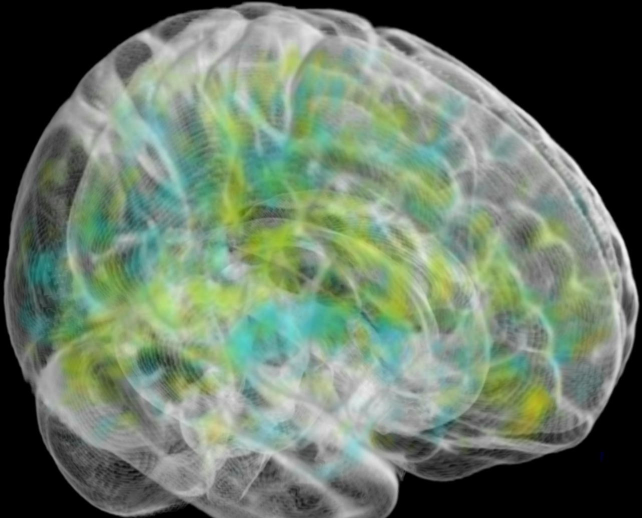
Tales from fMRI

Learning from limited labeled data

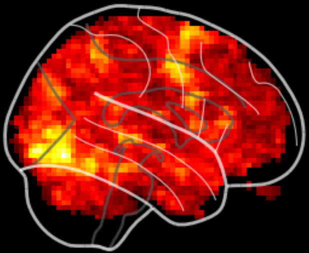
Gaël Varoquaux

Inria

PARIETAL



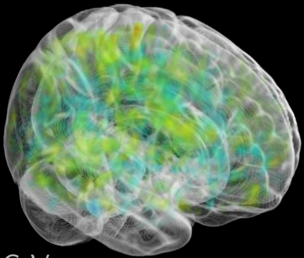
fMRI data



- $p \sim 100\,000$ voxels per map
- Heavily correlated $+$ structured noise
- Low SNR: $\sim 5\%$ $\quad \sim -13\,\text{dB}$

Brain response maps (activation)

- $n \sim$ Hundreds, maybe thousands

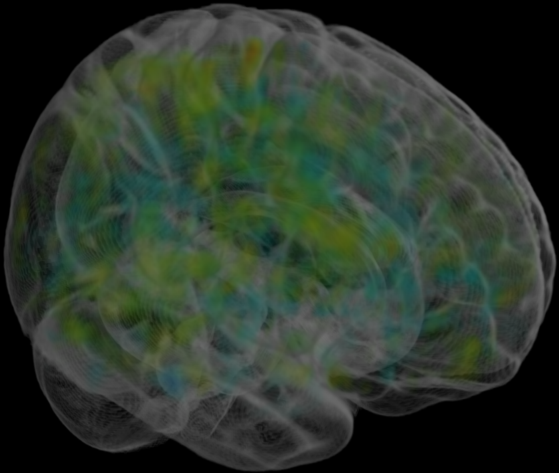


Resting-state (no cognitive labels)

- $n \sim$ 100–10 000 per subject
- Thousands of subjects
- No salient structure

TL;DR

- Estimators with small sample complexity
- Increasing the amount of data

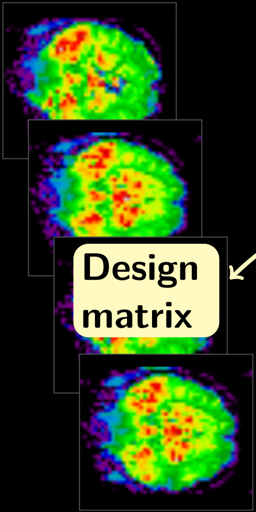


Outline of this talk

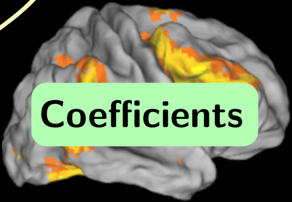
1. **Regularizing linear models**
2. **Covariance estimation**
3. **Merging data sources**

1 Regularizing linear models

$$\mathrm{sign}(\mathbf{X}\,\mathbf{w} + \mathbf{e}) = \mathbf{y}$$



$\times$

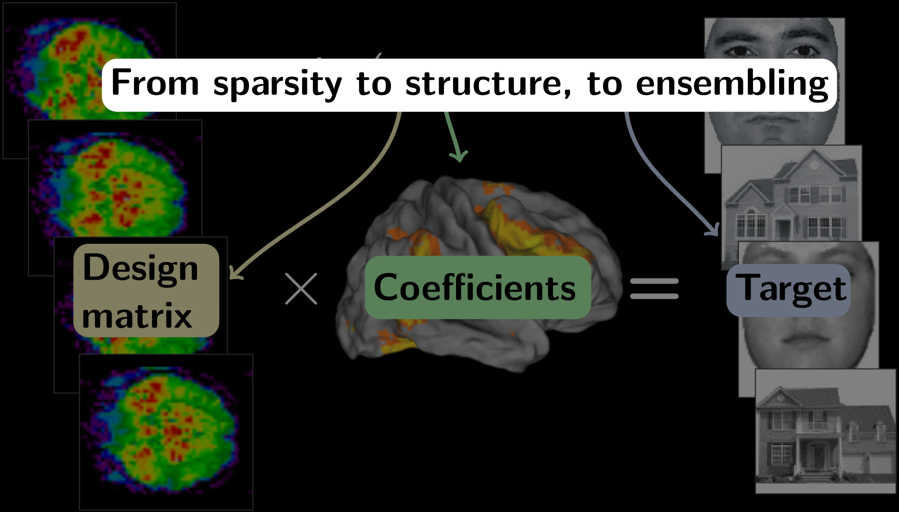


$=$

p $\sim$ **50 000**

n $\sim$ **100 per category**

1 Regularizing linear models



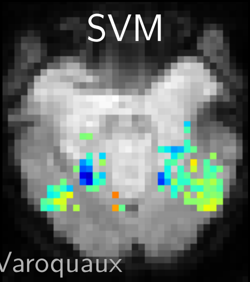
1 Sample complexity, ℓ_1 versus ℓ_2 regularization

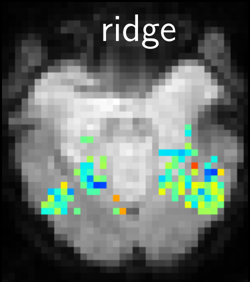
Def Sample complexity: n required for small error *w.h.p.*

Rotationally invariant estimators are data hungry

Thm For rotational invariant estimators,

$$\text{sample complexity} > \mathcal{O}(p)$$ [Ng 2004]





1 Sample complexity, ℓ_1 versus ℓ_2 regularization

Def Sample complexity: n required for small error *w.h.p.*

Rotationally invariant estimators are data hungry

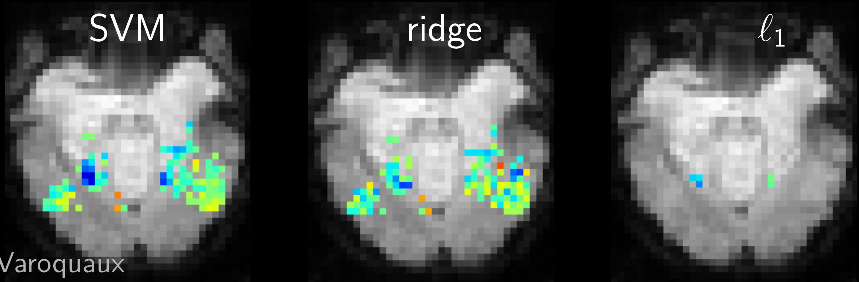
Thm For rotational invariant estimators,

sample complexity $> \mathcal{O}(p)$ [Ng 2004]

Sparsity, compressive sensing

To recover k non-zero coefficients, $n \sim k \log p$

[Wainwright 2009]



1 Sample complexity, ℓ_1 versus ℓ_2 regularization

Def Sample complexity: n required for small error *w.h.p.*

Rotationally invariant estimators are data hungry

Thm For rotational invariant estimators,

sample complexity $> \mathcal{O}(p)$ [Ng 2004]

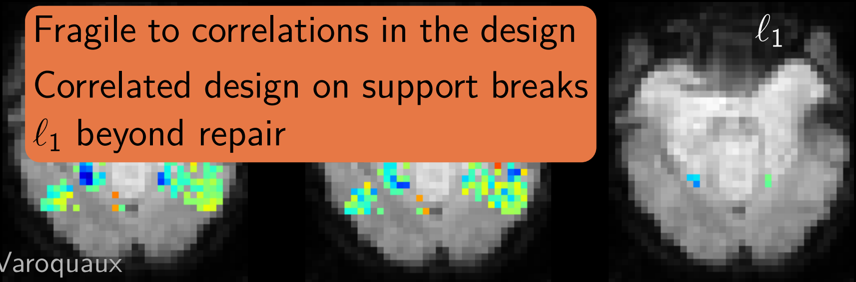
Sparsity, compressive sensing

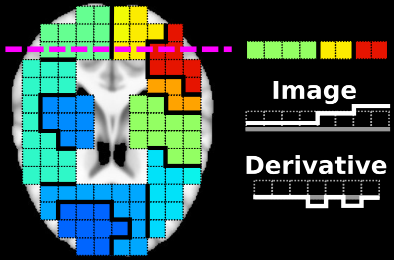
To recover k non-zero coefficients, $n \sim k \log p$

[Wainwright 2009]

Fragile to correlations in the design

Correlated design on support breaks ℓ_1 beyond repair



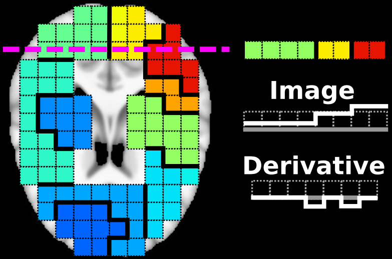


Total-variation penalization

Impose sparsity on the gradient of the image:

$$p(\mathbf{w}) = \ell_1(\nabla \mathbf{w})$$

In fMRI: [Michel... 2011]



Total-variation penalization

Impose sparsity on the gradient of the image:

$$p(\mathbf{w}) = \ell_1(\nabla \mathbf{w})$$

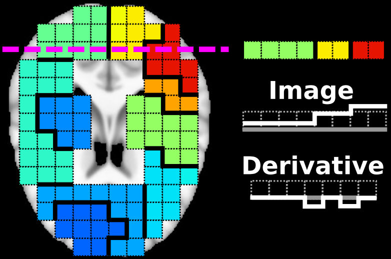
In fMRI: [Michel... 2011]

TV-ℓ_1: **Sparsity + regions**

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\text{argmin}}\, l(\mathbf{y} - \mathbf{X}\,\mathbf{w}) + \lambda\Big(\rho\,\ell_1(\mathbf{w}) + (1-\rho)\,TV(\mathbf{w})\Big)$$

l: data-fit term

[Baldassarre... 2012, Gramfort... 2013]



Total-variation penalization

Impose sparsity on the gradient of the image:

$$p(\mathbf{w}) = \ell_1(\nabla \mathbf{w})$$

In fMRI: [Michel... 2011]

Analysis sparsity: $||\mathbf{K}\mathbf{w}||_{21}$ [Eickenberg... 2015]

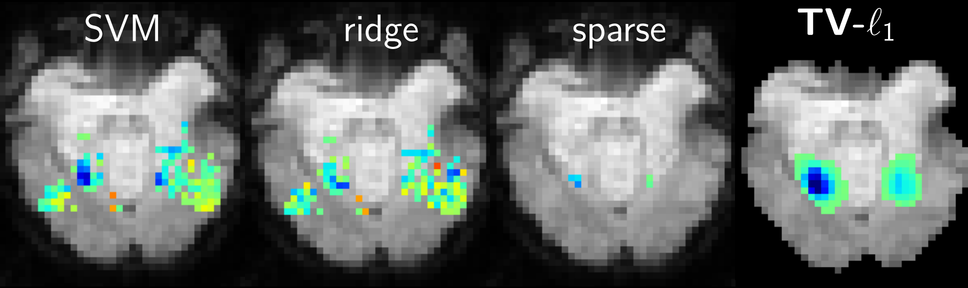
TV-ℓ_1: **Sparsity + regions**

$$\hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}}\, l(\mathbf{y} - \mathbf{X}\,\mathbf{w}) + \lambda\Big(\rho\,\ell_1(\mathbf{w}) + (1-\rho)\,TV(\mathbf{w})\Big)$$

l: data-fit term [Baldassarre... 2012, Gramfort... 2013]

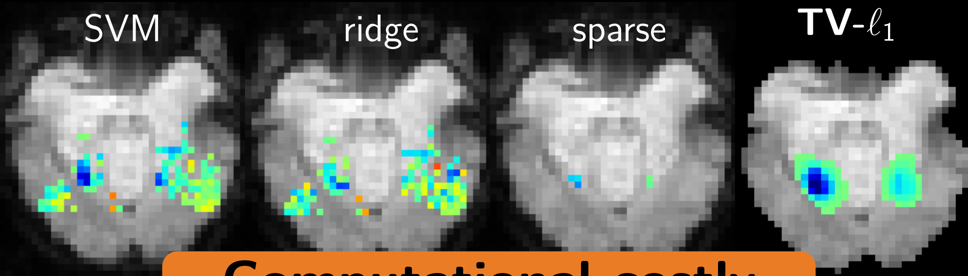
- **Good prediction performance**

- **Segment the relevant regions**



- **Good prediction performance**

- **Segment the relevant regions**



Computational costly

- Hyper-parameter selection brittle
- Tedious convergence

1 Hyper-parameter selection

Hyper-parameter setting important for ℓ_1 models

- Cross-validation, rather than hold-out, for small n

Cross-validation

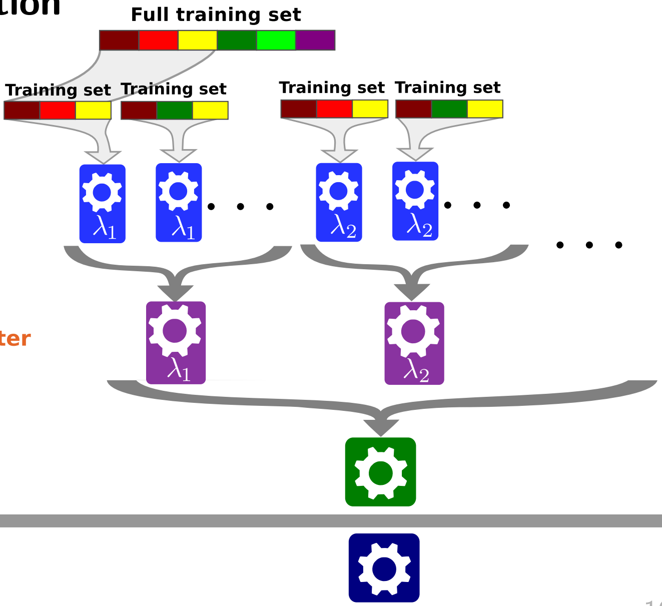
Subsampling

Fitting with each hyperparameter

Average models with the same hyperparameter

Select the best hyperparameter

Refit the model (final model)



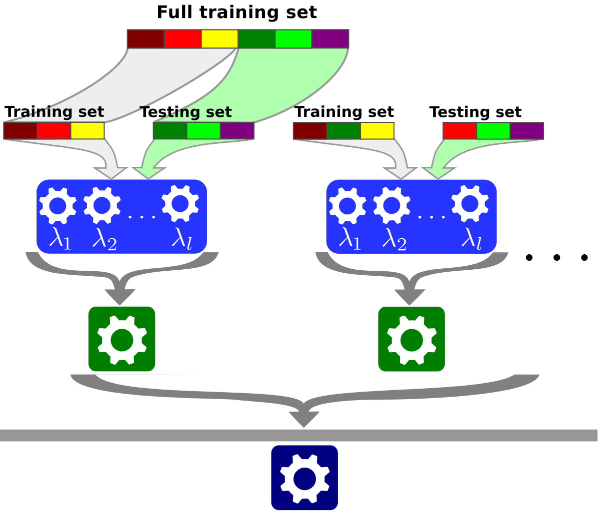
CV-bagging

Subsampling

Fitting with each hyperparameter

Select the best model per CV fold

Averaging (final model)

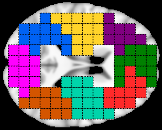


Bagging reduces variance

[Maillard... 2017, McInerney 2017]

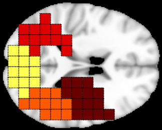
1 Fixing sparsity with clustering

Idea: cluster together correlated features



1 clustering to form reduced features

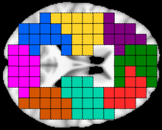
2 sparse linear model on reduced features



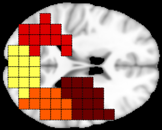
[Varoquaux... 2012]

1 Fixing sparsity with clustering and bagging

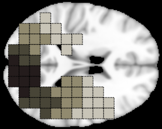
Idea: cluster together correlated features



1. `loop:` perturb randomly data
2. clustering to form reduced features



3. sparse linear model on reduced features
4. accumulate non-zero features

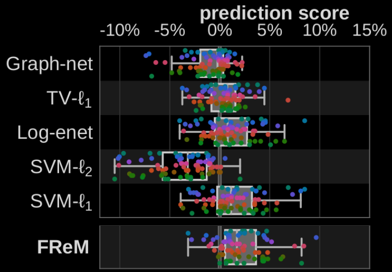


[Varoquaux... 2012]

Bagging:

- Clustering for dimension reduction
- Feature selection
- Linear model
- Hyper-parameter selection (CV-bagging)

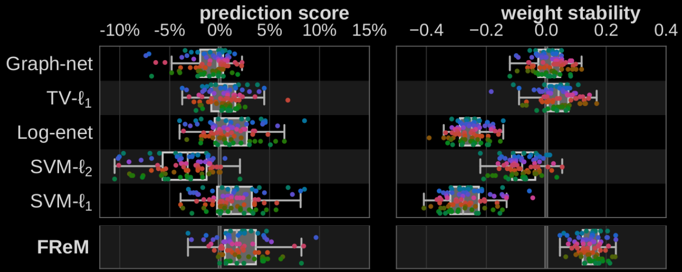
Empirical results



Bagging:

- Clustering for dimension reduction
- Feature selection
- Linear model
- Hyper-parameter selection (CV-bagging)

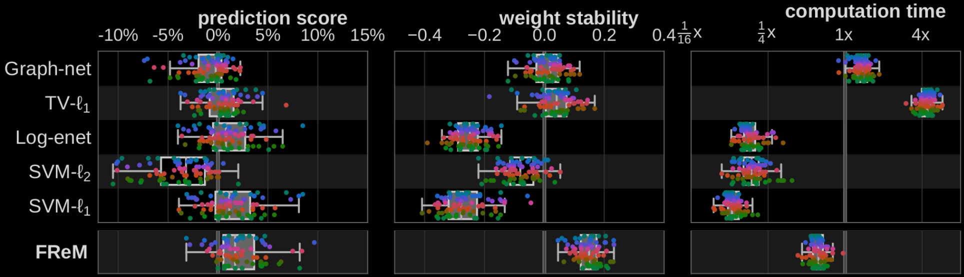
Empirical results



Bagging:

- Clustering for dimension reduction
- Feature selection
- Linear model
- Hyper-parameter selection (CV-bagging)

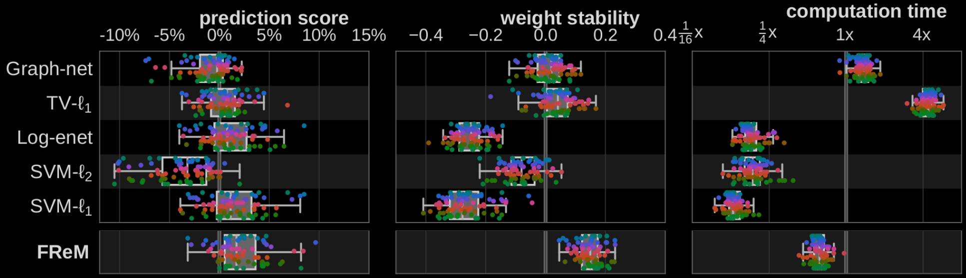
Empirical results



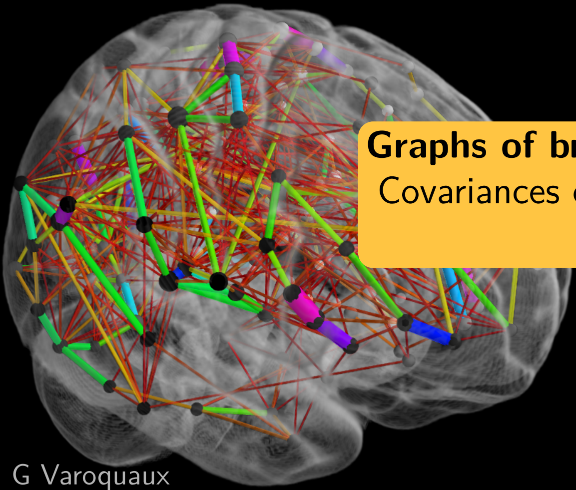
Lessons learned trying to regularize

- Sparsity is not enough: structure is needed
- Optimal sparse-structured is finicky and expensive
- Ensemble greedy approaches

Empirical results



2 Covariance estimation

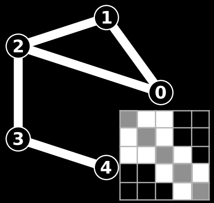


Graphs of brain function

Covariances capture interactions between regions

- Multivariate normal:

$$\mathcal{P}(\mathbf{X}) \propto \sqrt{|\mathbf{\Sigma}^{-1}|}e^{-\frac{1}{2}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}}$$



- Model parametrized by inverse covariance matrix, $\mathbf{K} = \mathbf{\Sigma}^{-1}$: ***conditional*** covariances

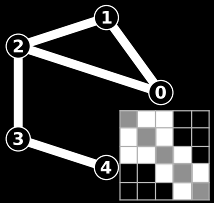
$$\mathbf{X}_i \perp\!\!\!\perp \mathbf{X}_j \Leftrightarrow \mathbf{K}_{i,j} = 0$$

- Graphical lasso: ℓ_1-penalized MLE
  Maximum-likelihood of $\mathbf{K}$ needs $\mathcal{O}(p^2)$ samples.
  ℓ_1 enables support recovery [Ravikumar... 2011]

2 Gaussian graphical models

- Multivariate normal:

$$\mathcal{P}(\mathbf{X}) \propto \sqrt{|\mathbf{\Sigma}^{-1}|}\, e^{-\frac{1}{2}\mathbf{X}^T\mathbf{\Sigma}^{-1}\mathbf{X}}$$



- Model parametrized by inverse covariance matrix,

Sample complexity of recovering s edges

$$n = \mathcal{O}((s+p)\log(p))$$

$$s = o(\sqrt{p})$$

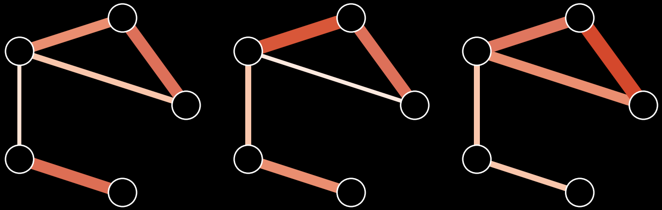
- Graphical lasso: ℓ_1-penalized MLE

Maximum-likelihood of $\mathbf{K}$ needs $\mathcal{O}(p^2)$ samples.

ℓ_1 enables support recovery [Ravikumar... 2011]

2 Larger n: multi-subject sparse covariance

Common independence structure but different connection values



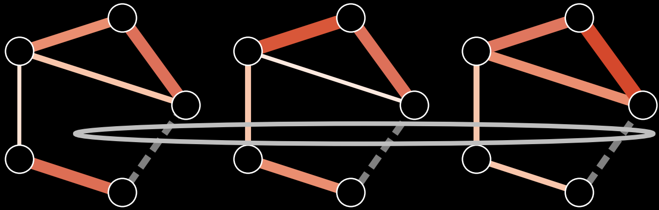
$$\{\mathbf{K}^s\} = \underset{\{\mathbf{K}^s \succ 0\}}{\operatorname{argmin}} \sum_s \mathcal{L}(\hat{\mathbf{\Sigma}}^s | \mathbf{K}^s) + \lambda \, \ell_{21}(\{\mathbf{K}^s\})$$

Multi-subject data fit, Likelihood

Group-lasso penalization

[Varoquaux... 2010]

Common independence structure but different connection values



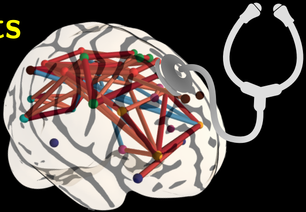
$$\{\mathbf{K}^s\} = \underset{\{\mathbf{K}^s \succ 0\}}{\operatorname{argmin}} \sum_s \mathcal{L}(\hat{\mathbf{\Sigma}}^s | \mathbf{K}^s) + \lambda\, \ell_{21}(\{\mathbf{K}^s\})$$

Multi-subject data fit, Likelihood

ℓ_1 on the connections of the ℓ_2 on the subjects

[Varoquaux... 2010]

Our goal may be to compare patients



- ℓ_1 recovery is unstable
- Brain graphs are not that sparse
  Between-subject differences may be sparse
  [Belilovsky... 2016]

Which risk should we minimize on the covariance?

James-Stein shrinkage

To estimate a mean θ:

$$\hat{\theta}_{JS} = (1 - \alpha)\,\theta_{\text{MLE}} \;+\; \alpha\,\theta_{\text{guess}} \quad \text{whith } \alpha \sim \frac{\sigma^2}{n\|\theta - \theta_{\text{guess}}\|}$$

$\hat{\theta}_{JS}$ dominates $\hat{\theta}_{MLE}$ for the MSE

Ledoit-Wolf covariance shrinkage estimator

$$\hat{\mathbf{\Sigma}}_{LW} = (1 - \alpha)\,\mathbf{\Sigma}_{\text{MLE}} \;+\; \alpha\,\text{trace}(\mathbf{\Sigma}_{\text{MLE}})\,\mathbf{I}$$

with α oracle for $n \to \infty, \frac{n}{p} \to \text{cst}$

$\hat{\mathbf{\Sigma}}_{LW}$ dominates $\hat{\mathbf{\Sigma}}_{MLE}$ for the MSE

[Ledoit and Wolf 2004]

2 James-Stein and Ledoit-Wolf

James-Stein shrinkage

To estimate a mean θ:

$$\hat{\theta}_{JS} = (1-\alpha)\,\theta_{\text{MLE}} \;+\; \alpha\,\theta_{\text{guess}} \quad \text{whith } \alpha \sim \frac{\sigma^2}{n\|\theta-\theta_{\text{guess}}\|}$$

For inter-subject comparison, Ledoit-Wolf performs as well as ℓ_1 estimators, but **faster** & **less brittle**.

Ledoit-Wolf covariance shrinkage estimator

$$\hat{\mathbf{\Sigma}}_{LW} = (1-\alpha)\,\mathbf{\Sigma}_{\text{MLE}} \;+\; \alpha\,\text{trace}(\mathbf{\Sigma}_{\text{MLE}})\,\mathbf{I}$$

with α oracle for $n \to \infty, \frac{n}{p} \to \text{cst}$

$\hat{\mathbf{\Sigma}}_{LW}$ dominates $\hat{\mathbf{\Sigma}}_{MLE}$ for the MSE

[Ledoit and Wolf 2004]

Shrinkage with order-2 moment

- Shrinkage = MMSE = Bayesian posterior mean for Gaussian distribution 4.1.2 [Lehmann and Casella 2006]

  $\Rightarrow$ Use prior $\mathcal{N}(\boldsymbol{\Sigma}_0, \boldsymbol{\Lambda}_0)$ learned on population

  * $\boldsymbol{\Lambda}_0$ is a **covariance on covariances**

[Rahim... 2017, 2018]

Shrinkage with order-2 moment

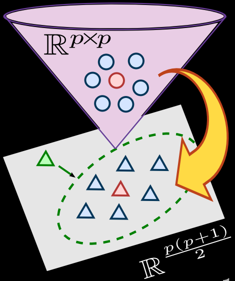
- Shrinkage = MMSE = Bayesian posterior mean
  for Gaussian distribution 4.1.2 [Lehmann and Casella 2006]
  $\Rightarrow$ Use prior $\mathcal{N}(\mathbf{\Sigma}_0, \mathbf{\Lambda}_0)$ learned on population

  * $\mathbf{\Lambda}_0$ is a **covariance on covariances**

Information geometry / covariance manifold

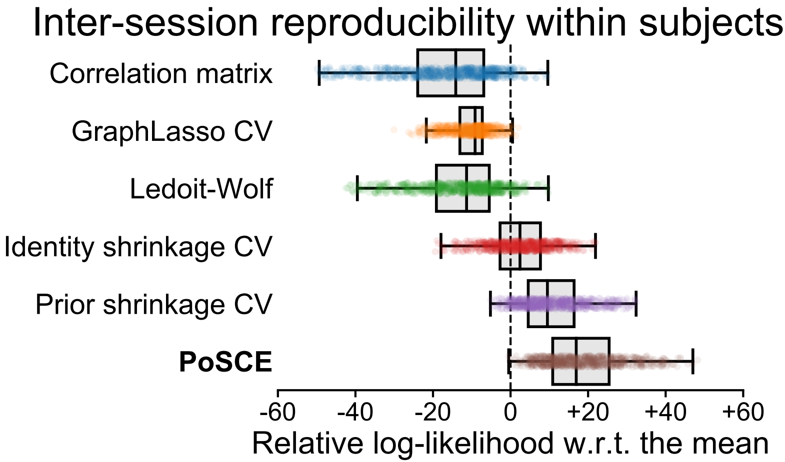
Covariances are not a vector space

- Computations on the manifold
- Turns MLE into an MSE



PoSCE: Population shrinkage of covariance

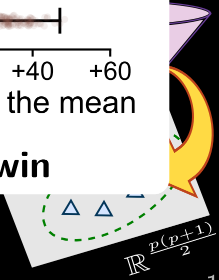
[Rahim... 2017, 2018]



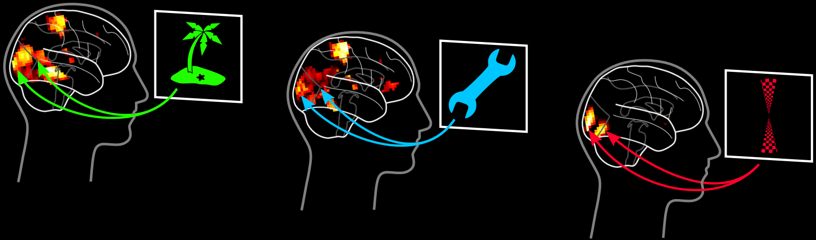
Anisotropic shrinkage for the win

PoSCE: Population shrinkage of covariance

[Rahim... 2017, 2018]



3 Merging data sources



More data trumps fancy regularizations

[Mensch... 2017]

3 There is plenty of fMRI data

Dozens of thousands of fMRI sessions,

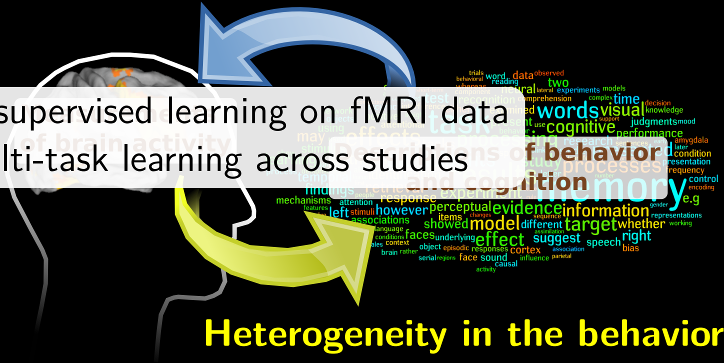
but terribly heterogeneous

Heterogeneity in the behavior

Formal modeling of behavior is a open knowledge representation problem

3 There is plenty of fMRI data

Dozens of thousands of fMRI sessions,
but terribly heterogeneous



- Unsupervised learning on fMRI data
- Multi-task learning across studies

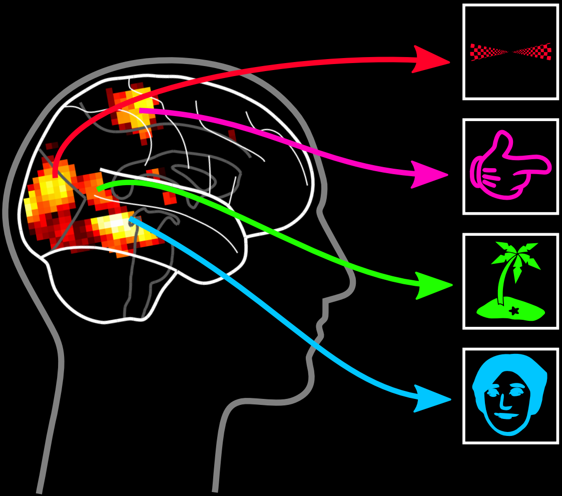
Heterogeneity in the behavior

Formal modeling of behavior is a open knowledge representation problem

3 Mapping cognition across studies labels

Cognitive label across many studies?

Very difficult to assign

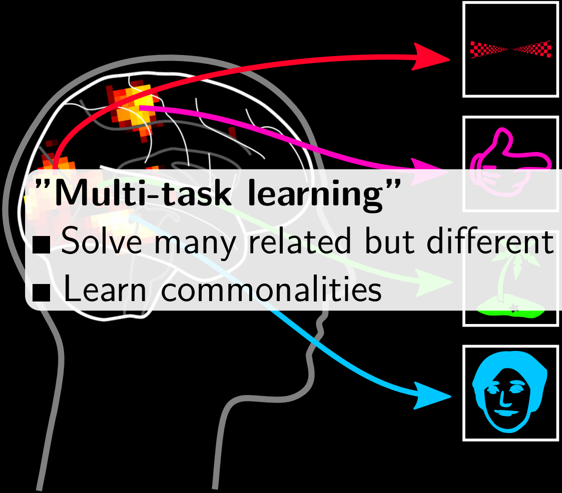


- Visual
- Auditory
- Foot
- Hand
- Calculation
- Reading
- Checkboard
- Face
- Place
- Object
- Digit
- Saccade
- ...

Mapping cognition across studies labels

Cognitive label across many studies?

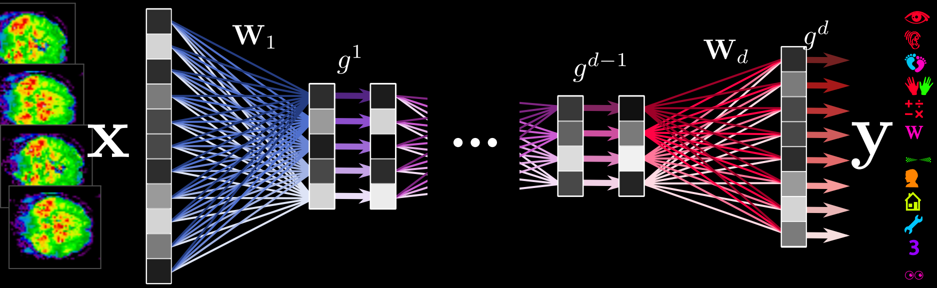
Very difficult to assign



"Multi-task learning"

- Solve many related but different problems
- Learn commonalities

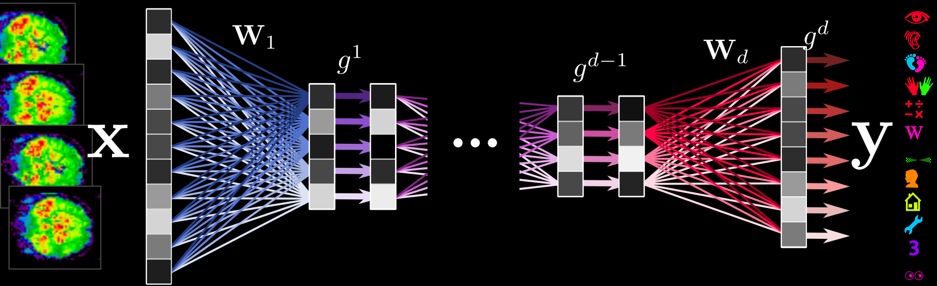
Visual
Auditory
Foot
Hand
Calculation
Reading
Checkboard
Face
Place
Object
Digit
Saccade
...



- Great for multiple output (tasks)

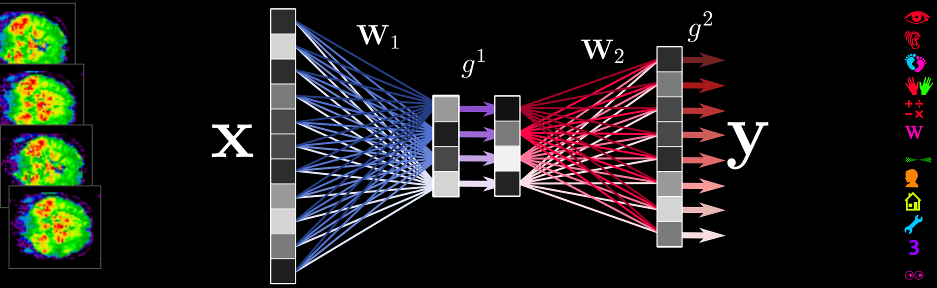
[Bengio 2009]

3 Sharing representations across tasks



- Great for multiple output (tasks)
- Millions of parameters, thousands of data points

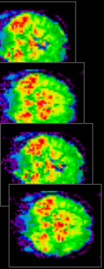
[Bengio 2009]

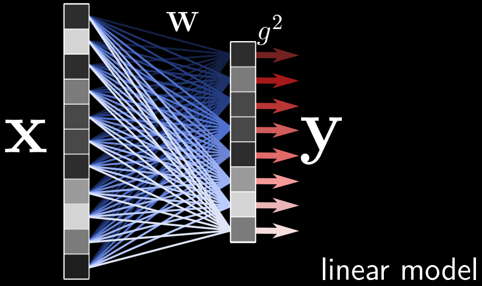


- Great for multiple output (tasks)
- Millions of parameters, thousands of data points

Simplify

Sharing representations across tasks

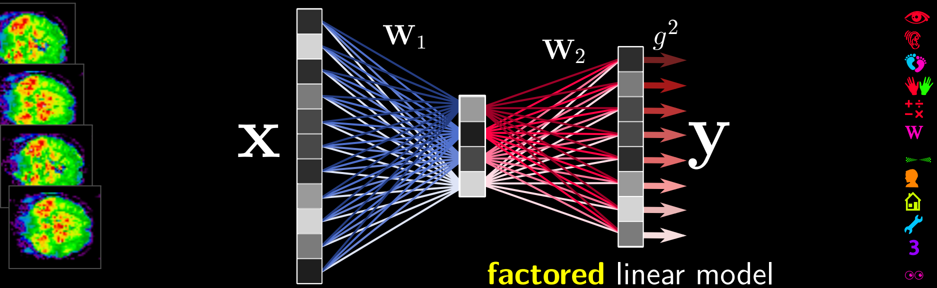




- Great for multiple output (tasks)
- Millions of parameters, thousands of data points

Simplify **simplify more**

Sharing representations across tasks



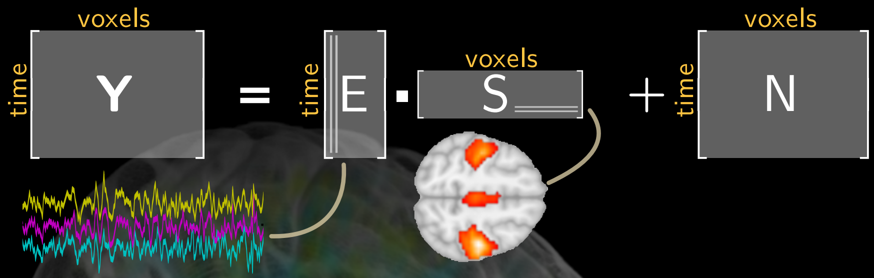
factored linear model

- Great for multiple output (tasks)
- Millions of parameters, thousands of data points

Simplify ~~**simplify more**~~

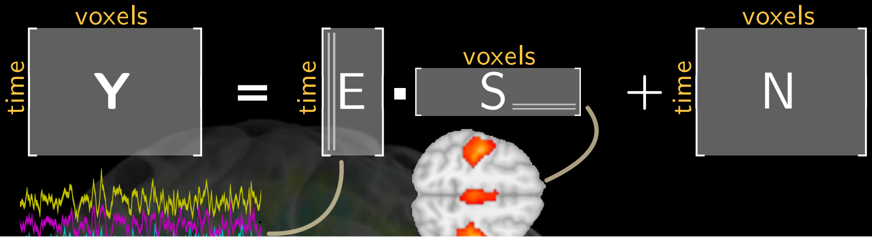
[Bzdok... 2015]

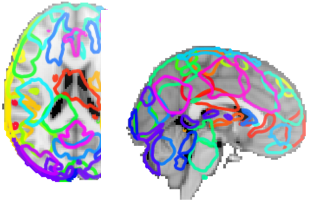
3 Unsupervised learning for spatial atoms



Decomposing time series into spatial maps
with sparsity to localize atoms

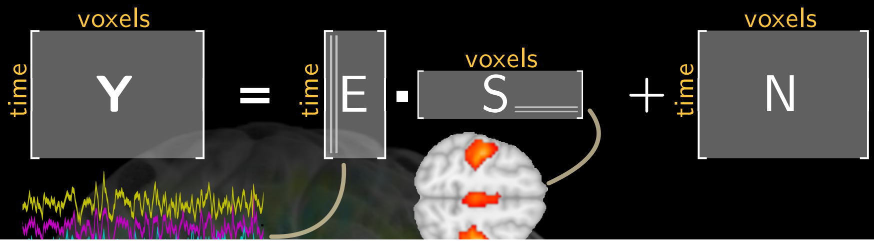
3 Unsupervised learning for spatial atoms

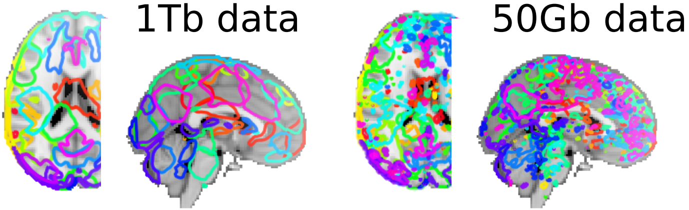




- Adapted representations that capture local correlations

3 Unsupervised learning for spatial atoms

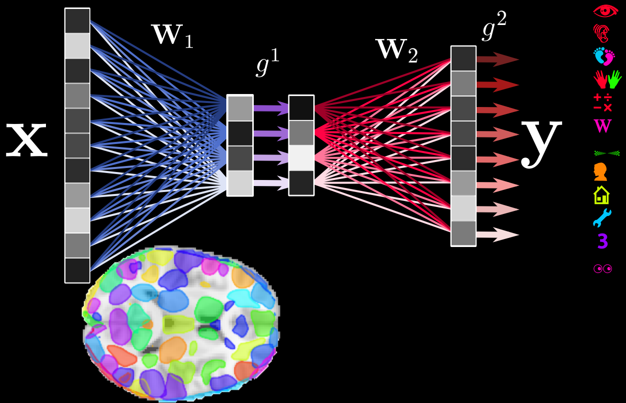




- Adapted representations that capture local correlations
- More data is always better

computational cost [Mensch... 2016]

3 Multi-task across studies

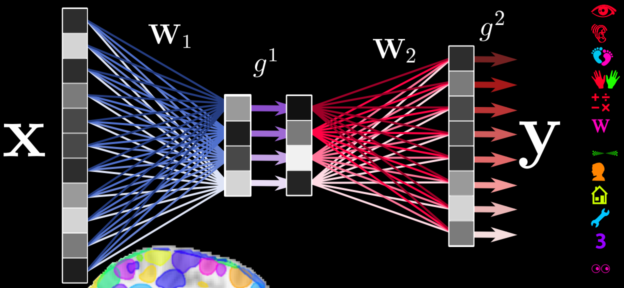


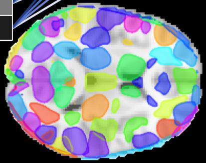
- Decode in each study
- But learn representations across

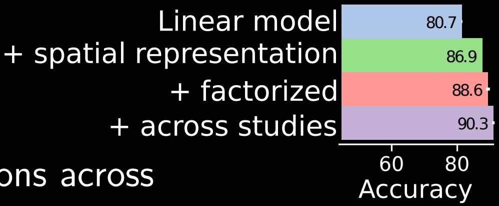
Loss-engineering & regularization

[Mensch... 2017]

3 Multi-task across studies







- Decode in each study
- But learn representations across

Loss-engineering & regularization

[Mensch... 2017]

Learning with limited labeled data: fMRI lessons

- Sparse models are unstable and need ensembling
- Parameter selection is unstable and needs ensembling
- ℓ_2 shrinkage is powerful, in particular with good mean & covariance
- Unsupervised learning of representations
- Multi-task to pool data

Learning with limited labeled data: fMRI lessons

- Sparse models are unstable and need ensembling
- Parameter selection is unstable and needs ensembling
- ℓ_2 shrinkage is powerful, in particular with good mean & covariance
- Unsupervised learning of representations
- Multi-task to pool data
- Software for machine learning in neuroimaging: http://nilearn.github.io

ni learn

@GaelVaroquaux

scikit learn

machine learning in Python

References I

L. Baldassarre, J. Mourao-Miranda, and M. Pontil. Structured sparsity models for brain decoding from fMRI data. In *PRNI*, page 5, 2012.

E. Belilovsky, G. Varoquaux, and M. B. Blaschko. Testing for differences in gaussian graphical models: applications to brain connectivity. In *Advances in Neural Information Processing Systems*, pages 595–603, 2016.

Y. Bengio. Learning deep architectures for ai. *Foundations and trends in Machine Learning*, 2:1–127, 2009.

D. Bzdok, M. Eickenberg, O. Grisel, B. Thirion, and G. Varoquaux. Semi-supervised factored logistic regression for high-dimensional neuroimaging data. In *Advances in neural information processing systems*, pages 3348–3356, 2015.

M. Eickenberg, E. Dohmatob, B. Thirion, and G. Varoquaux. Total variation meets sparsity: statistical learning with segmenting penalties. *MICCAI*, 2015.

References II

A. Gramfort, B. Thirion, and G. Varoquaux. Identifying predictive regions from fMRI with TV-L1 prior. In *PRNI*, page 17, 2013.

A. Hoyos-Idrobo, G. Varoquaux, Y. Schwartz, and B. Thirion. Frem – scalable and stable decoding with fast regularized ensemble of models. *NeuroImage*, 2017.

O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88: 365, 2004.

E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

G. Maillard, S. Arlot, and M. Lerasle. Cross-validation improved by aggregation: Agghoo. *arXiv preprint arXiv:1709.03702*, 2017.

J. McInerney. An empirical bayes approach to optimizing machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2709–2718, 2017.

References III

A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Dictionary learning for massive matrix factorization. In *International Conference on Machine Learning*, pages 1737–1746, 2016.

A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Learning neural representations of human cognition across many fMRI studies. In *NIPS*, 2017.

V. Michel, A. Gramfort, G. Varoquaux, E. Eger, and B. Thirion. Total variation regularization for fMRI-based prediction of behavior. *Medical Imaging, IEEE Transactions on*, 30:1328, 2011.

A. Ng. Feature selection, l1 vs. l2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78, 2004.

References IV

M. Rahim, B. Thirion, and G. Varoquaux. Population-shrinkage of covariance to estimate better brain functional connectivity. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 460–468. Springer, 2017.

M. Rahim, B. Thirion, and G. Varoquaux. PoSCE: Population shrinkage of covariance to estimate better brain functional connectivity. *submitted*, 2018.

P. Ravikumar, M. J. Wainwright, G. Raskutti, B. Yu, ... High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

G. Varoquaux and B. Thirion. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 3:28, 2014.

G. Varoquaux, A. Gramfort, J. B. Poline, and B. Thirion. Brain covariance selection: better individual functional connectivity models using population prior. In *NIPS*. 2010.

References V

G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. In *ICML*, page 1375, 2012.

M. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *Trans Inf Theory*, 55:2183, 2009.