HST.583 Functional Magnetic Resonance Imaging: Data Acquisition and Analysis
Fall 2008

Some Multivariate Methods for fMRI Data Analysis

Mark Vangel

Harvard Clinical and Translational Science Center

(Catalyst)

MGH & MIT

HST 583, November 2008

Some Multivariate Methods

- Functional and Effective Connectivity
- PCA
- ICA
- Granger Causality

Connectivity, Correlation and Association

- Functional vs Effective Connectivity
- Causality vs Association

Brain Networks: Functional Connectivity

- Brain structures as nodes in network.
- White matter tracts as (*undirected*) links connecting these nodes.
- We would like to determine which are important connections among the nodes.
- **Statistical Concepts:** correlation, partial correlation, coherence, partial coherence
- **Some Statistical Methods:**
  - For determining the links: correlation analysis, partial correlation analysis
  - For determining the nodes: *Neuroscience theory (a-priori)*, or PCA, ICA, and clustering methods.

Brain Networks: Effective Connectivity

- Brain structures as nodes in network.
- White matter tracts as (*directed*) links connecting these nodes.
- We would like to determine which are important *causal* connections among the nodes.
- **Some Statistical Methods:**
- For determining the links: SEM, DCM, Granger Causality
- For determining the nodes: same as for functional connectivity.

Association vs. Causality

- **Causality:** Predictability according to a law or set of laws.
- **Association:** Occurring together, with or without a causal relation (If A and B are associated, then perhaps A causes B, B causes A, or maybe C drives both A and B.)
- Methods for causal inference are invariably dependent on model assumptions.
- *definitive* answers require a combination of various kinds of evidence.

Principal Components Analysis (PCA)

- Introduction via a Simple Example
  - PCs and loadings
  - Interpretation in terms of eigenvectors and eigenvalues of the covariance or correlation matrix
  - Relationship to the Singular Value Decomposition (SVD) of the data matrix.
- PCA for fMRI
  - Why this is different from most “textbook” PCA.
  - Duality of spatial and temporal PCA
  - Expressions for PCS and loadings in terms of eigenvectors and eigenvalues
  - Some statistical inference

Turtle Shells: An Example of a PCA

- Males (X_m), 24 turtles by 3 variables

	Gender	Length	Width	Height
1	Male	93	74	37
2	Male	94	78	35
.	.	.	.	.
.	.	.	.	.
24	Male	135	106	47

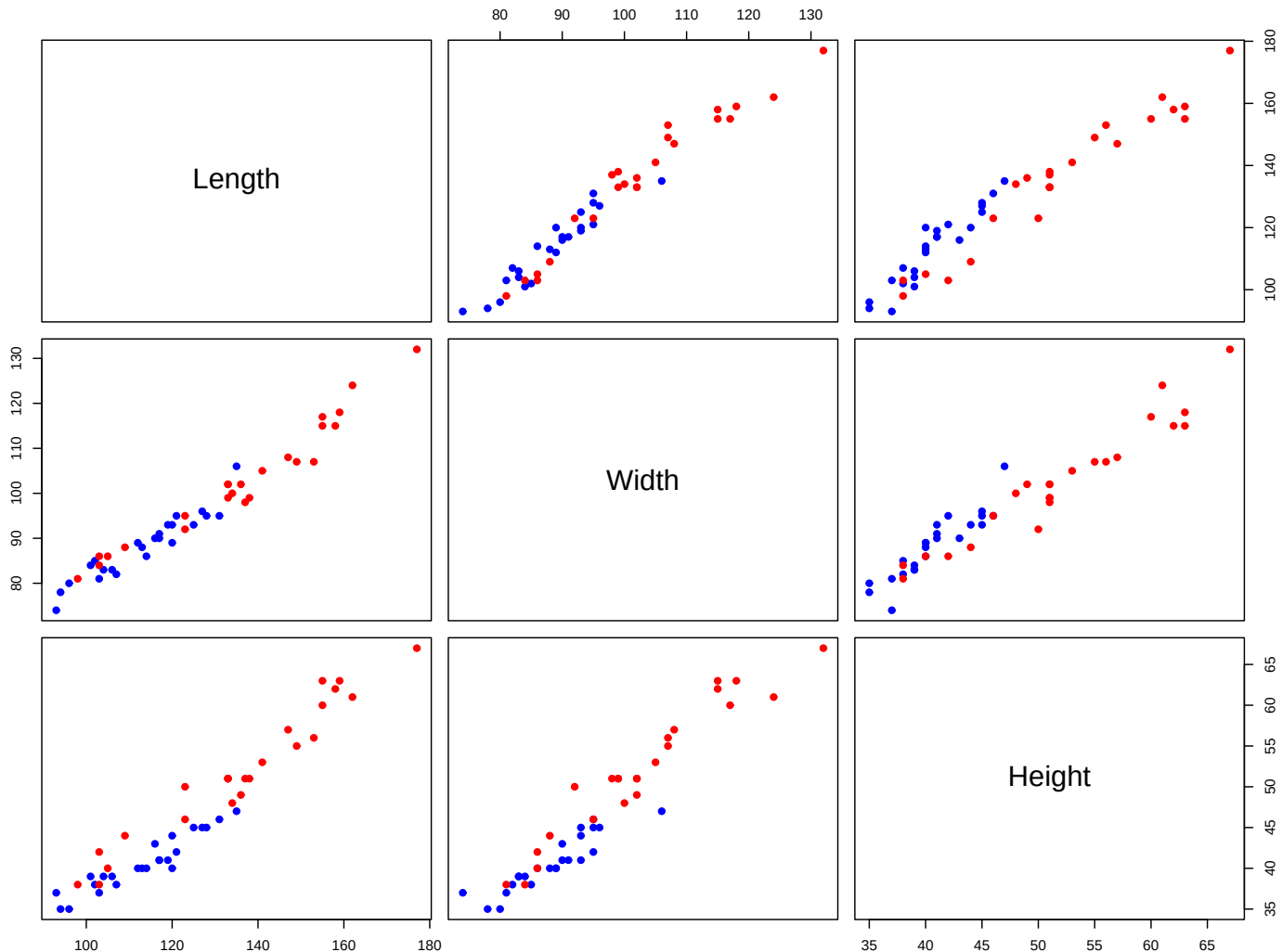
- Females (X_f), 24 turtles by 3 variables

	Gender	Length	Width	Height
25	Female	98	81	38
26	Female	103	84	38
.	.	.	.	.
.	.	.	.	.
48	Female	177	132	67

- Let $\tilde{X}_m$ and $\tilde{X}_f$ denote X_m and X_f, respectively, after centering.

Pairwise Plot of Variables

Blue plot character indicate males; red plot character indicates females.



Covariance and Correlation Matrices

- Covariance matrix (all turtles)

	Length	Width	Height
Length	419	254	166
Width	254	161	102
Height	166	102	70

- Correlation matrix (all turtles)

	Length	Width	Height
Length	1.00	0.98	0.96
Width	0.98	1.00	0.96
Height	0.96	0.96	1.00

- $$\text{Corr.} = \frac{\text{Cov}\,(x, y)}{\text{SD}\,(x)\text{SD}\,(y)}$$

Basic Idea of PCA

- Rotation of coordinates. First, find direction of greatest variability in the multidimensional cloud of data points.
- Now iterate: let $i+1$ st axis in rotated system be the direction of greatest variability *orthogonal* to all i previously determined directions.
- If X is the column-centered case-by-variables data matrix, then the rotation matrix is the (orthogonal) matrix of eigenvectors of $X^T X$. This is usually called the *loadings* matrix.
- The *principal components* (PCs, also called *scores*) are the coordinates of the data points with respect to the new coordinate system.
- The PCs are *uncorrelated*, with variances equal to the eigenvalues of the covariance matrix (i.e., $X^T X/(n-1)$).

PCA for Turtle Data: Loadings

- Loading matrix for males (A_m)

	PC1	PC2	PC3
Length	0.84	-0.49	-0.24
Width	0.49	0.87	-0.05
Height	0.23	-0.08	0.97

- Loading matrix for females (A_f)

	PC1	PC2	PC3
Length	0.81	-0.55	0.21
Width	0.50	0.83	0.25
Height	0.31	0.10	-0.95

- Note: PC2 and PC3 might help us discriminate genders.

PCA for Turtles: PCs

- Means for Males

  Length $=$ 113; Width $=$ 88; Height $=$ 41

- Principal components for first male turtle:

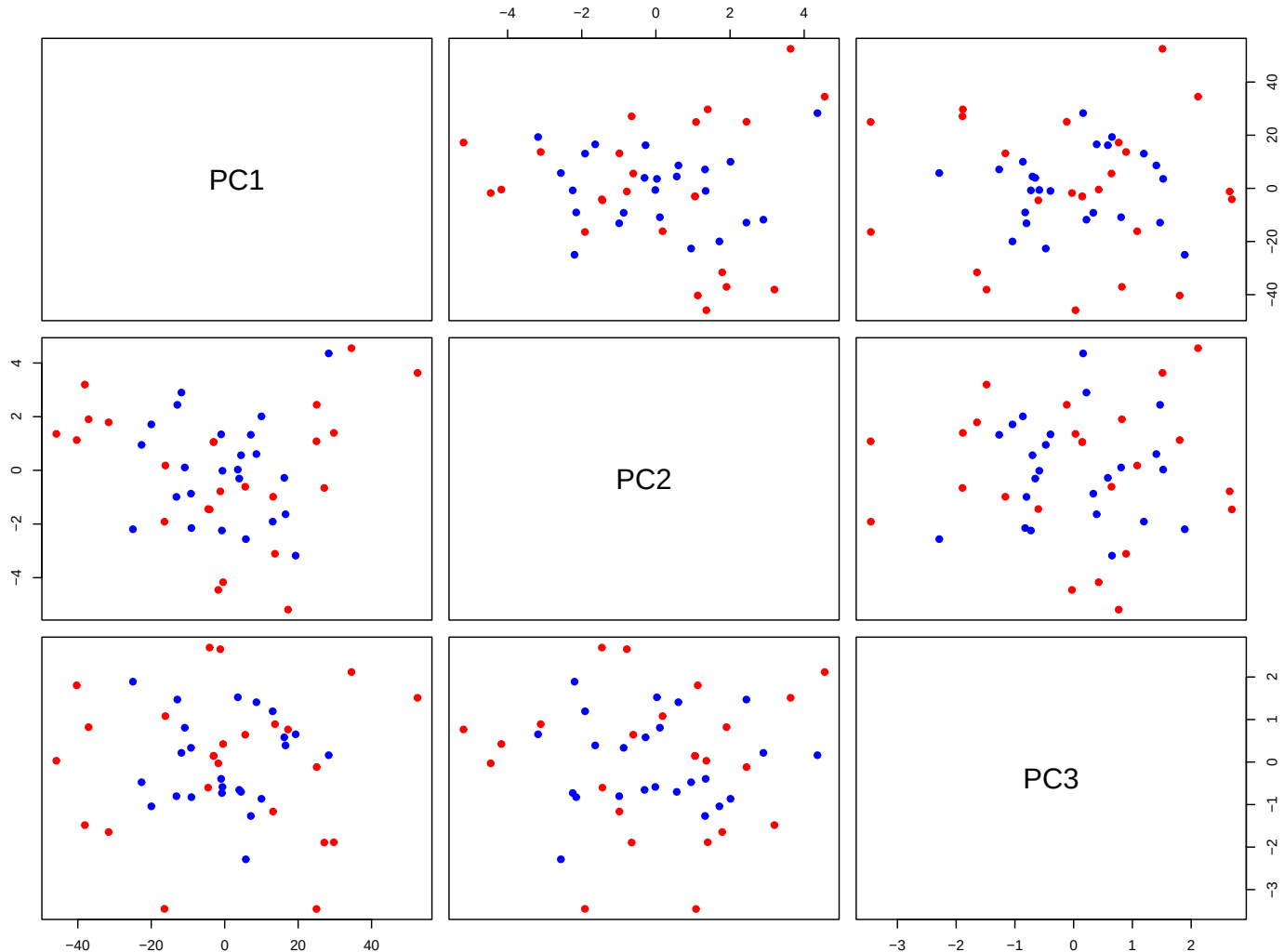
$$\begin{bmatrix} 0.84 & -0.49 & -0.24 \\ 0.49 & 0.87 & -0.05 \\ 0.23 & -0.08 & 0.97 \end{bmatrix}^T \begin{bmatrix} 93-113 \\ 74-74 \\ 37-41 \end{bmatrix} = \begin{bmatrix} -25.0 \\ -2.2 \\ 1.9 \end{bmatrix}$$

- Matrix equation for all male turtle data. Recall that $\tilde{X}_m$ is the centered data matrix for males:

$$\tilde{X}_m A_m = P_m$$

Pairwise Plot of PCs

Blue plot character indicate males; red plot character indicates females.



PCA on Correlations

- In addition to centering the columns of the data matrix, one can also scale these columns to have unit variance, prior to performing the PCA.
- This is equivalent to doing PCA on the correlation matrix, rather than the covariance matrix.
- It is often desirable to perform PCA on correlations rather than covariances, and it is nearly *essential* to do so when some of the variables are on different scales.

PCA for fMRI: References

- Following Andersen et al., *Magnetic Resonance in Imaging*, 17(6), 785-815, 1999. particularly pages 798-799. See also Bullmore et al., *NeuroImage*, 4, 16-33, 1996, for an applied article which clearly presents the basic methodological ideas.

How PCA in fMRI is Different

- Usual PCA scenario: Rows of $X_{n \times p}$ are a sample of size n of a p-variate Gaussian.
- In fMRI, $X_{n \times p}$ is a matrix of time $\times$ space, hence a sample of size one from a np-dimensional Gaussian.
- (Actually, one can also cast the X matrix for "typical" non-fMRI PCAs in this form, but the covariance matrix will be block diagonal with the $p \times p$ covariance matrix of the column variables repeated n times on the diagonal, because of the independent cases).

Covariance Matrices

- $X^T X$, $p \times p$, is the (singular) *spatial* covariance matrix
- $X X^T$, $n \times n$, is the *temporal* covariance matrix
- Duality: We can switch the roles of n and p (that is, work with X^T in the role of X, in which case the voxels are regarded as “cases” and time points as “variables”.)

Eigenvalues and Eigenvectors

- Eigenvalues of XX^T (temporal), and (nonzero) eigenvalues X^TX (spatial):

$$\lambda_1 > \lambda_2 > \lambda_3 > \ldots > \lambda_n$$

- Eigenvectors $\{w_k\}$ (spatial, $p \times 1$), and $\{u_k\}$ (temporal, $n \times 1$):

$$\begin{aligned} X^TXw_k &= \lambda_k w_k \\ XX^Tu_k &= \lambda_k u_k \end{aligned}$$

Decomposition of Variance

- Let $\sigma_{x_i}^2$ be the variance of x_i, the ith row of X, i.e., the variance of image at the ith time point. Then

$$\sum_{i=1}^{n} \sigma_{x_i}^2 = \sum_{i=1}^{n} \lambda_i = \text{tr}\,(XX^T)$$

- So the variance is preserved by the PCA.
- Hence, the normalized eigenvalues

$$\tau_k \equiv \frac{\lambda_k}{\sum_i \lambda_i}$$

can be interpreted as the proportion of variance attributable to the kth PCA.

Spatial-Temporal Duality

- The products $\{Xw_k\}$ yields the scores (PCs) for a decomposition in terms of the loadings $\{w_k^T\}$;
- The products $\{X^T u_k\}$ Yoda's the scores for a decomposition in terms of the loadings $\{u_k\}$.
- Since

$$u_k = Xw_k/\sqrt{\lambda_k}, \quad \text{and} \quad w_k = X^T u_k/\sqrt{\lambda_k}$$

Spatial-Temporal Duality, Continued

- Thus the *loadings* of the eigenimages of the spatial covariance matrix $X^T X$ are proportional to the *scores* for the dual analysis with time points as variables, and vice versa.
- Since $n << p$, it is computationally far more efficient to compute the eigenvectors of the $n \times n$ matrix XX^T. We can then use the duality to easily get the p-dimensional eigenvectors of $X^T X$.

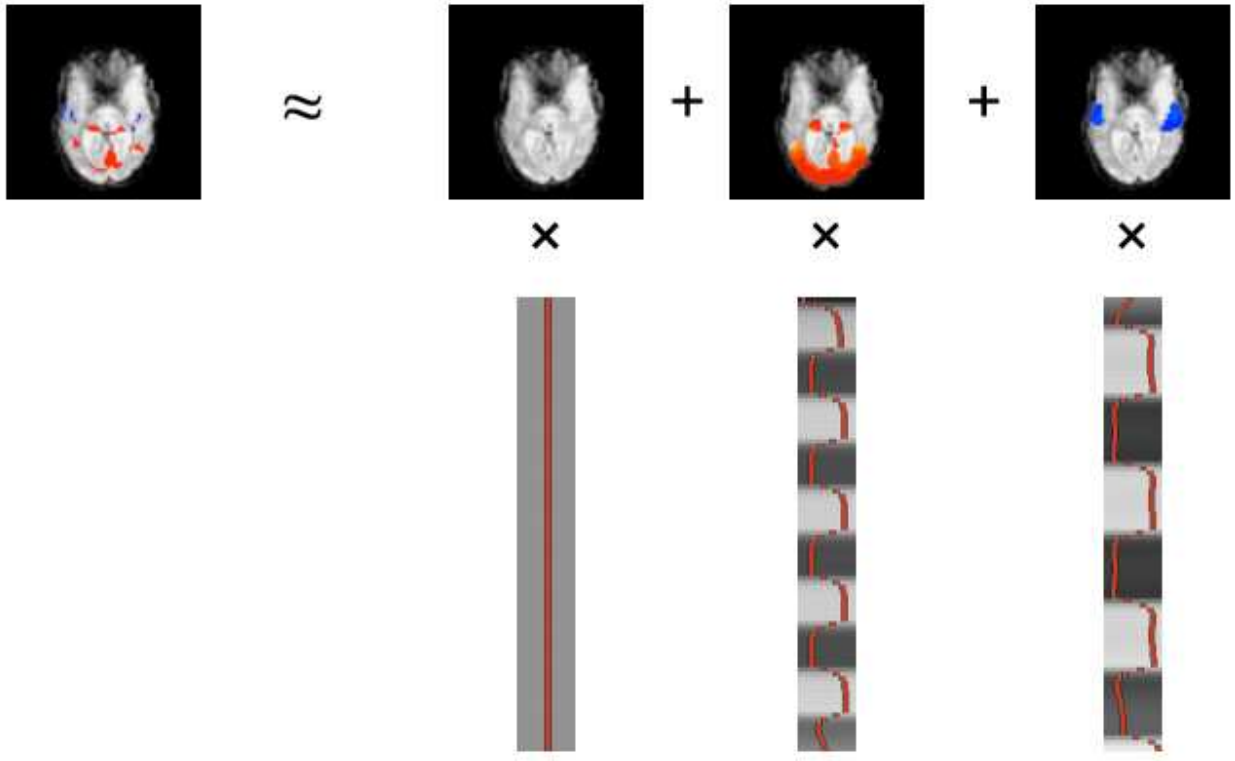
Decomposition of X

- Finally, we can use the SVD (interpreted via PCA) to express the space-time matrix X as a sum of p $n \times n$ rank-one matrices, each determined by a spatial eigenimage w_i and a temporal eigenvector u_i, weighted by decreasing eigenvalues λ_i:

$$X = \sum_{i=1}^{n} \sqrt{\lambda_i} u_i w_i^T$$

- Each term is the outer product of an eigenimage with a univariate timecourse. One can hope that the first few eigenimages are interpretable, and that they explain most of the variability in X.

Schematic of Decomposition



(Truncating the sum at $q \leq n$ terms provides the closest least-squares rank$-q$ approximation to X.)

Statistical Inference

- For the first time, we now use the Gaussian assumption of our data model, and not just the covariance matrices.

  We distinguish between estimates (e.g., $\hat{\lambda}_k$) and their corresponding population values.

$$\text{Var}\,(\hat{\lambda}_k) \approx 2\lambda_k^2/(p-1)$$

$$\text{Var}\,(\hat{u}_k) \approx \frac{\lambda_k}{p-1} \sum_{\substack{j=1 \\ j \neq k}} \frac{\lambda_j}{(\lambda_k - \lambda_j)^2} u_k u_k^T$$

- Asymptotically valid for large n and p. Substitute estimates for population values.
- Var $(\hat{w}_k)$ is of the same form as above, with obvious changes.

How Many PCs are different from noise?

- There is an extensive literature on this topic.
- Andersen et al. reference W.G. Mallows, Latent vectors of random symmetric matrices, *Biometrika*, 48, 133-149, 1960.
- In practice, one often empirically keeps a “few” eigenimages, such that a reasonably large proportion of the variance is explained, and for which the networks indicated by the images have reasonable interpretations.
- Replication over subjects is more convincing than any formal statistical test, since the tests are asymptotically valid, and even then only when the Gaussian assumption is satisfied.

How Many PCs: Common Approaches

- Plot the eigenvalues in decreasing order (a *scree plot*, and look for a break (“knee”), with the eigenvalues to the right of the knee being very small.
- Retain eigenvalues greater than the average. (Recall that the eigenvalues are the variances of the PCs).
- Formally test for the smallest eigenvalue equal to zero, then the last two eigenvalues equal to zero, etc., until the null hypothesis is rejected. This tends to retain more components than other approaches, in part because the tests are not independent.

Why do PCA?

- The first few components often account for most of the variability. We can keep only the components with “large” variances, that is, PCA can be used for dimension reduction.
- The loadings for at least some of the PCs might be interpretable, (e.g., first PC above might correspond roughly to the size of the turtle shell, or the first few eigenimages might correspond to neural networks which seem sensible given the experiment).

Independent Components Analysis (ICA)

- Correlation vs. Independence
- The blind-source separation problem
- A simple example with two components
- The importance of NonGausianity
  - The central limit theorem
  - Kurtosis
  - Negentropy
- Preprocessing
- Probabilistic ICA
- Comparison with PCA

Correlation

- If one plots observations of two random variables in a 2D scatterplot, then correlation measures the extent to which these points fall along a *straight line*.
- If X and Y are random variables, with expected values $E(X)$ and $E(Y)$, then the correlation ρ is defined as

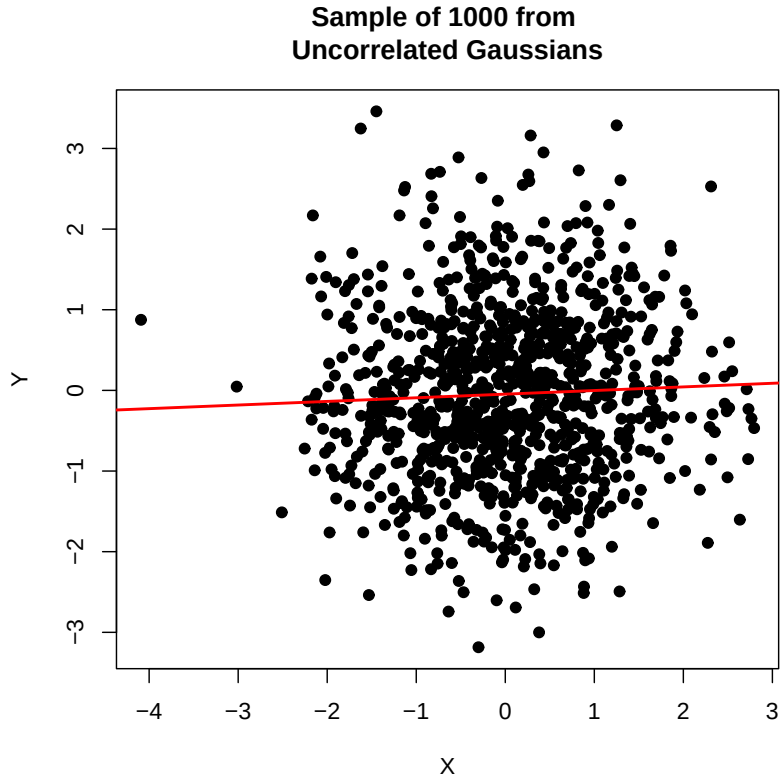
$$\rho = \frac{E[(X - E(X))(Y - E(Y))]}{\sqrt{E(X - E(X))^2}\sqrt{E(Y - E(Y))^2}}$$

- The sample form of this equation is

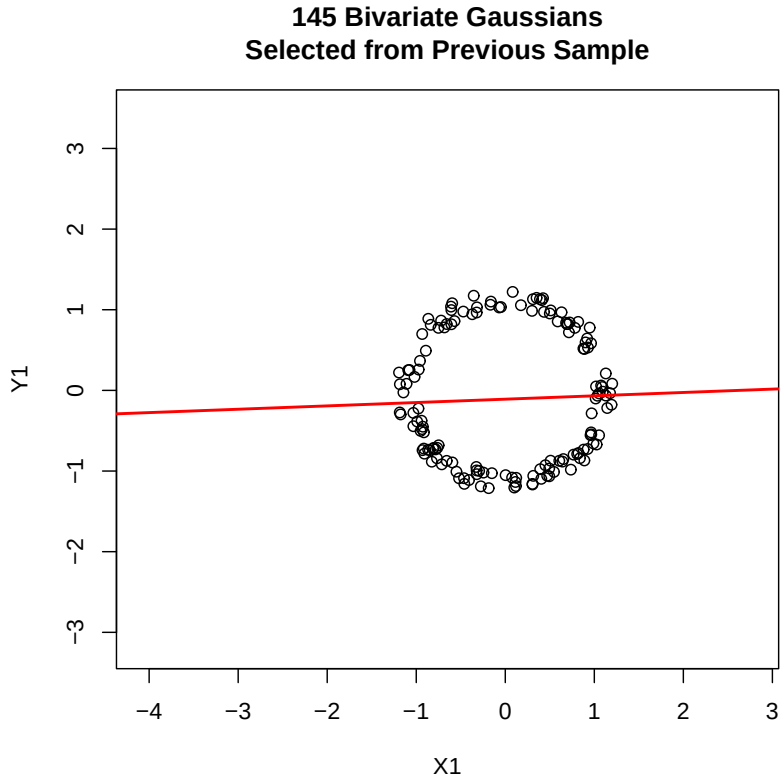
$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

- The numerator is the *covariance*, and the denominator is the product of standard deviations.

Uncorrelated Gaussians are Independent



Uncorrelated Non-Gaussians can be Dependent



ICA: Blind Source Separation

- Imagine that there are n time-varying sources $s_i(t)$.
- We cannot directly observe the s_i; instead we observe linear combinations

$$x_i(t) = a_{i1}s_1(t) + a_{i2}s_2(t) + \ldots + a_{in}s_n(t),$$

  or, in matrix form

$$X = AS$$

- (Note, by the way, that there is no error term above.)
- If the $s_i(t)$ are *independent* at each t, then one can, in principle, determine the s_i uniquely up to a multiplicative constant.

Example: Hyvärinen and Oja (2000)

- Two independent signals uniformly distributed on $[-\sqrt{3}, \sqrt{3}]$:

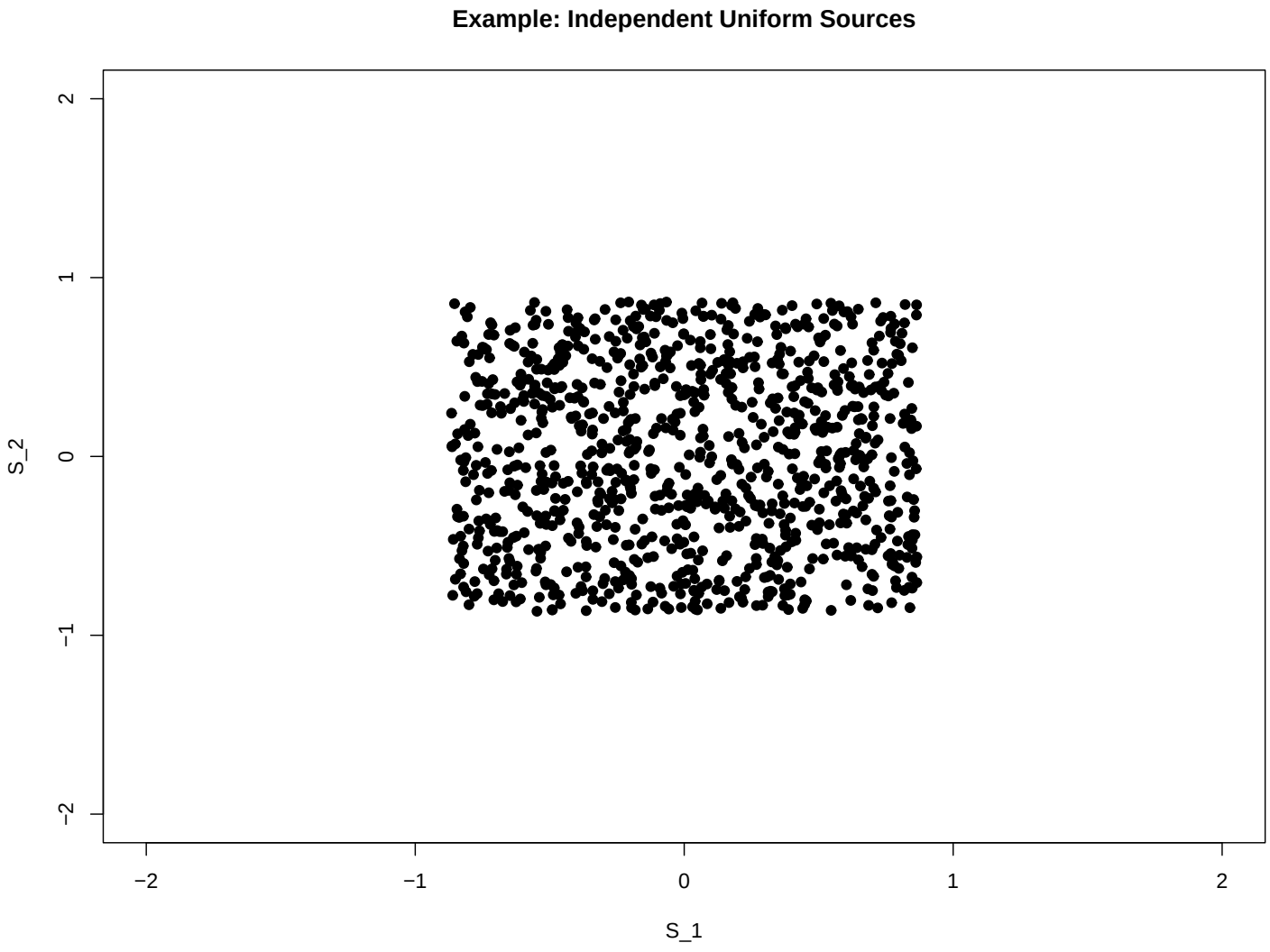
$$s_i \sim U(-\sqrt{3}, \sqrt{3}),$$

for $i = 1, 2$.

- We observe the following pair of *dependent* random variables, which is a linear mixture of the s_i:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$$

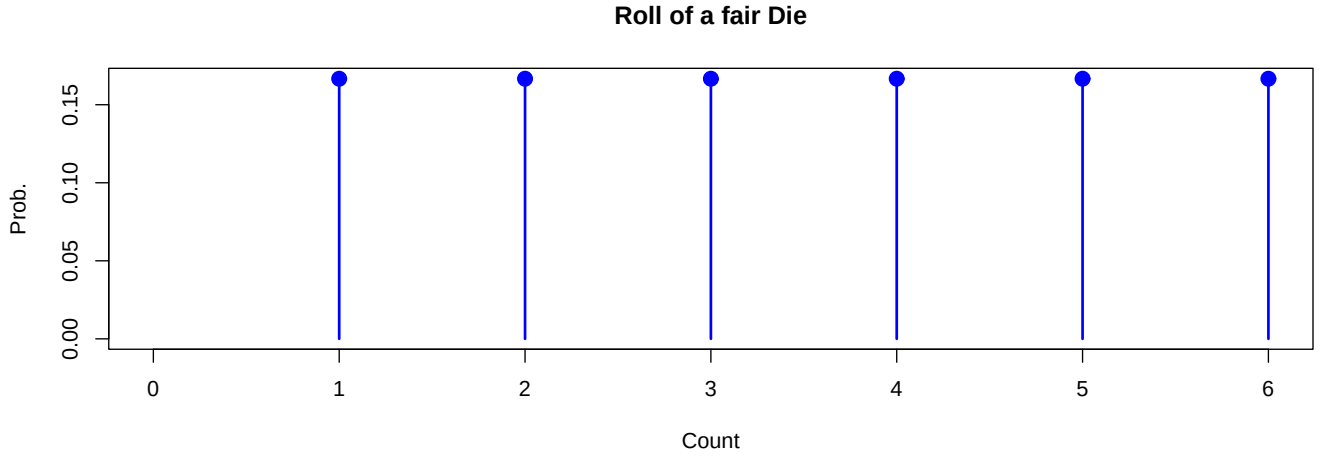
Example: Latent Uniform Sources

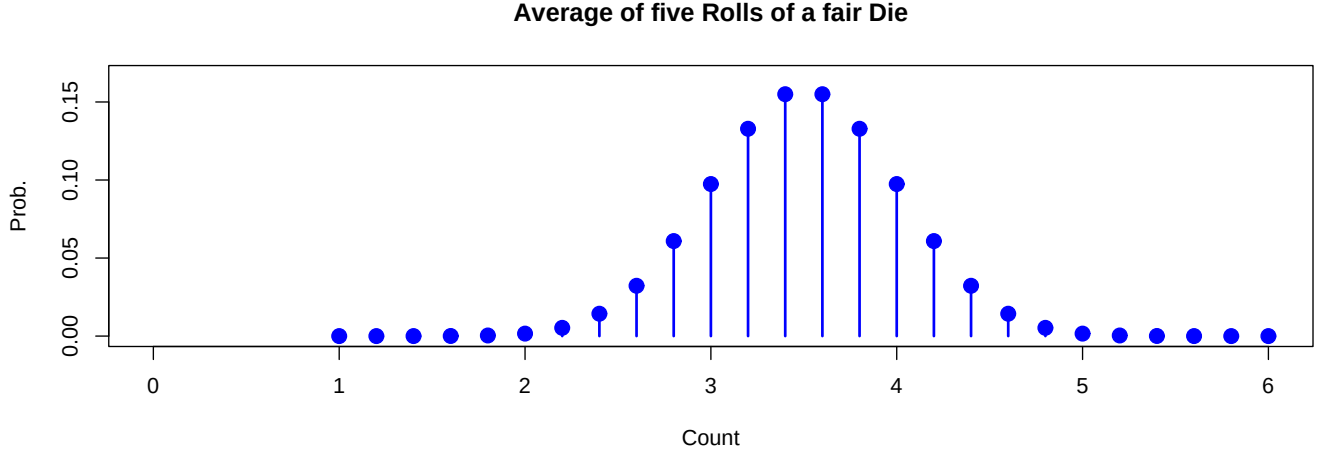


Example: Observed Mixture

The Central Limit Theorem

- Averages of random variables tend to be more Gaussian then the averaged components





The Basic Idea of ICA

- A linear combination of the x_is is necessarily also a linear combination of the s_is:

$$y = w^T x = w^T (As) \equiv z^T s$$

- We want to find a vector z for which y is as *non-Gaussian* as possible.
- There are $2n$ local solutions, equal to $\pm s_i(t)$, for $i = 1, \ldots, n$.
- If two or more of the s_i are Gaussian, then the sources *cannot* be found, since *any* linear combination of Gaussians is also Gaussian.

NonGaussianity Criteria: Kurtosis

- Kurtosis is the fourth central moment. The fourth moment is equal to 3 for Gaussians, leas than 3 for densities which peak near the center (*leptokurtic*, sub-Gaussian), and greater than 3 for densities flat near the center (*platykurtic*, super-Gaussian).
- Assume x and y have variance 1. Kurtosis is

$$\kappa(x) = \sum_i (x_i - \bar{x})^4 - 3$$

- Kurtosis is additive:

$$\kappa(c_1 x + c_2 y) = c_1^4 \kappa(x) + c_2^4 \kappa(y)$$

- Kurtosis is computationally simple, but highly sensitive to outliers.

NonGaussianity Criteria: Negentropy

- Entropy, $H(\cdot)$, measures the amount of information in a random variable; the more "random" (i.e., unpredictable), the higher the entropy:

$$H(Y) \equiv -\sum_i \Pr(Y = a_i) \log[\Pr(Y = a_i)],$$

  summed (or integrated) over all values in the support of Y.

- Among all random variables with the same variance, the Gaussian has the highest entropy.
- Let Z be Gaussian, and Y any other random variable with the same variance as Z. Negentropy is defined as follows

$$J(Y) \equiv H(Z) - H(Y)$$

- Negentropy is thus always non-negative.

Preprocessing

- Since ICA does not make use of covariances, it makes sense computationally to first center and pre-whiten the data.
- If this is done, then the mixing matrix A (analogous to the “eigentimecourses” in PCA) will be orthogonal.
- Thus, one might consider doing PCA first, then whitening (and thus discarding the covariance information on which PCA is based), and following with ICA.

Probabilistic ICA (MELODIC)

- In order to avoid overfitting, and also to be able to probabilistically rank, threshold, and select ICs, we need to extend the IC model to include noise.
- One approach is implemented in the FSL package MELODIC (Beckmann and Smith)
- The MELODIC model (i indexes voxels):

$$x_i = As_i + \mu + \eta_i$$

- x_i is $p \times 1$, A is $p \times q$, where $q < p$, μ is the mean, $\eta_i \sim \mathrm{N}(0, \sigma^2 \Sigma_i)$.
- Approximate negentropy is used to find the ICs
- A Gaussian mixture model is used to detect activation
- A Baysian approach to assign probabilities to components.

ICA vs PCA

- PCA and ICA are formally very similar: both produce a sequence of images and associated timecourses.
- PCA was developed for multivariate Gaussian data, for which uncorrelatedness implies independence.
- There is a natural ordering of principal components, based on the eigenvalues of the covariance or correlation matrix.
- PCA is often used to hopefully reduce the dimensionality of the problem, but replacing a high-dimensional dataset with a smaller number of orthogonal principal components.

ICA vs PCA (Continued)

- ICA *fails* for Gaussian data. It is an attempt to find *independent* non-Gaussian components. (In practice, one first centers and whitens the data, so that the data are uncorrelated from the beginning.)
- There is no natural ordering of independent components, so no natural way to reduce dimensionality.
- The PCA algorithm is essentially unique, based on the singular value decomposition.
- ICA is algorithm-dependent. And even repeated runs using the same algorithm can give different answers, since there are random choices made in the optimization algorithm (e.g., random starting vectors).

Granger Causality

- **Prototype Problem:** Time series for two nodes X and Y in a network, want to measure directed association $X \Rightarrow Y$.
- **Idea:**
  - Fit model. Estimate variance of forecast of Y given past of Y (σ_1^2), and forecast of Y given past of **both** X and Y (σ_2^2).
  - If $X \Rightarrow Y$, then one expects that $\sigma_2^2 < \sigma_1^2$; equivalently

$$F_{X \to Y} \equiv \log(\sigma_1^2/\sigma_2^2) > 0$$

- **Implementation:** Multivariate AR models. Reduces essentially to time series regression. Seminal work by econometrician Geweke in 1980s.

Granger Causality: Simulated Example

- Multivariate autoregressive model (here, two series):

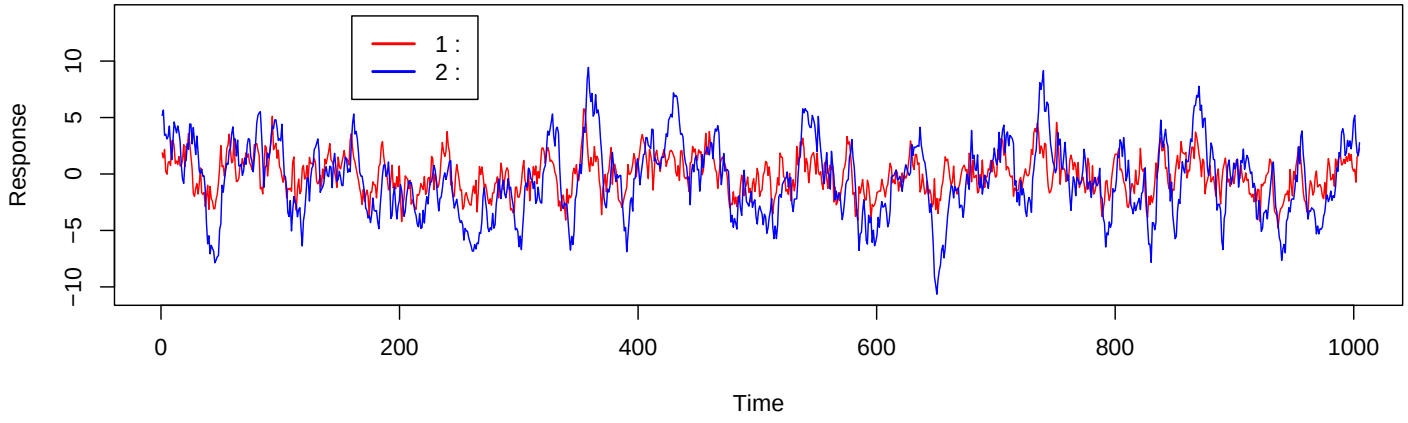
$$\begin{aligned} X_t^{(1)} &= 0.8X_{t-1}^{(1)} + 0.5X_{t-1}^{(2)} + \xi_t \\ X_t^{(2)} &= 0.8X_{t-1}^{(2)} + \eta_t \end{aligned}$$

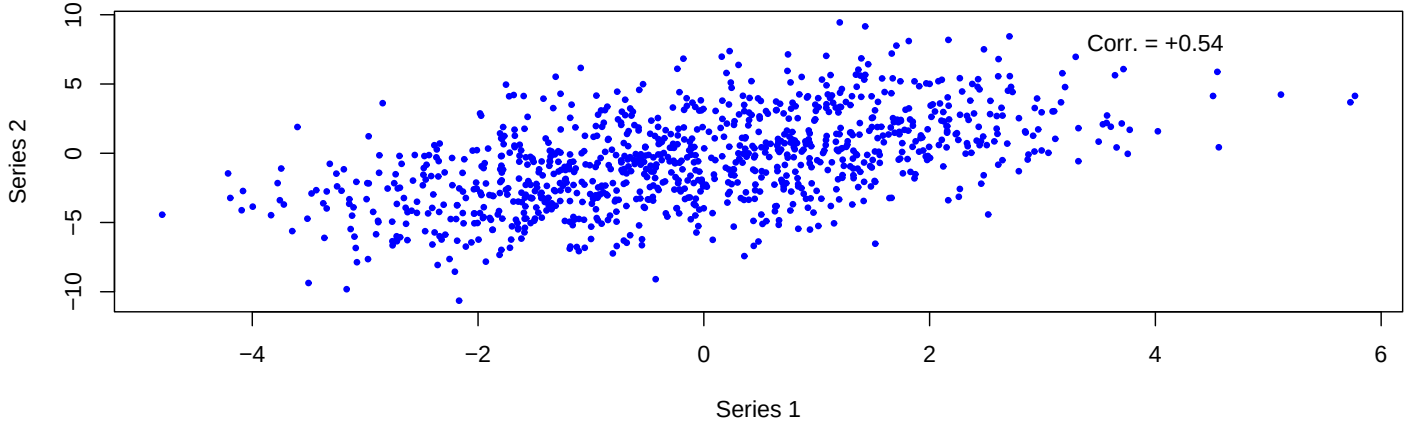
  Noise term is white, with unit variances.

- Granger causality is **defined** to be the log of the ratio of the prediction variance involving the past of only one series, to the corresponding variance using past of both series.

- We follow convention and refer to this as “causality”; Granger himself later suggested that “temporally related” would have been a better term.

Simulated Data





Model Fit and Granger Causality

- Coefficients and Coefficient Estimates:

Coef.	$a_{11} = 0.8$	$a_{12} = 0.5$	$a_{21} = 0$	$a_{22} = 0.8$
Est.	0.77	0.50	0.008	0.79

- Prediction variances (σ^2) and Granger causality (κ). Possible evidence of $2 \to 1$, but not for $1 \to 2$.

$\sigma^2_{X\|X} = 1.47$	$\sigma^2_{X\|X,Y} = 0.96$	$\kappa_{2\to 1} = 0.43$
$\sigma^2_{Y\|Y} = 1.13$	$\sigma^2_{Y\|X,Y} = 1.11$	$\kappa_{1\to 2} = 0.02$

Geweke *JASA*, (1982):

- Multivariate time series Z divided into k-dimensional component X and ℓ-dimensional component Y.
- X^-, X^+, X: Past, Past+Present, All of X
- Y^-, Y^+, Y: Past, Past+Present, All of Y
- We compare forecast variance of X_t given X^- with forecast variances which also include Y^-, Y^+, and all of Y. (Similarly for predicting Y_t).

Regressions, Forecast Variances

Dep. Variable	Covariates	Forecast Gen. Var.
X_t	X^-	$\lvert\Sigma_1\rvert$
X_t	X^-, Y^-	$\lvert\Sigma_2\rvert$
X_t	X^-, Y^+	$\lvert\Sigma_3\rvert$
X_t	X^-, Y	$\lvert\Sigma_4\rvert$
Y_t	Y^-	$\lvert T_1\rvert$
Y_t	Y^-, X^-	$\lvert T_2\rvert$
Y_t	Y^-, X^+	$\lvert T_3\rvert$
Y_t	Y^-, X	$\lvert T_4\rvert$

Regressions for X Influencing Y ($F_{X \to Y}$)

- $F_{X \to Y} = \log(|T_1|/|T_2|)$

	Past	Present	Future
X	Add		
Y	Given	*	

- $F_{X \to Y} = \log(|\Sigma_3|/|\Sigma_4|)$

	Past	Present	Future
X	Given	*	
Y	Given	Given	Add

Regressions for Y Influencing X ($F_{Y \to X}$)

- $F_{Y \to X} = \log(|\Sigma_1|/|\Sigma_2|)$

	Past	Present	Future
X	Given	*	
Y	Add		

- $F_{Y \to X} = \log(|T_3|/|T_4|)$

	Past	Present	Future
X	Given	Given	Add
Y	Given	*	

Instantaneous Influence Between X and Y ($F_{X.Y}$)

- $F_{X \cdot Y} = \log(|T_2|/|T_3|)$

	Past	Present	Future
X	Given	Add	
Y	Given	*	

- $F_{X \cdot Y} = \log(|\Sigma_2|/|\Sigma_3|)$

	Past	Present	Future
X	Given	*	
Y	Given	Add	

Overall Dependence Between X and Y ($F_{X,Y}$)

- $F_{X,Y} = \log(|T_1|/|T_4|)$

	Past	Present	Future
X	Add	Add	Add
Y	Given	*	

- $F_{X,Y} = \log(|\Sigma_1|/|\Sigma_4|)$

	Past	Present	Future
X	Given	*	
Y	Add	Add	Add

Decomposition of Dependence

- $F_{X,Y}$ was first defined by Gelfand and Yaglom (1959) as the "measure of information" between X and Y.
- Geweke (1982) decomposed $F_{X,Y}$ into a sum of three measures of linear feedback:

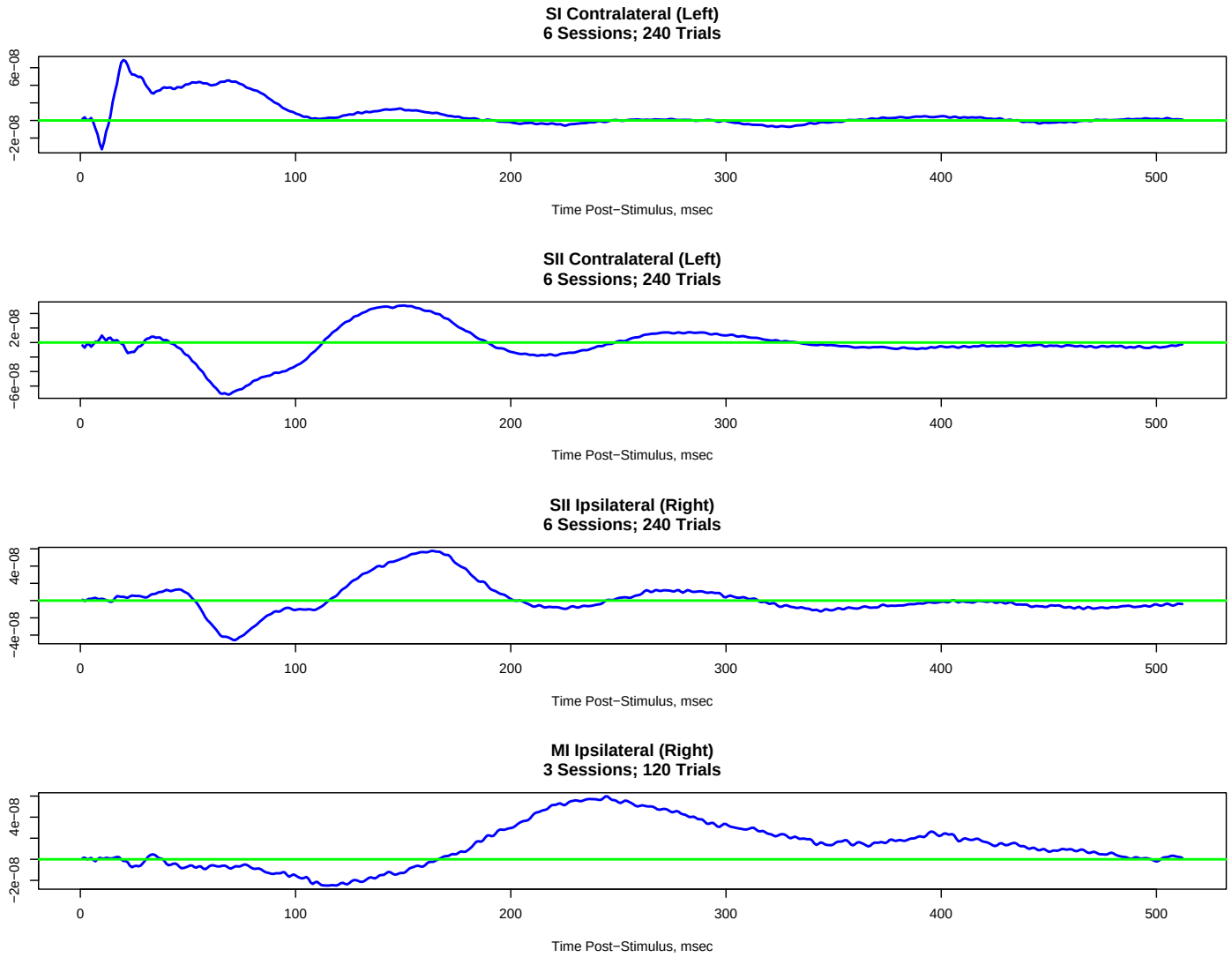
$$F_{X\to Y} + F_{Y\to X} + F_{X\cdot Y} =$$
$$\log(|\Sigma_3|/|\Sigma_4|) + \log(|\Sigma_1|/|\Sigma_2|) + \log(|\Sigma_2|/|\Sigma_3|) =$$
$$\log\left(\frac{|\Sigma_3||\Sigma_1||\Sigma_2|}{|\Sigma_4||\Sigma_2||\Sigma_3|}\right) =$$
$$\log(|\Sigma_1|/|\Sigma_4|) = F_{X,Y}$$

Hypothesis Testing

- The measures of linear feedback are each likelihood ratio statistics comparing a regression model with a nested sub-model.
- Likelihood ratio statistics are asymptotically χ^2-distributed. If k and ℓ denote the dimension of X and Y respectively, n is the number of time points, and p is the order of the autoregression, then:

$$\begin{aligned} n\hat{F}_{X\to Y} &\overset{a}{\sim} \chi^2(k\ell p) \\ n\hat{F}_{Y\to X} &\overset{a}{\sim} \chi^2(k\ell p) \\ n\hat{F}_{X\cdot Y} &\overset{a}{\sim} \chi^2(k\ell) \\ n\hat{F}_{X,Y} &\overset{a}{\sim} \chi^2[k\ell(2p+1)] \end{aligned}$$

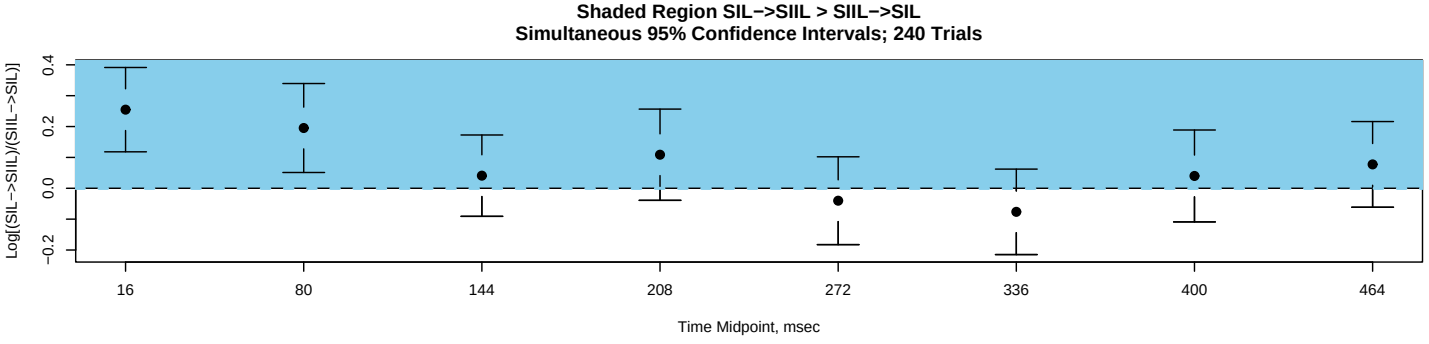
MEG Somatosensory Data

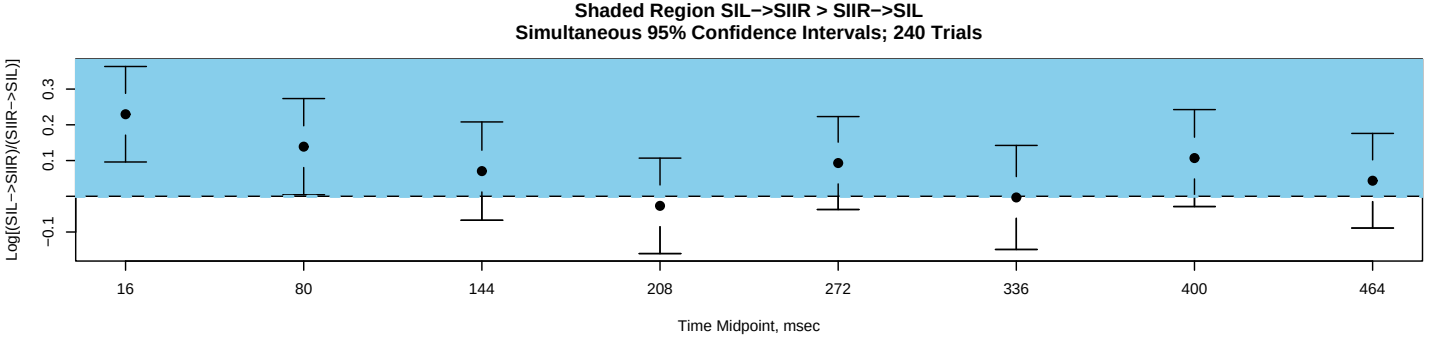


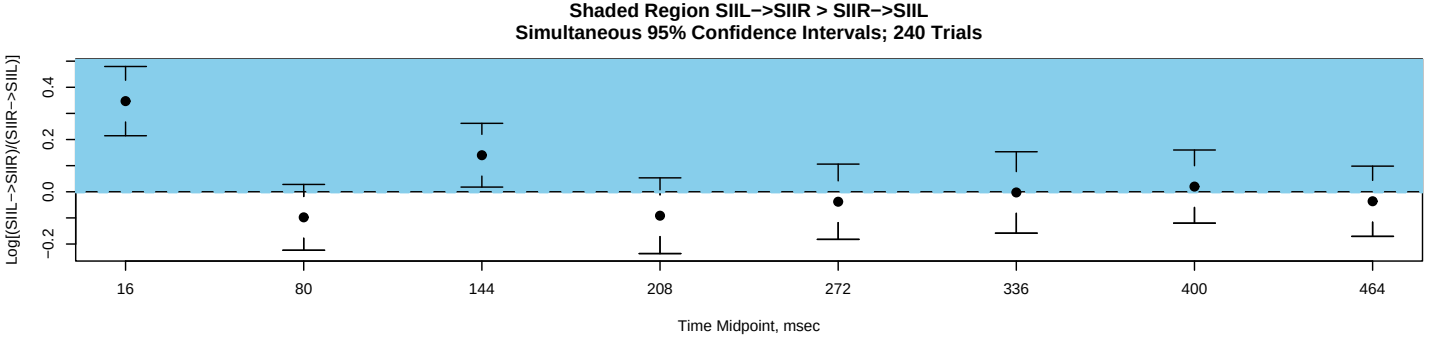
Time-Domain Analysis

- For each pair of nodes (X, Y), for each trial:
  1. Consider 512 msec post-stim; divide into eight 64 msec blocks.
  2. Fit multivariate AR model of order 5.
  3. Calculate
  $$G \equiv \log \left[\frac{F_{X \to Y}}{F_{Y \to X}} \right]$$
  4. If $X \Rightarrow Y$ is stronger (weaker) association than $Y \Rightarrow X$, then $G > 0$ $(G < 0)$.
  5. Use average and standard error *over trials* to determine simultaneous confidence intervals.

MEG Somatosensory Data







Granger Causality: fMRI Example

- Goebel R., Roebroeck A, Kim D-S, Formisano E. “Investigating directed cortical interaction in time-related fMRI data using autoregressive modeling and Granger causality mapping,” *Magnetic Resonance Imaging*, **21**, 1251-1261, 2003.
- Event related design. TR=1s. Two classes of objects are associated with L/R button presses. After some trials, the subject is cued to switch this association.
- Fit GLM to find “seed” ROI. For each ROI, create “Granger Causality Maps” over the whole brain for $F_{X\to Y}$, $F_{Y\to X}$, $F_{X\cdot Y}$, based on AR models with lag 1.
- Study appears to be proof-of-concept, claims that it is indeed possible to detect causal effects with fMRI.