Center for Lifespan Psychology

The best of both worlds: Towards a synthesis of theory-based and data-driven modeling

Andreas Brandmaier
Big Data: Volume, Velocity, Variety

Cattell’s Big Data Cube
Big Data: Surprise
Theory-Based Modeling
Data-Driven Modeling
Two Cultures

There are two cultures in the use of statistical modelling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown.

[...] If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

-- Breiman, 2001
Can we use machine-learning-inspired approaches to modify our initial theory? Would ML allow us to select among variables those who are most “important”?
Typical Questions Asked

Given a theory/model:

• “How can we best explain the heterogeneity/uncertainty in our data?”

• “What subset of variables is most predictive about my outcome(s)?”
THEORY-DRIVEN MODELING
Structural Equation Modeling

- Express relationships among observed and latent variables
- Encompass a variety of linear models:
  - t Test
  - Regression
  - (repeated measures) ANOVA
  - Mixed-effect models
  - Mediation models
  - Multilevel models
SEM = Measurement + Structure
What can SEM do for you? (the practical description)

• A universal language to formalize your natural language hypotheses
• Moving your hypotheses from an item to a construct level
• A framework to specify and also test models of your hypotheses (model comparison)
• A one-to-one mapping between formal languages of SEM (matrix algebra, sets of equations, computer programs) to diagrams
Example: Frontal Lobe Thinning

- Neuroscience in Psychiatry Network Data
- 176 participants
- Mean age = 18.84, range 14.3-24.9
- Outcome: Frontal pole volume
- Q: Are there sex differences in cortical development?

Using the likelihood ratio test, we observe significant decreases in model fit by constraining the mean of frontal lobe volume at T1 to be equal across the sexes ($\chi^2_{\Delta} = 38.01$, df$_{\Delta} = 2$, $p < 0.0001$). Inspection of parameter estimates shows, unsurprisingly, greater FP volume in males, compatible with either larger brains, delayed cortical thinning, or a combination of the two. Contrary to expectations, constraining the intercept of the change scores did not lead to a significant decrease in fit ($\chi^2_{\Delta} = 0.31889$, df$_{\Delta} = 2$, $p = 0.57$), indicating an absence of reliable differences in cortical thinning. However, constraining the variance of change scores to be equal did result in a significant drop in fit ($\chi^2_{\Delta} = 49.319$, df$_{\Delta} = 2$, $p < 0.0001$), with males showing greater individual differences in rates of thinning than females. Finally, constraining the covariance between frontal volume and change scores also led to a drop in model fit, with males showing a stronger (negative) association between volume at T1 and rate of change (compatible with the hypotheses of delayed development in males).

Figure 9: NSPN: Differential variability in frontal lobe thinning. Panel A shows longitudinal development in frontal structure. Panel B shows the model fit for the best model. Where parameters are different between groups we show male estimates in blue, female in red. Panel C shows the AIC and BIC of the free versus constrained models – in all cases only one parameter is constrained to equality and compared to the 'all free' model. Panel D shows the left and right frontal poles of the neuromorphometrics atlas used in our analysis. See Appendix B for more details on the imaging pipeline.

Kievit et al., Dev Cog Neuro, 2018
Example: Frontal Lobe Thinning

Scaling factor = .983. Next, we explored which (if any) of the four parameters above differed between the sexes. If a parameter is different between the groups, constraining it to be equal should result in a significant decrease in model fit. Using the likelihood ratio test, we observe significant decreases in model fit by constraining the mean of frontal lobe volume at T1 to be equal across the sexes ($\chi^2_{\Delta} = 38.01$, df$_{\Delta} = 2$, p < 0.0001). Inspection of parameter estimates shows, unsurprisingly, greater FP volume in males, compatible with either larger brains, delayed cortical thinning, or a combination of the two. Contrary to expectations, constraining the intercept of the change scores did not lead to a significant decrease in fit ($\chi^2_{\Delta} = 0.31889$, df$_{\Delta} = 2$, p = 0.57), indicating an absence of reliable differences in cortical thinning. However, constraining the variance of change scores to be equal did result in a significant drop in fit ($\chi^2_{\Delta} = 49.319$, df$_{\Delta} = 2$, p < 0.0001), with males showing greater individual differences in rates of thinning than females. Finally, constraining the covariance between frontal volume and change scores also led to a drop in model fit, with males showing a stronger (negative) association between volume at T1 and rate of change (compatible with the hypotheses of delayed development in males).

Figure 9: NSPN: Differential variability in frontal lobe thinning. Panel A shows longitudinal development in frontal structure. Panel B shows the model fit for the best model. Where parameters are different between groups we show male estimates in blue, female in red. Panel C shows the AIC and BIC of the free versus constrained models – in all cases only one parameter is constrained to equality and compared to the ‘all free’ model. Panel D shows the left and right frontal poles of the neuromorphometrics atlas used in our analysis. See Appendix B for more details on the imaging pipeline.

McArdle & Hamagami 2001, Kievit et al., Dev Cog Neuro, 2018
Surprise in Theory-Based Modeling

• Likelihood of data given model = surprise under data from the true distribution ("cross-entropy")

• A well-fitting data set is not surprising (most published results are not surprising in an information-theoretic sense)

• In theory-based modeling, we hope for low surprise!

Kievit et al., Dev Cog Neuro, 2018
Commercial Break: Ωnyx

Von Oertzen, Brandmaier, & Tsang, 2013
DATA-DRIVEN MODELING
The Decision Tree

Gigerenzer & Kurzenhäuser 2005, Green & Mehr 1997
Example: Should I go to this conference?

<table>
<thead>
<tr>
<th>Food quality</th>
<th>Scient. quality</th>
<th>Fees</th>
<th>Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>high</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>low</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>high</td>
<td>sad</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>low</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>low</td>
<td>happy</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>low</td>
<td>sad</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td>happy</td>
</tr>
</tbody>
</table>

Total Entropy = 0.56

Big Data in $\Psi$ 2018
**Example: Should I go to this conference?**

<table>
<thead>
<tr>
<th>Food quality</th>
<th>Scient. quality</th>
<th>Fees</th>
<th>Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>high</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>low</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>high</td>
<td>sad</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>low</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>low</td>
<td>sad</td>
</tr>
</tbody>
</table>

**Fees**

Total Entropy = 0.56

Entropy | high fees = 0.50
Entropy | low fees = 0.64

Entropy | fees = 0.55
Example: Should I go to this conference?

<table>
<thead>
<tr>
<th>Food quality</th>
<th>Scient. quality</th>
<th>Fees</th>
<th>Mood</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>high</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>low</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>high</td>
<td>sad</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td>happy</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>low</td>
<td>happy</td>
</tr>
<tr>
<td>low</td>
<td>low</td>
<td>low</td>
<td>sad</td>
</tr>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td>happy</td>
</tr>
</tbody>
</table>

Food

Entropy=0.56

Entropy | nice food=0
Entropy | bad food=0.64
Entropy | food=0.24
Decision Tree

Food

- great
  - Happy
- bad
  - Science
    - great
      - Happy
    - bad
      - Sad
A Geometric Perspective
(Toy Example)

Prediction of Happiness

Adapted from Brandmaier et al., book chapter, 2013
A Geometric Perspective (Toy Example)

Prediction of Happiness

Adapted from Brandmaier et al., book chapter, 2013
Interim Summary: Decision Trees

• **Reduce surprise** (by maximizing information gain)
• Using a greedy **heuristic**
• No distributional assumptions
• **Straightforward to understand** (equivalent to IF-THEN rules)
• Complexity depends on the data
• Uncover **non-linear** influences of predictors
• **Invariant to scaling** of predictors
• Detection of **complex interactions**
Why Predictive Modeling?

1) Uncover **surprising** causal/correlational associations and lead to the **generation of new hypotheses**.

2) Predictive models may serve as a **reference point** to existing explanatory models.

3) Predictive modeling can **suggest improvements** to existing explanatory models.
What if...

..we combined SEM and decision trees?
A Simple Example: Wechsler Intelligence Scale for Children

Brandmaier et al., Psychol Methods, 2013
A Simple Example: Wechsler Intelligence Scale for Children

Brandmaier et al., Psychol Methods, 2013
Split Candidate: Sex

Brandmaier et al., Psychol Methods, 2013
Split Candidate: Mother’s Education

Brandmaier et al., Psychol Methods, 2013
Two-Level Tree

Brandmaier et al., Psychol Methods, 2013
A Probabilistic Perspective
A Probabilistic Perspective
Likelihood Ratio Splitting = Surprise Minimization

$\chi^2$ distribution with df=5 and alpha=.05

$H_0$: "Split is uninformative. Information gain is zero"

mother's education

- No graduation
- High school graduation

verbal performance

$6 \ 7 \ 8 \ 9 \ 10 \ 11$

$10 \ 30 \ 50 \ 70$
TERMINAL DECLINE IN WELL-BEING
Terminal Decline in Well-Being

Gerstorf et al., Psychol Aging, 2016
Quadratic Growth Curve Models
Example: Terminal Decline in Well-being using SOEP

- 4,404 now-deceased participants of the nationwide German SOEP (age at death: M = 73.2 years; 17-102 years; SD = 14.3 years; 52% women)
- Terminal decline, all available observations obtained in the last 10 years of life realigned along a time-to-death time metric
- Outcome: “How satisfied are you currently with your life, all things considered?”, 11-point scale
- Predictors: socio-demographic (e.g., age at death, education, religion), health and burden (e.g., disability, unemployment, divorce), psychosocial (e.g., social participation, perceived control, life goals).
First Two Levels of the Well-Being Tree

Social Participation

Low  High

Disability  Hospital

No  Yes  No  Yes

Brandmaier et al., Dev. Psychol., 2017
Problems in Interpretation

• What about variables not in the tree? Important or unimportant?
• Single trees may be unstable. Slight changes may drastically change their structure
A Resampled Well-Being Tree

Disability

Success Goals

Family Goals

30 35 40 45 50 55

Distance − to − death

Life Satisfaction − 9 − 8 − 7 − 6 − 5 − 4 − 3 − 2 − 1 0

30 35 40 45 50 55

Distance − to − death

Life Satisfaction − 9 − 8 − 7 − 6 − 5 − 4 − 3 − 2 − 1 0

30 35 40 45 50 55

Distance − to − death

Life Satisfaction − 9 − 8 − 7 − 6 − 5 − 4 − 3 − 2 − 1 0

A Resampled Well-Being Tree
SEM FORESTS
SEM Forests

• Interpretation of variables in a tree as “important” is impeded by instability of trees (small changes in the sample may lead to different trees, e.g., two almost equally strong predictors)
• How to assess predictors left out (e.g., when strongly correlated with in-tree predictors)?

Remedy (SEM Tree + Random Forests):
• Draw bootstrap or subsamples from original datasets
• Fit a SEM Tree to each sub sample
• Draw subset of predictors when determining the next splitting variable

Brandmaier, Prindle, McArdle, & Lindenberger, 2016
Permutation Importance

A simple scheme:

1. Compute likelihood of each tree on the OOB
2. Permute one predictor at a time
3. Compute drop in likelihood as measure of importance
4. Average over all trees
Bootstrapped Tree #3 and #294

Disability

Success Goals

Family Goals

Success Goals

Education

Death Partner

Rather important

Rather not important
Variable Importance in Well-Being

- Disability
- Hospitalization
- Social Participation
- Perceived Control
- Social Goals
- Religion
- Unemployment
- Unemployment Partner
- Education
- Age at Death
- Sex
- Income Loss >1000
- Success Goals
- Disability of Partner
- Family Goals
- Death of Partner
- Hospitalization of Partner
- Income Loss >3000
- Divorce
- Death of Parent

Relative loss in fit
Summary: Variable Importance

• Generic, non-parametric approach (independent of method used) to assess importance
• VI summarizes main effects and interactions
• Does not require expensive re-training the forest

BUT
• Standard error depends on number of forests, so don't use NHST
• May have a bias for correlated predictors (but see conditional variable importance; Strobl et al., 2008)
Summary

SEM Trees and Forests

– combine **model-based** and **data-driven** modelling
– are tools to **recursively** identify **sub groups** and their predictors in the data
– explain **heterogeneity** in a sample
– by **reducing surprise** (increasing information)
– discover differences both on the **construct level** and on the **measurement level**
Summary

SEM Trees and Forests as a hybrid of two modeling cultures allows us:

• **Challenge established models** when comparing predictive accuracy *(hold out set!)*.

• Tree/forest may lead to a **revision of the substantial theory** and the formulation of a new parametric model and/or experiment

• Conclusion that **postulated model applies only to a limited range of subjects**
Idiographic versus Nomothetic

- May be useful as an approach to “develop the neglected territory between idiographic and nomothetic analytic approaches.”

(see Singer, Ryff, Carr, and Magee, 1998)
Caveats

Prediction ≠ Explanation

• No short-cut from data to theory or knowledge
• The model with best predictions may not be the true model
• Shmueli et al. (2010): parsimonious but less „true“ model can have a higher predictive validity than a „truer“ but more complex model, particularly when
  – Data are noisy
  – When the true effects of the left-out variables are small
  – Sample size is small
Caveats

Prediction ≠ Causation

• With correlational data: association learning and curve-fitting
• No causal claims
• No claims about (temporal) precedence of one predictor over the other
A Dark Prophecy

Reviewer #2’s (2012) damning verdict:

• „Psychologists, with their obsession with their theories, are going to find data mining hard to grasp.“

• „Data mining and SEM make strange bedfellows. Admittedly, bad data mining practices could take over-fitting to a new, horrific level.“
Overfitting / Generalization

The term “exploratory” is considered by many as less than an approach to data analysis and more a confession of guilt

— McArdle (2013)

“inductive-deductive spiral” is an essential process in a good science

— Cattell, 1966

“Confirm first, then explore!”

— McArdle (2013)
Overfitting / Generalization

- Make Observations
  What do I see in nature? This can be from one's own experiences, thoughts, or reading.

- Think of Interesting Questions
  Why does that pattern occur?

- Formulate Hypotheses
  What are the general causes of the phenomenon I am wondering about?

- Develop Testable Predictions
  If my hypothesis is correct, then I expect a, b, c, ...

- Refine, Alter, Expand, or Reject Hypotheses

- Gather Data to Test Predictions
  Relevant data can come from the literature, new observations, or formal experiments. Thorough testing requires replication to verify results.

- Develop General Theories
  General theories must be consistent with most or all available data and with other current theories.
THANKS!

Twitter @brandmaier
Website: http://www.brandmaier.de