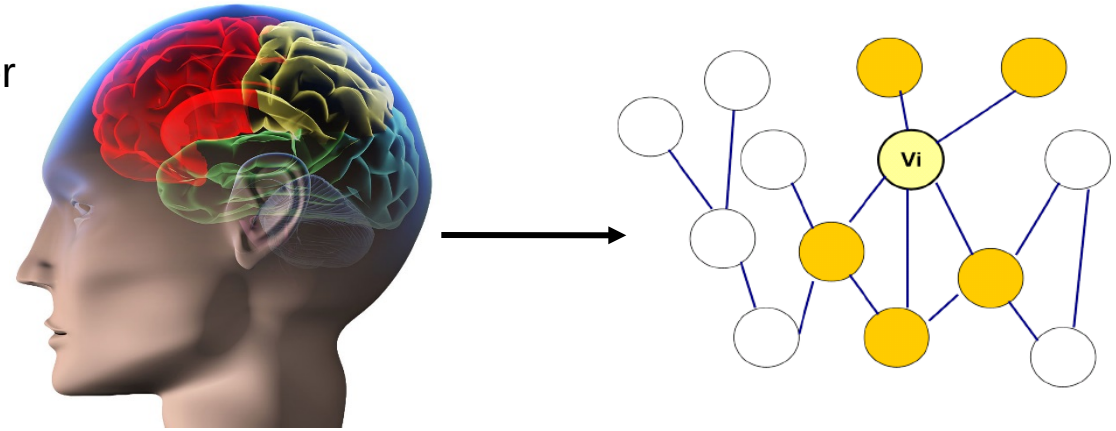
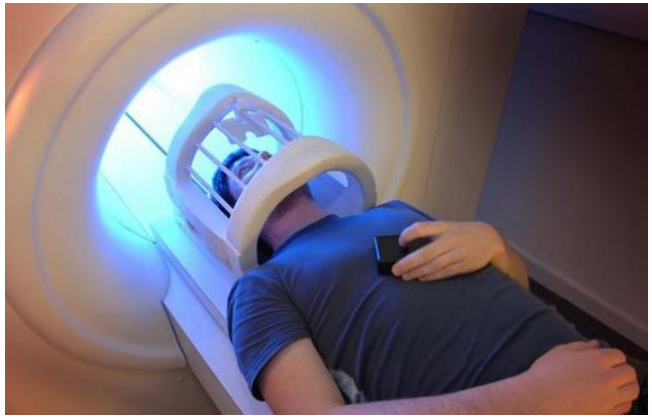
Network-Based Predictive Models

Irina Rish
IBM T.J Watson Research Center
Yorktown Heights, NY



Functional Magnetic Resonance Imaging (fMRI)



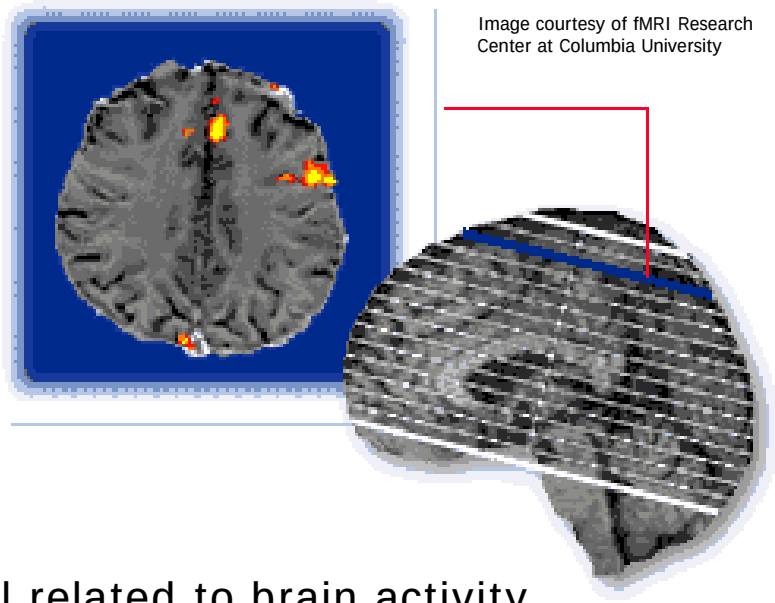


Image courtesy of fMRI Research Center at Columbia University

- Blood-oxygen-level-dependent (BOLD) signal related to brain activity while subject performs some task in scanner

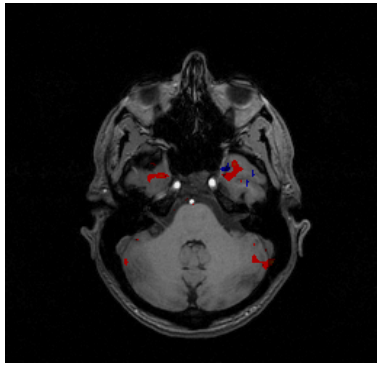
- 4D 'brain movie': a sequence of 3D brain volumes

  3D voxels ~ 3x3x3 mm, time repetitions (TR) ~1-2s

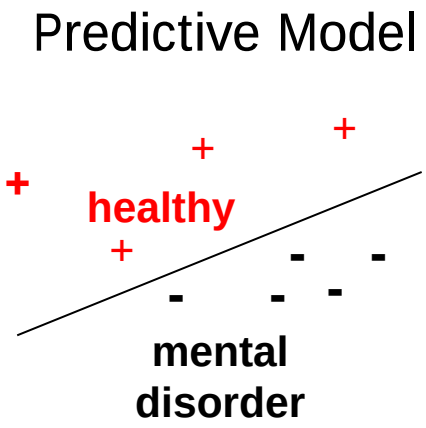
- Challenge: high-dimensional, small-sample data

  10,000 to 100,000 variables (voxels), but only 100s of TRs (samples), and less than 100 subjects

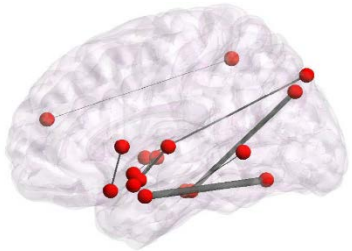
Why Network-Based Models?



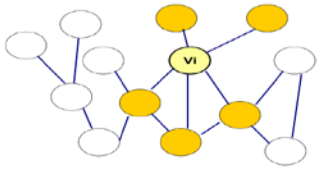
Predictive features (biomarkers)



- Brain is a network – so let's treat it like one! ☺



- Network types we consider here:
  - Functional networks (correlation matrix)
  - Markov networks (inverse covariance/precision matrix)



- Why networks?
  - rich source of predictive features!
  - discriminative information unavailable in task-based activations
  - interpretability: beyond brain areas, towards global brain functioning

Discriminative Network Models of Schizophrenia

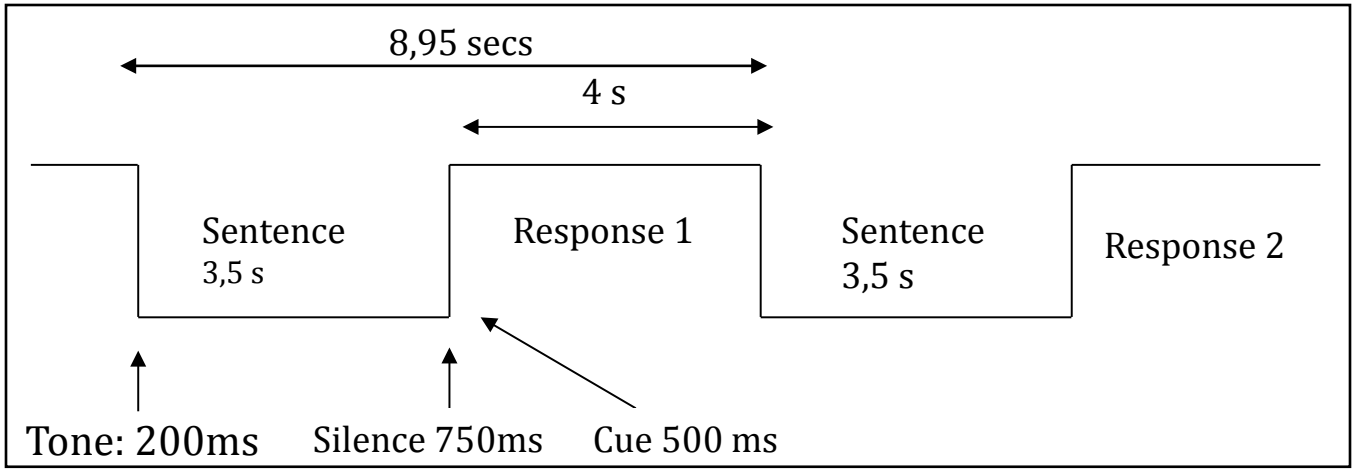
[Rish et al, PLoS One 2013]
[Cecchi et al, NIPS 2009]

- Schizophrenia is not a localized dysfunction, unlike depression, epilepsy, stroke, Parkinson's, etc.
- Hypothesized to be a disconnection syndrome [Wernicke1906; Bleuler, 1911; Friston & Frith, 1995]

Our focus: network-based 'biomarkers'

- What specific effects does schizophrenia have on functional networks as defined by fMRI?
- Are network disruptions explainable by area-specific, task-dependent linear disruptions?
- Is it possible to use functional networks to provide for consistent predictive modeling?

Experiment: Simple Auditory Task in fMRI Scanner*

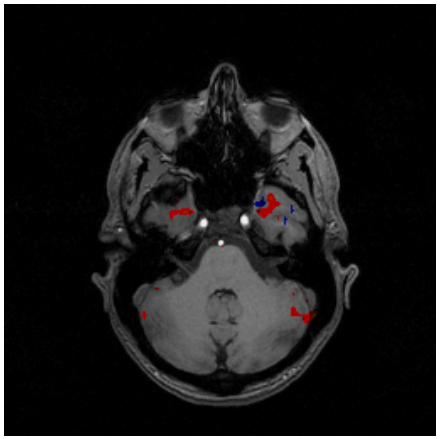


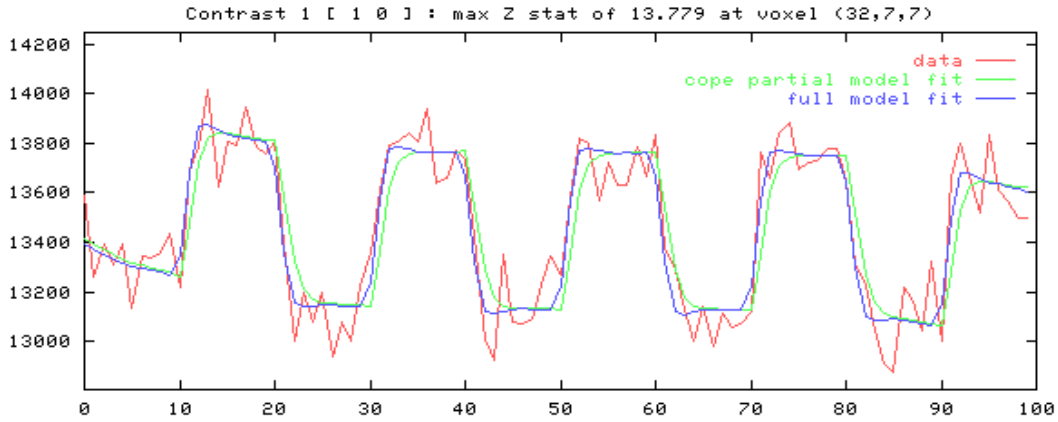
96 trials, with 32 sentences in French (native), 32 sentences in foreign languages, and 32 silence interval controls. Two runs.

- Patient Group (11 subjects)
  - Prone to auditory hallucinations
  - Native French speakers, right-handed, 3+ yrs. illness
- Normal Group (11 subjects)

*M. Plaze, et al., *Schizophrenia Research* (2006)

Standard Approach: Univariate, Task-Related Activations





fMRI activation image and time-course
courtesy of Steve Smith, FMRIB

- For each voxel, compute a score (e.g., correlation, or GLM coefficient) reflecting how well its activity matches the stimulus sequence
- Threshold the scores to select only statistically significant ones

However, no statistically significant differences were found across groups; also, classification based on activation features was not very accurate*

*G. Cecchi, et al., *Neural Information Processing Systems (NIPS-2009)*

ROI Analysis: Inter-Region Correlations

Linear activation identification

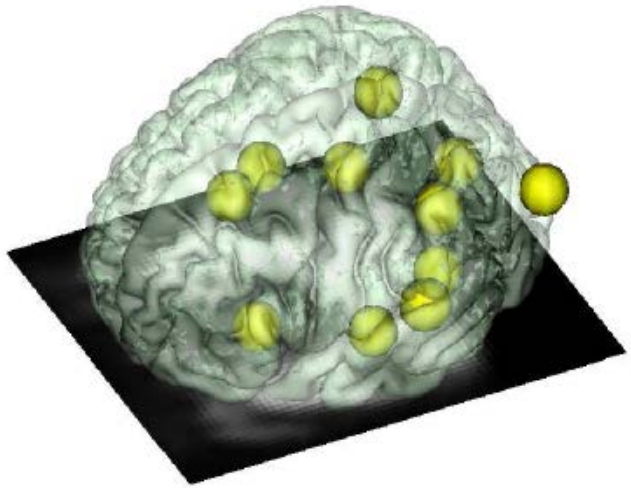
↓

Template-based ROI network identification

↓

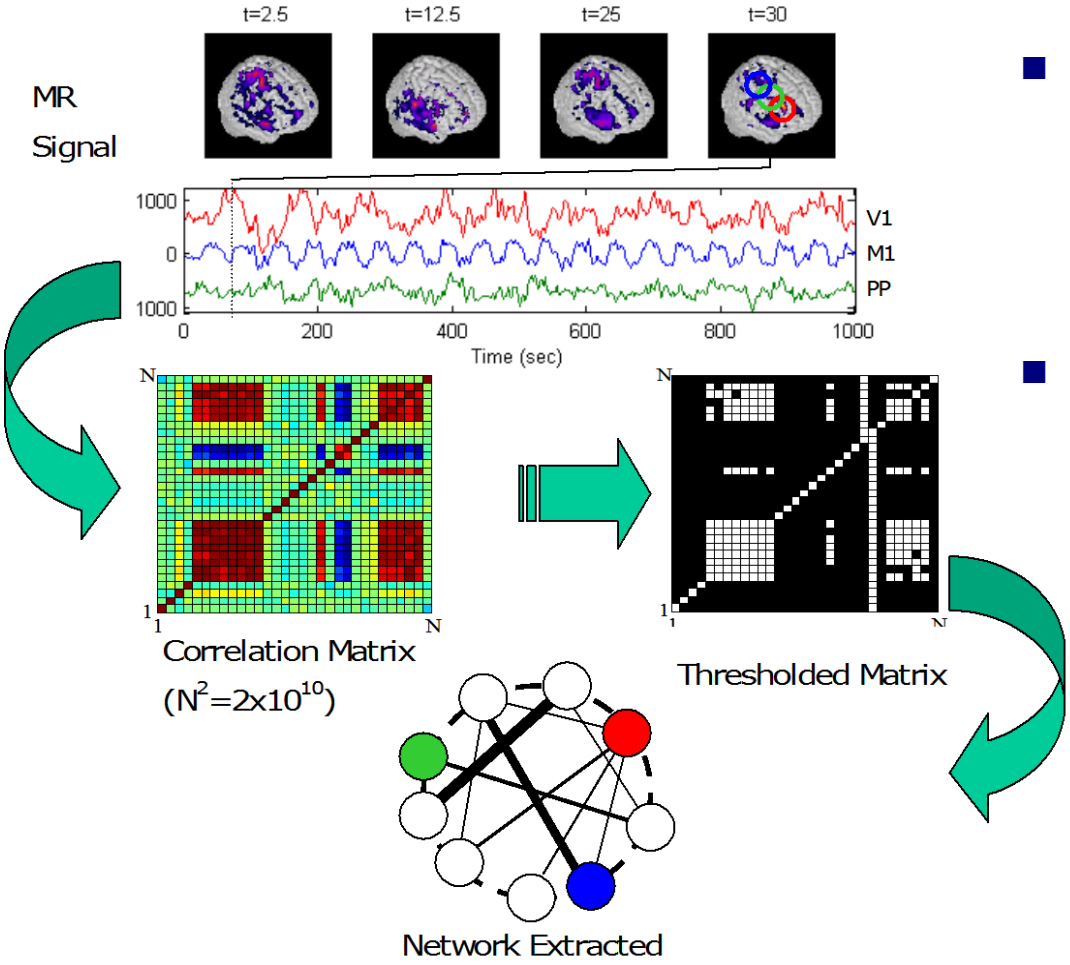
Group discrimination

	ROI name	(x,y,z) position	Anatomical position
1	'Temporal_mid_L'	-44,-48,4	Left temporal
2	'Temporal_mid_et_sup_L'	-56,-36,0	Middle and superior left temporal
3	'Frontal_inf_L'	-40,28,0	Left Inferior frontal
4	'cuneus_L'	-12,-72,24	Left cuneus
5	'Temporal_sup_et_mid_L'	-52,-16,-8	Middle and superior left temporal
6	'Angular_L'	-44,-48,32	Left angular gyrus
7	'Temporal_sup_R'	40,-64,24	Right superior temporal
8	'Angular_R'	40,-64,24	Right angular gyrus
9	'Cingulum_post_R'	4,-32,24	Right posterior cingulum
10	'ACC'	0,20,30	Anterior cingulated cortex



However, no statistically significant links were identified

Functional Networks: Voxel-Based Correlations



- Network link (i,j) ⇔ correlation between BOLD(i) and BOLD(j) is above a threshold (e.g.,0.7)
- Degree maps:

  degree(voxel i) = number of its neighbors in the network

Variety of degree maps:

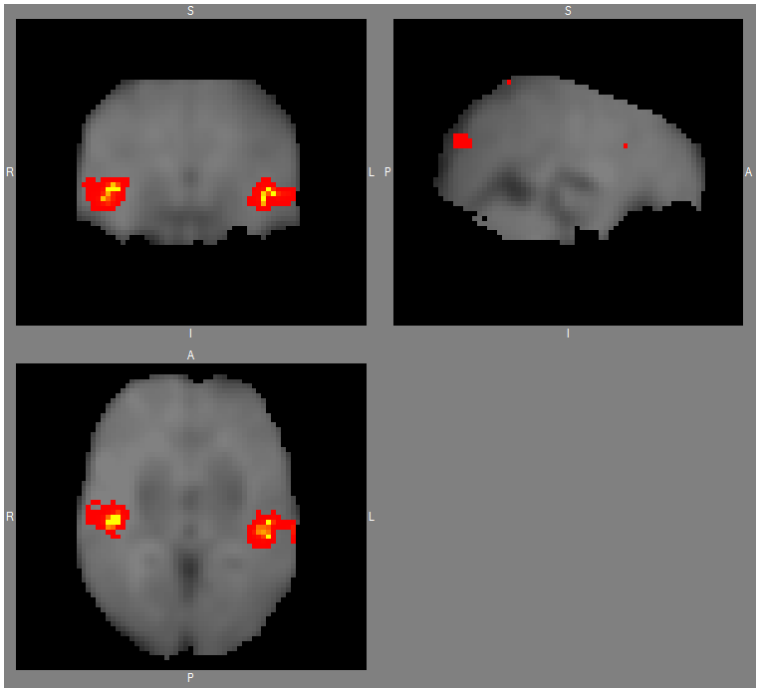
- Full degree maps
- Long-distance degree maps – non-local connections (> 5 voxels apart)
- inter-hemispheric degree maps – only links between the hemispheres

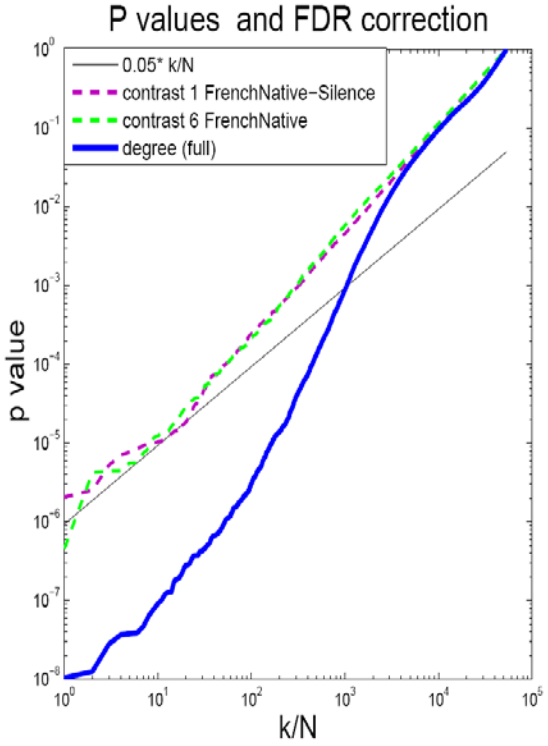
Degree Maps Reveal a Distinctive Pattern (Unlike Activation Maps)

- Degree maps show a clear pattern even after FDR correction for multiple comparisons
  - abnormal degree distribution in correlation networks for
  - schizophrenic patients such as lack of “hubs” in auditory cortex

- However, activation maps practically do not survive the correction!

FDR-corrected
(Normalized) Degree Maps





2-sample t-test performed for each voxel in degree and activation maps

Red/yellow: Normal subjects have *higher* values than Schizophrenics

Bonferroni (too strict):
- degree maps: 50 voxels
- activation maps: 0-1 voxels

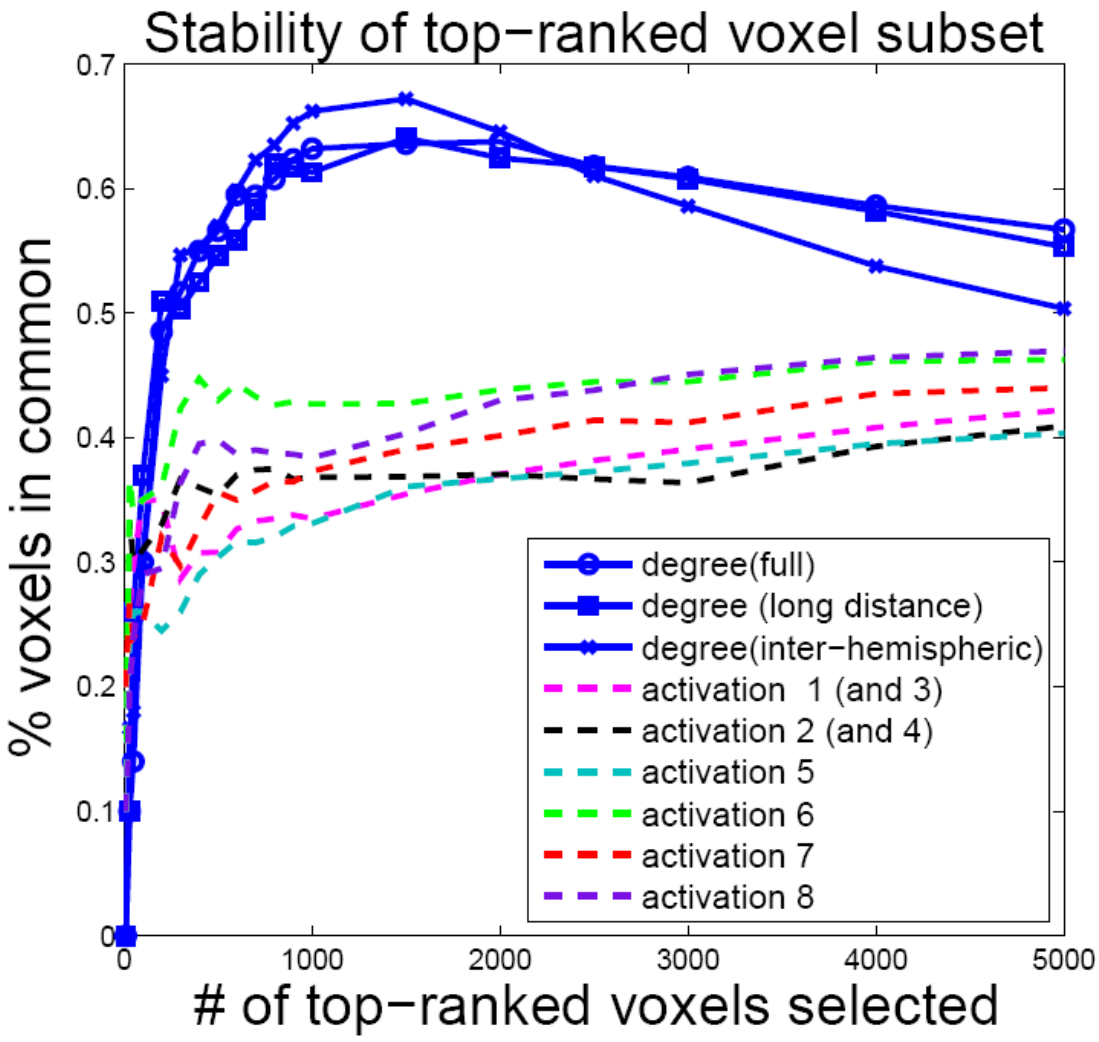
False-Discovery Rate (FDR):
- degree maps: 1033 voxels
- activation maps: 0-7 voxels

Predictive Modeling with Topological vs Activation Features

- 44 samples: (11 schizophrenics, 11 normals) x 2 runs
- 22-fold cross-validation
  - leave-one-subject-out (i.e., 2 runs for each subject),
    NOT leave-one-sample-out (samples are not i.i.d., subjects are)
- Classifiers:
  - Linear SVM
  - Naïve Bayes
  - Markov Networks (Markov Random Fields)
- Features: voxel activations vs. topological features (voxel degrees and 'global' topological features, e.g., mean degree, mean geodesic distance, giant component size, etc.)
- Which features – topological or activation – are more informative? More stable?

  Which ones allow to generalize better to unseen samples, as opposed to the standard hypothesis-testing framework?

Stability: Degrees Are More Stable Features than Activations

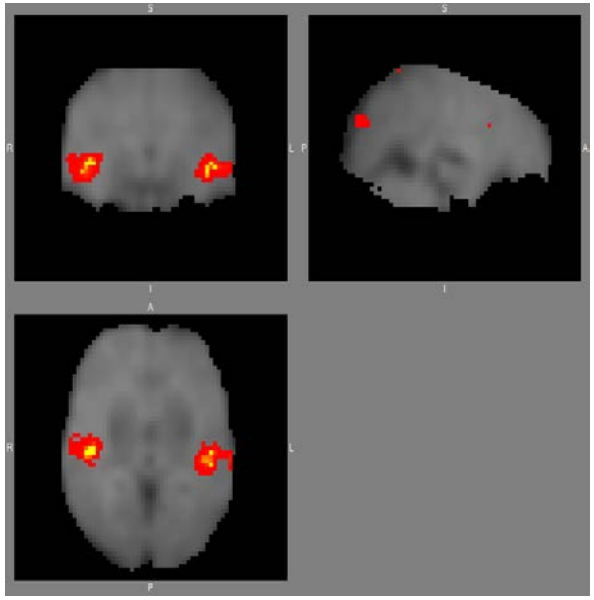


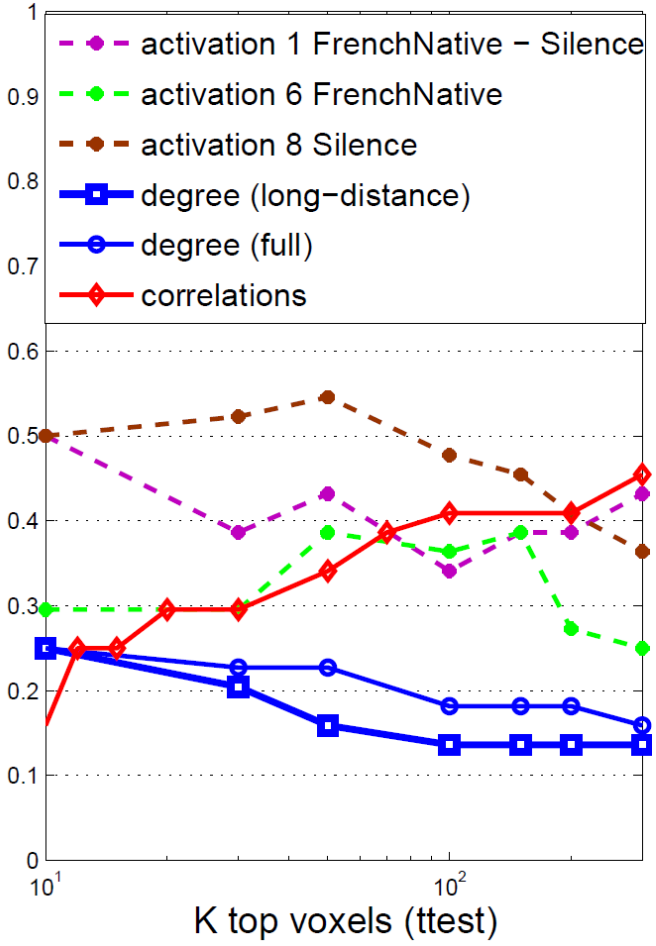
- When selecting top-K most significant voxels over data subsets in leave-subject-out cross-validation, degree maps yield higher overlap (~70% common voxels), unlike activation maps

Classification: Degrees Greatly Outperform Activations

[Cecchi et al, NIPS 2009]

Voxel degree features +
Sparse Markov Random Field*
Classifier = 86% accuracy





Statistically significant degree features (after FDR correction)

Areas: BA 22 and BA 21(auditory/temporal areas, possibly related to the auditory task)

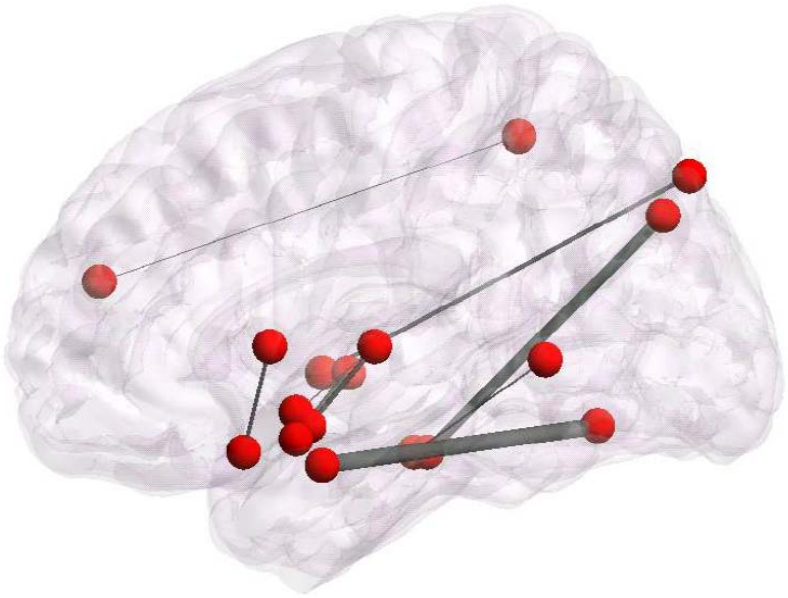
*Sparse MRF classifier is a probabilistic classifier that learns a sparse MRF model for each class, then assigns test sample to the most likely class

More on sparse MRFs – later in this talk

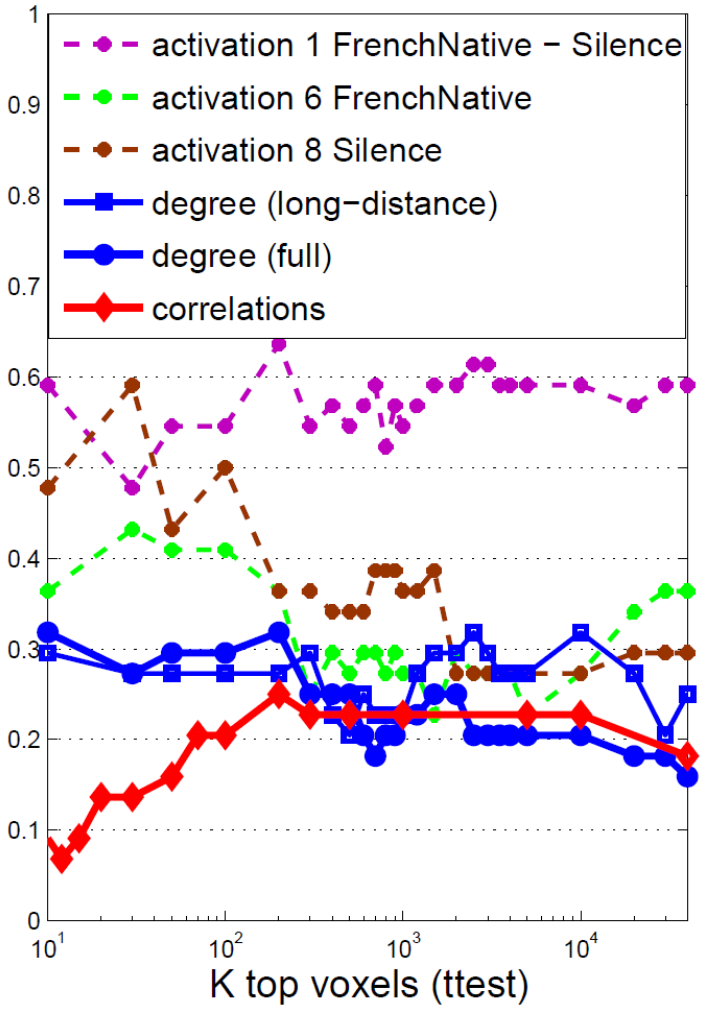
Even Better Results with Link Weights (Pairwise Correlations)

[Rish et al, PLOS One 2013]

Top cross-voxel correlations (link 'weights') + SVM classifier = 93% accuracy



Top statistically significant correlations: cerebellum (declive) and the occipital cortex (BA 19)



Network Features as Schizophrenia Markers

- Voxel degrees:
  degree(voxel i) = number of its neighbors in functional network

Showed significant statistical differences across two groups, and 86% accuracy in schizophrenia classification

- Correlations: link weights

Just a dozen of correlation features allows for even better 93% accuracy

- Voxel strength: sum of weights of links connected to the node. Absolute strength ⇔sum of abs. weights. Positive strength ⇔ sum of pos. only weights.

Variety of other network features, beyond degrees, show significant statistical differences and are predictive (though less than link weights and degrees) about schizophrenia [PLOS One'12]

- Clustering coefficient of a voxel: the fraction of node's neighbors that are neighbors of each other
- Local efficiency: global efficiency computed on voxel's neighborhoods. Global efficiency: average inverse shortest path length

Functional networks contain large amounts of schizophrenia-related information that may not be present in task-based activations

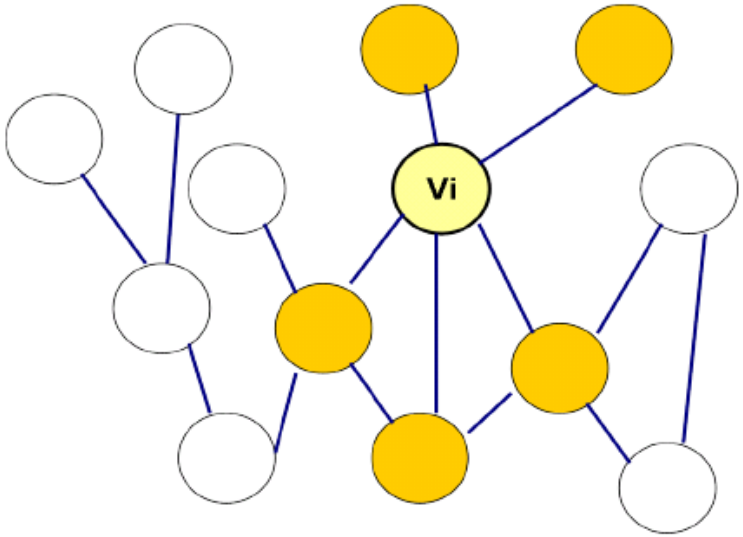
Part 2: Markov Networks (Markov Random Fields)

MRFs are probabilistic models that capture joint distribution over the voxels, rather then just pairwise correlations as in functional networks

Given joint P(X), one can make probabilistic queries: given activity in brain area A, what happens to brain area B, etc.

$$X = \{X_1, ..., X_p\}, \quad G = (V, E)$$

$$P(\boldsymbol{X}) = \frac{1}{Z} \prod_{C \in Cliques} \Phi_C(\boldsymbol{X}_C)$$

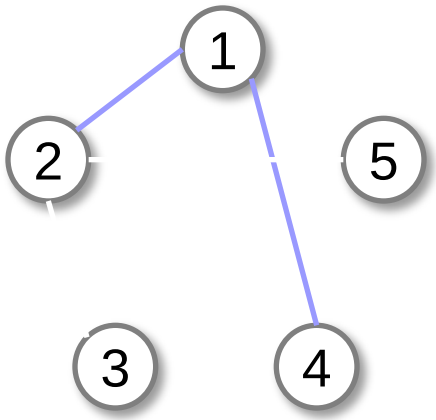


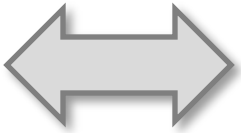
Lack of edge $(i, j) \rightarrow$
conditional independence $X_i \perp X_j | rest$

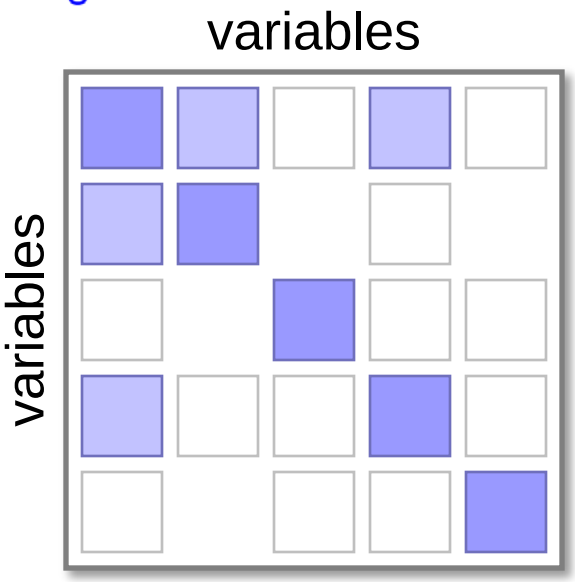
Gausian Markov Random Fields (GMRFs)

Markov random field of jointly Gaussian variables

- $P(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$
- Σ - covariance matrix, Σ^{-1} - precision (concentration) matrix
- Zeros in Σ: marginal independence
- Zeros in $\Sigma^{-1} \Leftrightarrow$ conditional independence $\Leftrightarrow$ lack of edge







Sparse $\Sigma^{-1} \Leftrightarrow$ sparse Markov network

Precision matrix
(Inverse covariance)

Maximum Likelihood Estimation of the Inverse Covariance Matrix

Assume the data $\mathbf{X}$ are centered to have zero mean. Then:

$$\hat{\Sigma}^{-1} = \arg\max_{C \succ 0} \log p(C|\mathbf{X}) = \arg\max_{C \succ 0} \log p(\mathbf{X}, C) =$$

$$\boxed{= \arg\max_{C \succ 0} \log\det(C) - \mathrm{tr}(SC)}$$

where $S = \frac{1}{N}\sum_{i=1}^{N} x_i^T x_i$ is the empirical covariance matrix (MLE of Σ)

Why not just use $\hat{\Sigma}^{-1} = S^{-1}$?

- in small-sample case ($n < p$), S may not be even invertible
- even if it is, S^{-1} almost never contains exact zeros
- needed: an explicit sparsity-enforcing constraint (regularization)

l_1-Regularized Maximum Likelihood Problem

Primal:

$$\hat{\Sigma}^{-1} = \arg\max_{C \succ 0} \log\det(C) - \mathrm{tr}(SC) - \lambda ||C||_1 \qquad (1)$$

Convex problem; unique optimum for any $\lambda > 0$ (Banerjee et al., 2008)

Various algorithms:

- Approximation: LASSO for each node (Meinshausen&Buhlman, 2006)
- Block-coordinate descent: COVSEL (Banerjee et al, 2006), glasso (Friedman et al, 2007)
- Projected gradient (Duchi et al, 2008)
- Greedy ascent SINCO (Scheinberg and Rish, 2010)
- Alternating Linearization Method (Scheinberg et al, 2011)
- several more recent efficient techniques available

Various sparse structure, besides basic edge-sparsity:

- diagonal structure (Levina et al., 2008)
- block structure for known block-variable assignments (Duchi et al., 2008)
- unknown block-variable assignments (Marlin & Murphy, 2009; Marlin et al., 2009)
- spatial coherence (Honorio et al, 2009)
- common structure among multiple tasks (Honorio et al, 2010)

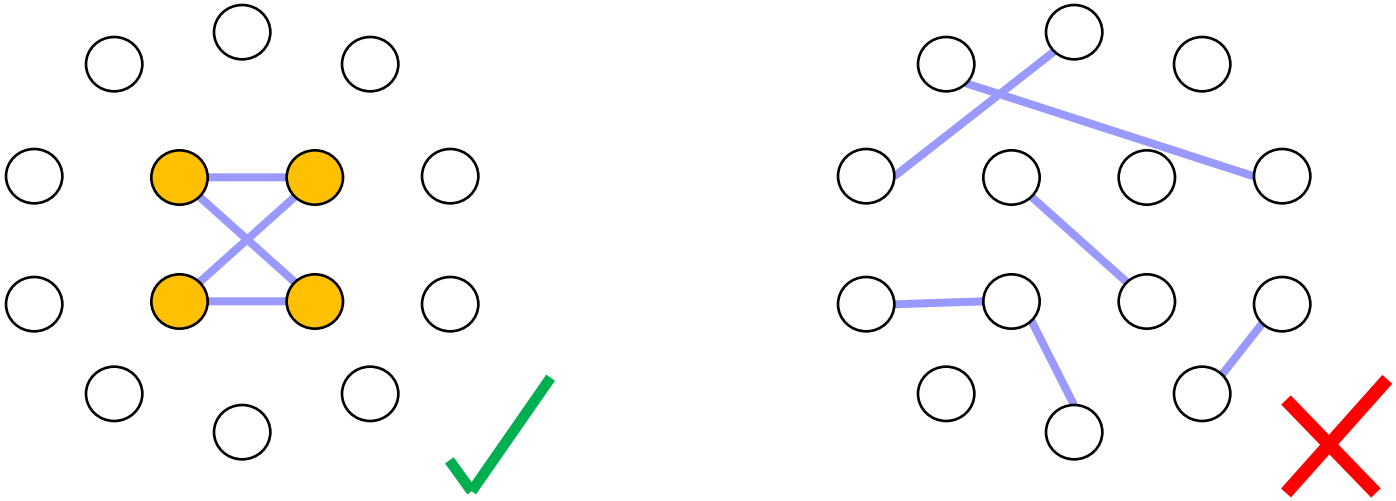
Variable Selection in Gaussian MRFs

[Honorio et al, AISTATS 2012]

- Datasets with thousands of variables:

  fMRI, gene expression, stock prices, world weather

- Hypothesis: often, only a relatively few variables are interacting with each other, forming network clusters; the rest are not relevant



- Goal: select these *important* nodes, and find their interaction pattern

Variable-Selection Regularizer: Block-Sparsity over Node's Neighbors

- An additional (last) term to encourage varia

A proxy for the number of connected nodes ➔ NP-hard learning

$$\max_{C \succ \mathbf{0}} \underbrace{\log\det C - tr(SC)}_{\text{log-likelihood of the dataset}} - \underbrace{\lambda||C||_1}_{\text{sparseness prior}} - \underbrace{\tau||C||_{1,p}}_{\text{our variable-selection prior}}$$

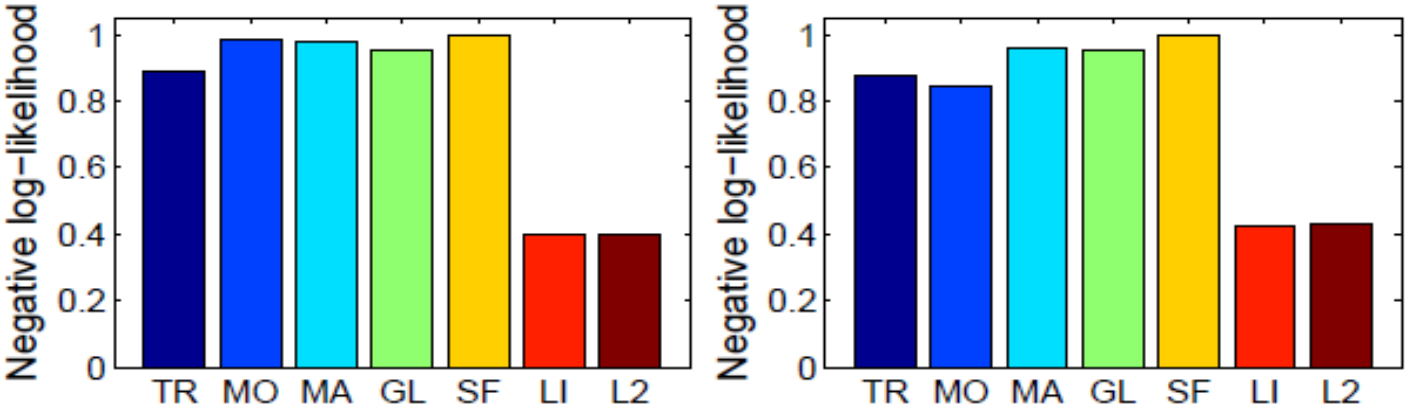
- Variable-selection prior: block l1/lp norm, for $p \in \{2, \infty\}$

$$||C||_{1,p} = \sum ||c_{n,1}, \ldots, c_{n,n-1}, c_{n,n+1}, \ldots, c_{n,N}||_p$$

- We use Block-Coordinate Descent (BCD) on the primal (not dual!): a *sequence of quadratic subproblems with closed form solutions, see* (Honorio et al, AISTATS 2012)

Example: Cocaine Addiction fMRI Data

- fMRI dataset previously collected by (Goldstein et al, 2007)
  - 15 cocaine addicted subjects and 11 control subjects
  - 87 scans/TRs (3.5 s), 53x63x46 voxels
- Subsampling to reduce dimensionality: 4x4x4 voxel cubes ⇔869 nodes
- Task: visual attention, with monetary reward (see [Honorio 2012])

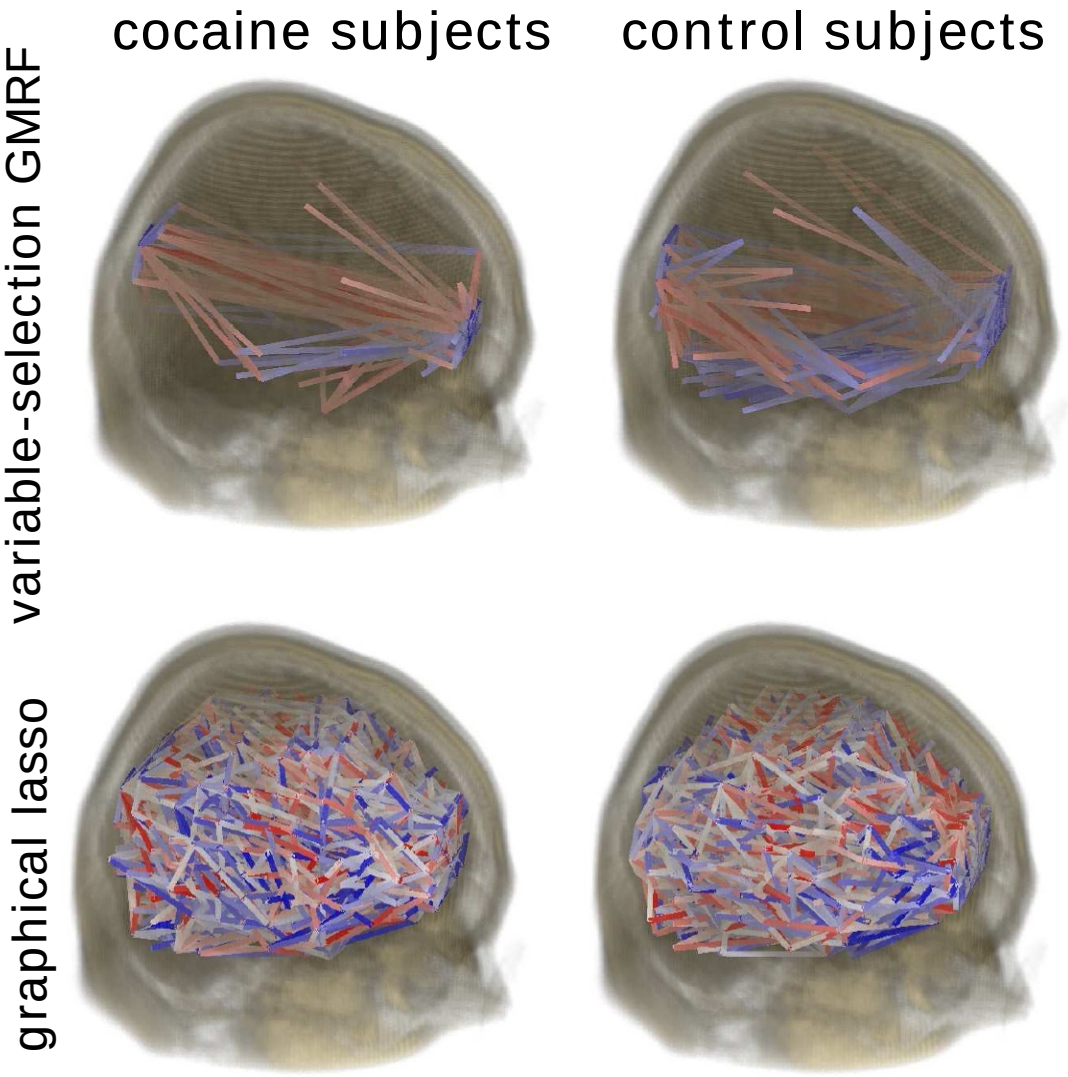


addicted subjects control subjects

Variable-selection GMRF (LI,L2) learn higher-likelihood models that its competitors including standard graphical lasso (GL), Meinshausen-Buhlmann (MO,MA) approach, scale-free networks (SF) and Tikhonov regularization (TR).

Variable-selection assumption seem to fit the data better than standard sparse GMRFs.

(Much) Better Interpretability Than Standard Sparse Gaussian MRFs



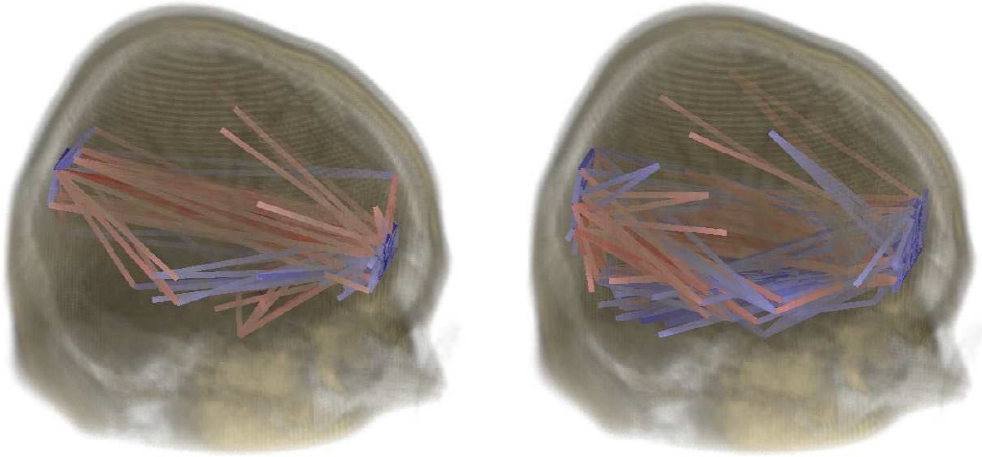
Blue - positive interactions
red - negative interactions

The variable-selection sparse GMRF approach learns networks that involve fewer connected variables (~***50 connected nodes***) and have higher log-likelihood than the standard edge-only sparse GMRF learning, 'graphical lasso'.

When performing classification of cocaine vs. control by using GMRFs, all methods obtain **84.6%** leave-one-subject-out accuracy

Discussion

cocaine subjects control subjects



Blue - positive interactions
red - negative interactions

In cocaine addicts, as compared to controls, we observe

- increased interactions between the visual cortex (left) and the prefrontal cortex (right)
- decreased density of interactions between the visual cortex with other brain areas

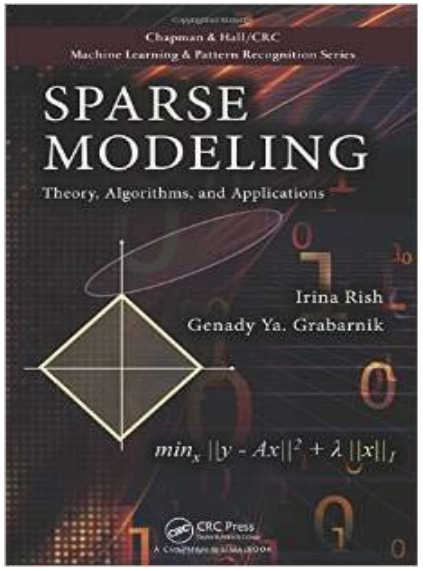
Note that the trigger for reward was a visual stimulus and that abnormalities in the visual cortex was reported in (Lee et al, 2003) when comparing cocaine abusers to control subjects

Also, prefrontal cortex is involved in decision making and reward processing, and abnormal monetary processing in the prefrontal cortex was reported in (Goldstein et al, 2009) when comparing cocaine addicted individuals to controls.

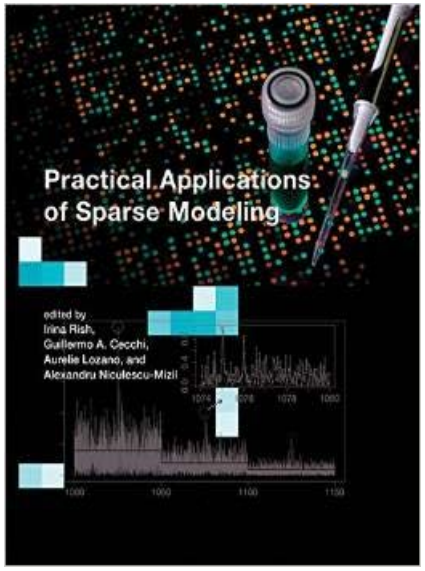
Summary

- Two types of networks: functional (covariance) and Markov (inverse covariance)
- Network-based features can be very informative even when GLM-based voxel activations are not (schizophrenia study)
- Markov networks (MRFs) , unlike functional, summarize joint probability distribution over the whole brain. We focus here on sparse Gaussian MRFs
- Variable-selection prior produces much more interpretable GMRFs
- To learn more about sparse models and their applications to functional MRI, see

Rish and Grabarnik,
Sparse Modeling: Theory, Algorithms, and Applications



Rish et al.
Practical Applications of Sparse Modeling



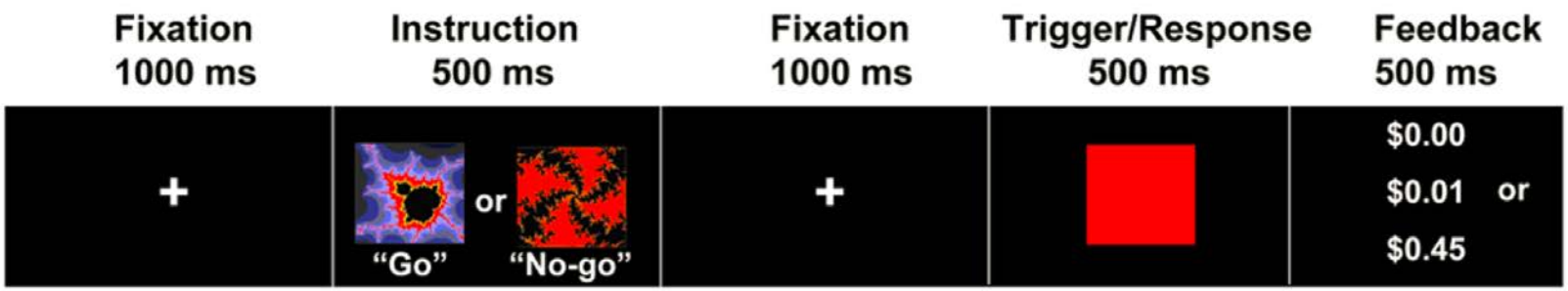
Extra stuff

Cocaine Addiction fMRI Data

- fMRI dataset previously collected by (Goldstein et al, 2007)
  - 15 cocaine addicted subjects and 11 control subjects
  - 87 scans/TRs (3.5 s), 53x63x46 voxels
- Subsampling to reduce dimensionality: 4x4x4 voxel cubes ⇔869 nodes
- Task: visual attention, with monetary reward

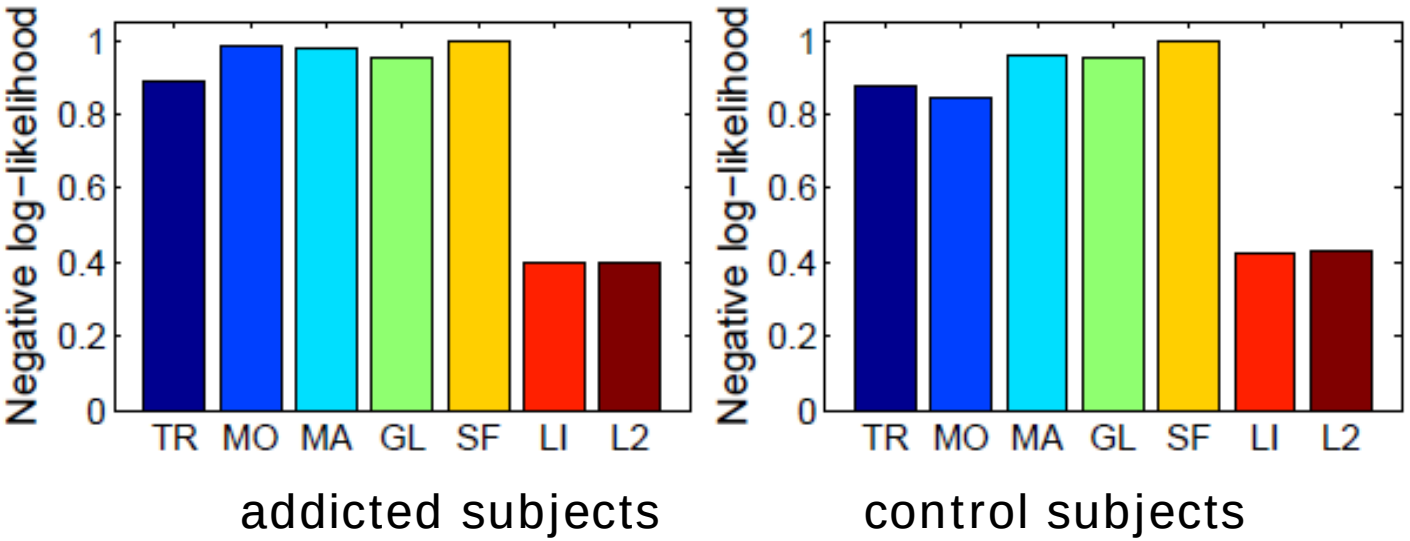
Instructions were to press a response button (using the thumb of the dominant hand) with speed and accuracy upon seeing the target (red square) after a "Go" but not after a "No-go" instruction stimulus.

B. Each 3.5-sec trial (81 go and 81 no-go trials in each block):



Results: Better Model Fit (Better Likelihood)

- 1/3 of the data was used for training, 1/3 for validation and 1/3 for testing
- Results comparing negative log-likelihood (lower ⇔ better) of networks learned using various methods:



Our methods (LI,L2) outperform competitors, e.g. Meinshausen-Buhlmann (MO,MA), graphical lasso (GL), scale-free networks (SF) and Tikhonov regularization (TR).

Variable-selection assumption seem to fit the data better than standard sparse GMRFs.