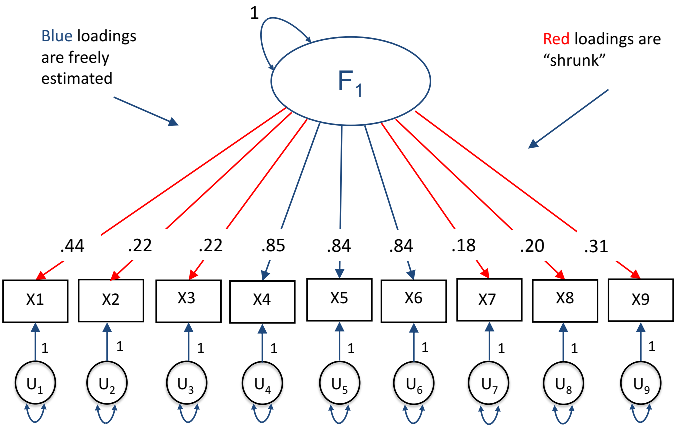
Flexible Specification of Large Structural Equation Models with Regularization

Ross Jacobucci
University of Notre Dame

University of Notre Dame

Introduction

Simple Motivating Example



Model Modification

Factor Models

- Model may not fit well
- Uncertainty exists regarding the inclusion of specific items/scales

SEM Models

- Too many predictors in MIMIC Model
- Multicollinearity among predictors
- Model that overfits – Has too many parameters
- Complex theoretical model but small sample size

Regularized Regression

$$Ridge : \hat{\beta}^{ridge} = argmin\Big\{ \overbrace{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2}^{OLS} + \lambda \overbrace{\sum_{j=1}^{p} \beta_j^2}^{penalty} \Big\} \quad (1)$$

$$Lasso : \hat{\beta}^{lasso} = argmin\Big\{ \sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\}. \quad (2)$$

λ is the shrinkage parameter

- λ controls the size of the coefficients
- λ controls the amount of **regularization**
- As $\lambda \downarrow 0$ least squares solution
- As $\lambda \uparrow \infty$ parameters go to zero

Comparison of Ridge & Lasso

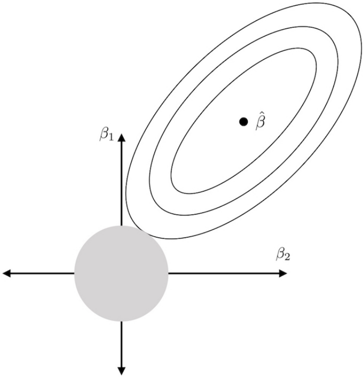


Figure 1: Ridge: $\beta_1^2 + \beta_2^2 \leq s$

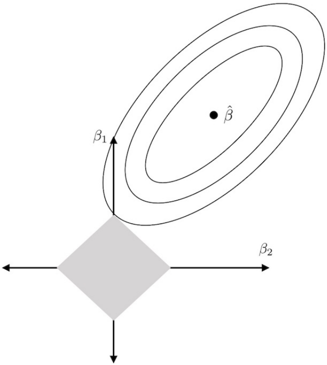


Figure 2: Lasso: $|\beta_1| + |\beta_2| \leq s$

Benefits of Regularized Regression

Lasso Regression

- Performs variable selection
- Estimate models with $P > N$

Ridge Regression

- Handles multicollinearity
- Estimate models with $P > N$
- Less bias with large coefficients

Many extensions of these that combine the benefits of each, along with producing sparser solutions

Regularized Structural Equation Modeling

RegSEM

Regularized structural equation modeling (RegSEM; Jacobucci, Grimm, & McArdle, 2016) builds upon the traditional maximum likelihood cost function for SEM models

$$F_{ML} = log(|\Sigma|) + tr(C * \Sigma^{-1}) - log(|C|) - p. \tag{3}$$

and builds in a separate element to the cost function

$$F_{regsem} = F_{ML} + \lambda P(\cdot) \tag{4}$$

where λ is the regularization parameter, and takes some value between zero and infinity.

This same method has been called Penalized Likelihood for SEM (Huang, Chen, & Weng, 2017)

Types of Penalties

$P(\cdot)$ Options

- Lasso $\lambda||\theta_{pen}||_1$
- Ridge $\lambda||\theta_{pen}||_2$
- Elastic Net $(1-\alpha)\lambda||\theta_{pen}||_2 + \alpha\lambda||\theta_{pen}||_1$
- Adaptive Lasso $\lambda||\theta_{ML}^{-1} * \theta_{pen}||_1$
- Smoothly Clipped Absolute Deviation (SCAD)
- Minimax Concave Penalty (MCP)
- Spike-and-Slab Lasso

- Difficult to pair desired sparsity with computational tractability.

RegSEM Equation Example

Using RAM notation, it is most common to penalize the directed paths (A matrix; regressions or factor loadings).

From the one-factor example, the A matrix is:

```
> extractMatrices(lav.out)$A_est
   x1 x2 x3 x4 x5 x6 x7 x8 x9             1        f1
x1  0  0  0  0  0  0  0  0  0 -9.679728e-17 0.4372132
x2  0  0  0  0  0  0  0  0  0 -3.598572e-16 0.2200393
x3  0  0  0  0  0  0  0  0  0  8.245646e-17 0.2224102
x4  0  0  0  0  0  0  0  0  0 -3.580857e-17 0.8462708
x5  0  0  0  0  0  0  0  0  0  1.165953e-16 0.8397105
x6  0  0  0  0  0  0  0  0  0 -1.550197e-16 0.8366859
x7  0  0  0  0  0  0  0  0  0 -3.025939e-16 0.1801725
x8  0  0  0  0  0  0  0  0  0  1.035388e-16 0.2008473
x9  0  0  0  0  0  0  0  0  0 -3.792083e-17 0.3062909
1   0  0  0  0  0  0  0  0  0  0.000000e+00 0.0000000
f1  0  0  0  0  0  0  0  0  0  0.000000e+00 0.0000000
```

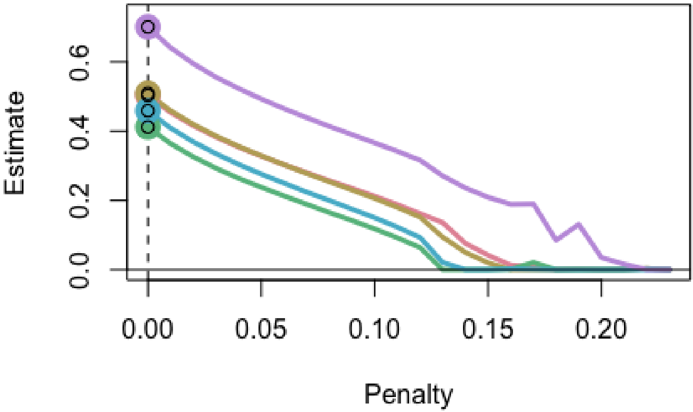
RegSEM Equation Continued

If we use Lasso penalties, with a penalty of 0.1, the equation becomes:

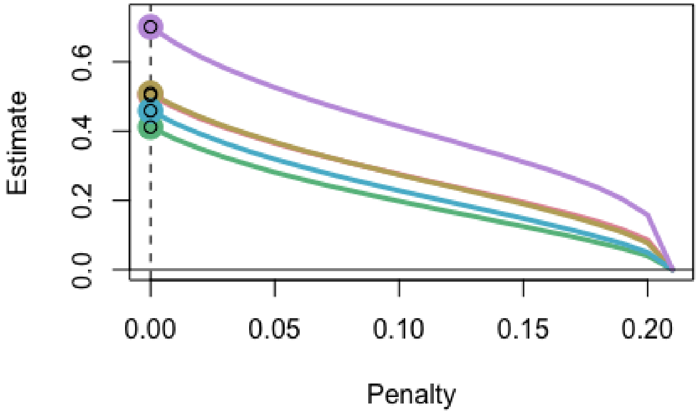
$$F_{regsem} = F_{ML} + 0.1 * \left\| \begin{matrix} .44 \\ .22 \\ .22 \\ .18 \\ .20 \\ .31 \end{matrix} \right\|_1 \tag{5}$$

With this, it is typical to use somewhere from 20-100 penalties, then selecting the model that fits best according to the BIC or bootstrapped chi-square.

Lasso Parameter Trajectories



Ridge Parameter Trajectories



Regularization Rationale

In Psychology, it is common to have complex hypotheses (models) and limited datasets. In these cases, our datasets may have contain enough information to test a complex models. Something has to give.

Regularized SEM, both frequentist and Bayesian forms, is one way to overcome this, by keeping the complex model, but reducing (or eliminating) the influence of parameters that are not of central interest.

Regularization Rationale

Regularization Fallacy:
Since regularization is from machine learning -> exploratory method

Key to applying regularization in SEM:

1. Specify what your theory as a model.
2. Identify where your uncertainty lies.

In confirmatory research – No uncertainty to model specification.
In exploratory research – Uncertainty dominates.

Where most research is – Some uncertainty and some theory.

Regularization Rationale Continued

Two Rationales for Regularization:

1. ***functional sparsity*** – Our aim is to create a more parsimonious model by removing parameters/variables.

Fundamental assumption of most regularization: Sparsity

This is unlikely in Psychology – e.g. "crud factor" Meehl (1990)
- Therefore we accept some **bias**.

2. **We have a mismatch between model complexity and sample size.**

Limit the influence of non-target parameters and improve the stability of the parameter estimates.

Specific Applications

Papers:

- EFA Subset Selection many papers circa 2012-2015
- CFA Subset Selection Jacobucci et al. (2016); Feng et al. (2017)
- Multiple Group Models Feng et al. (2018)
- Latent Interactions Brandt et al. (2018)
- Mediation Models Serang et al. (2017), Serang & Jacobucci (2019)
- Latent Change Score Models Jacobucci & Grimm (2018)
- Rotation in EFA Scharf & Nestler (2019)
- MIMIC Models **Jacobucci, Brandmaier, & Kievit, 2019**

Regularized MIMIC Models

Motivating Example

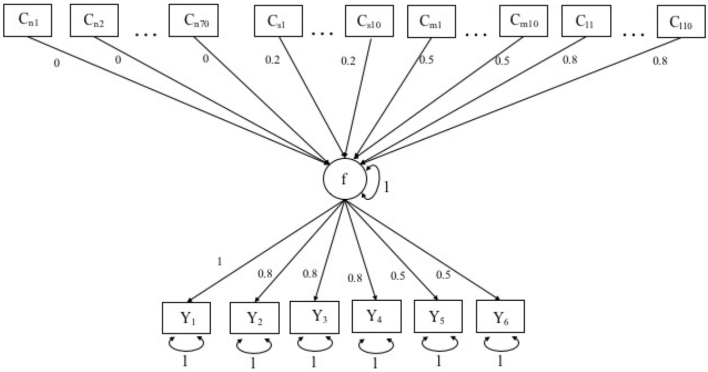
The sample consists of 627 participants from the Cambridge Study of Cognition, Aging and Neuroscience (Cam-CAN, www.cam-can.org).

Using a subset of the variables, we focused on a specific cognitive task (the visual short term memory task) and a common index of white matter microstructure (Fractional Anisotropy, FA; 48 indicators) for participants with complete data.

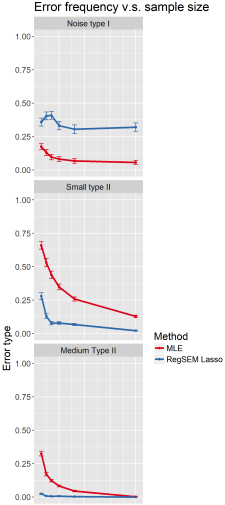
Problem: Could we keep short-term memory as a latent variable, deal with collinearity among a large number of predictors, and get the model to run?

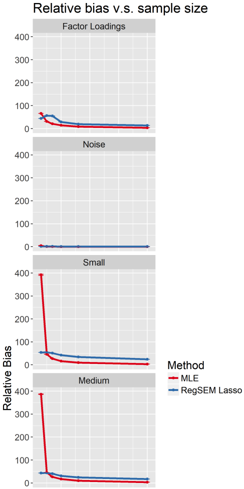
Basis for: Jacobucci, R., Brandmaier, A., & Kievit, R. (2019). A practical guide to variable selection in structural equation models with regularized MIMIC models. Advances in Methods and Practices in Psychological Science, 2, 55-76.

Model


C n1
C n2
...
C n70
C s1
...
C s10
C m1
...
C m10
C l1
...
C l10
0
0
0
0.2
0.2
0.5
0.5
0.8
0.8
f
1
1
0.8
0.8
0.8
0.5
0.5
Y1
Y2
Y3
Y4
Y5
Y6
1
1
1
1
1
1

Simulation Results





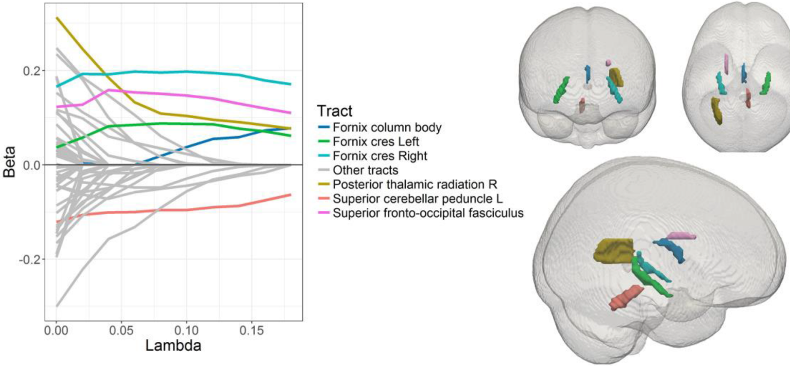


Figure 9: In the final model six non-zero tracts for this penalty are shown as individual colours (top left and top right panels) whereas the tracts regularized to 0 are shown in grey.

- Kep piece: Given simulation results, more confidence in these findings than if we used MLE.

Future

Regularization for Study Design

Study Design

In planning a study, we generally run a power analysis to determine the necessary sample size.
- Particularly challenging in the context of big data.

In contrast, secondary data analysis is becoming increasingly common.
- Here, **sample size is fixed**.

Pivot in aim: How to maximize *knowledge* extracted from data, *conditional* upon the sample size.
– Addditionaly, in big data we are chasing small effects.

Knowledge Extraction

One challenge with regularization: variable selection is dependent upon sample size and data quality.
- Data mining in general: Model selection is conditional upon data quality (reliability; Jacobucci & Grimm, under review).
- Larger the sample size = smaller penalty generally selected as *best* with the lasso.

Pairing regularization with cross-validation or information criteria allows researchers to maximally extract the most complex model that the data can afford.

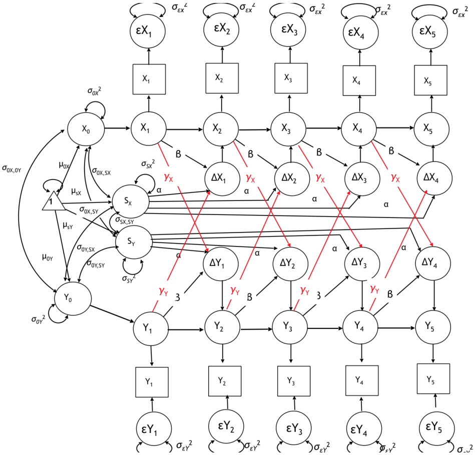
Analysis Steps

Given that sample size is fixed, we focus on the quality of the data.

Analysis Steps:

- Specify maximally complex model
  - Confirmatory: is data rich enough to fully evaluate?
  - Exploratory: Purely about extraction.
- Apply regularization to parameters of interest
  - SEM: directed paths
  - Theory testing: paths that are novel
- Select final model based on index that favors generalizability

Example


εX1
εX2
εX3
εX4
εX5
X1
X2
X3
X4
X5
X0
ΔX1
ΔX2
ΔX3
ΔX4
1
SX
SY
ΔY1
ΔY2
ΔY3
ΔY4
Y0
Y1
Y2
Y3
Y4
Y5
εY1
εY2
εY3
εY4
εY5
β
α
γX
γY
μ0X
μSX
μSY
μ0Y
σ0X,0Y
σ0X,SX
σ0X,SY
σSX,SY
σ0Y,SX
σ0Y,SY

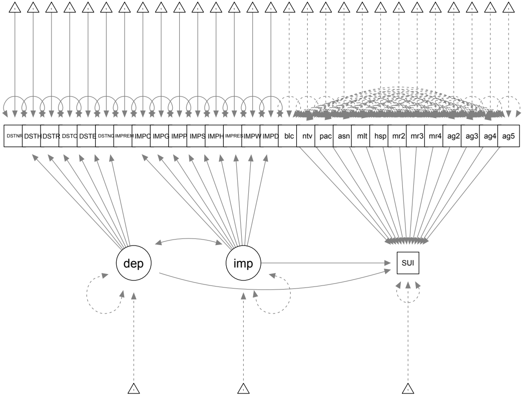
Large Sample Size

If the sample size is large enough, the optimal λ will often be chosen as 0.

Functional Sparsity:

- Goal is to reduce model complexity
  - Design short form scale for use in time-restricted context.
  - Confluence of small effects: Need better summarization.
- Choose λ based on aims of study.
  - Set number of desired indicators or predictors.
  - Increasing homogeneity of latent variables.
  - Increasing the explained variance of the latent variables.
- Fit is about ommission and may not be an appropriate metric in some studies.

Example


DSTNR
DSTH
DSTR
DSTC
DSTE
DSTNG
IMPREM
IMPC
IMPG
IMPP
IMPS
IMPH
IMPRES
IMPW
IMPD
blc
ntv
pac
asn
mlt
hsp
mr2
mr3
mr4
ag2
ag3
ag4
ag5
dep
imp
SUI

Stability Selection

Stability Selection

Main drawback of RegSEM Lasso: High False Positives.

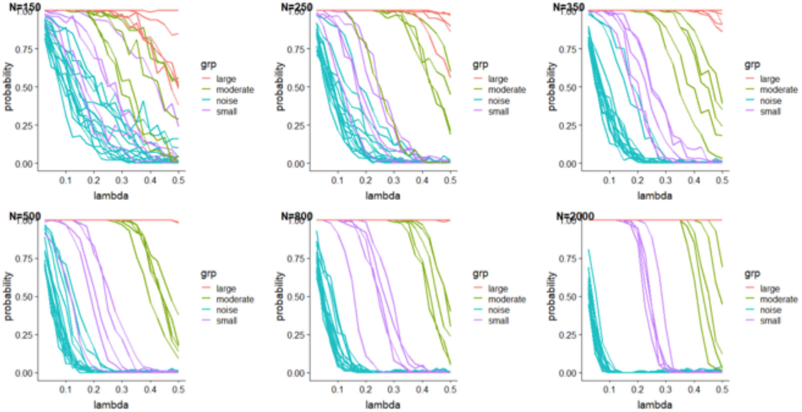
One proposal to address this in regression is stability selection (Meinshausen & Bühlmann, 2009).

Stability Selection:

- Subsampling + selection algorithms
- A general method, could have wide range of applications
- Provide finite sample control for false discovery rate
- Lasso regression + stability selection $\rightarrow$ consistent selection

Work is in preparation with Xiaobei Li
- Main takeaway (thus far) – Just bootstrapping doesn't work.

Preliminary Simulation



Handling Complex Data Features

Expanding Estimation

The `regsem` package uses covariance-based expectations, meaning that it can't handle:
- random effects
- missing data (FIML)
- multiple groups
- categorical indicators and outcomes
- sample weights
- mixtures
- others

Talk to me if you have an example that conforms to any of these: We're looking for good test cases.

`mxregsem` Package

In collaboration with Tim Brick and Joshua Pritikin, we are developing `mxregsem`:
Github repository: https://github.com/trbrick/mxregsem

Steps:

1. Specify base model in `OpenMx` (any type of model).
2. Use `regularizeMxModel` on base model.

Should be on CRAN by the end of the summer.

Truly Large SEM

SEM w/ 1000's of Variables

What if we wished to look at a more granular level, either in MRI or genetics research?

I have two main concerns:
1. Would we gain much by staying in SEM (vs. factor or summed scores)?
2. Will software allow us to run the models?

Current testing:
1. Mplus – ULS or Bayes will run, but no regularization
2. regsem – 100-200 variables possible.
3. lslx – $<$ 50 variables
4. JAGS – will run, but slow
5. Neural Net software – will work, but not enough testing done.

Software Idea #1: Stan

`Stan` blends automatic differentiation with MCMC to perform more efficient sampling of complex models.

Can use either matrix or factor score based SEM for model specification. - Pair with sparse priors

Drawback: Speed
Advantage: P $>$ N & flexibility

Haven't tested, but should be comparable: PyMC3

Software Idea #2: Tensorflow

van Kesteren & Oberski (2019): https://arxiv.org/abs/1905.04492

Uses matrix based SEM translated to `tensorflow`: a general neural network architecture.
- involves automatic differentiation = more general model specification.

Can also test through `greta` in `R`: provides easy to use interface to tensorflow.

Main challenge: **optimization**
- Regularization + model constraints (SEM) = difficult to find optima

SEM w/ 1000's of Variables – Interactions

In the Regularized MIMIC Model project, the MIMIC model had 48 predictors. What if we wanted to test for all possible interactions?

This has recently been established in lasso regression, termed the *hierarchical lasso* (e.g. Bien, Taylor, & Tibshirani, 2013).

Two proposals for SEM:
1. Two-step procedure: Only test those identified as main effects (Strong Heredity).
2. Test all possible: 1,177 predictors (48 main + interactions)

Option #1 can be tested with any software.
Option #2 only Stan has been tested.

Thanks!

And special thanks to Jack McArdle, Kevin Grimm, Sarfaraz Serang, Rogier Kievit, Andreas Brandmaier, Xiaobei Li, & Tim Brick for collaboration on this work!

If you want access to any material, email me at rjacobuc@nd.edu or visit rjacobucci.com/publications