Lecture 1.4: Theory-based approaches to reliability

- In Lec. 1.2, we used the ANOVA model to define **reliability** in terms of the ratio of MS_b, variance between groups, to MS_w, variance within groups.
- In the 1-way model with k obs per group, let σ_a^2 be the *true* variance between groups, and σ_e^2 be the error variance, or variance within groups. We defined:

$$\text{Reliability} \equiv ICC = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}$$

$$= \frac{F-1}{F+(k-1)}, \text{ where } F = \frac{MS_b}{MS_w}.$$

- In a 2-way ANOVA with *n* = 1 obs per cell, ***k* raters** (columns), each rating ***p* objects** (rows);
  let MS_p **= MS for Objects,** and MS_{res} **= MS residual.**
  Let the 'true' variance across Objects be σ_p^2**.**

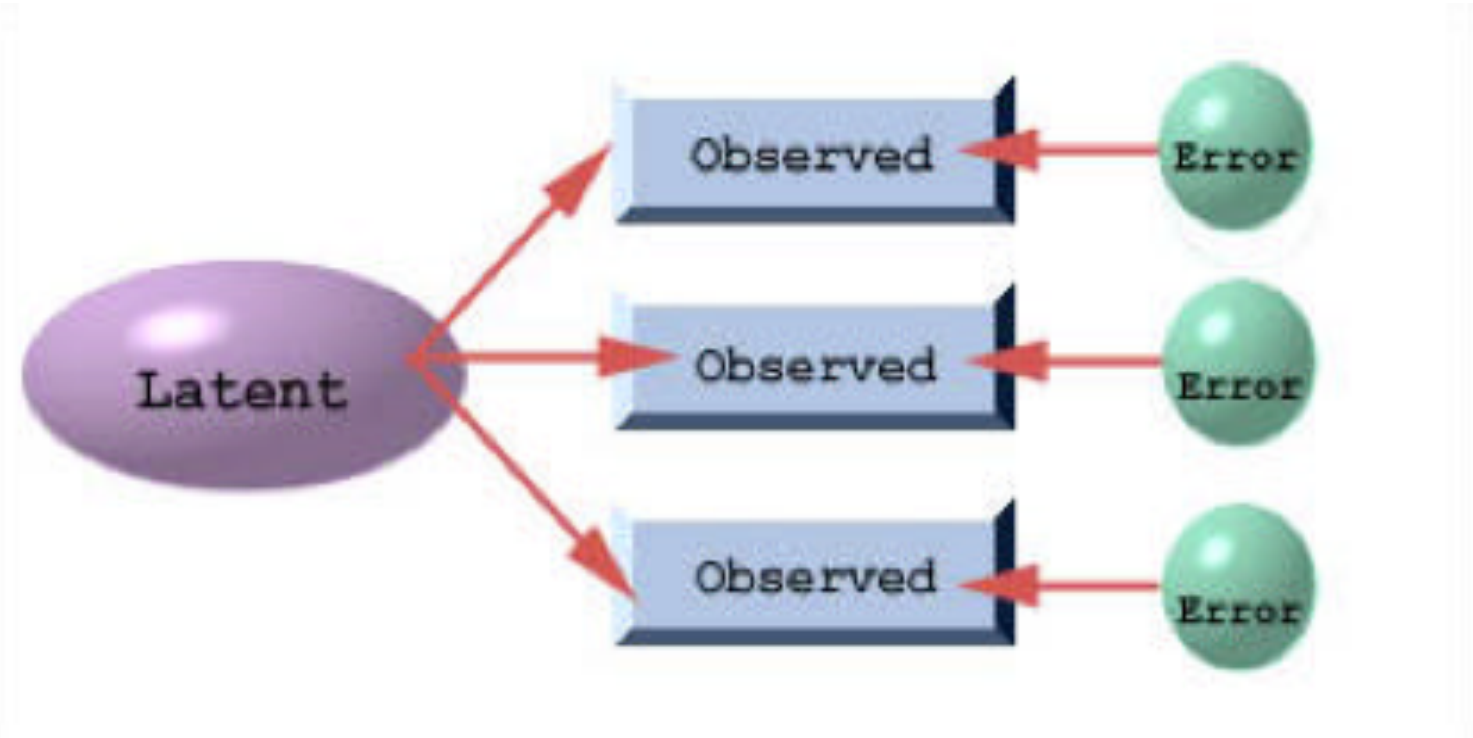
$$\text{Reliability} \equiv \text{Cronbach's } \alpha = \frac{\sigma_p^2}{\sigma_p^2 + MS_{res}}$$

$$= \frac{F-1}{F+(k-1)}, \text{ where } F = \frac{MS_p}{MS_{res}}.$$

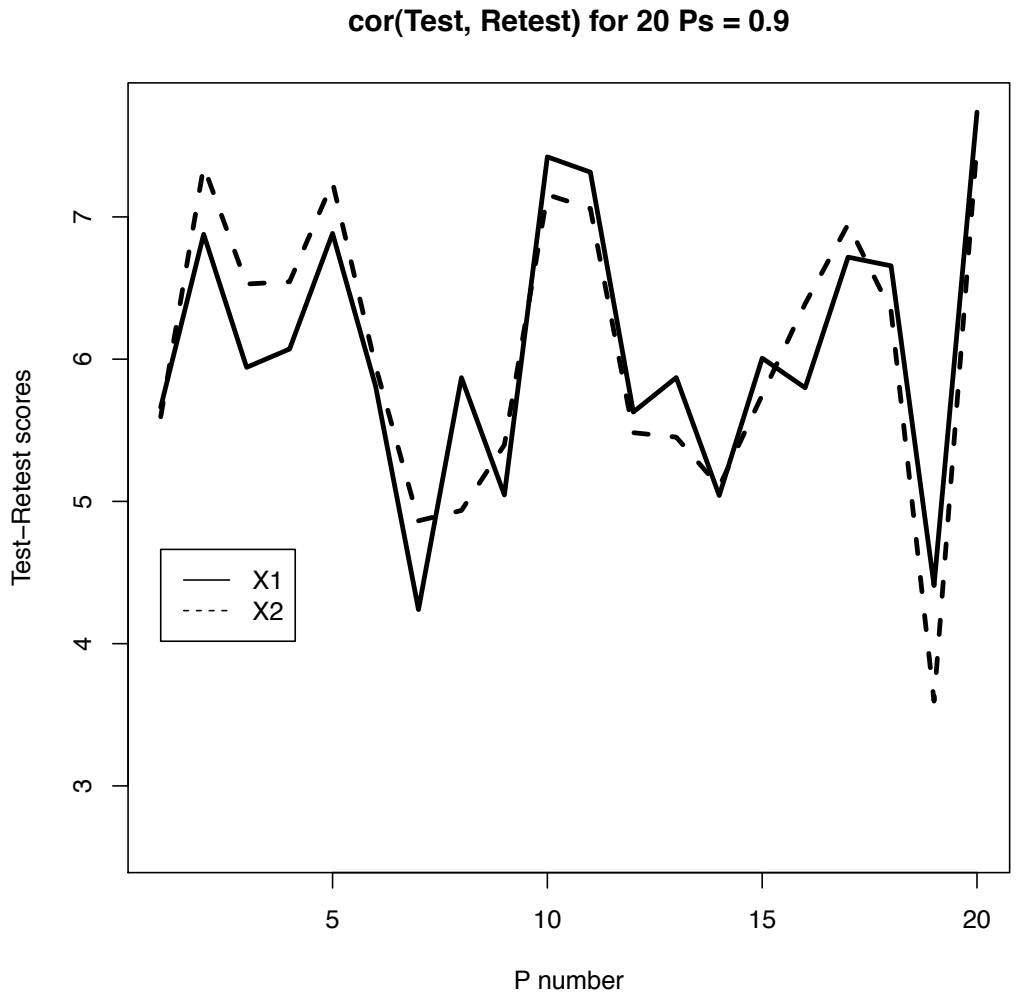
- We turn now to another, related definition of **reliability** in terms of certain **correlations**. The general approach is referred to as **Test Theory**.

Theory-based approaches to reliability

- Consider two **theory-based** approaches within a very general model, **Test Theory**. First, assume *observed*, *X*, is the sum of a *latent* (or *true*) score, *T*, plus an error, *e*: $X = T + e$, $X' = T + e'$. Second, assume that *X* is the sum of scores on *n* items.
- ***Test-Retest*** *reliability = corr(X, X')* across persons. I think of test-retest reliability as the **most concrete form** of the abstract concept. (See 2 displays below.)
- These test-theoretic models allow us to separate 'signal' (= 'structure') from 'noise' (= measurement error) when estimating substantive models.



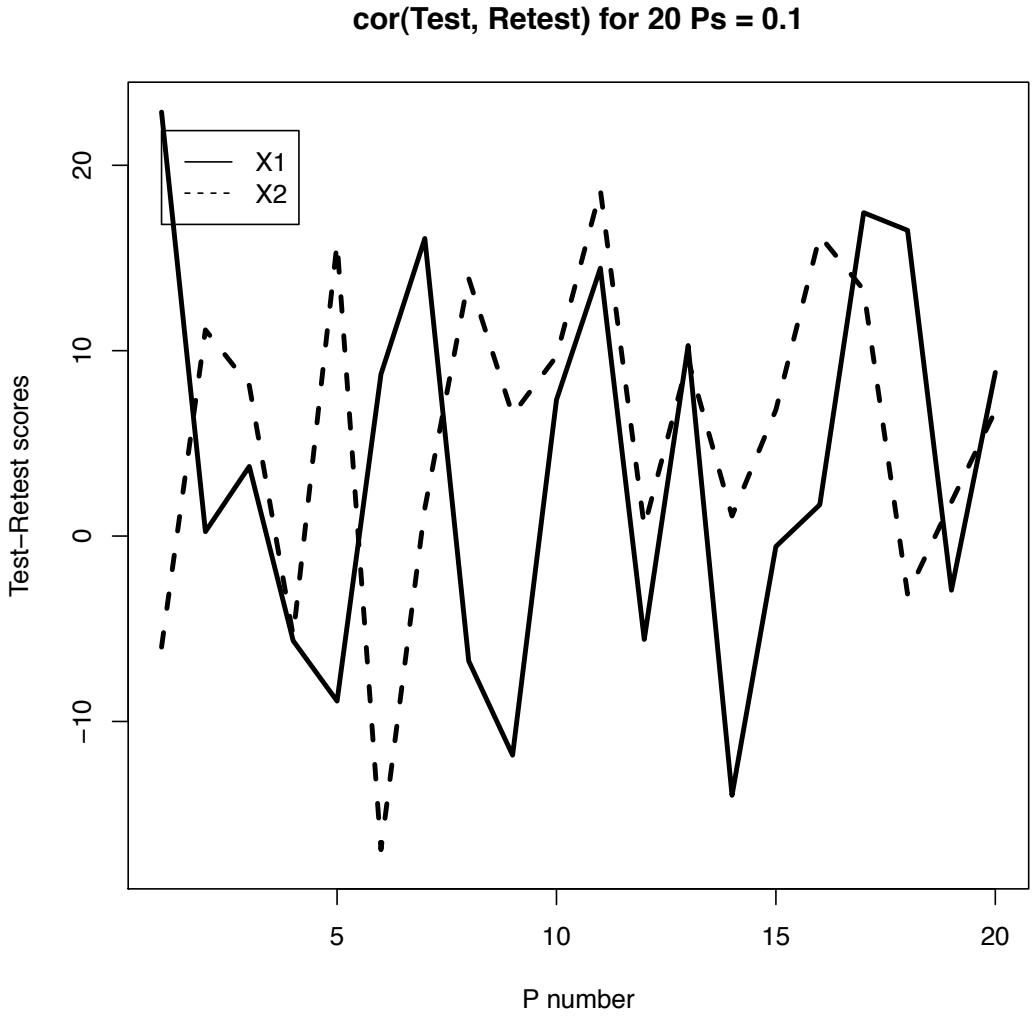
For a given object, *Latent* is **fixed**; successive measures of *Observed* (e.g., *X, X', X''*) on this object are different because of **independent** *Error* terms. Across objects, *Latent* varies, and this variation induces a **correlation**, $r_{XX'}$, between *X* and *X'*, known as the ***test-retest reliability***.



Think of 'test' score, X1 (or X) as Rater 1's score, and X2 (or X') as R2's score; P is the 'object' being rated.

$X = T + e$
is a special case of
$Y = a + bX + e$,
i.e., simple regression.

So r^2 *(=* $r(X,T)^2$*)* is the proportion of *X*-variance that is explained by *T*. This links 'correl' to 'reliability'.



In this 2nd case of low test-retest reliability, there is a significant **rater*object** interaction. This links **reliability** to ANOVA-type **interactions**.

Internal Consistency

- In the second theory-based approach, a **'test'** consists of *n* **'items'**, and test score, *X*, is the sum of the item scores, *Y*:

  $X = Y_1 + \ldots + Y_n$

  (instead of the sum of true score plus error:

  $X = T + e$)

- Define *reliability* as ***internal consistency*** of a 'test', which is based on the correlations among Y_i and Y_j across participants.

	Items			Total, X
	1	2	3 …	
P1	8.0	6.2	8.4	22.6
P2	7.4	9.8	8.5	25.7
P3	10.4	7.5	9.1	…
P4	4.6	3.7	5.6	
P5	9.5	9.7	9.9	
P6	9.2	8.6	8.6	
P7	5.3	4.7	5.5	
P8	6.6	7.9	6.9	
P9	4.2	6.9	3.6	
P10	5.4	6.4	6.9	
P11	6.9	6.9	10.2	
P12	10.6	10.6	11.6	
P13	11.5	11.6	11.3	
P14	6.3	8.2	6.8	
P15	7.5	9.7	9.0	26.2

The test is reliable if cor(1, 2), cor(1,3), etc. are high; i.e., if the items “hang together”.

Summary of Theory-based results

- Test-retest *reliability* = *var(T)/var(X)*, **proportion** of **observed score** variance explained by **true score** variance.
- Reliability of a test increases with the *length*, *n*, and with the *internal consistency*, *ρ*, of the test (**Spearman-Brown** formula), where *ρ* is the correl between items across persons.
- [To see the generality of this model, note we can replace 'test' by 'method', 'item' by 'rater', and 'person' by 'object'.]

Summary (cont'd)

- Reliability is reduced when we reduce the range (or variance) of *X.* The relevant formula allows us to answer questions like: **If the reliability of a test is .7 for 'all people'** (or all 'hi-school seniors'), **what is its reliability when used in a restricted population of 'patients'** (or all 'Stanford frosh')?
- The estimate, r_{XY}, of a model-based correlation, r_{AY}, between some 'interesting' *observed* variable, *Y*, and an 'interesting' *latent* construct, *A*, that is measured by *X*, **increases as the reliability of *X* increases**.

Comments on Causal Models

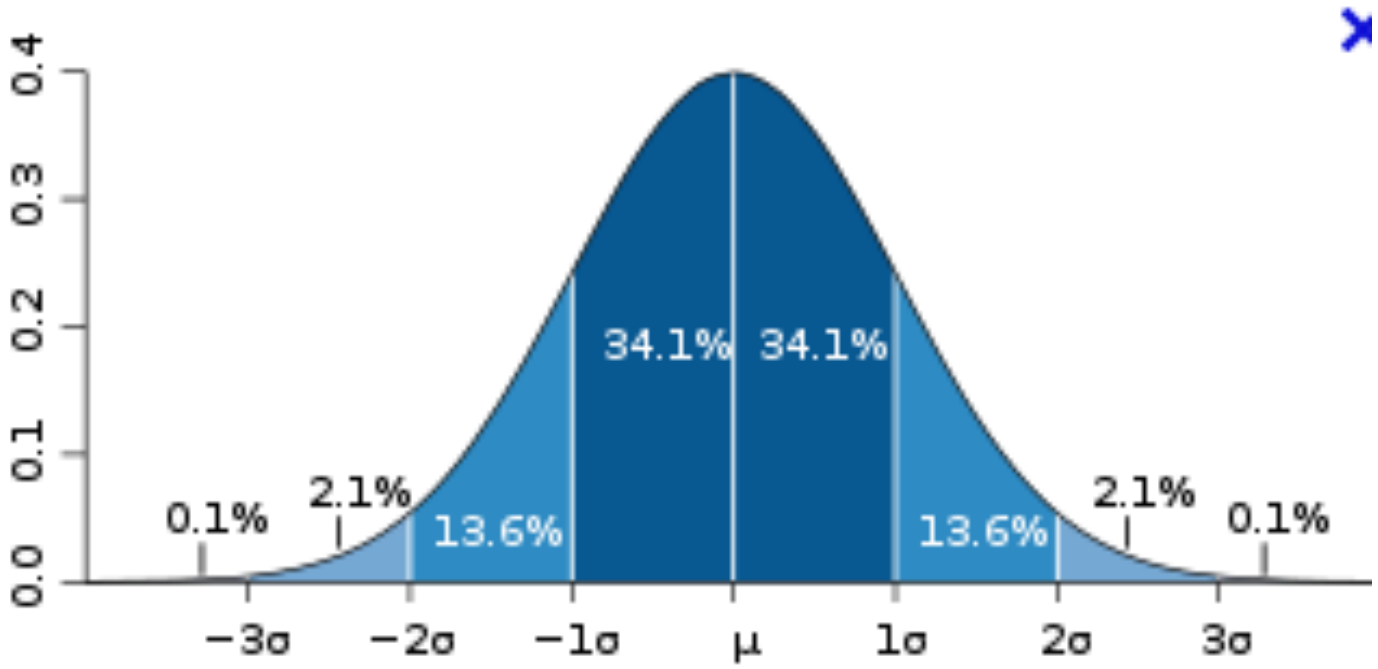
- But before the theories, let us consider the **purpose** of *causal models*, e.g., Test Theory and SEM, and *latent variables,* given that 'correlation doesn't imply causation'?
- We need causal models in theory and policy settings. This is partly because often we can't do experiments (which allow for strong inferences), and we must be content with correlational data.

The case for SEM by *Wiki*

- http://en.wikipedia.org/wiki/Structural_equation_modeling
- SEM: A statistical technique for testing and estimating *causal relations* using a combination of statistical data and qualitative causal assumptions
- Strengths include the inclusion of *latent variables*. These are variables which are not measured directly, but are estimated in the model from several measured variables, each of which 'taps into' the latent variables.
- This allows separate estimation of *unreliability of measurement* (noise) in the model, and the *structural relations* among latent variables.

- Caution: In *Applied Multivariate Data Analysis* (Chap 13), by BS Everitt and G Dunn (2001)
- Causal models "are best seen as convenient mathematical fictions which describe the investigator's belief about the causal structure of a set of variables. … Essentially, so-called causal models simply provide a parsimonious description of a set of correlations."

- “Latent variables are … hypothetical constructs invented by a scientist for the purpose of understanding some research area of interest, and for which there exists no operational method for direct measurement. … They serve to synthesize and summarize the properties of observed variables.
- Latent variables are as real as their predictive consequences are valid. … The justification for postulating latent variables is their **theoretical utility** rather than their reality.”



Model: "***X*** has mean, μ, and s.d., σ" is equivalent to "$X = \mu + \varepsilon$, where $\boldsymbol{\varepsilon}$ has mean, *0*, and s.d., σ".

We can call μ the *true score, T,* of *X*, and ε the *measurement error* in/of *X*. More typically,

$$X = T + \varepsilon,$$

where *T* varies across persons with some variance, σ_t^2.

- An easy case*:* *X* is a Rater's estimate of the **length** of an object. We have **objective** measures on this scale (ins, ft.), and can say: **High rater reliability = Low σ^2.** Can also define Rater **bias** as the difference between mean of *X* and objective length.
  - the meaning of 'low variance' depends on context.
- A hard case: *X* is a person's score on a personality test. Is the test reliable? We have no **objective** measures for this scale, and 'reliability' is more problematic.

- *X* is modeled as the sum of the person's true score, μ (typically written at t_i for *i*'th person), plus an **independent** error of measurement, ε: $X_i = t_i + \varepsilon_i$.
- High reliability = Low σ_e^2, **relative** to the variance of the true scores across persons.
- $Var(t_i) = \sigma_t^2$; $var(\varepsilon_i) = \sigma_e^2$;

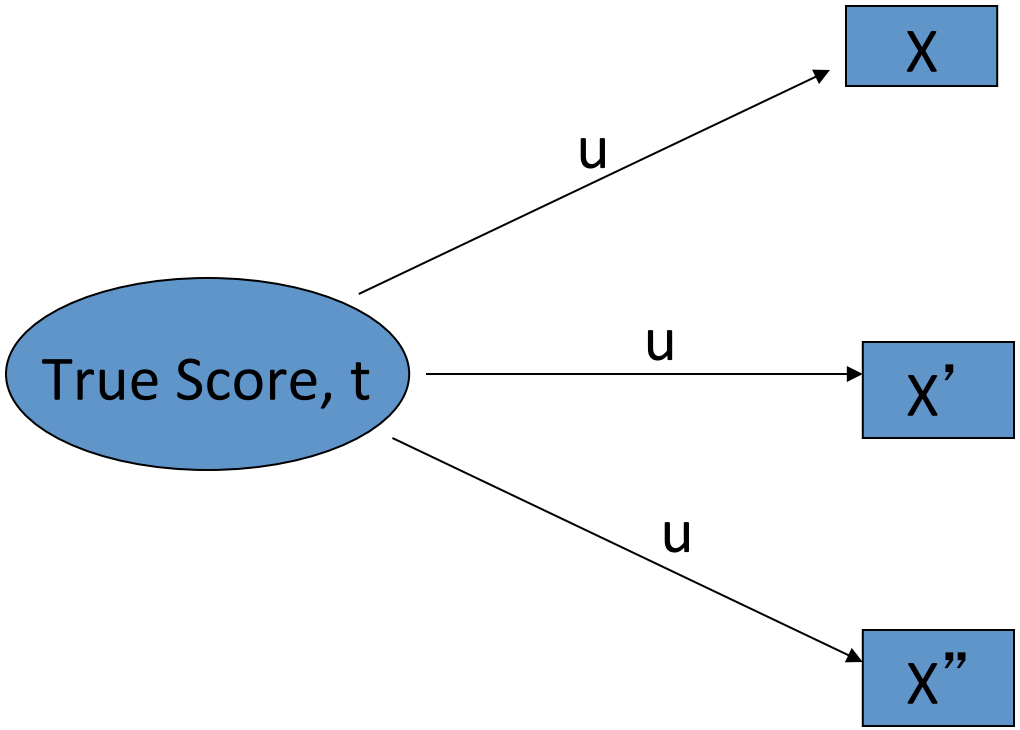
- Define *Reliability* as:

$$r = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}.$$

$$r = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}.$$

- We can't observe the true score or the error; so we cannot directly calculate the two variances. We need more data.
- Let *X* and *X'* be **two measures from the same person** on the test. $r_{XX'}$ is the correl across persons, and can be calculated. Relating $r_{XX'}$ to 'reliability' is the key insight of Test Theory.

Corr(X, X') = $u^2 \equiv r_{XX'}$ = Test-Retest reliability.
$u \equiv r_{Xt}$ ‘large’ implies ‘high’ reliability.



Theorem: If $t \rightarrow X$ *and* $t \rightarrow X'$, then $r_{XX'} = r_{tX} * r_{tX'} = u^2$

- **Proof**: Assume that X, X', t and e are **standardized** variables.
- The regression of X on t, is $X = ut + (1-u^2)^{1/2}e$.
- The regression of X' on t, is $X' = ut + (1-u^2)^{1/2}e'$.
- Multiplying these 2 eqns, we get:

  $XX' = u^2t^2 + ut*(1-u^2)^{1/2}e + \ldots$
- On taking expected values,

  $E(XX') = u^2E(t^2) + E[ut*(1-u^2)^{1/2}e + \ldots] = u^2,$

  because $E(t^2) = 1$, and the variables inside the *[]* brackets all have expected value = 0.

Completion of Proof

- Because X and X' are standardized,

  $r_{XX'} = E(XX') = u^2$, as stated earlier.

- This result will be used many times, e.g., in deriving correlations among observable variables in **Factor Analysis**.
- Note that $u = corr(t, X)$. Thus, in the 'regression',

  $X = bt + e'$,

- u^2 (i.e., "R^2") is the proportion of variance in X that is explained by t. Thus, **reliability, $r_{XX'}$, is the ratio of true score variance to total observed variance in *X***.
- Another proof follows.

- $X = t + \varepsilon$; $X' = t + \varepsilon'$.
- True score is the same, but error of meas differs across testing occasions. To simplify, assume
- $E(t) = 0$. This implies $E(X) = 0$, var(t) = E(t^2), and var(X) = E(X^2).
- var(X) = var(X') = var(t) + var(ε);
- $XX' = t^2 + t\varepsilon + t\varepsilon' + \varepsilon\varepsilon'$. So
- Cov(X, X') = E(XX') = E(t^2) = var(t) = σ_t^2

$$r_{XX'} = \frac{\mathrm{cov}(X, X')}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(X')}} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

$$r_{XX'} = \frac{\mathrm{cov}(X, X')}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(X')}} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}$$

- This shows that, even if we can't observe true scores and errors, we can estimate the reliability if we have test and re-test data.
- The various definitions of 'reliability' can all be related to this fundamental definition.
- Next we turn to the issue of **attenuation of reliability** (*cf*. attenuation of a correlation).

Two examples of Attenuation

- Restricting the 'range', e.g., 'variance', of *X* **reduces** ('attenuates') the reliability, $r_{XX'}$.

- Suppose *X* measures a latent trait, *A*, that varies across persons. We have a theory that predicts a positive correl between *A* and some other variable, *Y*, i.e., $r_{AY} > 0$. The observable, r_{XY}, is **less** than the unobservable, r_{AY}. This is attenuation.

- **Restricting the range of X reduces $r_{XX'}$** . To model this effect, assume σ_e^2, is the **same** whether the range (e.g., variance) of *X* is small or large.
- Let σ_x^2 denote the variance of *X*, which is equal to the variance of *X'*. From the formula above,

$$r_{XX'} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2} = 1 - \frac{\sigma_e^2}{\sigma_x^2};\ \text{implying}$$

$$\sigma_e^2 = \sigma_x^2(1 - r_{XX'}).$$

- Hence, if σ_x^2 is reduced, $(1 - r_{XX'})$ must increase, and, therefore, $r_{XX'}$ must decrease.

- **Example**.
- In one sample (e.g. admitted students), it was found that $r_{XX'}$ = 0.8, and the variance of *X* was 10.5.
- In a second sample (e.g., all who took the admissions test), the variance of *X* is known to be 18.5.
- What is the estimated reliability of the test for the second sample, assuming that the error of measurement is the same in the two samples?

Restricting the range of *X* reduces $r_{XX'}$.

- **Ans**. Let r_1 and r_2 be the reliability for the 1st and 2nd samples. Then

$$\sigma_e^2 = \sigma_{x1}^2(1-r_1) = \sigma_{x2}^2(1-r_2); \quad hence$$

$$10.5(1-0.8) = 18.5(1-r_2), \; and \; 1-r_2 = \frac{2.1}{18.5} = 0.114; \quad so \quad r_2 = 0.886.$$

- Restricting the range from a variance of 18.5 to a variance of 10.5 reduces the reliability from 0.886 to 0.800.

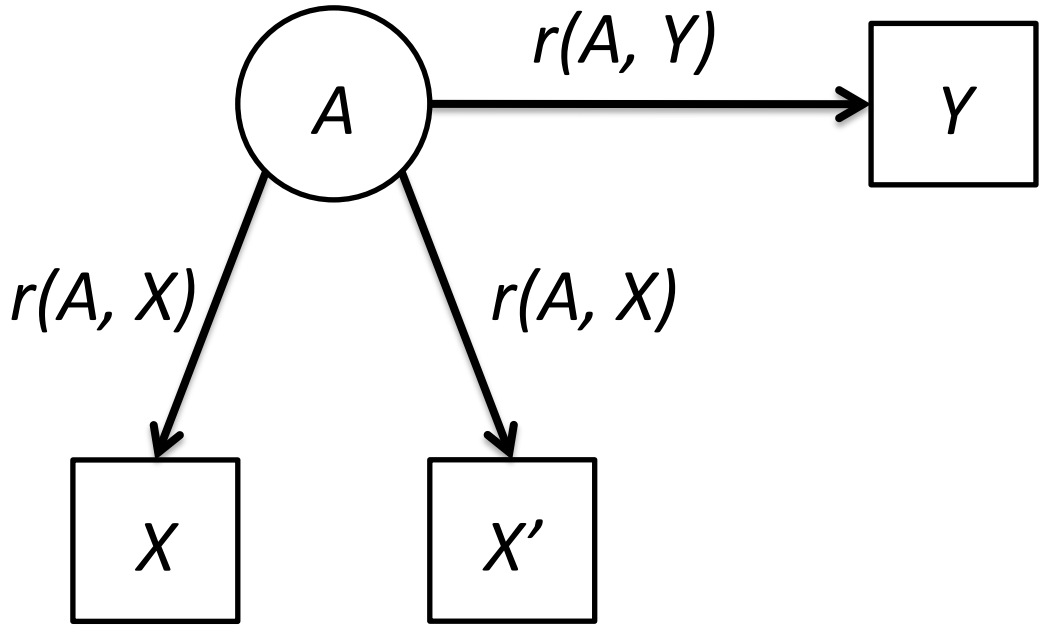
Reliability & Validity: Impact of Attenuation

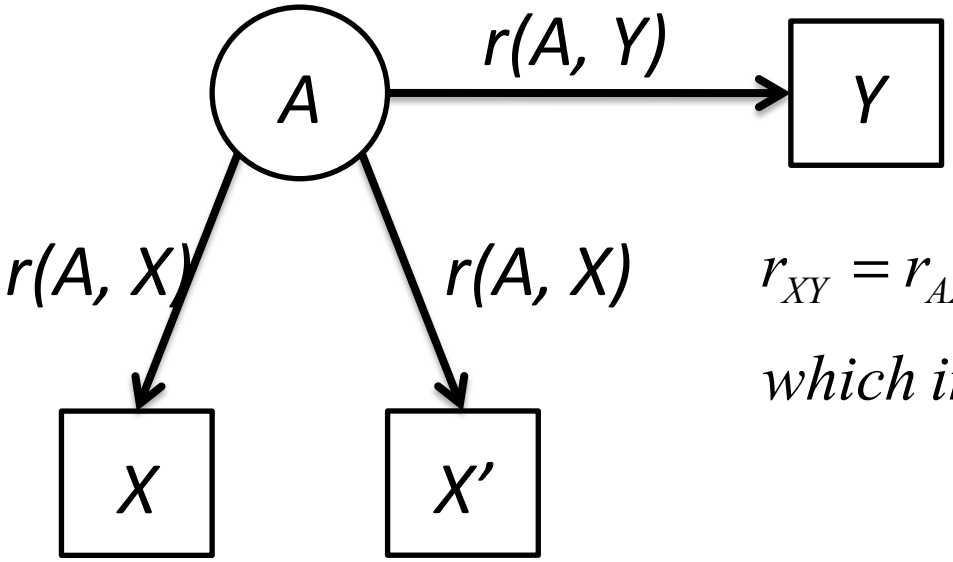
- A test, *X*, measures an underlying construct, *A* (e.g., 'anxiety'). To establish the external validity of *X*, find an important real-world manifestation, *Y*, of A. 'Validity' is indexed by r_{XY} = cor(*X, Y*); is it at least moderate?
- r_{XY} decreases as the reliability of *X*, $r_{XX'}$, decreases. This too is 'attenuation.'

Attenuation and Disattenuation of a correlation

- We want corr(*Y, A)*, but we cannot measure *A* perfectly. E.g., Y = ‘ideal affect’, A = ‘cultural orientation’.
- Our best measure, *X*, of *A* has a reliability, $r_{XX'}$. E.g., X = ‘interdependence’ or ‘influence’. We can observe *X*, not *A*, so we calculate r_{XY}, and we wish to use r_{XY} to estimate r_{AY}, the quantity in which we are really interested.
- r_{XY} is called an *attenuated* correlation, and replacing it by r_{AY} is called *disattenuating* the observed correlation.

- Consider the following causal diagram in which *A* is a latent variable, and *X, X'* and *Y* are observable, and the path coefficients are as shown.





$$r_{XY} = r_{AX} r_{AY}; \quad and \quad r_{XX'} = r_{AX} r_{AX} = r_{AX}^2,$$

$$which\ implies \quad r_{AX} = \sqrt{r_{XX'}}$$

$$r_{AY} = \frac{r_{XY}}{r_{AX}} = \frac{r_{XY}}{\sqrt{r_{XX'}}}$$

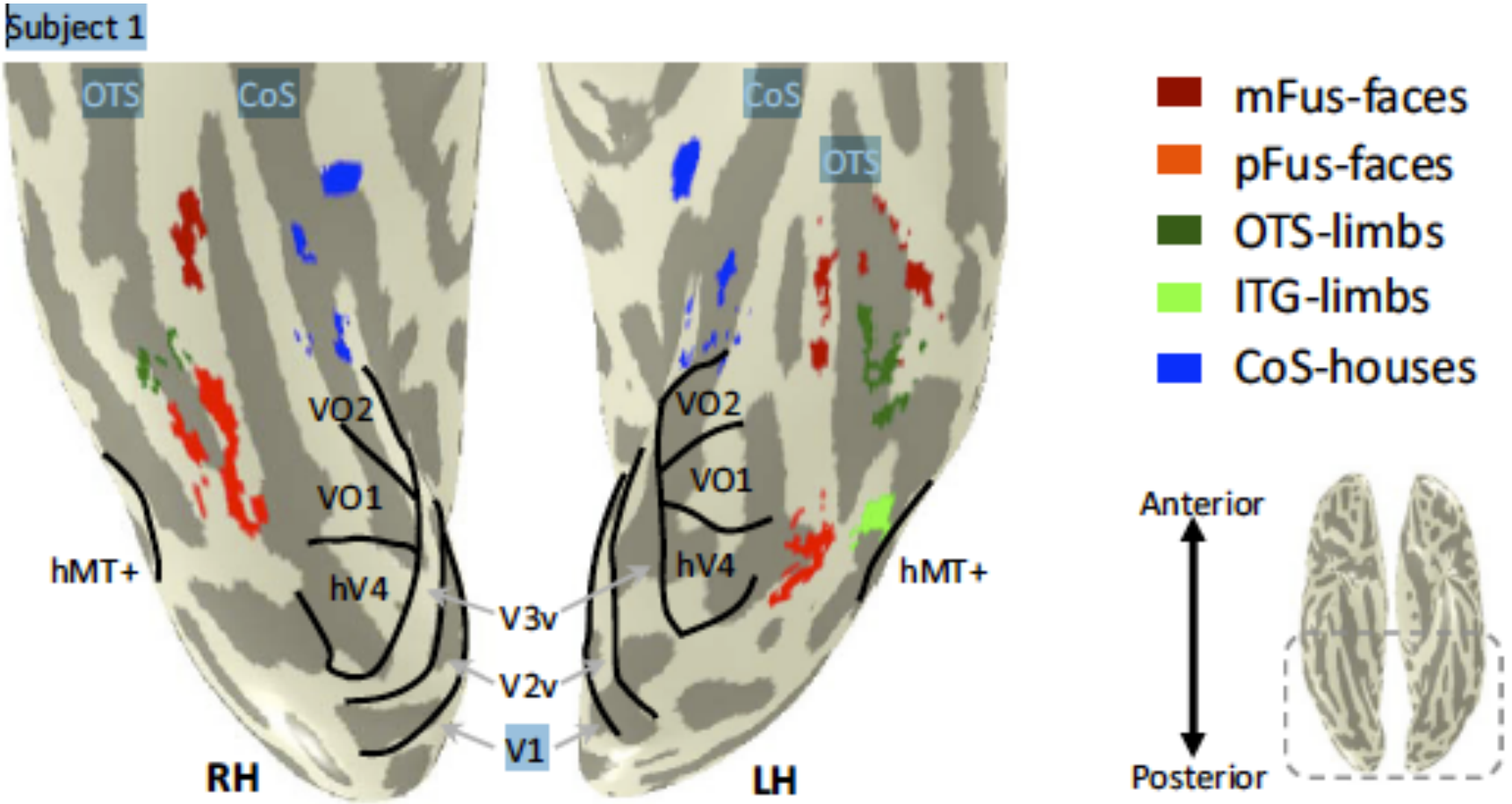
- This is the disattenuated correlation, and, because $r_{XX'}$ is less than 1, $r_{AY} > r_{XY}$. That is, the disattenuated correlation is always **larger** than the observed correlation.

Attenuation **in the measurement of brain** ***adaptation***

- *fMRI-Adaptation and Category Selectivity in Human Ventral Temporal Cortex*, by **Kevin S. Weiner**, Rory Sayres, Joakim Vinberg, and Kalanit Grill-Spector. In *J Neurophysiol* 103: 3349–3365, 2010.
- "When stimuli are repeated, cortical responses in high-level visual cortex generally decrease. When the responses are measured with fMRI, this **reduction** in activation is labeled *fMRI-Adaptation (fMRI-A)."*

- “This reduction, *fMRI-A*, has been observed in many spatial and temporal scales, … and is thought to be a marker of experience-dependent changes in high-level visual cortex.”
- “Results show that repetition produces **proportional** *fMRI-A*. Thus, in the regression analysis ***of repeated response, Y***, on ***non-repeated response, X***, a **slope equal to one** indicates **no** *fMRI-A*, a **slope less than one** indicates *fMRI-A*, and a slope more than one indicates response enhancement. ***However, noise (i.e., measurement error) can produce* a bias *in the slope estimation*, making it lower; i.e., attenuating it** (Frost and Thompson 2000).”
- Thus, *fMRI-A* is a substantively interesting *attenuation* (pun intended!). But to estimate it accurately (i.e., without bias), we need to ‘control for’ measurement-related *attenuation*!

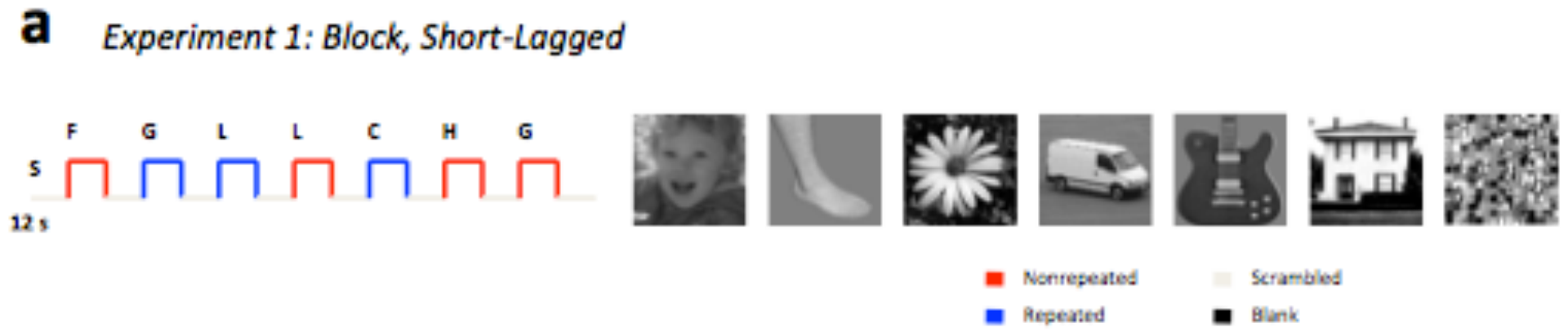
Category-selective regions of interest (ROIs)



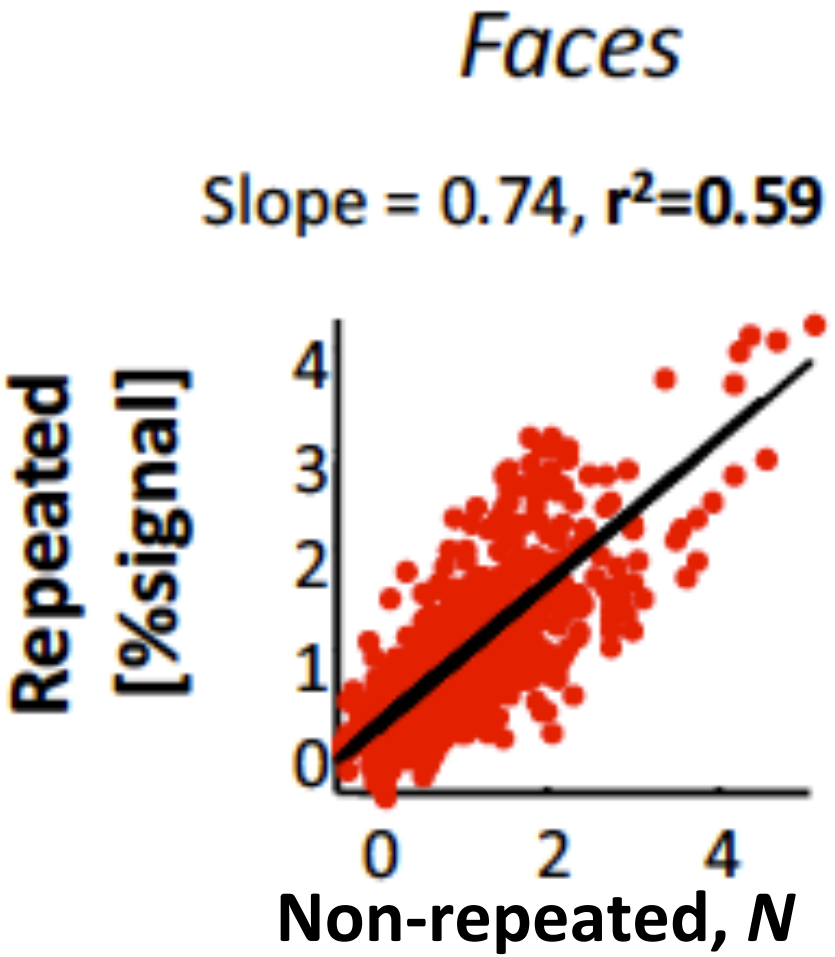
Stimuli

A block of different Faces (non-repeated), followed by a block of Guitars with repetitions (repeated), followed by a block of Limbs with repetitions (repeated), …

At each voxel, obtain average response, *N*, to non-repeated stimuli, and that, *R*, to repeated stimuli.



This plot of *R* against *N* has a slope less than 1, which is defined as *fMRI-A*. The line goes through the origin, showing **proportional** *fMRI-A*, within an ROI. This slope is **attenuated** by measurement error in *R* & *N*.



Study details

- Define regions of interest (ROI) in ventral occipito-temporal cortex for which category selectivity is seen in (a) regional preferences for particular object categories, e.g., faces, limbs, flowers, etc., and (b) distributed patterns of activity across regions.
- Some stimuli are presented once (**non-repeated**). For **repeated** stimuli, the interval between stimuli is varied from .5-3 secs (short-lagged), in one condition, to 1-174 secs (long-lagged) in another.

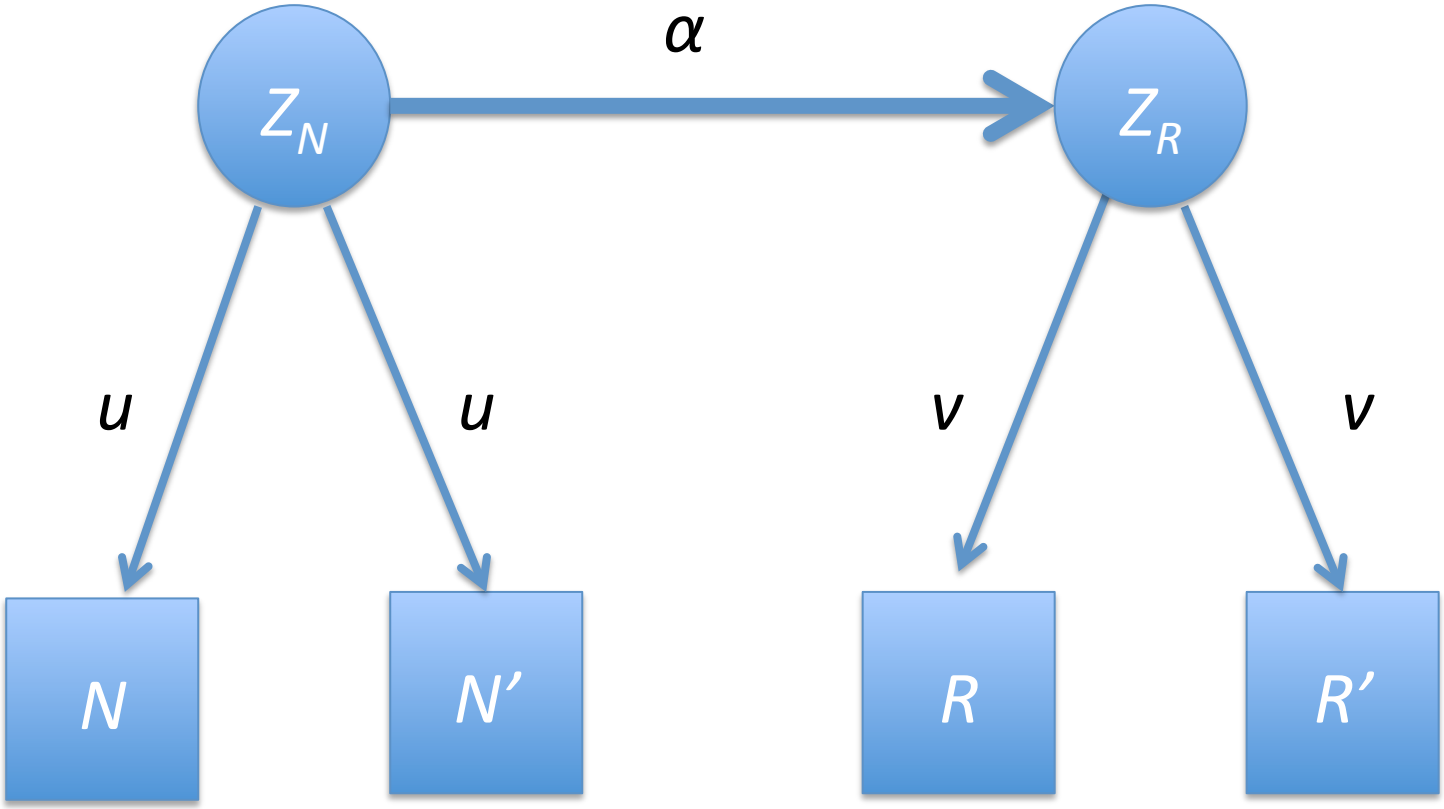
Study details

- For each ROI, each category (e.g., ROI = face-selective area, category = faces or limbs), and each voxel (= 1.5x1.5x3 mm of tissue), measure the BOLD response amplitude for non-repeated (*N*) and repeated (*R*) objects.
- Regress *R* on *N* and test if slope = 1. Compare slopes (= *fMRI-A*) between 'interesting' conditions, e.g., 'faces' vs. 'limbs' in face-area.
- But slope differences might be due to differences in *noise*, **not** in *category selectivity*! How to dispel this possibility?

Split-Half approach

- Divide expt into 'even' runs, with measures *N* and *R* for each voxel, and 'odd' runs, with measures *N'* and *R'* for each voxel.
- Noise in both variables attenuates the regression slope, *m*, of *R* on *N,* "as measured by comparing the slope with 1.0. To show that 1.0 is the appropriate standard, regress *N* on *N'* and show that this slope, m_{SH}, equals 1."
- Note, however, that noise also causes $m_{SH} < 1$, so that 1.0 is not the appropriate standard. The relevant test is whether $m < m_{SH}$, as can be seen in the following causal model.

SEM approach (see HW-2)



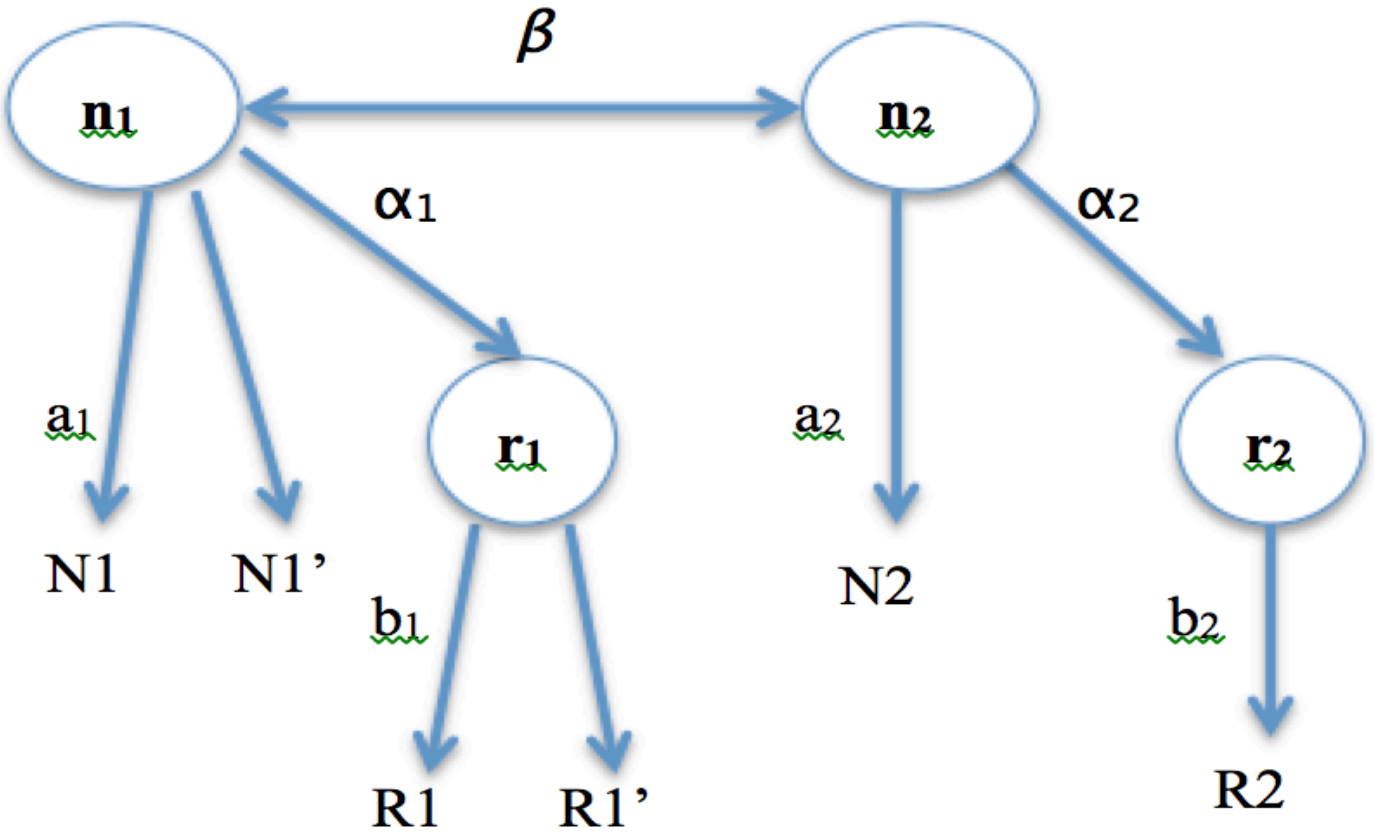
$r_{NN'} = u^2$, $r_{RR'} = v^2$; $r_{NR} = u\alpha v$. Hence,

$\boldsymbol{\alpha = r_{NR}/uv = r_{NR}/\sqrt{[r_{NN'}r_{RR'}]}}$, which can be estimated.

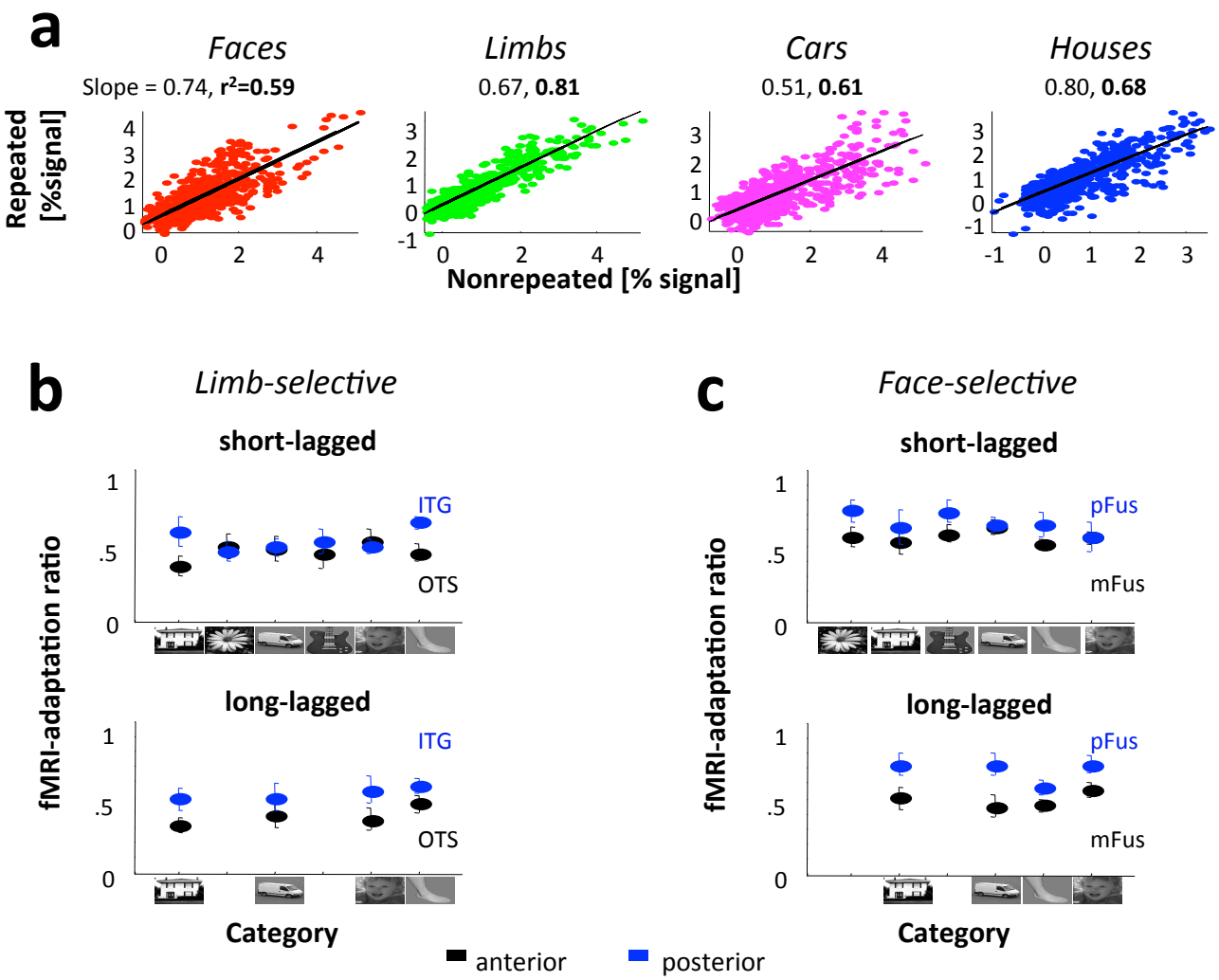
Ex: **Category 1:** $r_{NN'} = .62$, $r_{RR'} = .51$, and $r_{NR} = .36$.

Category 2: $r_{NN'} = .12$, $r_{RR'} = .11$, and $r_{NR} = .07$.

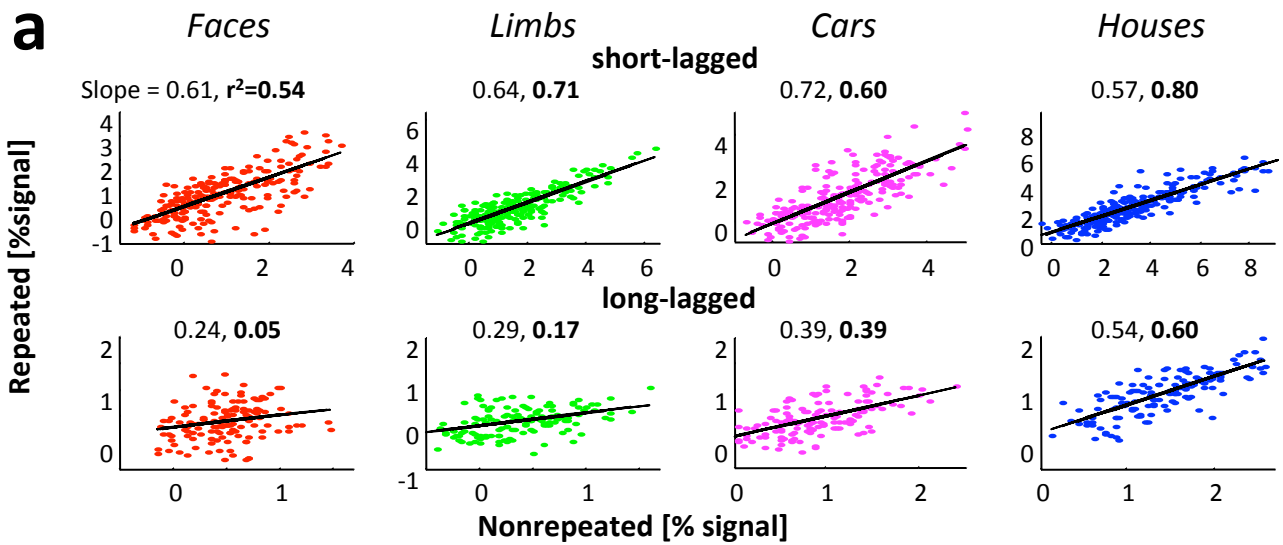
Ex. Correlations between Categories 1 & 2:
corr(*N1, N2*) = .25, and corr(*R1, R2*) = .085.

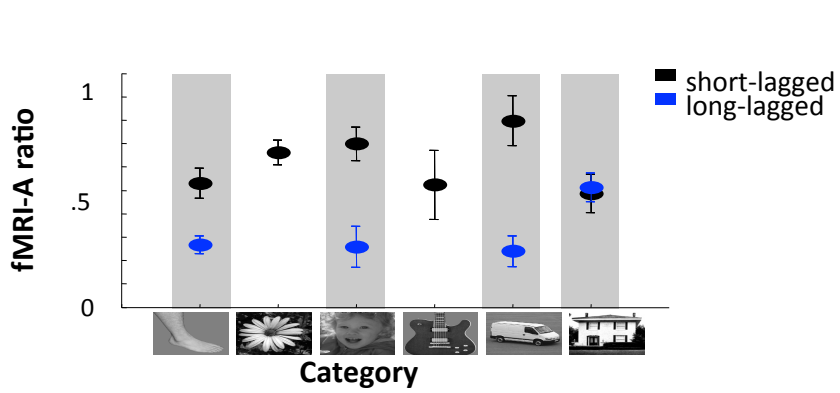


Corr(*N1, N2*) = $a_1 * \beta * a_2$; *corr(R1, R2) = …*



Exp 1: Demonstration of *fMRI-A* in different conditions





Exp 2: Demonstration of *fMRI-A* in different conditions

Test Construction

- In constructing a test made up of *n* items, **which items** from the preliminary studies to include, which to exclude?
- Look at **inter-item correls**, and **item-total correls**; also, whether **items** are too easy or too difficult.
- How large should *n* be so that the test is reliable? Larger *n* = higher reliability.
- **'Tests' inter-correlate more highly than do 'items'**. Use artificial data set to show this.

Data: Create set of *latent* '*True*' scores. q1, q2,… are *observed* '*item*' scores. *Xn.1, Xn.2* are 2 different scores on an *n*-item test. '*Critern*' is score on a 'criterion'. '*Normscr*' is a **normed** test score with mean = 100 and sd = 15:

- **True q1 q2 .. q6 Critern Normscr**
  69 75 62 86 8.2 117
  .. ..

- **Test and re-test scores on 2-item and 3-item tests:**
  X2.1 X2.2 X3.1 X3.2
  159 165 240 251
  .. ..

Correlations among scores on q1, q2, q3, q4, q5, q6, are modest (Median = .4)

	q1	q2	q3	q4	q5
q2	0.60**				
q3	0.04	0.09			
q4	0.31	0.36	0.33		
q5	0.45*	0.41	0.22	0.40	
q6	0.36	0.50*	0.56*	0.42	0.36

- For *n* = 1, median corr(qi, qj) = 0.4.
- For *n* = 2, corr(*X2.1, X2.2*) = 0.514.
- For *n* = 3, corr(*X3.1, X3.2*) = 0.71.

- Substantive item analysis: face and construct validity; difficulty; reverse coding
- Statistical item analysis in SPSS: **Analyze > Scale > Reliability Analysis**.
- **Test validity**. This is the correlation between test and **criterion**. Here, corr(*Normscr, Critern*) = 0.477 (p = 0.033).

Reliability & Test Length

- The variance-covariance matrix, **V**, for $\{Y_1, Y_2, \ldots, Y_n\}$
- A pretext for a digression on the *positive definiteness* of a square matrix, very important in iterative procedures
- Use **V** to calculate var(X), where $X = Y_1 + Y_2 + \ldots + Y_n$
- Use the resulting expressions for var(X) and E(XX') to derive Spearman-Brown formula

Dependence of reliability on test length: '*n*-item test model'

- To study the effect of test length, we use a more 'superficial' model involving only **observable** test and item scores. A test consists of *n* items; Y_j is the score on the *j*' th item, *X* = test score, *X*' = retest score. $r_{XX'}$ is the test-retest reliability.

$$X = Y_1 + Y_2 + \ldots + Y_n;\ X' = Y_1' + Y_2' + \ldots + Y_n'.$$

- The item scores are **correlated** with each other, so var(*X*) is **not** the sum of the item score variances; it depends on the correlation between the item scores. This inter-item correlation is expected to be **positive** if the items 'hang together', as we expect.

Dependence of reliability on test length

Rescale all variables to have a mean of 0. Then

$$\mathrm{var}(X) = \mathrm{var}(X') = E(X^2)\ and\ \mathrm{cov}(X, X') = E(XX').$$

$$r_{XX'} = \frac{\mathrm{cov}(X, X')}{\sigma_X \sigma_{X'}} = \frac{E(XX')}{\sigma_X^2}.$$

We now express $E(X^2)$, i.e., var(X), and $E(XX')$ in terms of var(Y_i) and cov(Y_i, Y_j). [Note that cov(Y_i, Y_i) = var(Y_i), showing that 'covariance' is a generalisation of 'variance' to the case of 2 variables.]

The variance-covariance matrix, V

- Consider a set of random variables, $\{Y_1, Y_2, \ldots, Y_n\}$. **V** is [note that $\sigma_{ij} = \sigma_{ji} = cov(Y_i, Y_j)$]:

	Y_1	Y_2	…	Y_n
Y_1	σ_1^2	σ_{12}	…	σ_{1n}
Y_2	σ_{21}	σ_2^2	…	σ_{2n}
…	…	…	…	…
Y_n	σ_{n1}	σ_{n2}	…	σ_n^2

Positive definite matrix

- Matrices have many properties that are analogous to those of real numbers, and one such property is 'positivity'. Positive numbers are special (e.g., they can be interpreted as 'size' indices). **Positive definite matrices** are also special, and play an important role in Statistics and Optimization problems.
- A matrix, **A,** is **positive definite** iff its **determinant** is positive, i.e., iff

$$\det(\mathbf{A}) \cong |\mathbf{A}| > 0.$$

Positive definite matrix

- A square matrix, **A,** has an inverse, $\mathbf{A}^{-1}$**,** iff $|\mathbf{A}| \neq 0$.
- Many computations, e.g., parameter estimation (say, of MLE's), are iterative, such that the 'adjustment' at each iteration depends on inverting some matrix, **A**.
- If **A** is defined so that it is always **positive definite,** then, *a fortiori*, we know that $|\mathbf{A}| \neq 0$ and that matrix inversion is always possible, and our iterative process will not 'blow up'.
- In ML estimation, the step-size at each iteration requires the inversion of a var-covar matrix.

The variance-covariance matrix, V, is *positive definite*

In the simple case of 2 variables, **V** is a 2x2 matrix, and $r = \sigma_{12}/\sigma_1\sigma_2$.

$|\mathbf{V}| = \sigma_1^2\sigma_2^2 - (\sigma_{12})^2 = \sigma_1^2\sigma_2^2[1 - (\sigma_{12}/\sigma_1\sigma_2)^2]$

$= \sigma_1^2\sigma_2^2[1 - r^2] > 0.$ QED

	Y_1	Y_2
Y_1	σ_1^2	σ_{12}
Y_2	σ_{21}	σ_2^2

Conversely, any positive definite matrix is the variance-covariance matrix for some multivariate distribution.

$$\text{If } X = Y_1 + Y_2 + \ldots + Y_n,$$
$$\text{var}(X) = \text{sum}(\boldsymbol{V})$$

- This is a very useful approach to finding the variance of the **sum of correlated random variables**.
- The proof is included in the derivation of the Spearman-Brown formula below.

Spearman-Brown formula: All variables have mean = 0

$$X = Y_1 + Y_2 + \ldots + Y_n;\ X' = Y_1' + Y_2' + \ldots + Y_n'.$$

$$Let\ \mathrm{var}(Y_i) = E(Y_i^2) = \mathrm{var}(Y_j') = \sigma_Y^2;$$

$$corr(Y_i, Y_j) = \rho,\ for\ i \neq j;\ and\ corr(Y_i, Y_j') = \rho.$$

$$Thus,\ for\ i \neq j,\ E(Y_i Y_j) = \mathrm{cov}(Y_i, Y_j) = \rho\sigma_Y^2.$$

$$Also,\ E(Y_i Y_j') = \mathrm{cov}(Y_i, Y_j') = \rho\sigma_Y^2.$$

$$XX' = Y_1Y_1' + Y_1Y_2' + \ldots + Y_nY_n' \ (for\, n^2\, terms);\ so$$

$$E(XX') = E(Y_1Y_1') + E(Y_1Y_2') + \ldots + E(Y_nY_n') = n^2\rho\sigma_Y^2.$$

$$X^2 = \sum_{i=1}^{n} Y_i^2 + \sum_{i\neq j} Y_iY_j \ [with\ n(n-1)\ terms\ in\ 2nd\ sum]$$

$$E(X^2) = \sum_{i=1}^{n} E(Y_i^2) + \sum_{i\neq j} E(Y_iY_j) = n\sigma_Y^2 + n(n-1)\rho\sigma_Y^2$$

$$= \sigma_X^2,$$

recalling that var(X) = $E(X^2)$.

Dependence of reliability on test length

$$r_{XX'} = \frac{\mathrm{cov}(X, X')}{\sigma_X \sigma_{X'}} = \frac{E(XX')}{\sigma_X^2} = \frac{n^2 \rho \sigma_Y^2}{n\sigma_Y^2 + n(n-1)\rho\sigma_Y^2}$$

$$= \frac{n^2 \rho \sigma_Y^2}{n\sigma_Y^2[1 + (n-1)\rho]} = \frac{n\rho}{1 + (n-1)\rho}.$$

ρ is the inter-item correlation, the degree to which items ‘hang together’, and n is the number of items comprising the test. This is the general form of the Spearman-Brown formula.

- **Dependence of reliability on test length**.

$$r_{XX'} = \frac{n\rho}{1+(n-1)\rho}$$

- **Example**. ρ = 0.1 (small), *n* = 20 (long),

$$r_{XX'} = \frac{20(.1)}{1+(19)(.1)} = 0.69 \quad \text{(decent).}$$

Spearman-Brown formula (cont'd)

Regard Y_j as the score on a test of "unit" length. Then X is the score on a test of length n, and $\rho = \text{corr}(Y_j, Y_k)$ becomes the test-retest reliability of a test of **unit** length. Let us denote ρ by r_{11}, and r_{XX} by r_{nn}. Then

$$r_{nn} = \frac{nr_{11}}{1+(n-1)r_{11}}$$

Spearman-Brown formula (cont'd)

- **Split-half reliability**. Now let *Y* be the total score on "even" items, and *Y'* be the total score on "odd" items. Then $r_{YY'}$ is the test-retest reliability of a "test" of some length, say, 1. $X = Y + Y'$ is then the score on a test of length 2. That is, if $r_{YY'} \equiv r_{11}$, then

$$r_{XX'} \equiv r_{22} = \frac{2r_{11}}{1+r_{11}} = \frac{2r_{YY'}}{1+r_{YY'}}$$

Spearman-Brown formula (cont'd)

- Cronbach's *alpha,* for a **1-item test**, is defined as:

$$\alpha = \frac{\sigma_p^2}{\sigma_p^2 + MS_{res}}$$

$$Compare: \quad r_{XX'} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_e^2}.$$

- Cronbach's *alpha,* for an ***n*-item test**, can then be found from the Spearman-Brown formula above.
- SPSS gives coefficient alpha, as shown in Chap. 36 of the SPSS manual. (**Analyze > Scale > Reliability Analysis,** etc.)

Alternative definition of *alpha*

- Suppose there are *n* items on a test; and *p* persons take the test. The scores, Y_i, on item *i* have some variance, var(Y_i), between persons. The total score, *X* ($X = Y_1 + Y_2 + \ldots + Y_n$), has variance, var(*X*), between persons.
- Cronbach's *alpha* **for the *n*-item test** is defined as:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_i \text{var}(Y_i)}{\text{var}(X)}\right).$$

A brief digression - from HW-2

Reliability of 1- vs *k*-item tests

- What motivates this distinction?
- Consider a reliability analysis based on a *p* (patients) x *k* (raters) matrix of scores. After this analysis, the rating method is applied to 'many' patients. Is a patient to be rated by 1 clinician? OR, is a patient rated by *k* = 2 or 4 clinicians, whose ratings are 'pooled' to assess the patient's status?
- The former leads us to report the reliability of a 1-item test; the latter that of a *k*-item test.

Using SPSS Output:

- The ratio, $\{\sum \text{var}(Y_i)\}/\text{var}(X)$, in the expression for ***alpha*** can be computed by requesting item statistics and the ANOVA table in SPSS:
- $\{\sum \text{var}(Y_i)\}/\text{var}(X)$ = {Mean of Item Variances}/{Mean Square Between Persons}
- Other formulae give slightly differing values.

Corollary: ρ has a lower bound!

- Consider the model in which all pairs of $\{Y_j\}$, j = 1, 2, …, n, have the same intercorrelation, ρ.
- The **variance** of the **sum,** X, of the $\{Y_j\}$ has to be positive, because all variances are positive. But we just proved that

$$\mathrm{var}(X) = n\sigma_Y^2 + n(n-1)\rho\sigma_Y^2 > 0.$$

- Dividing by $n\sigma_Y^2$, we get: $1 + (n-1)\rho > 0$.
- Thus $\rho > -1/(n-1)$. ρ **cannot** be too large negative! (The enemy of my enemy is my friend!)

An application

- For a voxel in the **face area**, we have N_0 and R_0, the responses to non-repeated and repeated **faces**. Suppose N_1 and R_1 are that voxel's responses to non-repeated and repeated **houses.**
- In Weiner et al., cor(N_0, N_1) < 0.
- According to our Corollary above, if all n correls, cor(N_i, N_j) = ρ, then $\boldsymbol{\rho}$ **cannot be less than -1/(*n*-1)**! This puts a mild constraint on our models.