fgroup, 1, sort)
ling$family.group = apply(fgroup, 2, paste, collapse=":")
agroup = cbind(ling$area1, ling$area2)
agroup = apply(agroup, 1, sort)
ling$area.group = apply(agroup, 2, paste, collapse=":")

ling$rho.center = scale(ling$rho)

```

Each observation is now associated with a language family pair:

```
head(ling[,c("l1", "l2", "rho", "family.group")])
```

```

##      l1 l2      rho      family.group
## 1  be es 0.3248551 Indo-European:Indo-European
## 2  bg be 0.3464910 Indo-European:Indo-European
## 4  bg cs 0.4091899 Indo-European:Indo-European
## 7  bg en 0.3145696 Indo-European:Indo-European
## 8  bg es 0.3686999 Indo-European:Indo-European
## 10 bg fr 0.3742889 Indo-European:Indo-European

```

And the same is true for area:

```
tail(ling[,c("l1", "l2", "rho", "area.group")])
```

```

##      l1 l2      rho      area.group
## 375 ru he 0.2888911 Greater Mesopotamia:Inner Asia
## 377 ru ja 0.2721048      Inner Asia:N Coast Asia
## 379 ru kk 0.2931095      Inner Asia:Inner Asia
## 380 ru la 0.2125645      Europe:Inner Asia
## 381 ru lt 0.3635630      Inner Asia:Inner Asia

```

## LMER models

Mixed effects model, predicting linguistic distances from cultural distances, with random intercept for family and area and random slope for cultural distance for family and area.

We compare a null model to a model with a fixed effect for cultural distance, with random intercepts for family and area, and random slopes for cultural distance by both.

```
m0 = lmer(
  rho.center ~ 1 +
    (1 + cult.dist.center | family.group) +
    (1 + cult.dist.center | area.group),
  data = ling
)
m1 = lmer(
  rho.center ~ 1 +
    cult.dist.center +
    (1 + cult.dist.center | family.group) +
    (1 + cult.dist.center | area.group),
  data = ling
)
anova(m0,m1)

## refitting model(s) with ML (instead of REML)

## Data: ling
## Models:
## m0: rho.center ~ 1 + (1 + cult.dist.center | family.group) + (1 +
## m0:      cult.dist.center | area.group)
## m1: rho.center ~ 1 + cult.dist.center + (1 + cult.dist.center | family.group) +
## m1:      (1 + cult.dist.center | area.group)
##      Df      AIC      BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## m0  8 289.70 311.00 -136.85  273.70
## m1  9 284.74 308.71 -133.37  266.74 6.9602      1  0.008334 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cultural distance is significantly correlated with linguistic distance. Here are the model estimates:

```
summary(m1)

## Linear mixed model fit by REML ['lmerMod']
## Formula:
## rho.center ~ 1 + cult.dist.center + (1 + cult.dist.center | family.group) +
##      (1 + cult.dist.center | area.group)
##      Data: ling
##
## REML criterion at convergence: 271.9
##
## Scaled residuals:
##      Min      1Q  Median      3Q      Max
## -2.3839 -0.5481  0.1305  0.5749  2.6242
##
## Random effects:
```

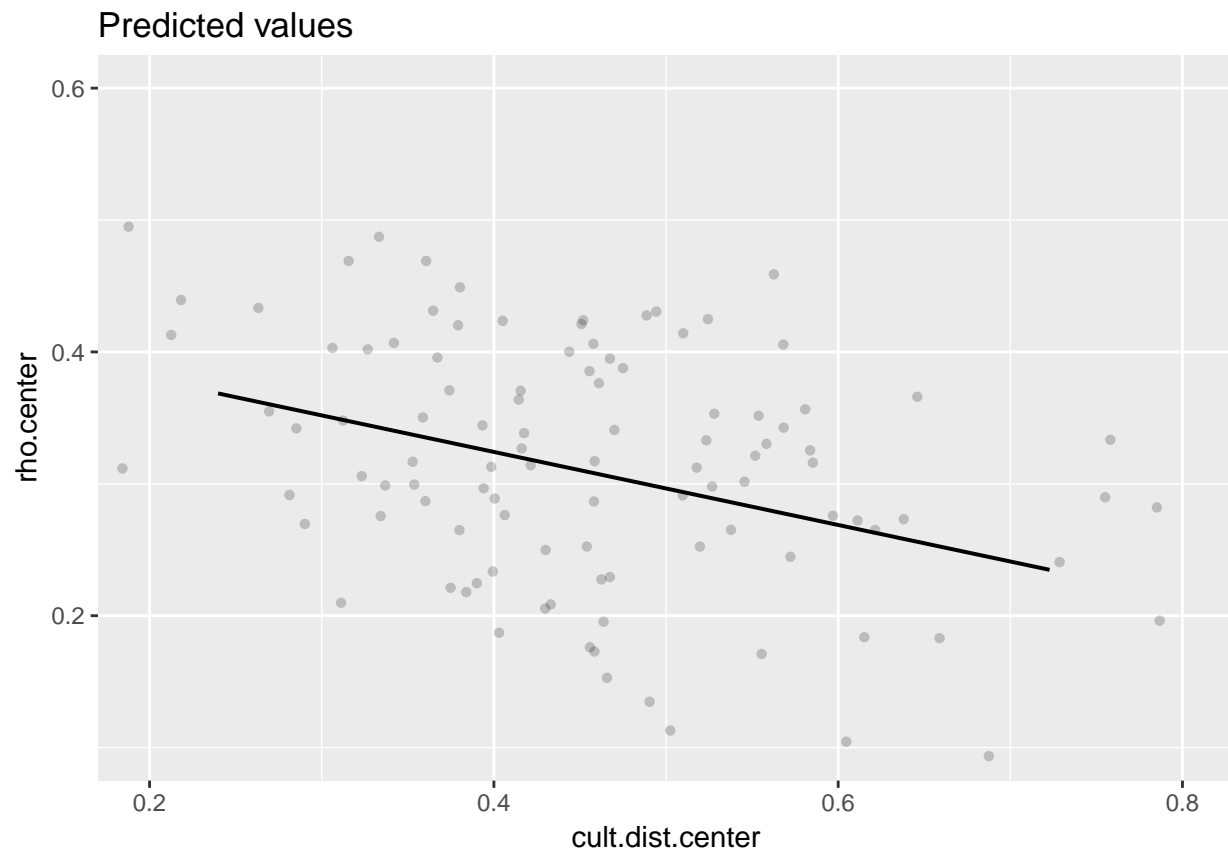
```
## Groups      Name      Variance Std.Dev. Corr
## family.group (Intercept) 0.000e+00 0.000e+00
##              cult.dist.center 2.227e-17 4.719e-09 NaN
## area.group   (Intercept) 5.688e-02 2.385e-01
##              cult.dist.center 6.269e-02 2.504e-01 1.00
## Residual                    6.821e-01 8.259e-01
## Number of obs: 106, groups: family.group, 11; area.group, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   -0.08275    0.11976  -0.691
## cult.dist.center -0.62755    0.12624  -4.971
##
## Correlation of Fixed Effects:
##              (Intr)
## clt.dst.cnt 0.557
```

Plot the estimates, rescaling the variables back to the original units:

```
gx = sjp.lmer(m1, 'pred', 'cult.dist.center', prnt.plot = F)
```

```
## Warning: package 'bindrcpp' was built under R version 3.3.2
```

```
gx$plot$data$y = gx$plot$data$y *
  attr(ling$rho.center, "scaled:scale") +
  attr(ling$rho.center, "scaled:center")
gx$plot$data$res.p.y = gx$plot$data$res.p.y *
  attr(ling$rho.center, "scaled:scale") +
  attr(ling$rho.center, "scaled:center")
gx$plot$data$x = gx$plot$data$x *
  cdc.s + cdc.c
gx$plot + coord_cartesian(ylim=c(0.1,0.6),
                          xlim=c(0.2,0.8))
```

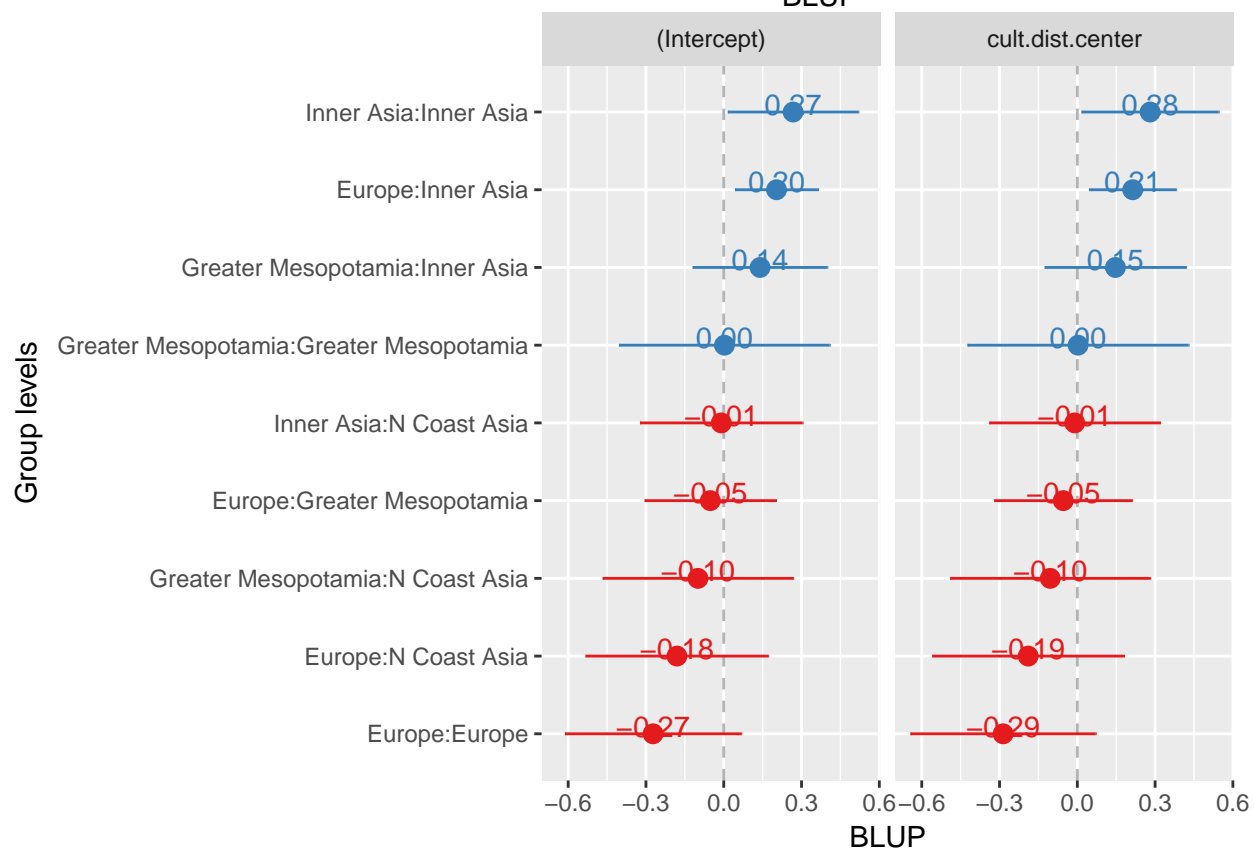
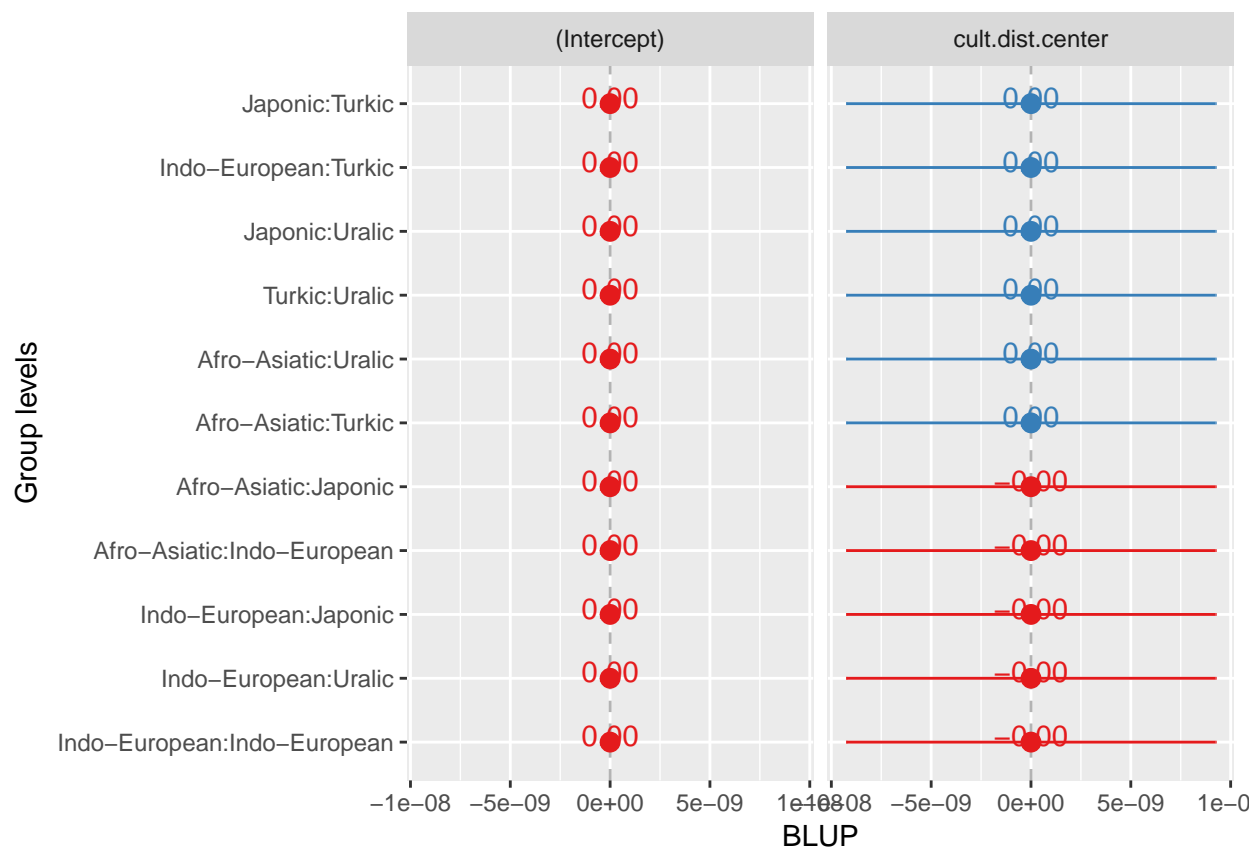


Plot the random effects:

```
sjp.lmer(m1, 're', sort.est = "cult.dist.center")
```

```
## Plotting random effects...
```

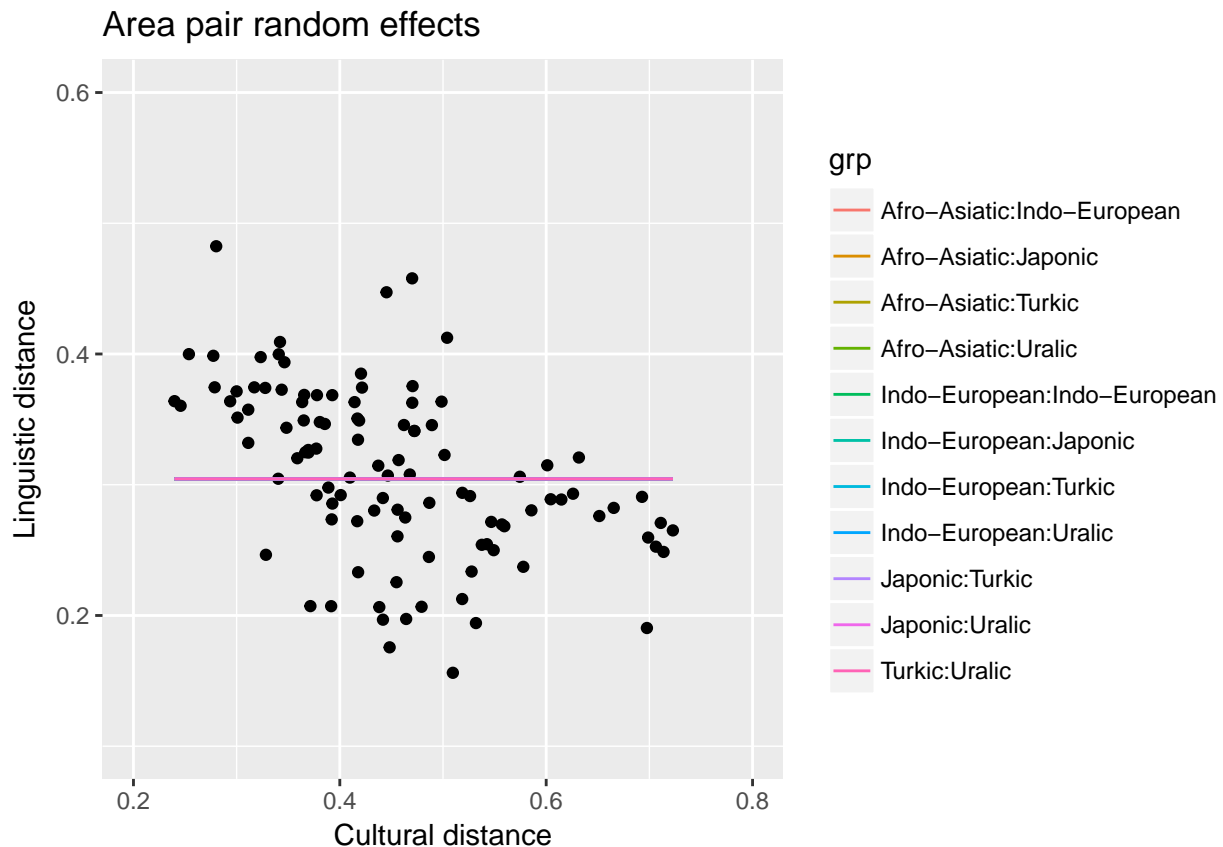
```
## Plotting random effects...
```



```

px = sjp.lmer(m1, 'rs.ri', prnt.plot = F)
dx = px$plot[[1]]$data
dx$x = dx$x * cdc.s + cdc.c
dx$y = dx$y *
  attr(ling$rho.center, "scaled:scale") +
  attr(ling$rho.center, "scaled:center")
ggplot(dx, aes(x, y)) +
  geom_point(data=ling,
             mapping=aes(x=as.numeric(cult.dist),
                        y=as.numeric(rho))) +
  geom_line(mapping = aes(colour=grp)) +
  xlab("Cultural distance")+
  ylab("Linguistic distance") +
  ggtitle("Area pair random effects") +
  coord_cartesian(ylim=c(0.1,0.6),
                 xlim=c(0.2,0.8))

```

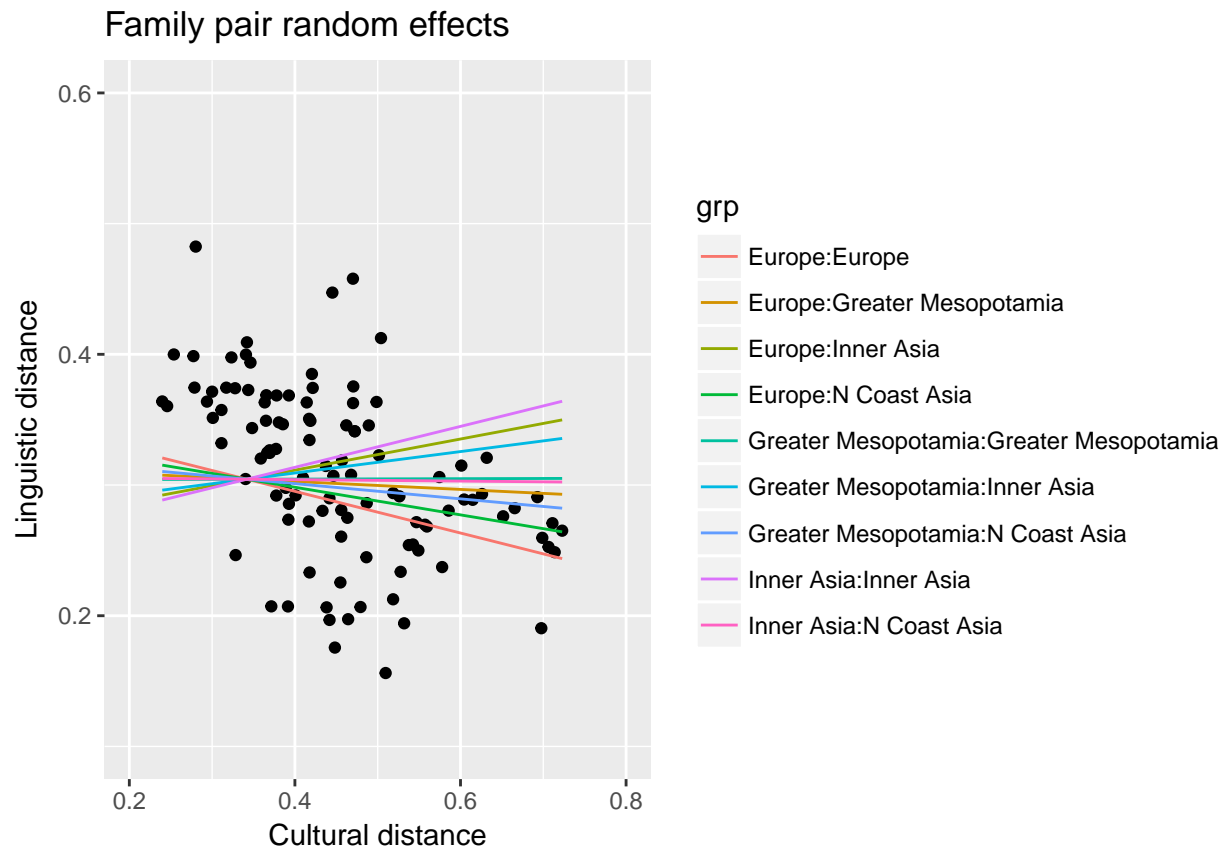


```

dx = px$plot[[2]]$data
dx$x = dx$x * cdc.s + cdc.c
dx$y = dx$y *
  attr(ling$rho.center, "scaled:scale") +
  attr(ling$rho.center, "scaled:center")
ggplot(dx, aes(x, y)) +
  geom_point(data=ling,
             mapping=aes(x=as.numeric(cult.dist),
                        y=as.numeric(rho))) +

```

```
geom_line(mapping = aes(colour=grp)) +
xlab("Cultural distance")+
ylab("Linguistic distance") +
ggtitle("Family pair random effects") +
coord_cartesian(ylim=c(0.1,0.6),
                xlim=c(0.2,0.8))
```



Note that the random slopes for family are set to 0. We can check whether taking language family out makes a difference:

```
m0b = lmer(
  rho.center ~ 1 +
    (1 + cult.dist.center | area.group),
  data = ling
)
m1b = lmer(
  rho.center ~ 1 +
    cult.dist.center +
    (1 + cult.dist.center | area.group),
  data = ling
)
anova(m0b,m1b)
```

```
## refitting model(s) with ML (instead of REML)
## Data: ling
## Models:
## m0b: rho.center ~ 1 + (1 + cult.dist.center | area.group)
```



```
## m1b: rho.center ~ 1 + cult.dist.center + (1 + cult.dist.center | area.group)
##      Df      AIC      BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## m0b  5 289.45 302.77 -139.73  279.45
## m1b  6 278.74 294.72 -133.37  266.74 12.715      1 0.0003628 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(m1b)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## rho.center ~ 1 + cult.dist.center + (1 + cult.dist.center | area.group)
##      Data: ling
##
## REML criterion at convergence: 271.9
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -2.3839 -0.5481  0.1305  0.5749  2.6242
##
## Random effects:
##      Groups      Name                Variance Std.Dev. Corr
## area.group (Intercept)      0.05688  0.2385
##                cult.dist.center 0.06269  0.2504  1.00
## Residual                    0.68214  0.8259
## Number of obs: 106, groups: area.group, 9
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   -0.08275    0.11976  -0.691
## cult.dist.center -0.62755    0.12624  -4.971
##
## Correlation of Fixed Effects:
##              (Intr)
## clt.dst.cnt 0.557
```

The model is numerically almost exactly the same, so the p-value from the model comparison is just lower.

## Tests within domains

Load distances for specific domains and match up to language family and area:

```
ling.dom = read.csv(
  "../results/EA_distances/All_Domains_with_ling.csv",
  stringsAsFactors = F)

ling.dom = ling.dom[!is.na(ling.dom$cult.dist),]

ling.dom$family1 = l[match(ling.dom$l1, l$iso2),]$family
ling.dom$family2 = l[match(ling.dom$l2, l$iso2),]$family
ling.dom$area1 = l[match(ling.dom$l1, l$iso2),]$autotyp.area
ling.dom$area2 = l[match(ling.dom$l2, l$iso2),]$autotyp.area

# Paste language family names together,
# but order shouldn't matter, so sort first
fgroup = cbind(ling.dom$family1, ling.dom$family2)
fgroup = apply(fgroup, 1, sort)
ling.dom$family.group = apply(fgroup, 2, paste, collapse=":")

agroup = cbind(ling.dom$area1, ling.dom$area2)
agroup = apply(agroup, 1, sort)
ling.dom$area.group = apply(agroup, 2, paste, collapse=":")
```

Center the data:

```
ling.dom$cult.dist.center = scale(ling.dom$cult.dist)
ling.dom$rho.center = scale(ling.dom$rho)
```

## LMER models

Test whether random slopes are warranted for family:

```
mD0 = lmer(
  rho.center ~ 1 +
    (1 | family.group) +
    (1 | area.group) +
    (1 | imputed_semantic_domain),
  data = ling.dom)
mD1 = lmer(
  rho.center ~ 1 +
    (1 | family.group) +
    (0 + cult.dist.center | family.group) +
    (1 | area.group) +
    (1 | imputed_semantic_domain),
  data = ling.dom)
mD2 = lmer(
  rho.center ~ 1 +
    (1 + cult.dist.center | family.group) +
    (1 | area.group) +
    (1 | imputed_semantic_domain),
  data = ling.dom)
```

```
anova(mD0,mD1,mD2)
```

```
## refitting model(s) with ML (instead of REML)

## Data: ling.dom
## Models:
## mD0: rho.center ~ 1 + (1 | family.group) + (1 | area.group) + (1 |
## mD0:   imputed_semantic_domain)
## mD1: rho.center ~ 1 + (1 | family.group) + (0 + cult.dist.center |
## mD1:   family.group) + (1 | area.group) + (1 | imputed_semantic_domain)
## mD2: rho.center ~ 1 + (1 + cult.dist.center | family.group) + (1 |
## mD2:   area.group) + (1 | imputed_semantic_domain)
##      Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## mD0  5 2784.1 2808.7 -1387.0  2774.1
## mD1  6 2772.6 2802.1 -1380.3  2760.6 13.4788      1 0.0002413 ***
## mD2  7 2774.3 2808.8 -1380.1  2760.3  0.2902      1 0.5901181
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Random slopes for family improves the fit of the model, but adding the correlation coefficient does not.

Test the same for area:

```
mD3 = lmer(
  rho.center ~ 1 +
    (1 | family.group) +
    (0 + cult.dist.center | family.group) +
    (1 | area.group) +
    (0 + cult.dist.center | area.group) +
    (1 | imputed_semantic_domain),
  data = ling.dom)
anova(mD1,mD3)
```

```
## refitting model(s) with ML (instead of REML)

## Data: ling.dom
## Models:
## mD1: rho.center ~ 1 + (1 | family.group) + (0 + cult.dist.center |
## mD1:   family.group) + (1 | area.group) + (1 | imputed_semantic_domain)
## mD3: rho.center ~ 1 + (1 | family.group) + (0 + cult.dist.center |
## mD3:   family.group) + (1 | area.group) + (0 + cult.dist.center |
## mD3:   area.group) + (1 | imputed_semantic_domain)
##      Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## mD1  6 2772.6 2802.1 -1380.3  2760.6
## mD3  7 2774.2 2808.7 -1380.1  2760.2 0.3702      1 0.5429
```

Random slopes for area do not improve the fit of the model.

```
mD4 = lmer(
  rho.center ~ 1 +
    (1 + cult.dist.center | family.group) +
    (1 | area.group) +
    (1 | imputed_semantic_domain),
  data = ling.dom)
```

And, in fact, area random intercepts are all set to 0:

```
ranef(mD4)$area.group
```

```
##                                (Intercept)
## Europe:Europe                    0
## Europe:Greater Mesopotamia      0
## Europe:Inner Asia                0
## Europe:N Coast Asia              0
## Greater Mesopotamia:Greater Mesopotamia 0
## Greater Mesopotamia:Inner Asia    0
## Greater Mesopotamia:N Coast Asia  0
## Inner Asia:Inner Asia            0
## Inner Asia:N Coast Asia          0
```

So we'll take them out.

Now we test the main effect of cultural distance:

```
mD5 = lmer(
  rho.center ~ 1 +
    (1 + cult.dist.center || family.group) +
    (1 + cult.dist.center | imputed_semantic_domain),
  data = ling.dom)
mD6 = update(mD5, ~.+cult.dist.center)
anova(mD5,mD6)
```

```
## refitting model(s) with ML (instead of REML)

## Data: ling.dom
## Models:
## mD5: rho.center ~ 1 + ((1 | family.group) + (0 + cult.dist.center |
## mD5:      family.group)) + (1 + cult.dist.center | imputed_semantic_domain)
## mD6: rho.center ~ (1 | family.group) + (0 + cult.dist.center | family.group) +
## mD6:      (1 + cult.dist.center | imputed_semantic_domain) + cult.dist.center
##      Df    AIC    BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## mD5  7 2757.2 2791.7 -1371.6  2743.2
## mD6  8 2756.7 2796.1 -1370.4  2740.7 2.4866      1    0.1148
```

Summary of the final model, with random effects plot:

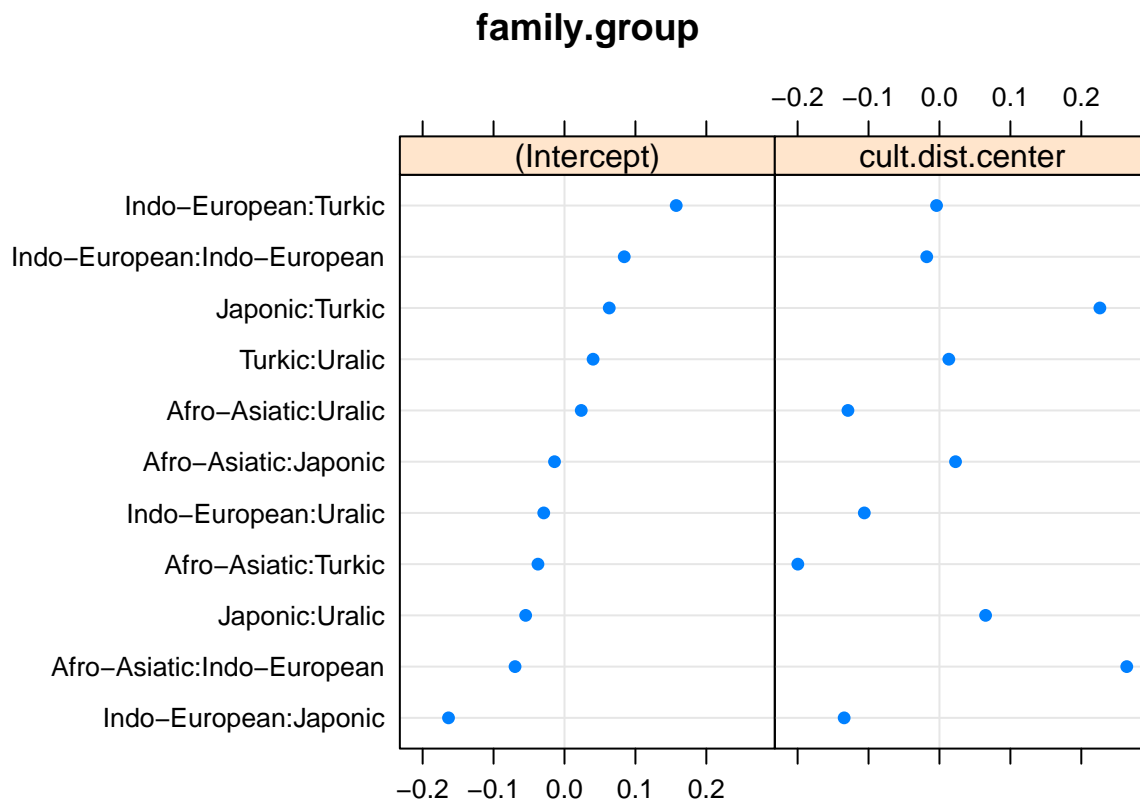
```
summary(mD6)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula:
## rho.center ~ (1 | family.group) + (0 + cult.dist.center | family.group) +
##      (1 + cult.dist.center | imputed_semantic_domain) + cult.dist.center
##      Data: ling.dom
##
## REML criterion at convergence: 2745.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -6.6194 -0.4231  0.0558  0.4703  4.1688
##
## Random effects:
##      Groups              Name                Variance Std.Dev. Corr
##      family.group          (Intercept)         0.02071  0.1439
##      family.group.1      cult.dist.center  0.04098  0.2024
```

```
## imputed_semantic_domain (Intercept)      0.13203  0.3634
##                                     cult.dist.center 0.03394  0.1842  -0.36
## Residual                                0.81293  0.9016
## Number of obs: 1020, groups:
## family.group, 11; imputed_semantic_domain, 10
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)   -0.08693    0.13504  -0.644
## cult.dist.center -0.16863    0.10501  -1.606
##
## Correlation of Fixed Effects:
##              (Intr)
## clt.dst.cnt -0.248
```

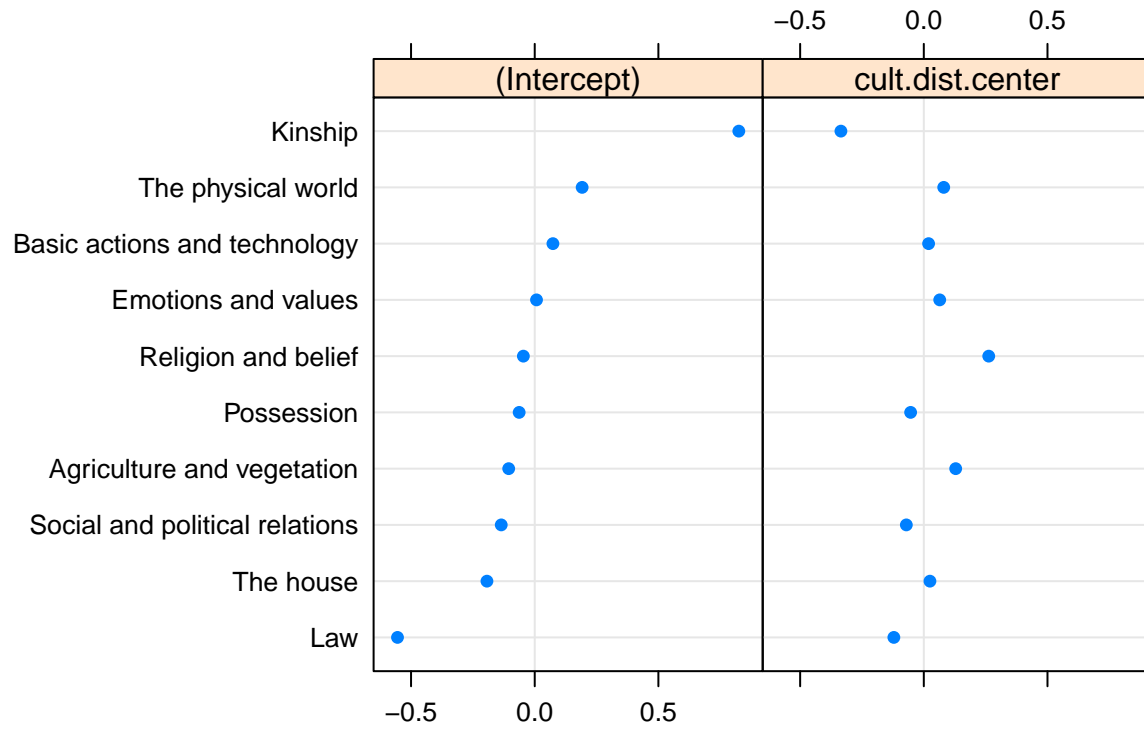
```
dotplot(ranef(mD6))
```

```
## $family.group
```



```
##
## $imputed_semantic_domain
```

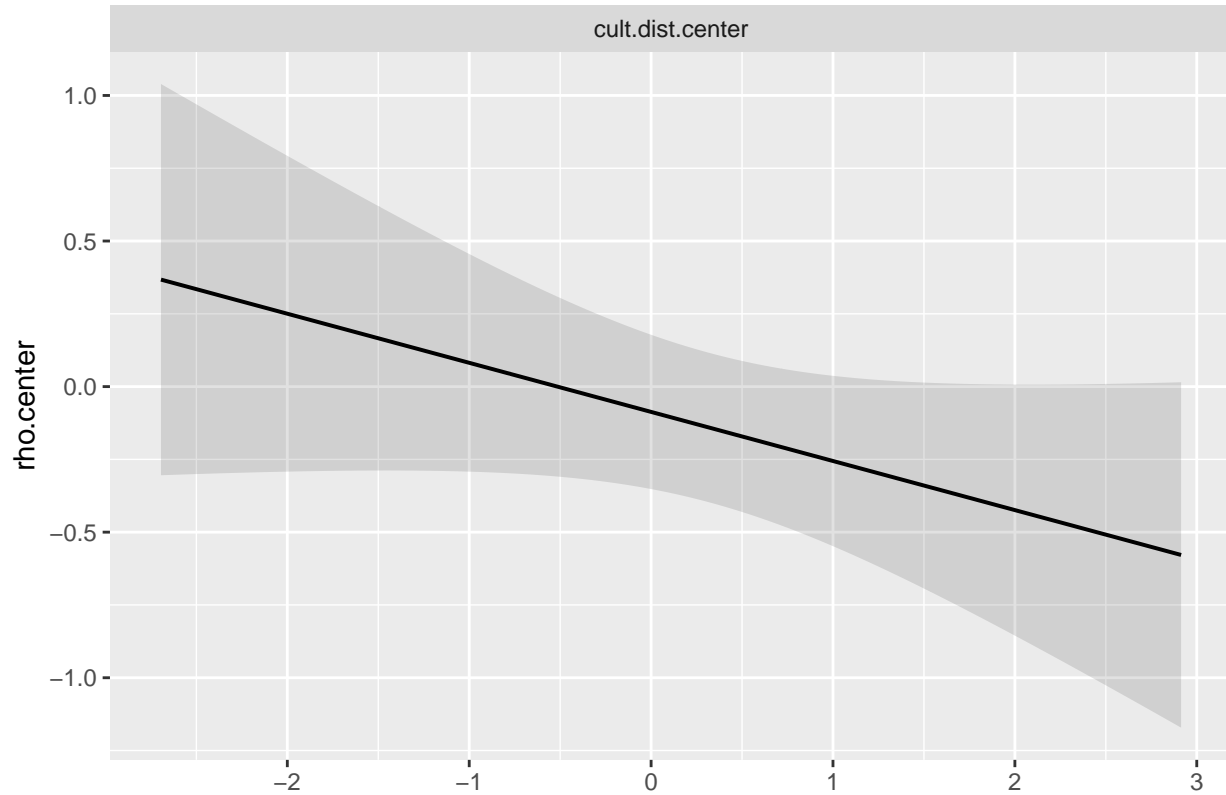
## imputed\_semantic\_domain



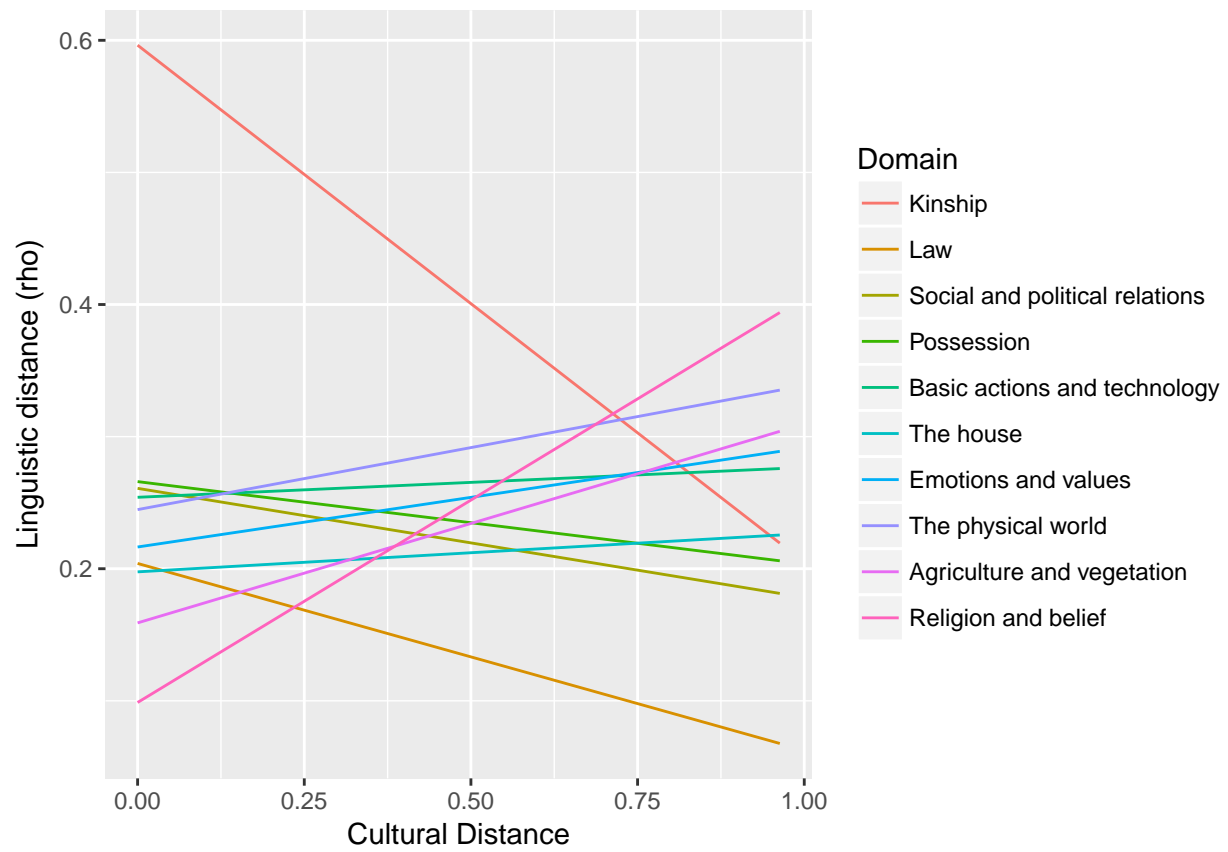
Plot the predicted relationships for each domain. The domains in the legend are sorted by the slope for cultural distance (greatest negative slope to greatest positive slope):

```
sjp.lmer(mD6,'eff', show.ci = T)
```

### Marginal effects of model predictors



```
dom.order = ranef(mD6)$imputed_semantic_domain
dom.order = rownames(dom.order[order(dom.order$cult.dist.center),])
px = sjp.lmer(mD6,'rs.ri', show.ci = T, prnt.plot = F)
pdx = px$plot[[2]]$data
pdx$Domain = factor(pdx$grp, levels = dom.order)
pdx$x = pdx$x *
  attr(ling.dom$cult.dist.center,"scaled:scale") +
  attr(ling.dom$cult.dist.center,"scaled:center")
pdx$y = pdx$y *
  attr(ling.dom$rho.center,"scaled:scale") +
  attr(ling.dom$rho.center,"scaled:center")
ggplot(pdx,
  aes(x,y,colour=Domain)) +
  geom_line() +
  xlab("Cultural Distance") +
  ylab("Linguistic distance (rho)")
```



“Religion and belief” seems to be working differently from the other domains.



## Mantel tests

Read the historical distances for Indo-European, based on the phylogenetic distances.

### Data prep

Load historical distances:

```
hist = read.csv("../data/trees/IndoEuropean_historical_distances.csv", stringsAsFactors = F)
hist = hist[!duplicated(hist[,1]),!duplicated(hist[,1])]
rownames(hist) = hist[,1]
hist = hist[,2:ncol(hist)]
hist.m = as.matrix(hist)
colnames(hist.m) = rownames(hist.m)
hist.m = hist.m/max(hist.m)
```

Read the cultural distances as a matrix:

```
cult.m = read.csv("../results/EA_distances/CulturalDistances.csv", stringsAsFactors = F)
rownames(cult.m) = cult.m[,1]
cult.m = cult.m[,2:ncol(cult.m)]
```

Convert the linguistic distances to a matrix. This uses `igraph` to make an undirected graph from the long format with `rho` as the edge weights, then output a matrix of adjacencies.

```
grph <- graph.data.frame(ling[,c("l1", "l2", "rho")], directed=FALSE)
# add value as a weight attribute
ling.m = get.adjacency(grph, attr="rho", sparse=FALSE)
rownames(ling.m) = 1[match(rownames(ling.m), l$iso2),]$Language2
colnames(ling.m) = 1[match(colnames(ling.m), l$iso2),]$Language2
```

Match the distance matrices

```
in.analysis = intersect(rownames(ling.m), rownames(cult.m))
in.analysis = intersect(in.analysis, rownames(hist.m))
cult.m2 = cult.m[in.analysis, in.analysis]
ling.m2 = ling.m[in.analysis, in.analysis]
hist.m2 = hist.m[in.analysis, in.analysis]
```

Note that there are only 11 languages with data on linguistic, cultural and historical distance.

### Tests

Simple correlation without control for history:

```
ecodist::mantel(as.dist(cult.m2) ~
               as.dist(ling.m2),
               nperm = 10000)
```

```
##      mantelr      pval1      pval2      pval3  llim.2.5% ulim.97.5%
## -0.3361506  0.8837000  0.1164000  0.1688000 -0.6866126 -0.1283738
```

Run a mantel test comparing the linguistic distances to the cultural distances, controlling for the historical distance between languages:

```
ecodist::mantel(as.dist(cult.m2)~  
                as.dist(ling.m2) +  
                as.dist(hist.m2),  
                nperm = 10000)
```

```
##      mantelr      pval1      pval2      pval3  llim.2.5%  ulim.97.5%  
## -0.31463547  0.86780000  0.13230000  0.20350000 -0.63040164 -0.03908519
```

## References

Bouckaert, Remco, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson (2012). Mapping the origins and expansion of the Indo-European language family. *Science*, 337(6097), 957-960.