Measures of pattern similarity: cross-validation and prewhitening

mike freund

May 13, 2020

In analyses within this repo I have considered 12 different forms of similarity.

- 3 **measures** of similarity
  - linear **correlation**
  - **euclidean** distance
  - **standardized euclidean** distance
- 2 **methods** of estimation
  - **vanilla RSA**: assessing similarity of patterns across scanning runs
  - **"cross-validated" RSA**: tweaking cross-run estimation procedure so that measures become *unbiased*
- 2 **normalizing** transforms
  - **"raw"**, or un-normalized
  - spatially **prewhitened**, or "multivariate noise normalized"

Below, these things are described and references linked.

measures of similarity and "methods of estimation"

A general note about 'cross-validation' and unbiasedness.

All of these measures — correlation, euclidean, standardized euclidean — involve a *quadratic form*, that is, a term that is multiplied to itself. In correlation, for example, this form is the variance of each measure: $\text{Var}(x) = \text{Cov}(x, x) = \mathbb{E}[(x - \bar{x})^2]$. In euclidean distance, this form is the square of a difference vector: $(\mathbf{x} - \mathbf{y}) \cdot (\mathbf{x} - \mathbf{y}) = (\mathbf{x} - \mathbf{y})^2$.

Within the context of classical test theory, a measure x is composed of true score and error components. Because these similarity measures involve mutliplying x to itself, both the true score and error components are squared, and both contribute to the expected value of the estimate. This is what makes these similarity measures 'biased'.

"Cross-validated" versions of these measures are unbiased: insensitive, in terms of expected value, to the error component. To make each of these measures unbiased, the same 'trick' is used. This trick is to "swap out" one of these x terms within the quadratic form with an independent "copy" of itself. In cross-validated correlation, for example, the quadratic form for $\text{Cov}(x, x)$ is substituted with two independent copies of x, obtained from different scanning runs: $\text{Cov}(x_{(1)}, x_{(2)})$. This swapping out creates an unbiased measure because the error terms cancel out (because of independence; see Alink et al, 2015 for the algebra).

linear correlation

vanilla method

The correlation between two pattern vectors x, y can be written

$$r = \text{cor}(x, y) = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)\text{Var}(y)}}$$

i.e., the covariance of a measure standardized by the (rooted product of the) variances.

This is equivalent to writing

$$r = \frac{\mathrm{Cov}(x, y)}{\sqrt{\mathrm{Cov}(x, x)\mathrm{Cov}(y, y)}}$$

because the covariance of a measure with itself is the variance.

In the context of "vanilla" / cross-run RSA, the correlation of x and y is estimated *across scanning runs* — e.g., x from run 1 and y from run 2:

$$r_{vanilla} = \mathrm{cor}(x_{(1)}, y_{(2)}) = \frac{\mathrm{Cov}(x_{(1)}, y_{(2)})}{\sqrt{\mathrm{Cov}(x_{(1)}, x_{(1)})\mathrm{Cov}(y_{(2)}, y_{(2)})}} \tag{1}$$

Estimating across runs gives robustness to potential design artifacts stemming from temporal autocorrelation (e.g., Alink et al, 2015, Cai et al., 2019.

cross-validated method

To make the linear correlation unbiased, the Spearman's correction for attenuation is applied.

The idea is that the maximum observable correlation between $x_{(1)}$ and $y_{(2)}$ is bounded by the (root product of their) reliabilities: lower reliability, lower observed correlation. So, to correct for measurement error, the observed correlation values are scaled by their (root product of their) reliabilities. Here, reliability is estimated as the cross-run covariance.

Essentially, this amounts to swapping one out of each of the measures in the denominator of Eq. (1), $x_{(1)}$ and $y_{(2)}$, with an independent 'copy' of itself, $x_{(2)}$ and $y_{(1)}$:

$$r_{cv} = \frac{\mathrm{Cov}(x_{(1)}, y_{(1)})}{\sqrt{\mathrm{Cov}(x_{(1)}, x_{(2)})\mathrm{Cov}(y_{(1)}, y_{(2)})}} \tag{2}$$

- see Appendix of Alink et al, 2015 for expanded definition of cross-run correlation.
- because a covariance between independent measurements can be ≤ 0, the denominator of the cross-validated correlation can be undefined (because of the root).
- this property is problematic for this measure, as even moderate amounts of noise can make this measure unstable and cause 'missing' data.
- I included the measure in my analyses for completeness but it was indeed very unstable.

euclidean distance

vanilla

The vanilla / cross-run squared euclidean distance can be written

$$d^2_{vanilla} = \sum_{v=1}^{V} (x_{v(1)} - y_{v(2)})^2$$

i.e., the sum of squared differences between voxels v in $1, \ldots V$ in pattern x from run 1 and pattern y from run 2.

The subtraction operation above can also be thought of in vector form, where x and y are vectors $\mathbf{x}$ and $\mathbf{y}$, and $\mathbf{a} = \mathbf{x} - \mathbf{y}$ is the **difference vector**, or vector pointing from $\mathbf{y}$ to $\mathbf{x}$. The squared euclidean distance can then be thought of as the squared *length* of this difference vector. (The squared length of a vector is just a vector multiplied by itself.) So:

$$d^2_{vanilla} = \mathbf{a}^2 = \mathbf{a} \cdot \mathbf{a} = \sum_{v=1}^{V} a_v^2 = \sum_{v=1}^{V} (x_{v(1)} - y_{v(2)})^2 \tag{3}$$

cross-validated

To make euclidean distance unbiased, the trick is the same one as before: swap out one of the terms within the quadtratic form with an independent 'copy' of itself. Here, the quadratic form is $\mathbf{a}^2$, which is replaced with $\mathbf{a}_{(1)} \cdot \mathbf{a}_{(2)}$. Essentially this amounts to performing the subtraction *within-run*, then multiplying the difference vectors *between-run*. Because they are from independent runs, the two terms in the multiplication have independent errors. Therefore these error terms cancel in the multiplication, meaning that the resulting product (distance) reflects only the true distance.

$$d^2_{cv} = \mathbf{a}_{(1)} \cdot \mathbf{a}_{(2)} = \sum_{v=1}^{V} a_{v(1)} a_{v(2)} = \sum_{v=1}^{V} (x_{v(1)} - y_{v(1)})(x_{v(2)} - y_{v(2)}) \tag{4}$$

If there is no consistent difference between x and y across scanning runs, then the expected value of d^2_{cv} is zero. (Note that, because d^2_{cv} can be negative, it is not square-rooted, but left as a squared euclidean distance.)

standardized euclidean distance

vanilla

If the patterns x and y are z-score standardized, the squared euclidean's distance between them will be equivalent, within a scaling factor, to linear correlation.

Let the tilde denote this standardization, e.g.,

$$\widetilde{x}_{v(1)} = \frac{x_{v(1)} - \bar{x}_{v(1)}}{\mathrm{sd}(x_{v(1)})}$$

so that

$$r_{vanilla} \propto \widetilde{d}^2_{vanilla} = \sum_{v=1}^{V} (\widetilde{x}_{v(1)} - \widetilde{y}_{v(2)})^2 \tag{5}$$

For intuition, consider that the correlation is sensitive to the pattern "shape" (or, in vector space, the *angle*); scale (vector length) and mean differences (vector length along the unity line), however, are removed. Likewise, the euclidean distance is sensitive to "shape" — but also to scale, and to mean differences. Removing scale and mean differences by z-score normalizing renders euclidean and correlation sensitive to idenditcal information.

cross-validated

Standardized euclidean distance can be cross-validated in the same manner as euclidean's distance:

$$\widetilde{d}^2_{cv} = \sum_{v=1}^{V} (\widetilde{x}_{v(1)} - \widetilde{y}_{v(1)})(\widetilde{x}_{v(2)} - \widetilde{y}_{v(2)}) \tag{6}$$

Perhaps an intuitive way of thinking about the cross-validated standardized euclidean is that it indicates how correlated the difference between two conditions was across scanning runs. If two conditions have *consistently different* pattern 'shapes' across scanning runs (i.e., different in a consistent way), they will have $\widetilde{d}^2_{cv} > 0$.

prewhitening

Any of these measures can be computed on spatially 'prewhitened' patterns.

background: LDA and mahalanobis distances

Spatial prewhitening is a normalization procedure that shrinks the patterns along dimensions of strong noise variance within high-D voxel space, thus stretching, relatively, along less noisy ones.

In the context of classification and Fisher's linear discriminant analysis, this whitening takes the form of normalizing by dividing patterns by the within-class (noise) covariance matrix, $\mathbf{\Sigma}$ (e.g.). $\mathbf{\Sigma}$ indicates how the class exemplars are distributed about their class centroids. (In LDA, all classess are assumed to share a common within-class covariance matrix.) Fisher

proved that using this transform maximized the between-class variance (the length of the difference vector between class centroids) relative to the within-class variance—and he actually did so nonparametrically (ESLII, p 110). After applying this transform, Euclidean distances between observations become Mahalanobis distances. The decision rule in Fisher's LDA is based on mahalanobis distances: a test observation is assigned to the class for which it has the smallest mahalanobis distance from centroid (see, e.g., ESLII, p 108, eq 4.9).

intuition

One way of thinking about this whitening transform is as a *sphering*. For a given class (experimental condition), trial-level patterns are distributed about their class centriod (mean pattern) within high-dimensional voxel space. This cloud of points is the distribution of noise. It has some shape. If the noise is *uncorrelated* across voxels, and all voxels had equal variance, the shape would be a (hyper)sphere. If the noise was correlated across voxels, the shape would be a non-spherical ellipsoid. The Fisher LDA whitening transform makes the noise structure spherical.

estimation

In the context of rapid event-related fMRI designs, problems arise in estimating $\mathbf{\Sigma}$, which has dimension $V \times V$. There aren't enough exemplars (pattern estimates) of each class (condition) to get a good estimate of the across-voxel correlation structure. But, a trick here is to use the residual timecourses from the GLM, not the within-class voxel-by-exemplar data matrices, to estimate $\mathbf{\Sigma}$ (e.g., Misaki et al., 2010).[1] Given that there are typically more TRs within a run than voxels within a parcel, estimating $\mathbf{\Sigma}$ is now much more tractable.

To perform prewhitening in my analyses, I followed procedures outlined in Diedrichsen et al (2016).

Let ϵ_r represent the matrix of residual timecourses for a given run r, parcel, and subject. I.e., ϵ_1 is $TR \times V$ and corresponds to run 1. The vertex-by-vertex covariance matrix of ϵ_1 is given by $\mathbf{\Sigma}_1 = \mathrm{Cov}(\epsilon_1) = \epsilon_1^T \epsilon_1$.

- to curb overfitting (and potentially enable inversion), $\mathbf{\Sigma}_1$ is regularized by shrinking it towards a diagonal matrix, $\mathbf{D}$ by a factor λ, i.e.:

  $$\mathbf{\Sigma}_1^* = \lambda \mathbf{D} + (1 - \lambda)\mathbf{\Sigma}_1$$

  Regularization of $\mathbf{\Sigma}$ is a common procedure for LDA, which greatly aids performance in high-dimensional scenarios (see, e.g. Guo, Hastie, Tibshirani, 2006]).
- The optimal λ can be estimated (Ledoit & Wolf, 2003). However, for a first-pass analysis, I used a healthy factor of $\lambda = 0.4$
- Once estimated, $\mathbf{\Sigma}_1^*$ is inverted, e.g. $\mathbf{\Sigma}_1^{*-1}$, to perform the matrix division.
- This estimation and inversion was done separately for both runs. Matrices were then averaged across run $\mathbf{\Sigma}^{*-1} = (\mathbf{\Sigma}_1^{*-1} + \mathbf{\Sigma}_2^{*-1})/2$, yielding a single matrix $\mathbf{\Sigma}^{*-1}$, to be applied to patterns of both runs.
  - however, an alternative method, which may give a lower variance estimate of $\mathbf{\Sigma}^{*-1}$, would be to concatenate ϵ_1 and ϵ_2 along the time dimension, $\epsilon = \binom{\epsilon_1}{\epsilon_2}$, then use the concatenated matrix to estimate a single covariance matrix $\mathbf{\Sigma} = \mathrm{Cov}(\epsilon)$ across all timepoints from run 1 and run 2. I'll leave this for future exploration.

application

Prewhitening can be incorporated into any of the estimation procedures outlined above. For example:

$$r_{vanilla,\ prw} = \mathrm{cor}(x_{(1)}\mathbf{\Sigma}^{*-1/2}, y_{(2)}\mathbf{\Sigma}^{*-1/2})$$

$$d^2_{vanilla,\ prw} = \mathbf{a}\mathbf{\Sigma}^{*-1}\mathbf{a}$$

$$d^2_{cv,\ prw} = \mathbf{a}_{(1)}\mathbf{\Sigma}^{*-1}\mathbf{a}_{(2)}$$

[1] Note that this is not the exact form of prewhitening derived by Fisher for LDA. Nevertheless, the *prewhitened cross-validated euclidean distance* is what Walther et al. (2016) refer to as the *cross-validated mahalanobis* ... which they also refer to as the "linear discriminant contrast". What they refer to as "linear discriminant t-value", is closely related. It's simply a linear discriminant contrast divided by its standard error. (See Appendix of Walther et al. 2016 for discussion.)