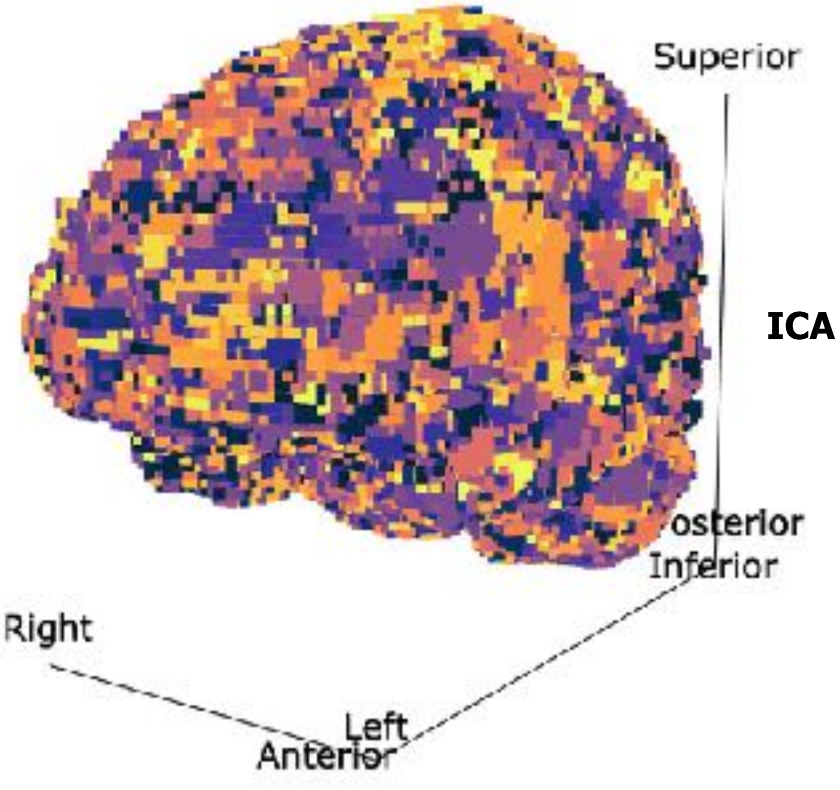
Fast structure learning with modular regularization

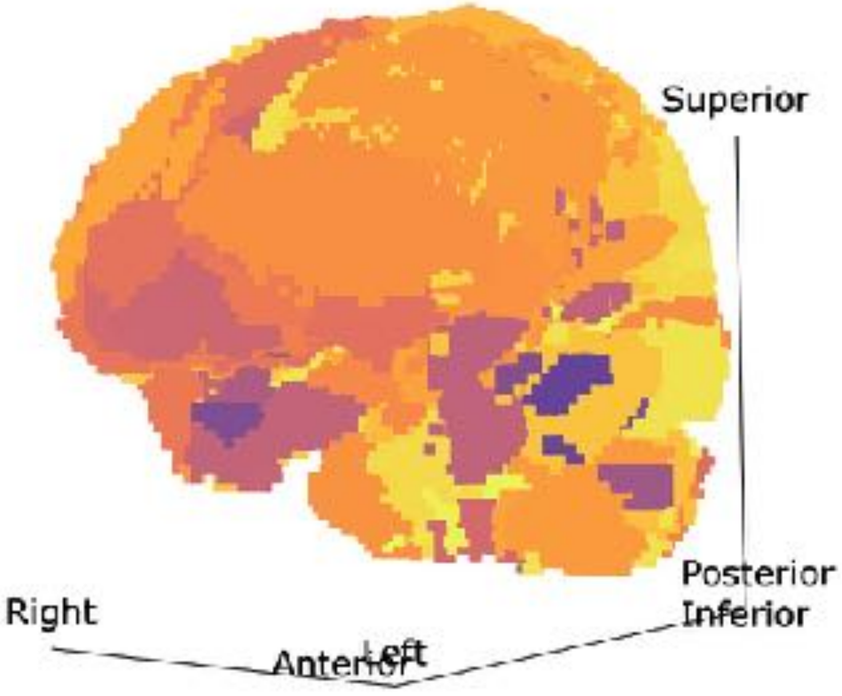
Greg Ver Steeg, Hrayr Harutyunyan, Daniel Moyer, Aram Galstyan

arxiv:1706.03353, NeurIPS 2019

USC Viterbi
School of Engineering
Information Sciences Institute



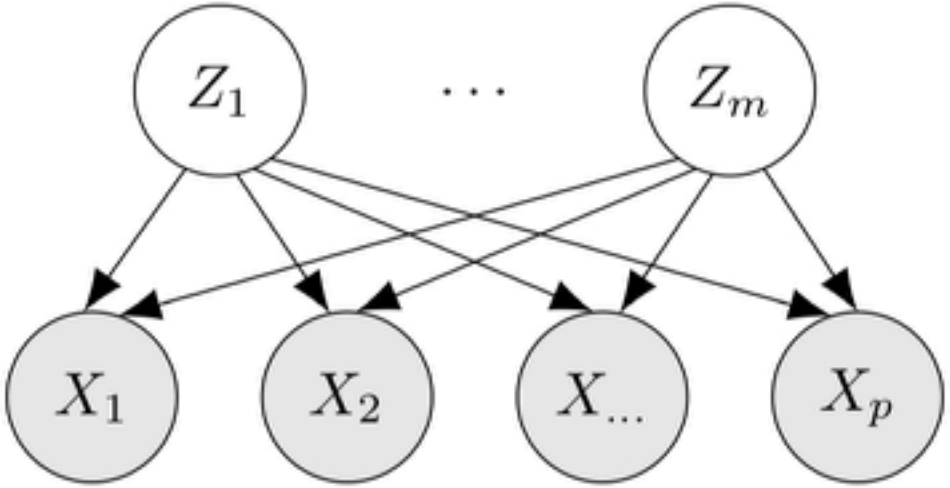
ICA



Proposed: Latent factor discovery with modular regularization

Mathematical principle for efficient modularity regularization

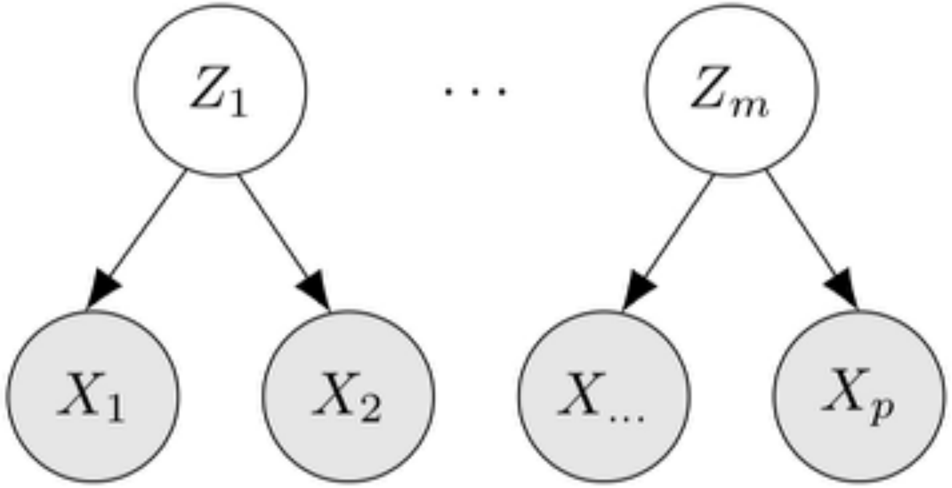
Unconstrained latent factor model



$$\Updownarrow$$

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

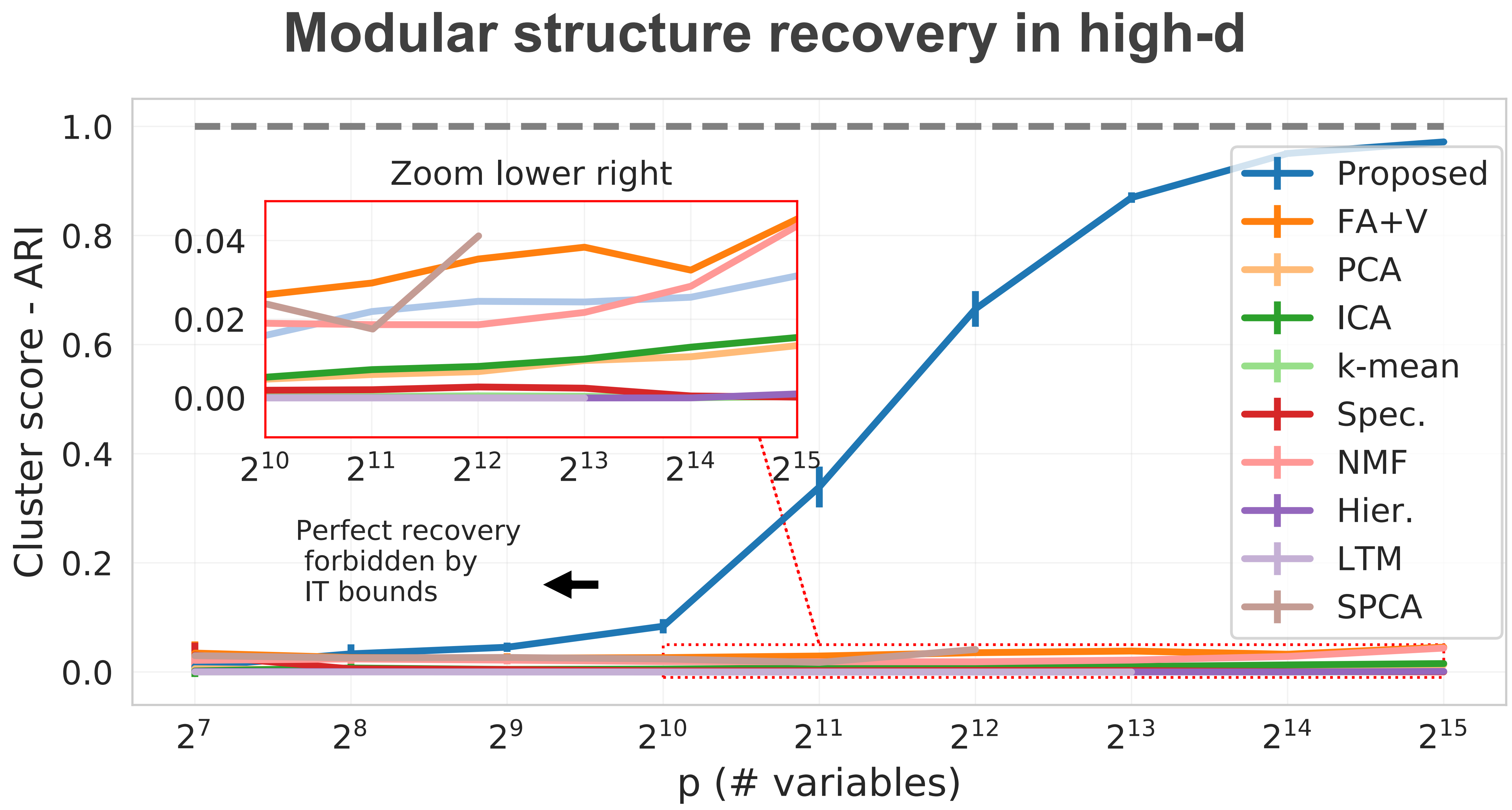
Modular latent factor model



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

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

- Suppose that variables approximately cluster into modules, one latent factor per module
- Combinatorial search for the best model would be infeasible: *exponentially* many models
- We re-formulate the learning problem as an *unconstrained* optimization whose global optima correspond to structured latent factor models



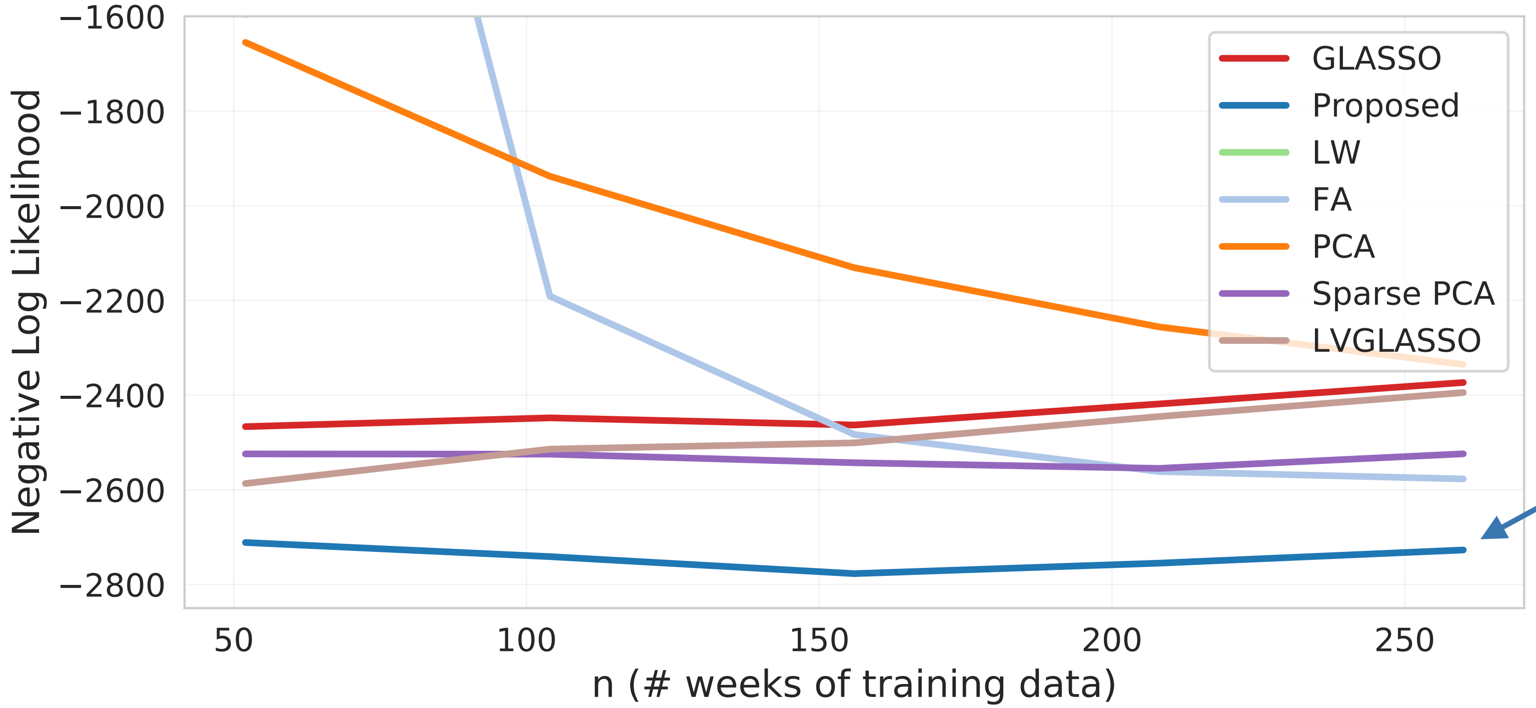
Modular structure recovery in high-d
Zoom lower right
Cluster score - ARI
1.0
0.8
0.6
0.4
0.2
0.0
0.04
0.02
0.00
2^{10} 2^{11} 2^{12} 2^{13} 2^{14} 2^{15}
Perfect recovery
forbidden by
IT bounds
2^{7} 2^{8} 2^{9} 2^{10} 2^{11} 2^{12} 2^{13} 2^{14} 2^{15}
p (# variables)
Proposed
FA+V
PCA
ICA
k-mean
Spec.
NMF
Hier.
LTM
SPCA

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*

Our approach outperforms GLASSO and Sparse PCA for covariance estimation on 50 out of 51 under-sampled datasets from OpenML

Estimating covariance on stock market data

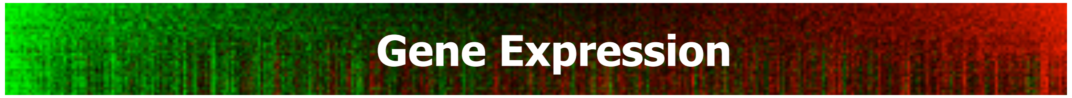


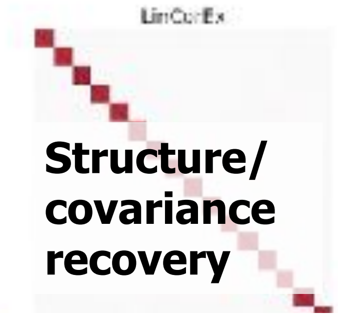
Proposed method gets significantly better log likelihood on test data with few weeks of training data

Some example factors that appear in stock market data

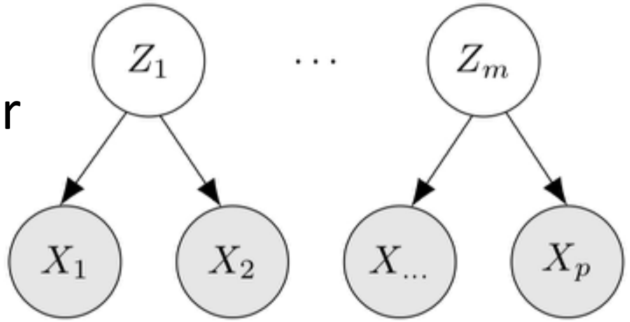
Factor	Stock ticker	Sector/Industry
0	RF, KEY, FHN	Bank holding (NYSE, large cap)
1	ETN, IEX, ITW	Industrial machinery
2	GABC, LBAI, FBNC	Bank holding (NASDAQ, small cap)
3	SPN, MRO, CRZO	Oil & gas
4	AKR, BXP, HIW	Real estate investment trusts
5	CMS, ES, XEL	Electric utilities
6	POWI, LLTC, TXN	Semiconductors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

Conclusion





- Connection between information-theoretic objective and structured latent factor models makes it efficient to add modular regularization
- Applications in latent factor discovery and covariance estimation in domains like: neuroscience, finance, and gene expression
- Theoretical bounds on sample complexity show that our approach is the only one to realize a "blessing of dimensionality", recovering latent factors better as the number of variables increases.



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

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

Paper: arxiv:1706.03353, NeurIPS 2019

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

Code: https://github.com/gregversteeg/LinearCorex, https://github.com/hrayrhar/T-CorEx (tensorflow)