Statistical Rethinking

Week 9:
Multilevel Models II
Adventures in Covariance

Richard McElreath

Varying slopes by dept

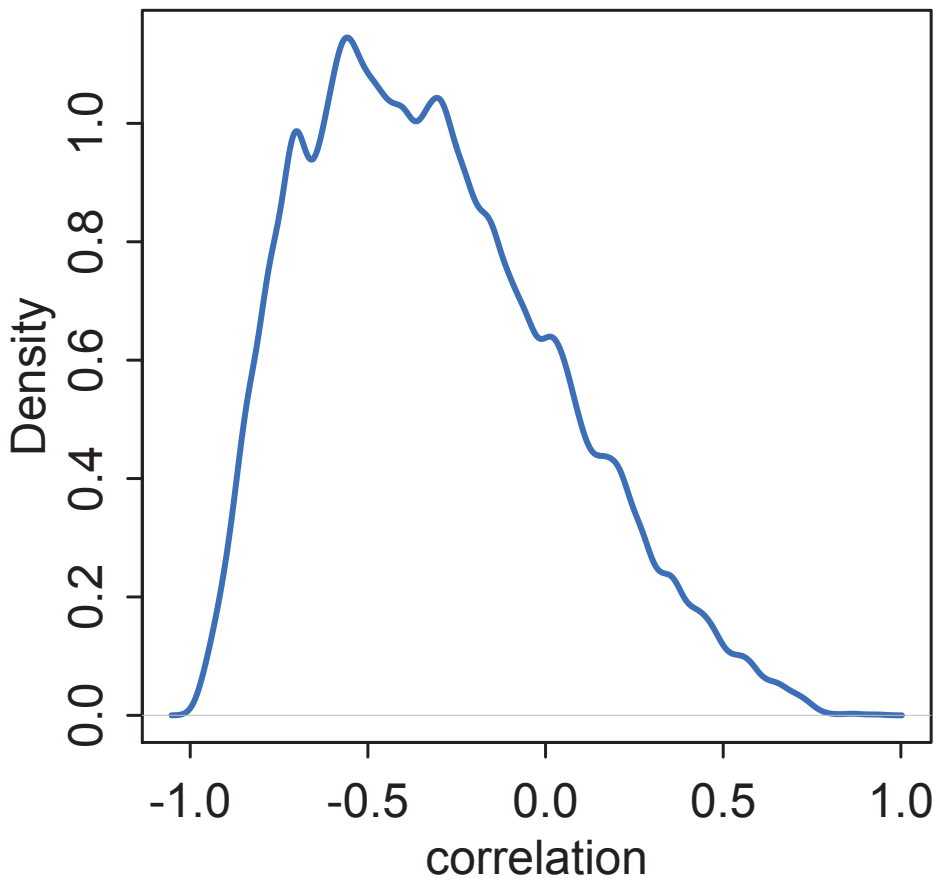
$$
\begin{aligned}
A_i &\sim \text{Binomial}(n_i, p_i) \\
\text{logit}(p_i) &= \alpha_{\text{DEPT}[i]} + \beta_{\text{DEPT}[i]} m_i \\
\begin{bmatrix} \alpha_{\text{DEPT}} \\ \beta_{\text{DEPT}} \end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \mathbf{S} \right) \\
\mathbf{S} &= \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \mathbf{R} \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta &\sim \text{Normal}(0, 1) \\
(\sigma_\alpha, \sigma_\beta) &\sim \text{HalfCauchy}(0, 2) \\
\mathbf{R} &\sim \text{LKJcorr}(2)
\end{aligned}
$$

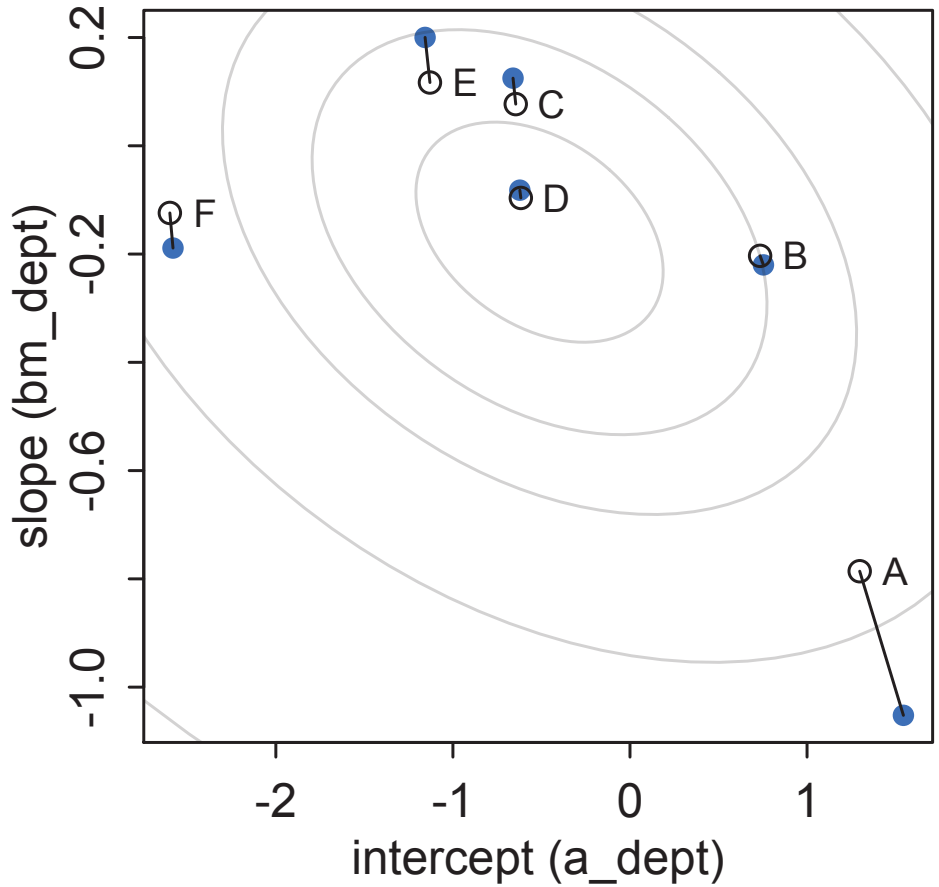
Varying slopes by dept

R code 13.19

```
m13.3 <- map2stan(
    alist(
        admit ~ dbinom( applications , p ),
        logit(p) <- a_dept[dept_id] +
                    bm_dept[dept_id]*male,
        c(a_dept,bm_dept)[dept_id] ~ dmvnorm2( c(a,bm) , sigma_dept , Rho ),
        a ~ dnorm(0,10),
        bm ~ dnorm(0,1),
        sigma_dept ~ dcauchy(0,2),
        Rho ~ dlkjcorr(2)
    ) ,
    data=d , warmup=1000 , iter=5000 , chains=4 , cores=3 )
```

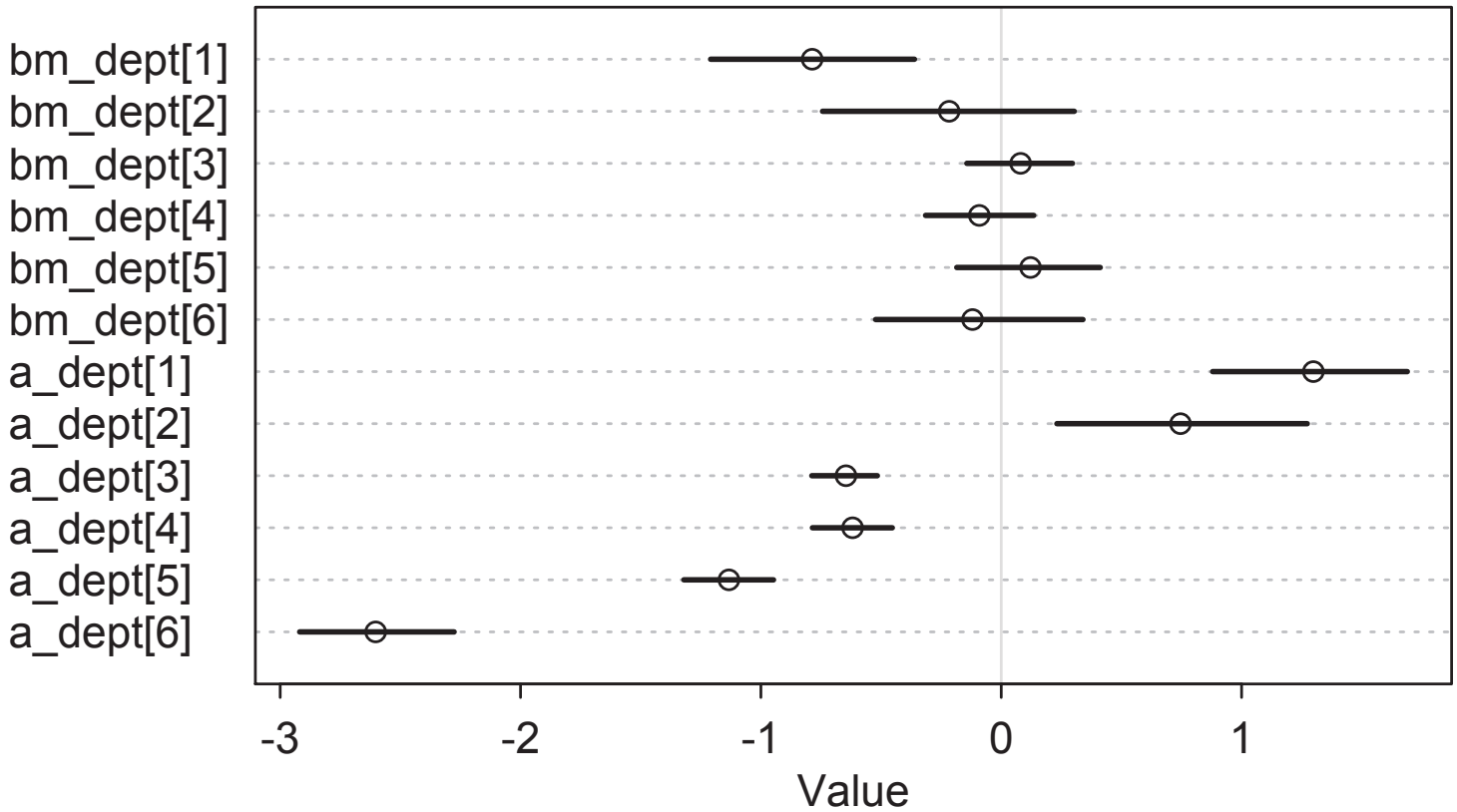
Correlated effects





Average effects can be a decoy

- Average slope not necessarily of interest
- Average can be zero, even when predictor very important for prediction



Cross-classified varying slopes

- More slopes: Higher dimension covariance matrix
- More clusters: More than one multivariate prior
- Reconsider chimpanzees data

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \mathcal{A}_i + (\mathcal{B}_{P,i} + \mathcal{B}_{PC,i} C_i) P_i \\
\mathcal{A}_i &= \alpha + \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]} \\
\mathcal{B}_{P,i} &= \beta_P + \beta_{P,\text{ACTOR}[i]} + \beta_{P,\text{BLOCK}[i]} \\
\mathcal{B}_{PC,i} &= \beta_P + \beta_{PC,\text{ACTOR}[i]} + \beta_{PC,\text{BLOCK}[i]}
\end{aligned}
$$

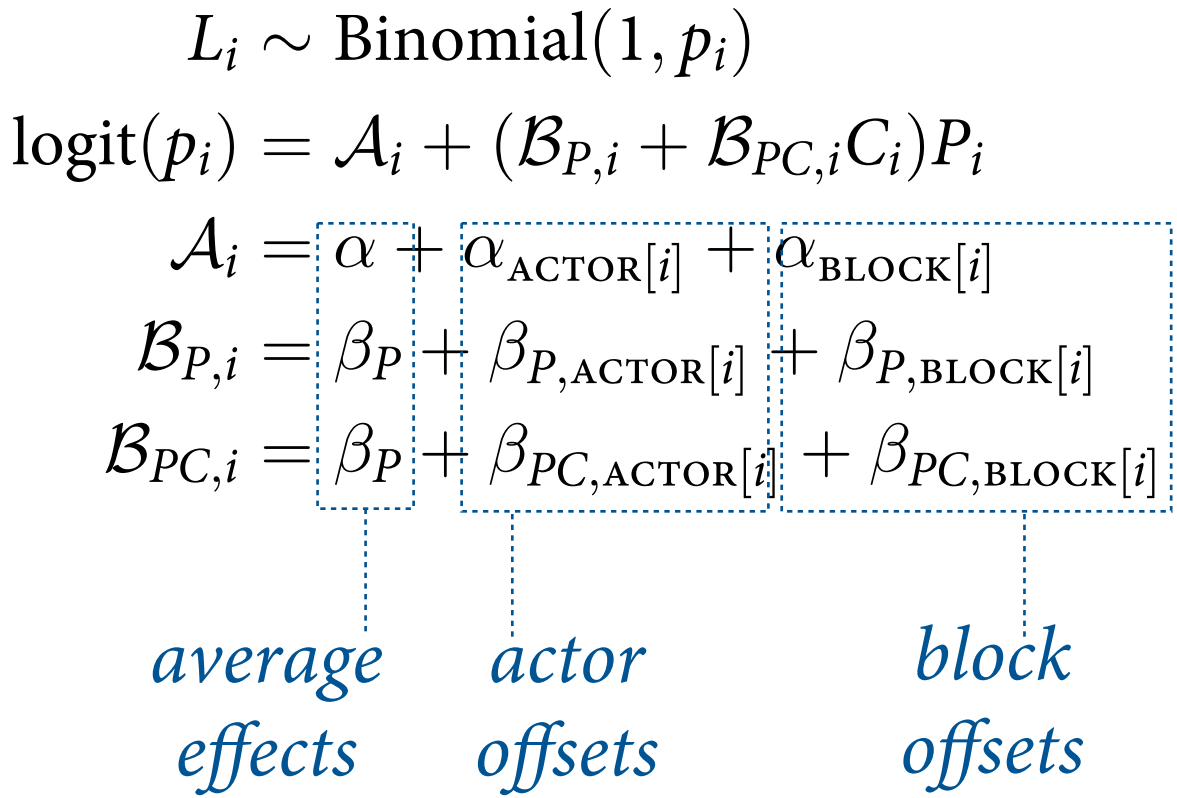
$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \mathcal{A}_i + (\mathcal{B}_{P,i} + \mathcal{B}_{PC,i} C_i) P_i \\
\mathcal{A}_i &= \alpha + \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]} \\
\mathcal{B}_{P,i} &= \beta_P + \beta_{P,\text{ACTOR}[i]} + \beta_{P,\text{BLOCK}[i]} \\
\mathcal{B}_{PC,i} &= \beta_P + \beta_{PC,\text{ACTOR}[i]} + \beta_{PC,\text{BLOCK}[i]}
\end{aligned}
$$

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \mathcal{A}_i + (\mathcal{B}_{P,i} + \mathcal{B}_{PC,i} C_i) P_i \\
\mathcal{A}_i &= \alpha + \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]} \\
\mathcal{B}_{P,i} &= \beta_P + \beta_{P,\text{ACTOR}[i]} + \beta_{P,\text{BLOCK}[i]} \\
\mathcal{B}_{PC,i} &= \beta_P + \beta_{PC,\text{ACTOR}[i]} + \beta_{PC,\text{BLOCK}[i]}
\end{aligned}
$$

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \mathcal{A}_i + (\mathcal{B}_{P,i} + \mathcal{B}_{PC,i} C_i) P_i \\
\mathcal{A}_i &= \alpha + \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]} \\
\mathcal{B}_{P,i} &= \beta_P + \beta_{P,\text{ACTOR}[i]} + \beta_{P,\text{BLOCK}[i]} \\
\mathcal{B}_{PC,i} &= \beta_P + \beta_{PC,\text{ACTOR}[i]} + \beta_{PC,\text{BLOCK}[i]}
\end{aligned}
$$

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \mathcal{A}_i + (\mathcal{B}_{P,i} + \mathcal{B}_{PC,i} C_i) P_i \\
\mathcal{A}_i &= \alpha + \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]} \\
\mathcal{B}_{P,i} &= \beta_P + \beta_{P,\text{ACTOR}[i]} + \beta_{P,\text{BLOCK}[i]} \\
\mathcal{B}_{PC,i} &= \beta_P + \beta_{PC,\text{ACTOR}[i]} + \beta_{PC,\text{BLOCK}[i]}
\end{aligned}
$$

average effects


$L_i \sim \text{Binomial}(1, p_i)$
$\text{logit}(p_i) = \mathcal{A}_i + (\mathcal{B}_{P,i} + \mathcal{B}_{PC,i} C_i) P_i$
$\mathcal{A}_i = \alpha + \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]}$
$\mathcal{B}_{P,i} = \beta_P + \beta_{P,\text{ACTOR}[i]} + \beta_{P,\text{BLOCK}[i]}$
$\mathcal{B}_{PC,i} = \beta_P + \beta_{PC,\text{ACTOR}[i]} + \beta_{PC,\text{BLOCK}[i]}$
average effects
actor offsets
block offsets

Cross-classified varying slopes

- Need two multivariate priors: actors and blocks
- Each 3-dimensional with own covariance matrix

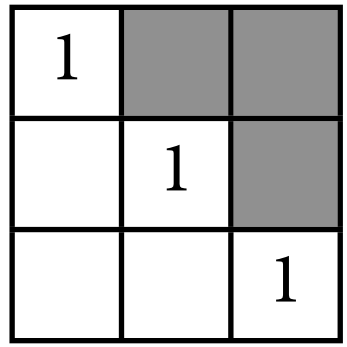
$$\begin{bmatrix} \alpha_{\text{ACTOR}} \\ \beta_{P,\text{ACTOR}} \\ \beta_{PC,\text{ACTOR}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{S}_{\text{ACTOR}} \right)$$

$$\begin{bmatrix} \alpha_{\text{BLOCK}} \\ \beta_{P,\text{BLOCK}} \\ \beta_{PC,\text{BLOCK}} \end{bmatrix} \sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \mathbf{S}_{\text{BLOCK}} \right)$$

```
m13.6 <- map2stan
    alist(
        #likelihood
        pulled_left ~ dbinom(1,p),

        #linear models
        logit(p) <- A + (BP + BPC*condition)*prosoc_left,
        A <- a + a_actor[actor] + a_block[block_id],
        BP <- bp + bp_actor[actor] + bp_block[block_id],
        BPC <- bpc + bpc_actor[actor] + bpc_block[block_id],

        # adaptive priors
        c(a,bp,bpc) ~ dnorm(0,1),
        c(a_actor,bp_actor,bpc_actor)[actor] ~
                                dmvnorm2(0,sigma_actor,Rho_actor),
        c(a_block,bp_block,bpc_block)[block_id] ~
                                dmvnorm2(0,sigma_block,Rho_block),

        # fixed priors
        sigma_actor ~ dcauchy(0,2),
        sigma_block ~ dcauchy(0,2),
        Rho_actor ~ dlkjcorr(4),
        Rho_block ~ dlkjcorr(4)
    ) , data=d )
```

Cross-classified varying slopes

- 54 parameters
  - 3 average effects
  - 3x7 = 21 varying effects on actor
  - 3x6 = 18 varying effects on block
  - 6 standard deviations
  - 6 free correlation parameters
- WAIC says $p_{\text{WAIC}} \approx 18$ (sigmas are small)




```
precis( m13.6NC , depth=2 , pars=c("sigma_actor","sigma_block") )
```

R code 13.25

```
                Mean StdDev lower 0.89 upper 0.89 n_eff Rhat
sigma_actor[1] 2.33   0.90       1.12       3.46  3296    1
sigma_actor[2] 0.46   0.36       0.00       0.88  5677    1
sigma_actor[3] 0.52   0.49       0.00       1.08  5868    1
sigma_block[1] 0.22   0.20       0.00       0.46  5809    1
sigma_block[2] 0.57   0.40       0.00       1.03  3931    1
sigma_block[3] 0.51   0.42       0.00       1.01  5834    1
```

Non-centered form

dmvnormNC usually samples more efficiently

R code 13.23

```
m13.6NC <- map2stan(
    alist(
        pulled_left ~ dbinom(1,p),
        logit(p) <- A + (BP + BPC*condition)*prosoc_left,
        A <- a + a_actor[actor] + a_block[block_id],
        BP <- bp + bp_actor[actor] + bp_block[block_id],
        BPC <- bpc + bpc_actor[actor] + bpc_block[block_id],
        # adaptive NON-CENTERED priors
        c(a_actor,bp_actor,bpc_actor)[actor] ~
                                dmvnormNC(sigma_actor,Rho_actor),
        c(a_block,bp_block,bpc_block)[block_id] ~
                                dmvnormNC(sigma_block,Rho_block),
        c(a,bp,bpc) ~ dnorm(0,1),
        sigma_actor ~ dcauchy(0,2),
        sigma_block ~ dcauchy(0,2),
        Rho_actor ~ dlkjcorr(4),
        Rho_block ~ dlkjcorr(4)
    ) , data=d , iter=5000 , warmup=1000 , chains=3 , cores=3 )
```

Non-centered form

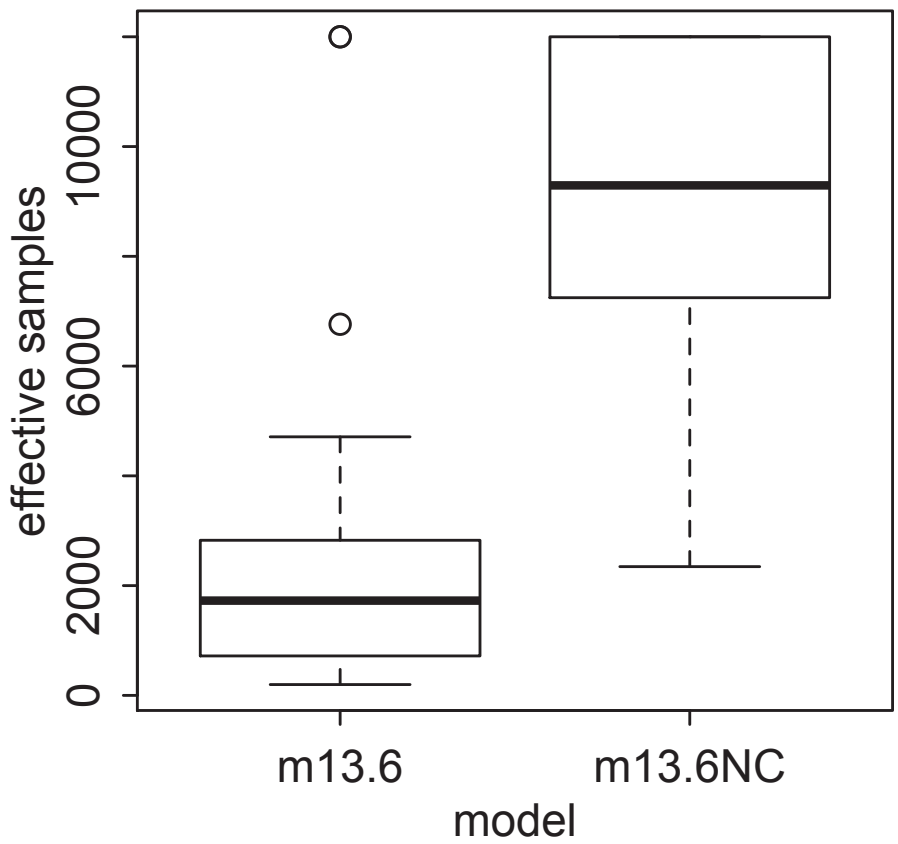


FIGURE 13.7. Distributions of effective samples, `n_eff`, for the ordinary and non-centered parameterizations of the cross-classified varying slopes model, `m13.6` and `m13.6NC`, respectively. Both models arrive at equivalent inferences, but the non-centered version samples much more efficiently.

Non-centered form

- Goal: Every dimension (parameter) in posterior shall be Normal(0,1)
- Once you start embedding parameters inside priors, this is hard — Normal(*a* , *sigma*) e.g.
- Solution: Factor things out of the prior

$$y \sim \text{Normal}(\mu, \sigma)$$

$$y = \mu + z\sigma$$
$$z \sim \text{Normal}(0, 1)$$

Non-centered form

- Simple case: Varying intercepts
- Factor the mean and sigma out of the prior
- Centered form:

```
mu <- a_actor[actor] + {stuff}
a_actor[actor] ~ normal( a , sigma )
```

- Non-centered form:

```
mu <- a + z_actor[actor]*sigma + {stuff}
z_actor[actor] ~ normal( 0 , 1 )
```

- See page 408. NB lower=0 constraint on sigma

Non-centered form

- What about varying slopes?
- Now need to factor correlation matrix out of the prior and smuggle into linear model
- Can be done: Cholesky!
- See page 409

André-Louis Cholesky (1875–1918)

Cholesky magic

```
N <- 1e4
sigma1 <- 2
sigma2 <- 0.5
rho <- 0.6
z1 <- rnorm( N )
z2 <- rnorm( N )
a1 <- z1 * sigma1
a2 <- ( rho*z1 + sqrt( 1-rho^2 )*z2 )*sigma2
```

```
> cor(z1,z2)
[1] -0.0005542644
> cor(a1,a2)
[1] 0.5999334
> sd(a1)
[1] 1.997036
> sd(a2)
[1] 0.4989456
```

http://elevanth.org/blog/2017/09/07/metamorphosis-multilevel-model/

Metamorphosis and the Multilevel Model

***Abstract**: The same multilevel model can be written different ways. Algebra don't mind. But the differences do matter to the computer.*

A terrible fact about applied statistical inference is that the same statistical model can be expressed in many different ways. This truth contains two aspects.

The first aspect is the cunning of algebra. Mathematical expressions can be arranged and rearranged into many equivalent forms. But these equivalent forms can sometimes look very different, and we can exploit this to aid our understanding.

The second aspect is the terror of implementation. To give life to a statistical model, we now use computers. The computer implementation typically makes one or another mathematically equivalent form of a model superior. So it can really pay to understand how to rearrange a model—how to **reparameterize** it.

I want to briefly explain by example of **multilevel models**. By "multilevel model," I mean a model with partially pooled parameters—random effects and the like. These models are not exotic. But they can be factored and expressed in many different,

Multilevel horoscopes

- Begin with "empty" model with no predictors, but with varying intercepts on clusters of interest
- Standardize all predictors
- Use regularizing priors
- Add in predictors and vary their slopes
- Can drop varying effects with tiny sigmas
- Consider two sorts of posterior prediction
  - Same units: What happened in these data?
  - New units: What might we expect for new units?
- Your knowledge of domain trumps all

ARIES (March 21–April 19)
You have more than one fresh start ahead of you—don't be afraid to reboot Vista often. Your lucky numbers for today are: 3.428, 1.417, 1.155, 1.096, and 1.043.

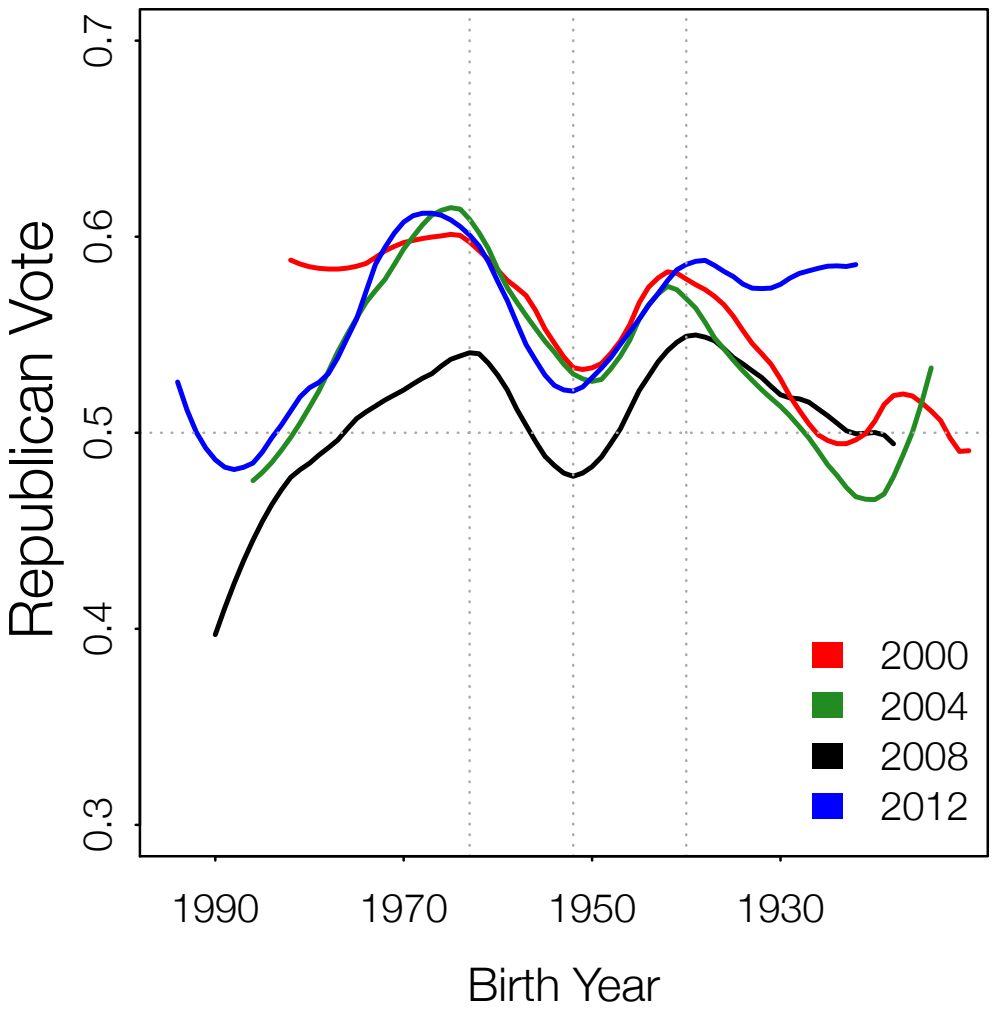
TAURUS (April 20–May 20)
There is harmony in the universal machinery that regulates the heavens. Get that filing in now!

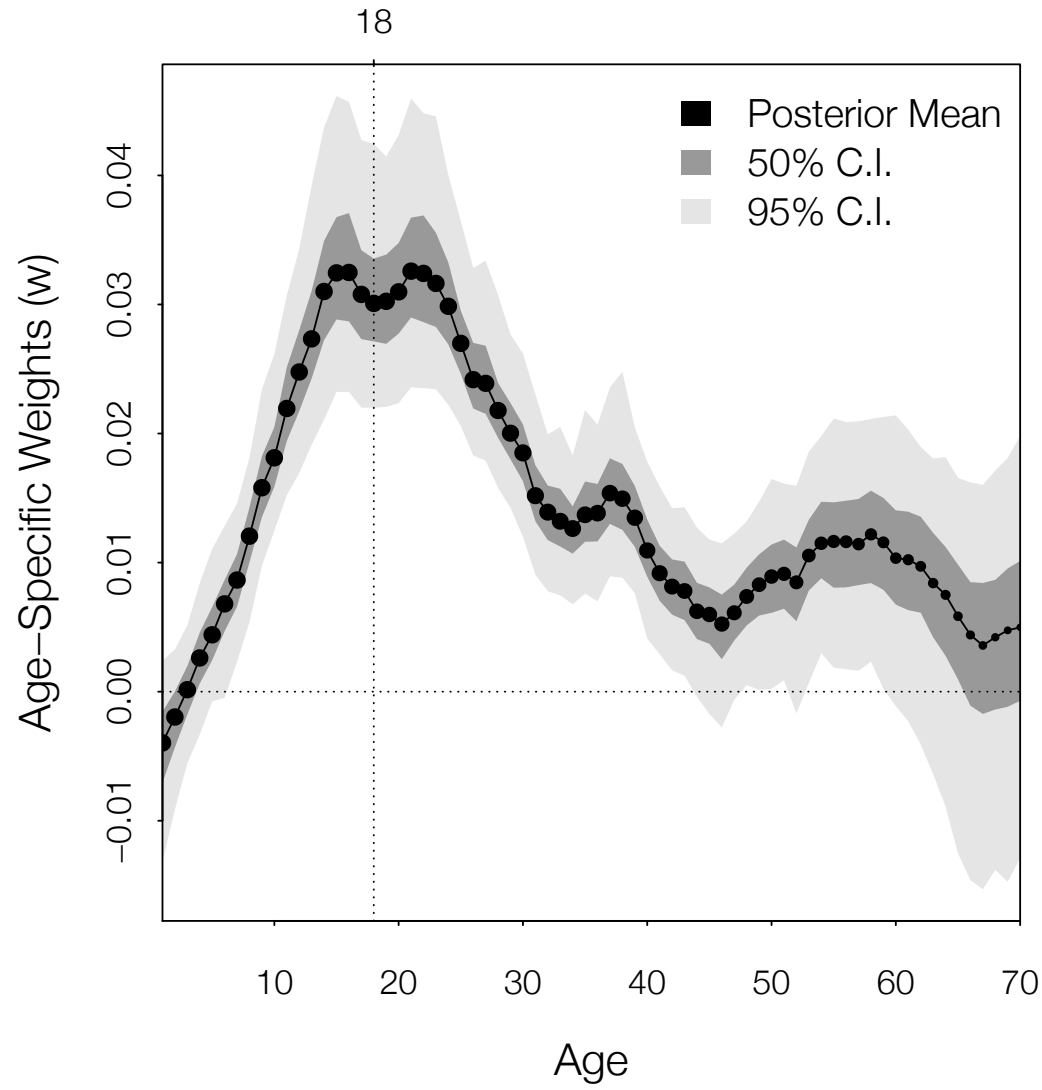
GEMINI (May 21–June 21)
You learn that your coworkers are more or less of one mind—that you need to get the team moving and on to new projects. Focus them on personal hygiene.

CANCER (June 22–July 22)
CFOs are reawakening their chakras. Channel this energy to strengthen reserves.

LEO (July 23–August 22)
Your negative energy is blocking your ability to utilize MS Office fully. Look to leverage pre-existing analyses and presentations. Don't forget to update those headers and footers!

VIRGO (August 23–September 22)
Triangles are aligning with the 5th moon of Neptune. Multiplicative methods beware! Cape Cod is more than a summer vacation destination!





The Great Society, Reagan's Revolution, and Generations of Presidential Voting

Yair Ghitza* Andrew Gelman†

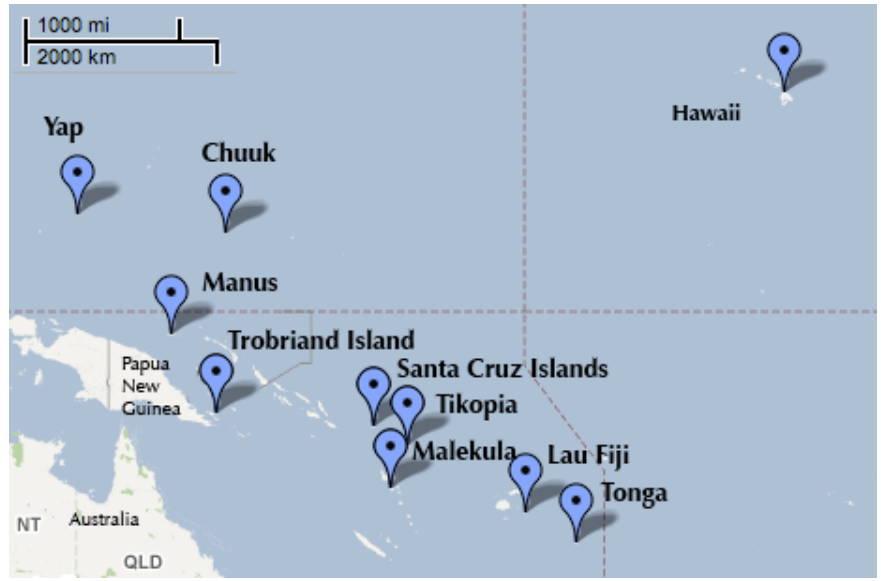
July 7, 2014

Continuous categories

- Traditional clusters discrete, unordered => every category equally different from all others (in prior)
- Continuous dimensions of difference:
  - Age, income, location, phylogenetic distance, social network distance, many others
  - No obvious cut points in continuum, but close values share common exposures/covariates/interactions
- Would like to exploit pooling in these cases as well
- Common approach: **Gaussian process regression**

GP e.g.: Spatial autocorrelation

- Relationship between tool complexity and population
- Close societies may share tools because of contact or similar geology/ecology
- Use space as proxy
- Spatial autocorrelation



	Ml	Ti	SC	Ya	Fi	Tr	Ch	Mn	To	Ha
Malekula	0.0	0.5	0.6	4.4	1.2	2.0	3.2	2.8	1.9	5.7
Tikopia	0.5	0.0	0.3	4.2	1.2	2.0	2.9	2.7	2.0	5.3
Santa Cruz	0.6	0.3	0.0	3.9	1.6	1.7	2.6	2.4	2.3	5.4
Yap	4.4	4.2	3.9	0.0	5.4	2.5	1.6	1.6	6.1	7.2
Lau Fiji	1.2	1.2	1.6	5.4	0.0	3.2	4.0	3.9	0.8	4.9
Trobriand	2.0	2.0	1.7	2.5	3.2	0.0	1.8	0.8	3.9	6.7
Chuuk	3.2	2.9	2.6	1.6	4.0	1.8	0.0	1.2	4.8	5.8
Manus	2.8	2.7	2.4	1.6	3.9	0.8	1.2	0.0	4.6	6.7
Tonga	1.9	2.0	2.3	6.1	0.8	3.9	4.8	4.6	0.0	5.0
Hawaii	5.7	5.3	5.4	7.2	4.9	6.7	5.8	6.7	5.0	0.0

distances in thousand km

Familiar likelihood

$$T_i \sim \text{Poisson}(\lambda_i)$$

$$\log \lambda_i = \alpha + \gamma_{\text{ISLAND}[i]} + \beta_P \log P_i$$

common mean

island offset

fixed log pop

Unfamiliar prior

- Gaussian process prior:
  - Multivariate Gaussian
  - Means all zero (usually)
  - Model the covariance matrix using pairwise distances

vector of offsets → $\gamma \sim \text{MVNormal}\big([0, \ldots, 0], \mathbf{K}\big)$ ← *covariance matrix*

$$\mathbf{K}_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + \delta_{ij}\sigma^2$$

Modeling covariance

covariance btw islands i & j

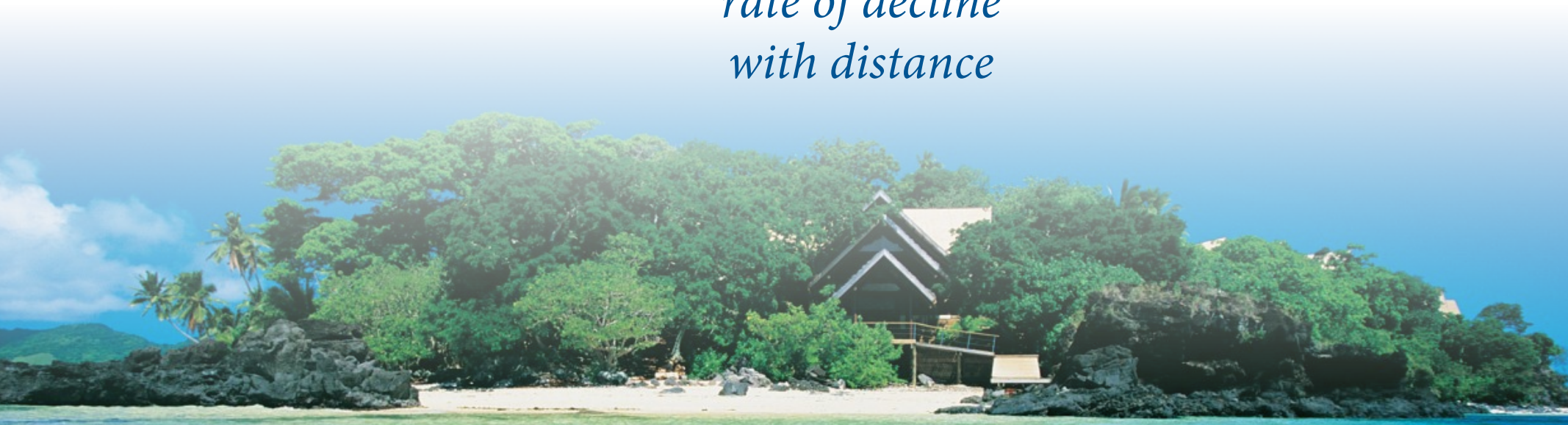
squared distance

$$\mathbf{K}_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + \delta_{ij}\sigma^2$$

max cov

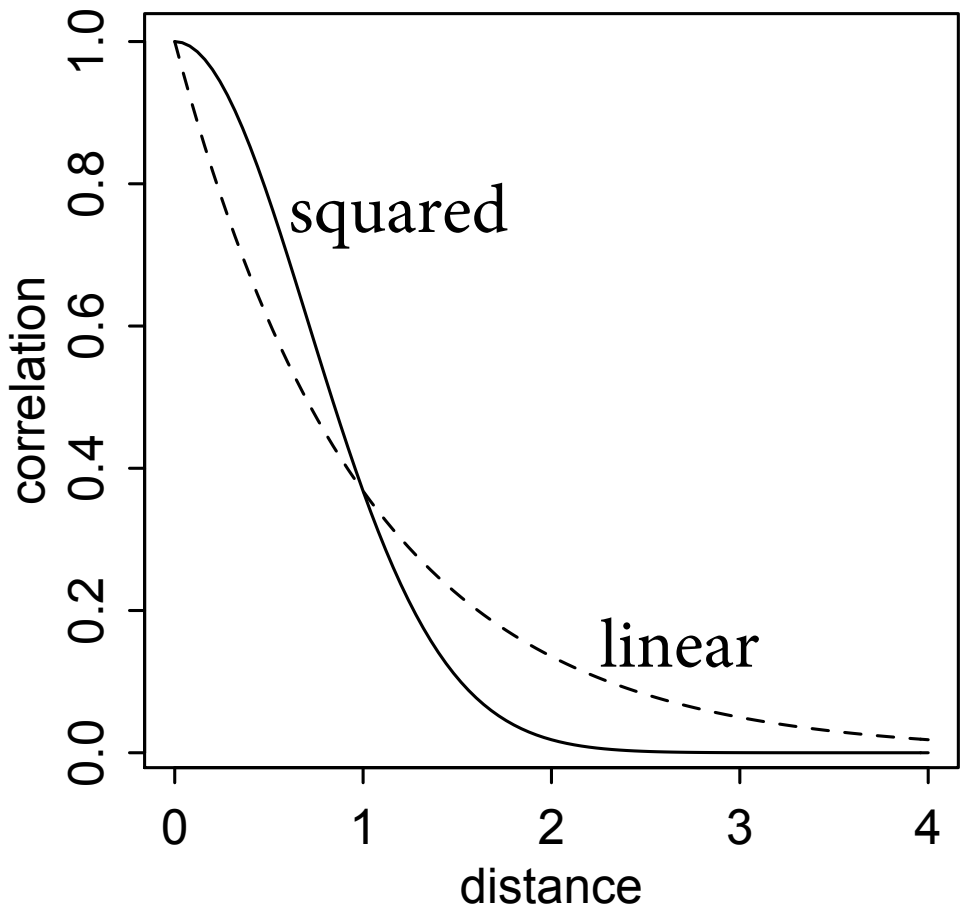
rate of decline with distance

"jigger"



Modeling covariance

$$\mathbf{K}_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + \delta_{ij}\sigma^2$$

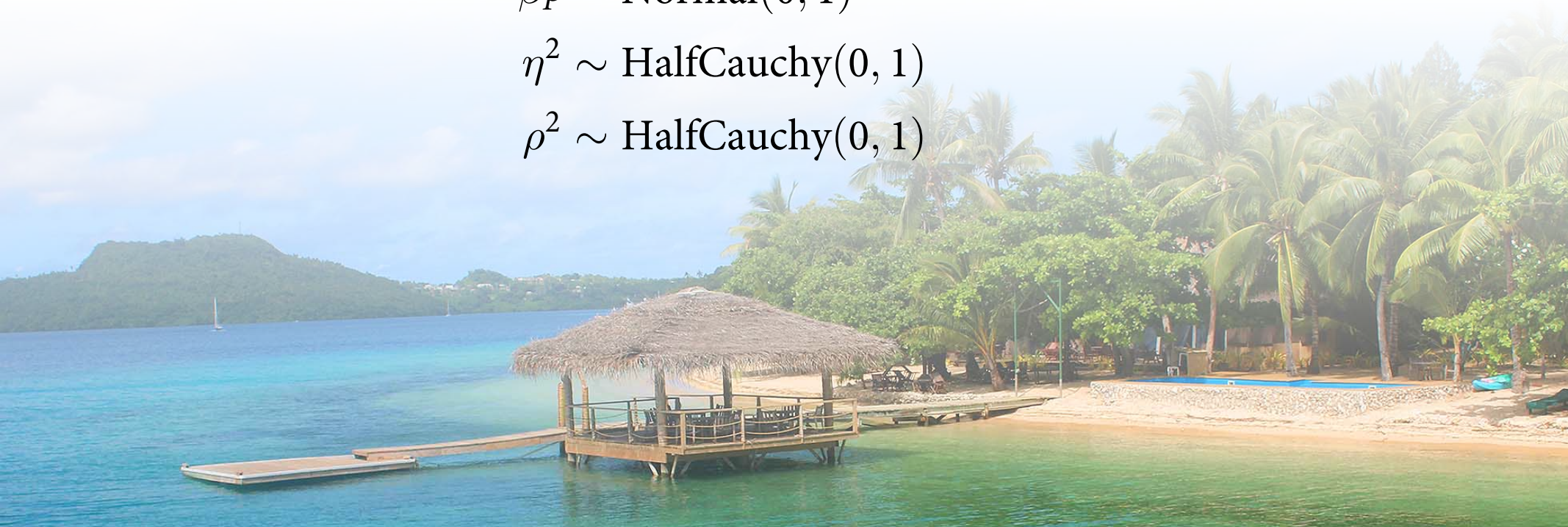


Linear: Cov declines fastest at near distances.

Squared: Cov declines fastest at intermediate distances.

Putting it all together

$$
\begin{aligned}
T_i &\sim \text{Poisson}(\lambda_i) \\
\log \lambda_i &= \alpha + \gamma_{\text{ISLAND}[i]} + \beta_P \log P_i \\
\gamma &\sim \text{MVNormal}\big((0, \ldots, 0), \mathbf{K}\big) \\
\mathbf{K}_{ij} &= \eta^2 \exp(-\rho^2 D_{ij}^2) + \delta_{ij}(0.01) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta_P &\sim \text{Normal}(0, 1) \\
\eta^2 &\sim \text{HalfCauchy}(0, 1) \\
\rho^2 &\sim \text{HalfCauchy}(0, 1)
\end{aligned}
$$



Fitting

```
m13.7 <- map2stan(
    alist(
        total_tools ~ dpois(lambda),
        log(lambda) <- a + g[society] + bp*logpop,
        g[society] ~ GPL2( Dmat , etasq , rhosq , 0.01 ),
        a ~ dnorm(0,10),
        bp ~ dnorm(0,1),
        etasq ~ dcauchy(0,1),
        rhosq ~ dcauchy(0,1)
    ),
    data=list(
        total_tools=d$total_tools,
        logpop=d$logpop,
        society=d$society,
        Dmat=islandsDistMatrix),
    warmup=2000 , iter=1e4 , chains=4 )
```

$$\gamma \sim \text{MVNormal}\big((0,\ldots,0),\mathbf{K}\big)$$

$$\mathbf{K}_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + \delta_{ij}(0.01)$$

Marginal posterior

- Coefficients on log scale, so a bit opaque

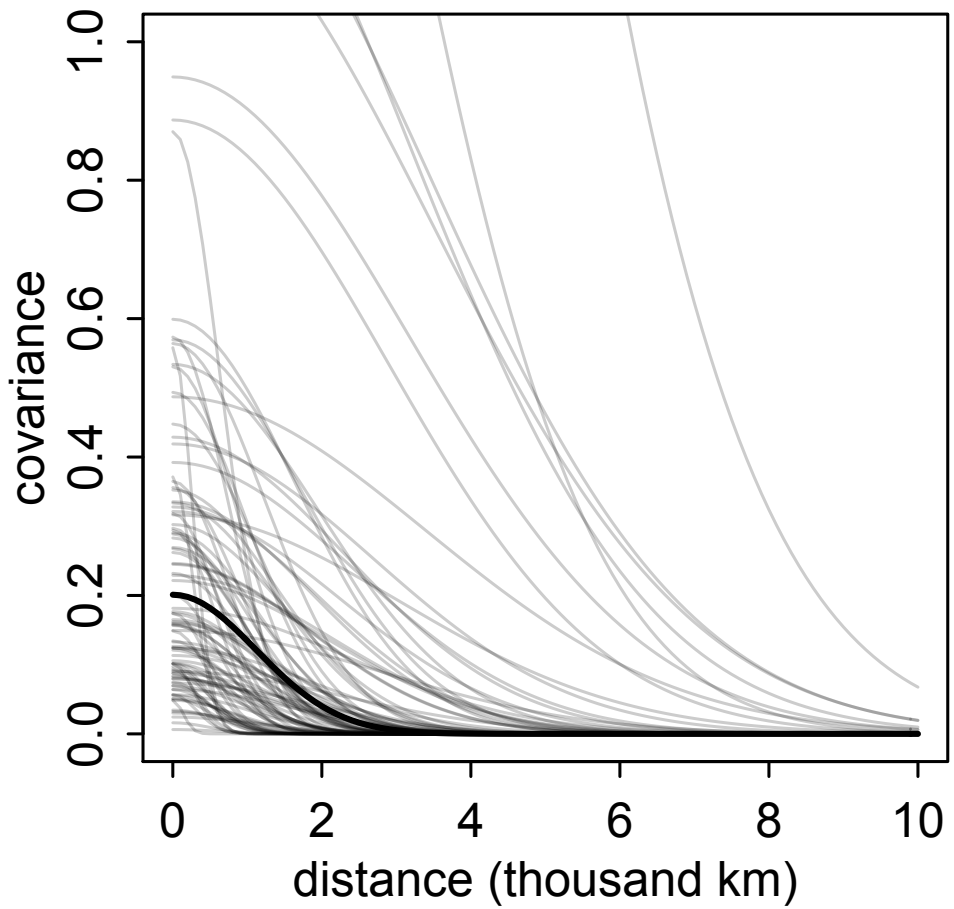
```
precis(m13.7,depth=2)
```

	Mean	StdDev	lower 0.95	upper 0.95	n_eff	Rhat
g[1]	-0.27	0.46	-1.30	0.62	2131	1
g[2]	-0.12	0.45	-1.09	0.77	2008	1
g[3]	-0.17	0.44	-1.08	0.70	1954	1
g[4]	0.30	0.39	-0.53	1.08	1991	1
g[5]	0.02	0.39	-0.79	0.80	2005	1
g[6]	-0.46	0.40	-1.30	0.27	2138	1
g[7]	0.09	0.38	-0.70	0.88	1978	1
g[8]	-0.27	0.39	-1.05	0.51	2113	1
g[9]	0.23	0.36	-0.53	0.94	2115	1
g[10]	-0.13	0.47	-1.09	0.83	4172	1
a	1.29	1.19	-1.06	3.79	3331	1
bp	0.25	0.12	0.02	0.49	5005	1
etasq	0.36	0.63	0.00	1.15	4343	1
rhosq	1.63	17.11	0.00	4.23	8111	1

Covariance function

$$\mathbf{K}_{ij} = \eta^2 \exp(-\rho^2 D_{ij}^2) + \delta_{ij}(0.01)$$

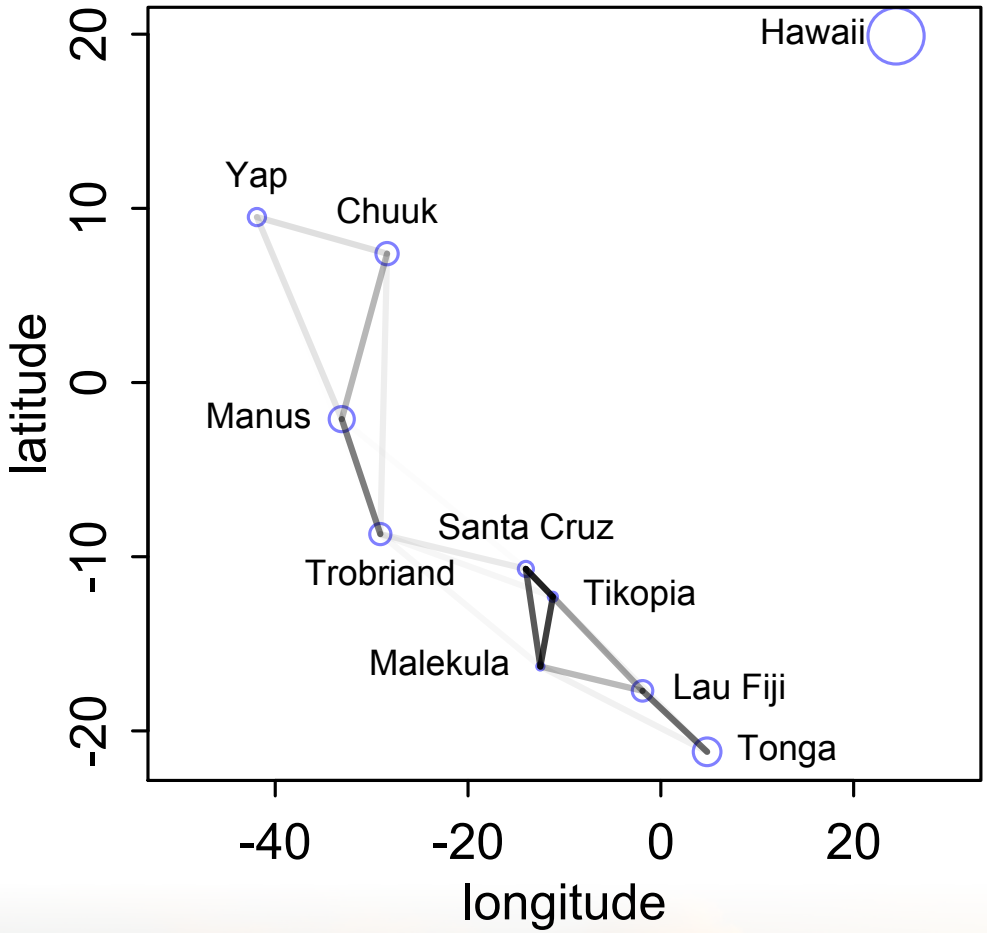
- Combination of *eta* and *rho* implies a covariance function **K**
- Draw samples from posterior and plot variation in these functions
- Yes, a posterior distribution of covariance functions

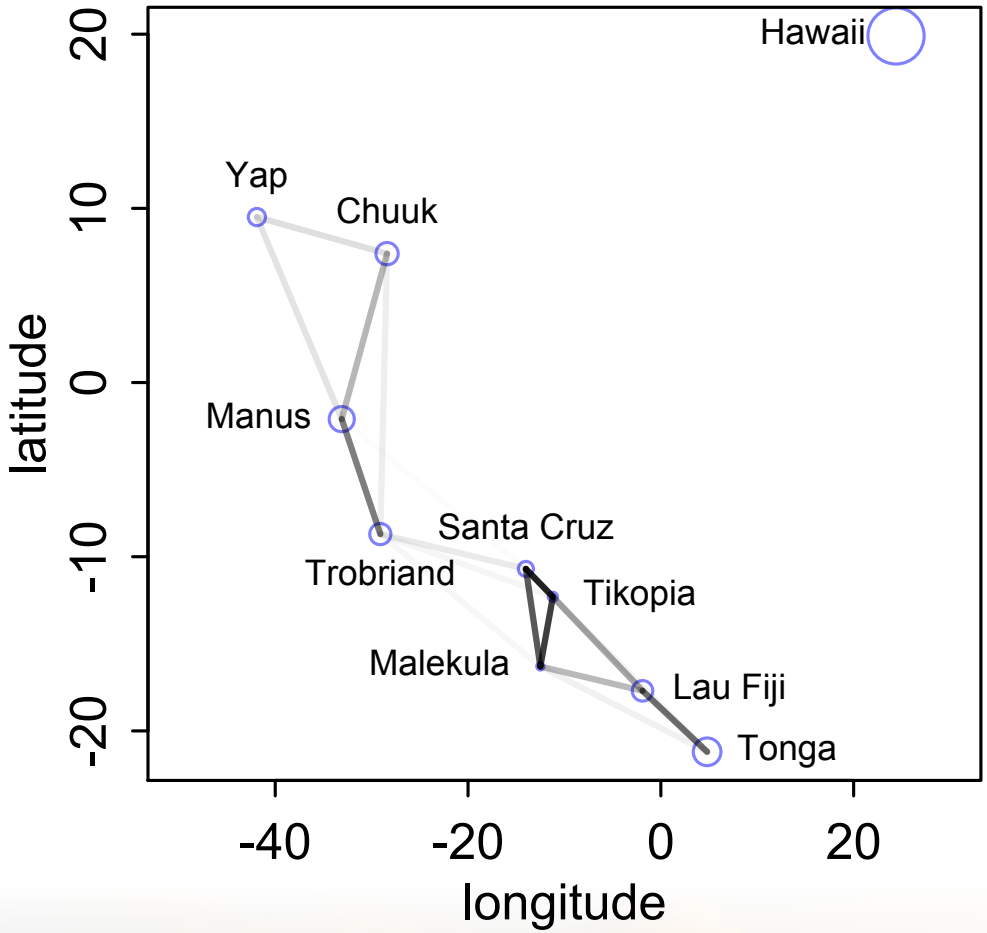


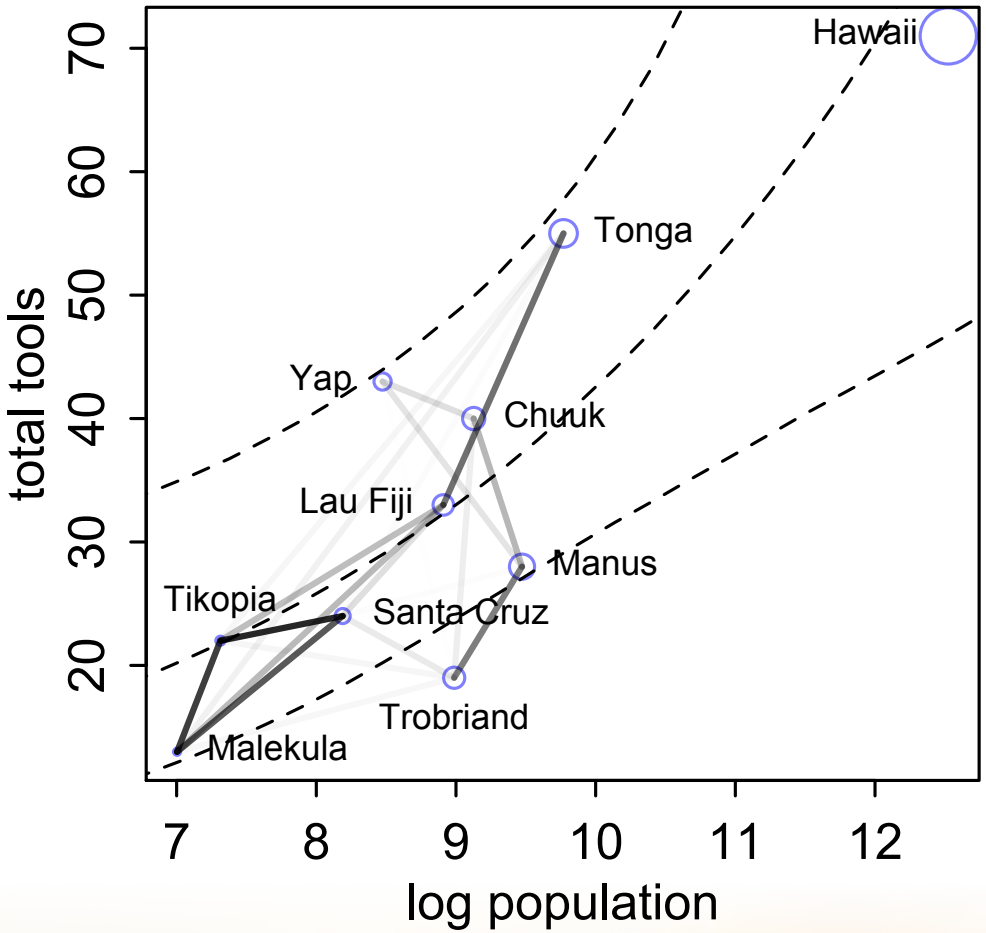
Implied correlations

- Covariance (and variance) on log scale, so hard to understand
- Compute correlations at posterior median:

	Ml	Ti	SC	Ya	Fi	Tr	Ch	Mn	To	Ha
Ml	1.00	0.87	0.82	0.00	0.52	0.19	0.02	0.04	0.24	0
Ti	0.87	1.00	0.92	0.00	0.52	0.19	0.04	0.06	0.21	0
SC	0.82	0.92	1.00	0.00	0.37	0.30	0.07	0.11	0.12	0
Ya	0.00	0.00	0.00	1.00	0.00	0.09	0.37	0.34	0.00	0
Fi	0.52	0.52	0.37	0.00	1.00	0.02	0.00	0.00	0.76	0
Tr	0.19	0.19	0.30	0.09	0.02	1.00	0.26	0.72	0.00	0
Ch	0.02	0.04	0.07	0.37	0.00	0.26	1.00	0.53	0.00	0
Mn	0.04	0.06	0.11	0.34	0.00	0.72	0.53	1.00	0.00	0
To	0.24	0.21	0.12	0.00	0.76	0.00	0.00	0.00	1.00	0
Ha	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1


Hawaii
Yap
Chuuk
Manus
Santa Cruz
Trobriand
Tikopia
Malekula
Lau Fiji
Tonga
latitude
longitude
20
10
0
-10
-20
-40
-20
0
20


Hawaii
Yap
Chuuk
Manus
Santa Cruz
Trobriand
Tikopia
Malekula
Lau Fiji
Tonga
latitude
longitude
20
10
0
-10
-20
-40
-20
0
20


Hawaii
Tonga
Yap
Chuuk
Lau Fiji
Manus
Tikopia
Santa Cruz
Trobriand
Malekula
total tools
log population
70
60
50
40
30
20
7
8
9
10
11
12

Gaussian process regression

- Many applications, many covariance functions
  - Periodic functions of time (seasonality)
  - Phylogenetic (*patristic*) distance => phylogenetic regression
  - Social networks
  - Non-parametric splines on any predictor
- Can use multiple dimensions in covariance, "automatic relevance determination"

$$\mathbf{K}_{ij} = \eta^2 \exp\left(-\left(\rho_D^2 D_{ij}^2 + \rho_P^2 (\log P_i - \log P_j)^2\right)\right) + \delta_{ij}\sigma^2$$

Next week, *dénouement*

- Homework: 13M3, 13M4, 13H1
- Next week:
  - Missing data
  - Measurement error
  - Enlightenment