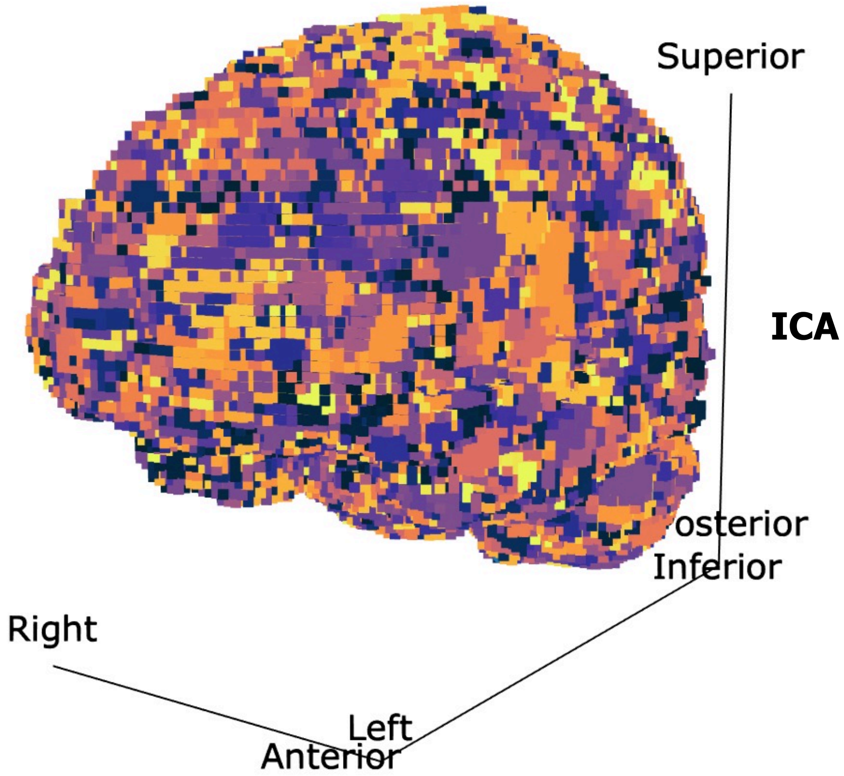
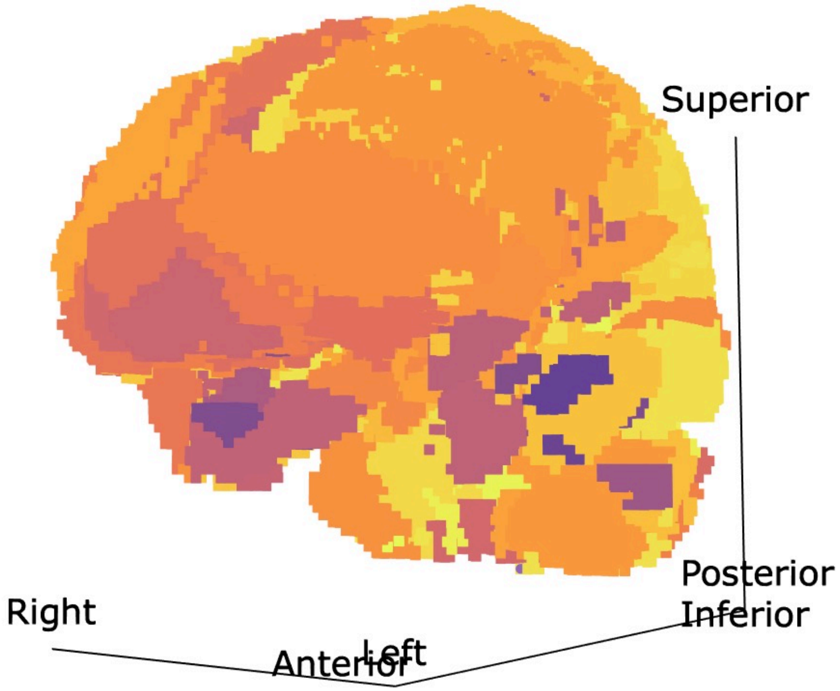
Fast structure learning with modular regularization

Greg Ver Steeg, Hrayr Harutyunyan, Daniel Moyer, Aram Galstyan

USC Viterbi School of Engineering
Information Sciences Institute



ICA

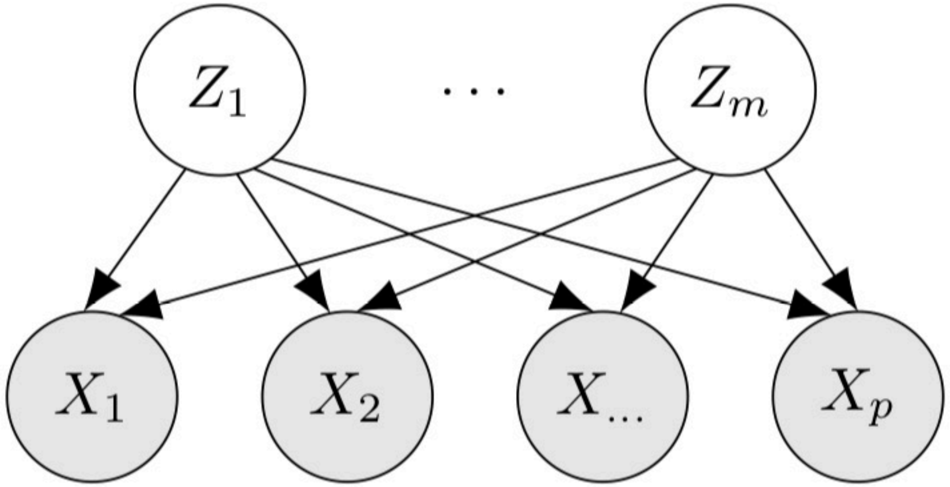


Proposed:

Latent factor discovery with information-theoretic modularity regularization

Information-theoretic idea for efficient modularity regularization

Unconstrained latent factor model

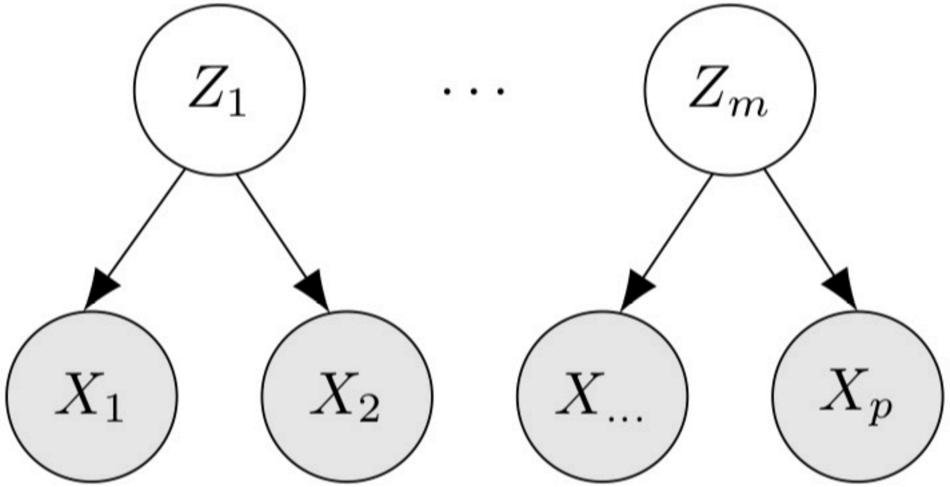


$$\Updownarrow$$

$$TC(X \mid Z) + TC(Z) = 0$$

(Related to VAE/ELBO: arXiv:1802.05822)

Modular latent factor model



$\Downarrow$ (for any distribution) $\Uparrow$ (for Gaussians)

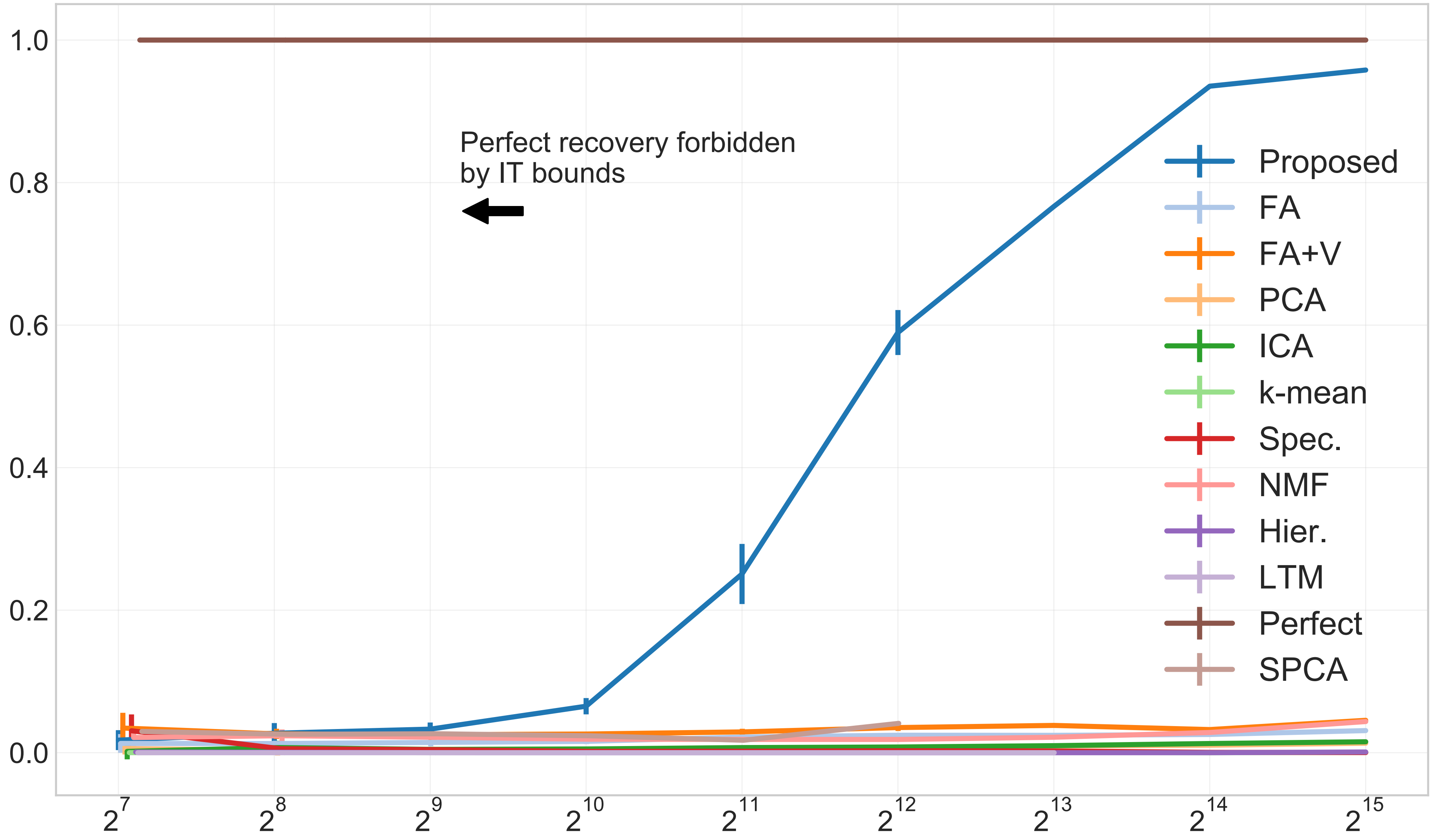
$$TC(X \mid Z) + TC(Z) = 0, \ \& \ \forall i, \underline{TC(Z \mid X_i) = 0}$$

Suppose that variables approximately cluster into modules, one latent factor per module:

- Combinatorial search for the best structured model would be infeasible: *exponentially* many
- We re-formulate the learning problem as an **unconstrained optimization** whose **global optima correspond to structured latent factor models**

Modular structure recovery in high-d (with 300 samples)

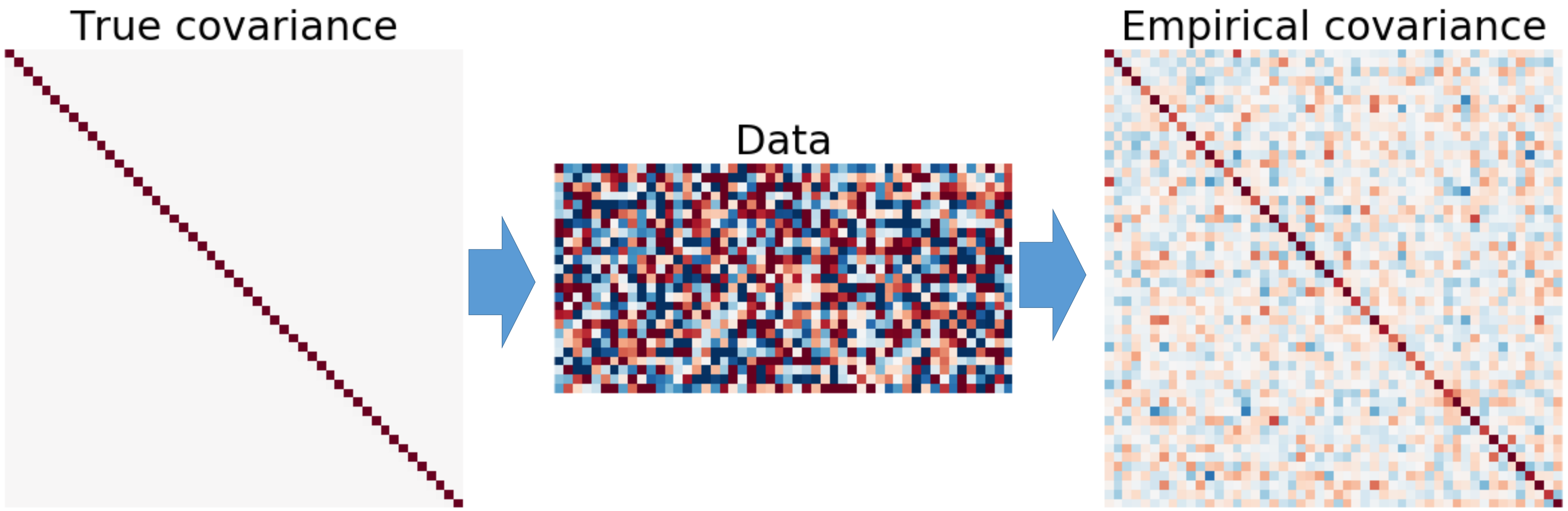
Structure recovery score (Adj. Rand Index)



Variables increasing →

Covariance estimation

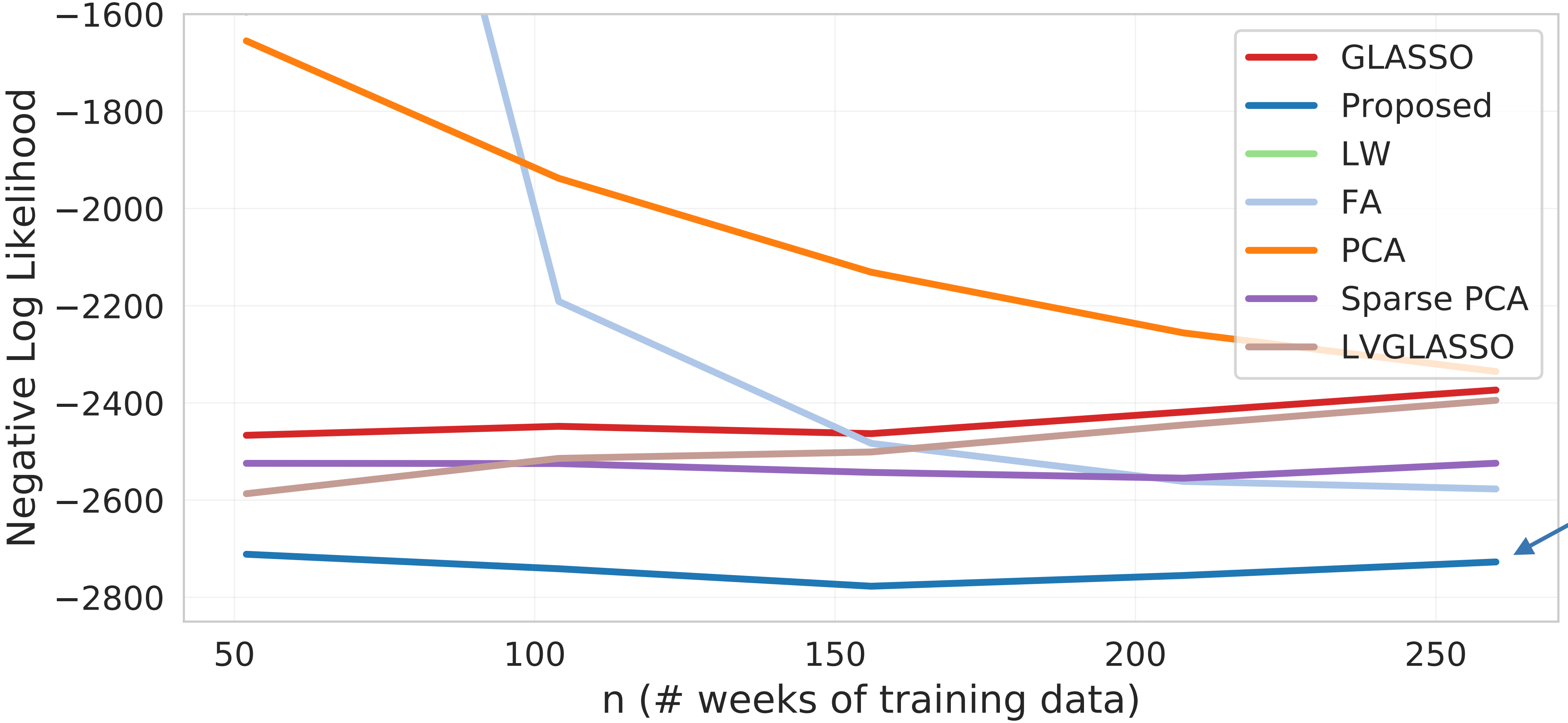
- If n (samples) < p (variables), empirical covariance is a *terrible, terrible estimate*
- But we can do better through priors: sparsity, independence, dim. red., *modularity*



wins on 51 real datasets from OpenML (best log-likelihood on test)

This work	**32/51**
Ledoit-Wolf	18/51
Sparse PCA	1/51
Factor Analysis	1/51
GLASSO (BigQUIC)	0/51

Estimating covariance from under-sampled stock market data

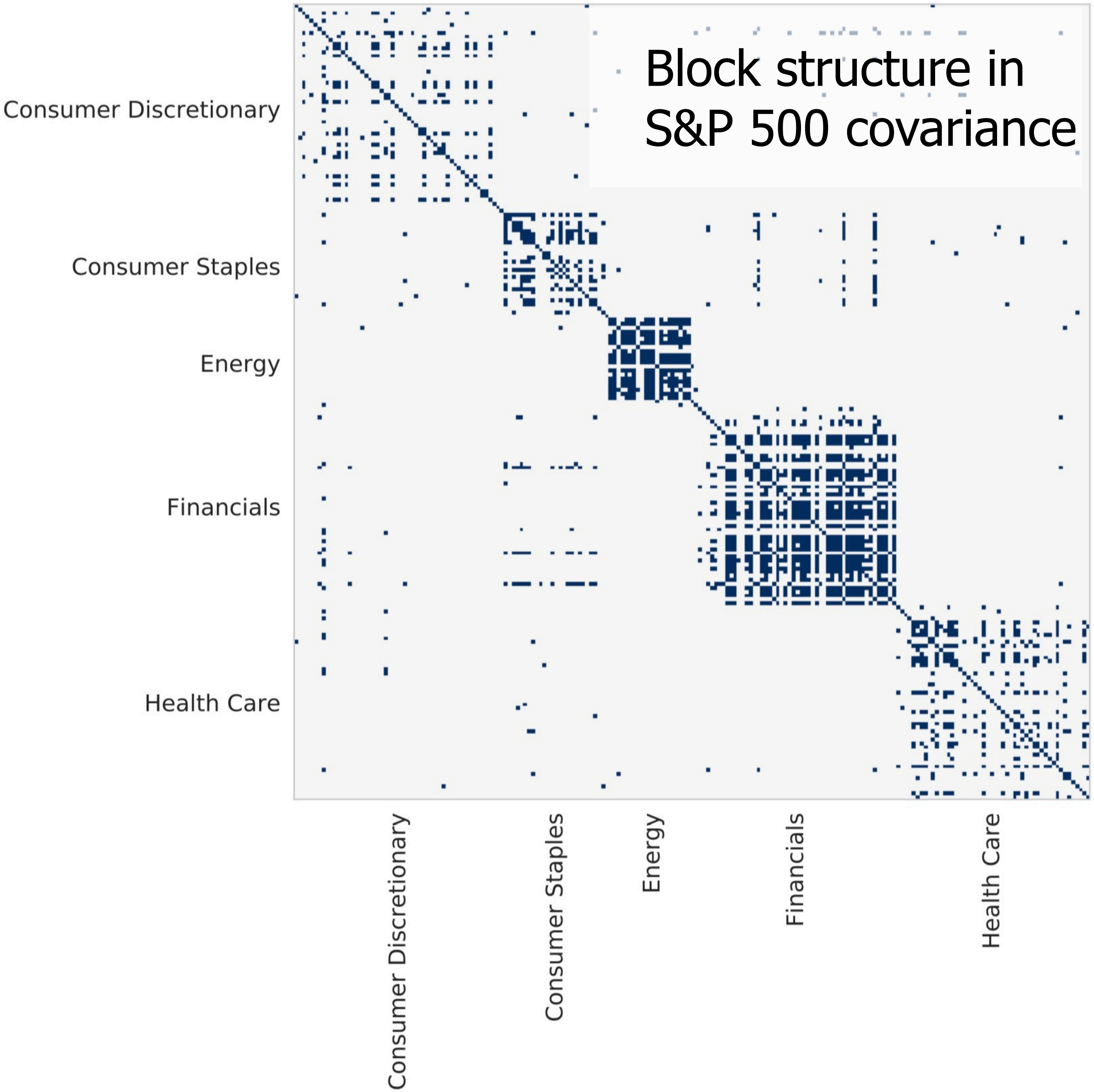


Proposed method gets significantly better log likelihood on test data with few weeks of training data

Interpretable modular structure

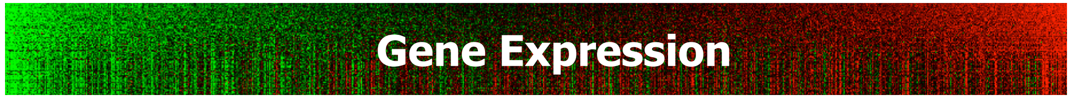
Factor	Stock ticker	Sector/Industry
0	RF, KEY, FHN	Bank holding (NYSE, large cap)
1	ETN, IEX, ITW	Industrial machinery
2	GABC, LBAI, FBNC	Bank holding (NASDAQ, small ca
3	SPN, MRO, CRZO	Oil & gas
4	AKR, BXP, HIW	Real estate investment trusts
5	CMS, ES, XEL	Electric utilities
6	POWI, LLTC, TXN	Semiconductors
7	REGN, BMRN, CELG	Biotech pharmaceuticals
8	BKE, JWN, M	Retail, apparel
9	DHI, LEN, MTH	Homebuilders

Example latent factors appearing in stock market data



Block structure in S&P 500 covariance

Conclusion



- Introduced an *information-theoretic optimization* to tractably discover *structured latent factor models*
- Theoretical bounds on sample complexity suggests a “blessing of dimensionality”, recovering latent factors better in higher-d.
- Applications in latent factor discovery and covariance estimation useful in many domains: *neuroscience, finance,* and *gene expression*

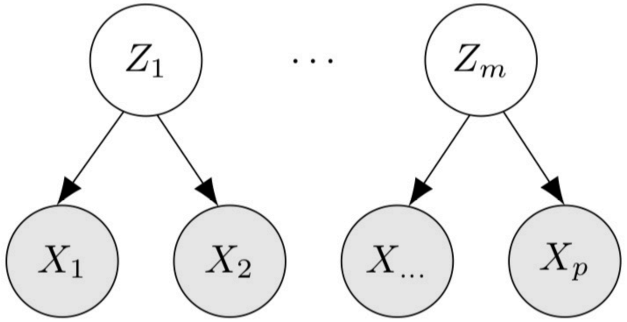
Poster 16 - in a few minutes

Paper: arxiv:1706.03353, NeurIPS 2019

Contact: hrayrh@isi.edu, gregv@isi.edu

Code:

https://github.com/gregversteeg/LinearCorex (numpy),
https://github.com/hrayrhar/T-CorEx (PyTorch)



$\Downarrow$ (for any distribution) $\Uparrow$ (for Gaussians)

$$TC(X \mid Z) + TC(Z) = 0, \;\&\; \forall i, TC(Z \mid X_i) = 0$$