Day 3: Missing Data in Longitudinal and Multilevel Models

by Levente (Levi) Littvay
Central European University
Department of Political Sciece
levente@littvay.hu

Multilevel and Longitudinal Models

- Longitudinal SEM (Latent Growth Curve)
  - Structural Equation Models
  - Most approaches that work with SEMs work
  - There are model size and identification issues
  - (Traditionally use) Direct Estimation
- Multilevel / Mixed / Random Effect Models
  - Pattern problems
  - Level problems
  - What to model and what not to model issues
  - (Traditionally use) Imputation

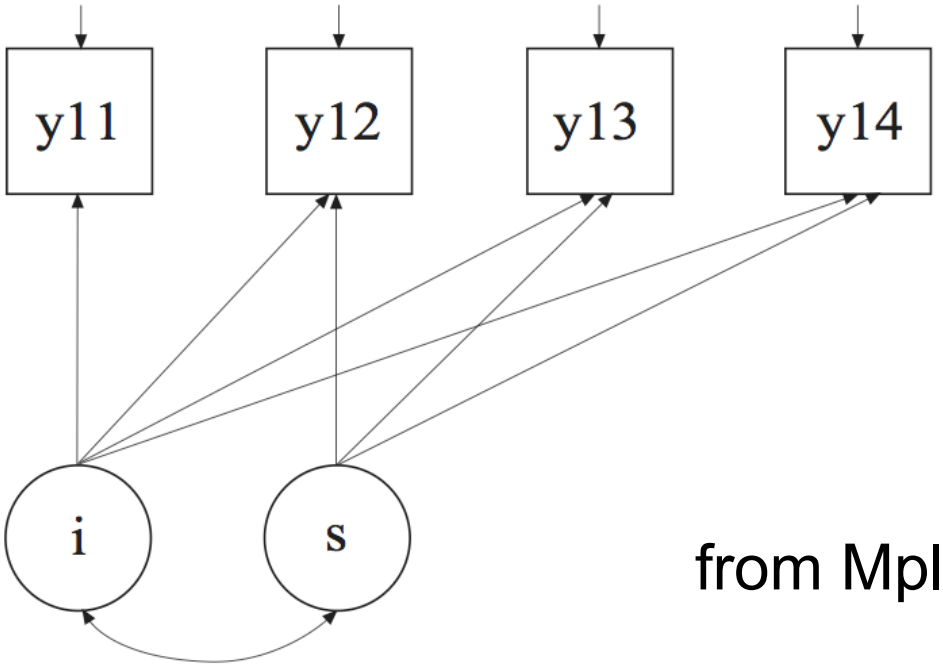
Missing Data in Longitudinal Structural Equation Models

Missing Data in SEMs

- Same approaches work
- Direct Estimation
  - More Common Approach
  - Missing can only be on the DV
    (usually not an issue with longitudinal models)
- Imputation
  - Can impute with an unstructured model
  - AMOS can impute using the analysis model
    (If no missing on the exogenous variables)

Longitudinal SEM

- Example - Latent Growth Curve
- It is just a structural equation model
- All observed variables are DVs



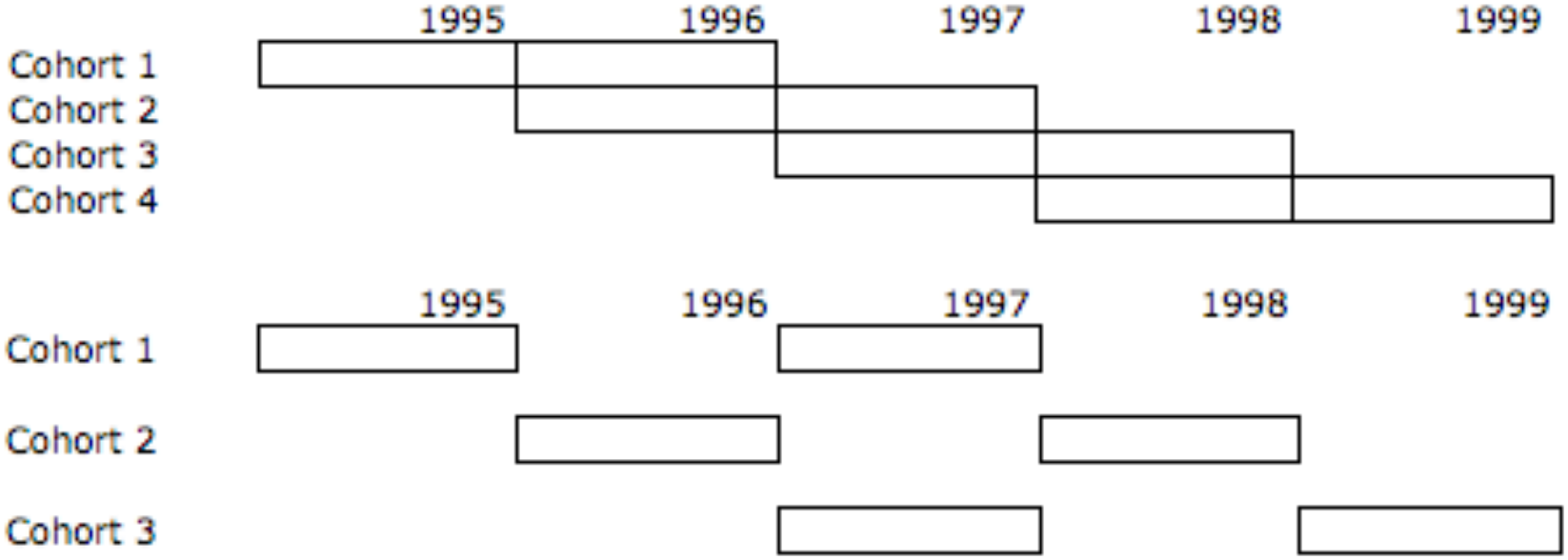
from Mplus Manual (ex 6.1)

Auxiliary Variables

- Just include them as you would otherwise
  - MI: include them in the imputation model
  - Direct estimation: correlate them with each other and all other observed variables
- Practical Issues
  - Can get out of hand
    - Imputation: Convergence + Model Size
    - Direct Estimation: Model Size + Convergence
  - Identification issues correlation of ~1 is not a unique information in the correlation matrix
  - Could collapse (if it still informs missingness)

Planned Missing

- Rolling Panel
  - You return to each person twice
  - You measure over a longer period of time
  - Can reduce panel effect



  - Always test power and convergence

Attrition

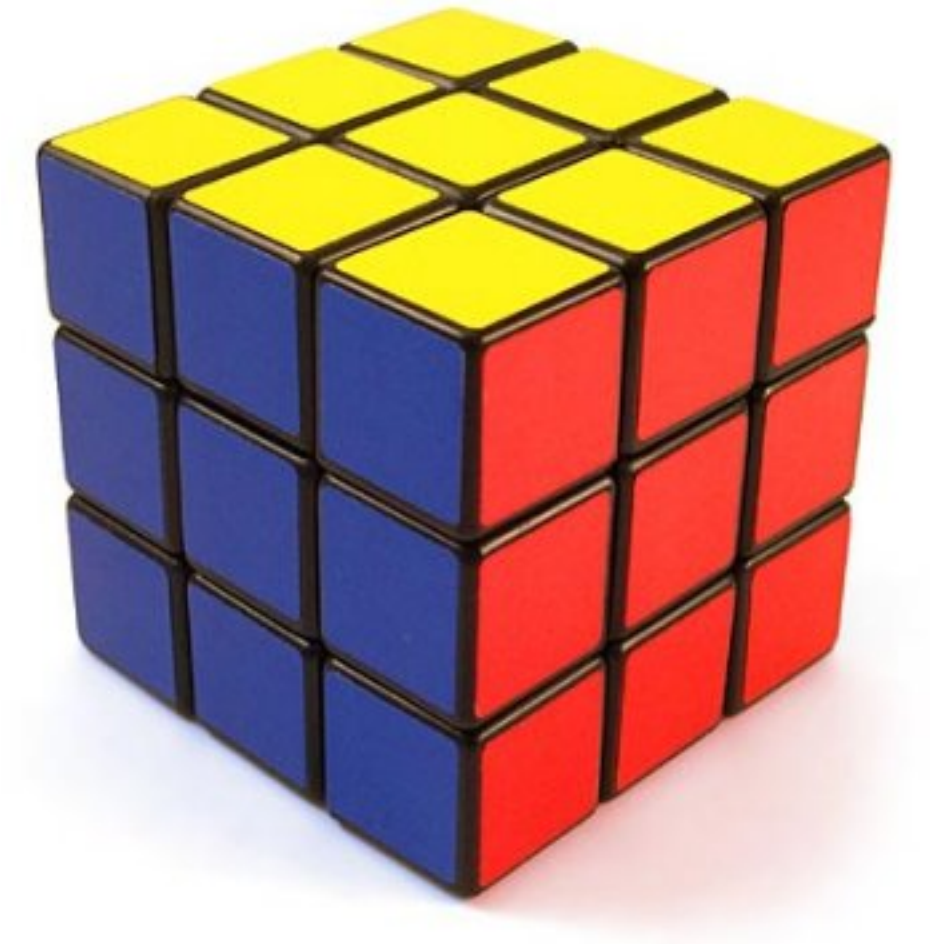
- If attrition is MAR you are fine
  - Ask questions like how likely are you to come back next time. etc.
- If not NMAR you are not fine

Extension of the Heckman Model

- The analytical model is estimated simultaneously with the model of missingness
- Mplus Mailing List (Moh-Yin Chang - SRAM)
- Model Dropout (with a Survival Model) simultaneously with the Longitudinal Model
- Let Residuals Correlate
- Pray that it Runs

Multilevel Models

Stacked Dataset Patterns



Example (My Dissertation)

- Over time data on 186 countries (1984-2004)
- Item Missing (Hungary Trade Volume 1991)
- A variable missing for a whole country
  (Had corruption data for 143 countries.)
- No data at all on Afghanistan, Cuba and North Korea (Unit Missing?)
- No data on energy consumption for 2004
- No data on West Germany after 1989
  (Should that even be treated as missing?)

MLM Missing Data

- You are OK with MAR missing on the DV
- You are OK with MAR wave missing
  - But if you have any information on the wave it will not be incorporated in the model
  - It is better to incorporate all info to help satisfy the MAR assumption

Multiple Imputation for Multilevel Models

MLM Imputation Procedures

- OK for Level 1 Missing Data
  - PAN (Schafer, Bayesian, S-Plus/R module)
  - MlWin (Implemented Schafer's PAN - Better)
  - WinMICE (Chained Equations)
  - Amelia II (Not true multilevel model)

- Upcoming: Shrimp (Yucel)

Imputation Model (Level 1)

- Thinking about the missing data model for multilevel models. (Conceptually Difficult)
  - Conventional Wisdom: Missing data model should be the same as the analysis model plus auxiliary variables.
  - Unstructured Model
- Issues
  - Inclusion of random effects for aux variables
  - Centering
  - Interactions

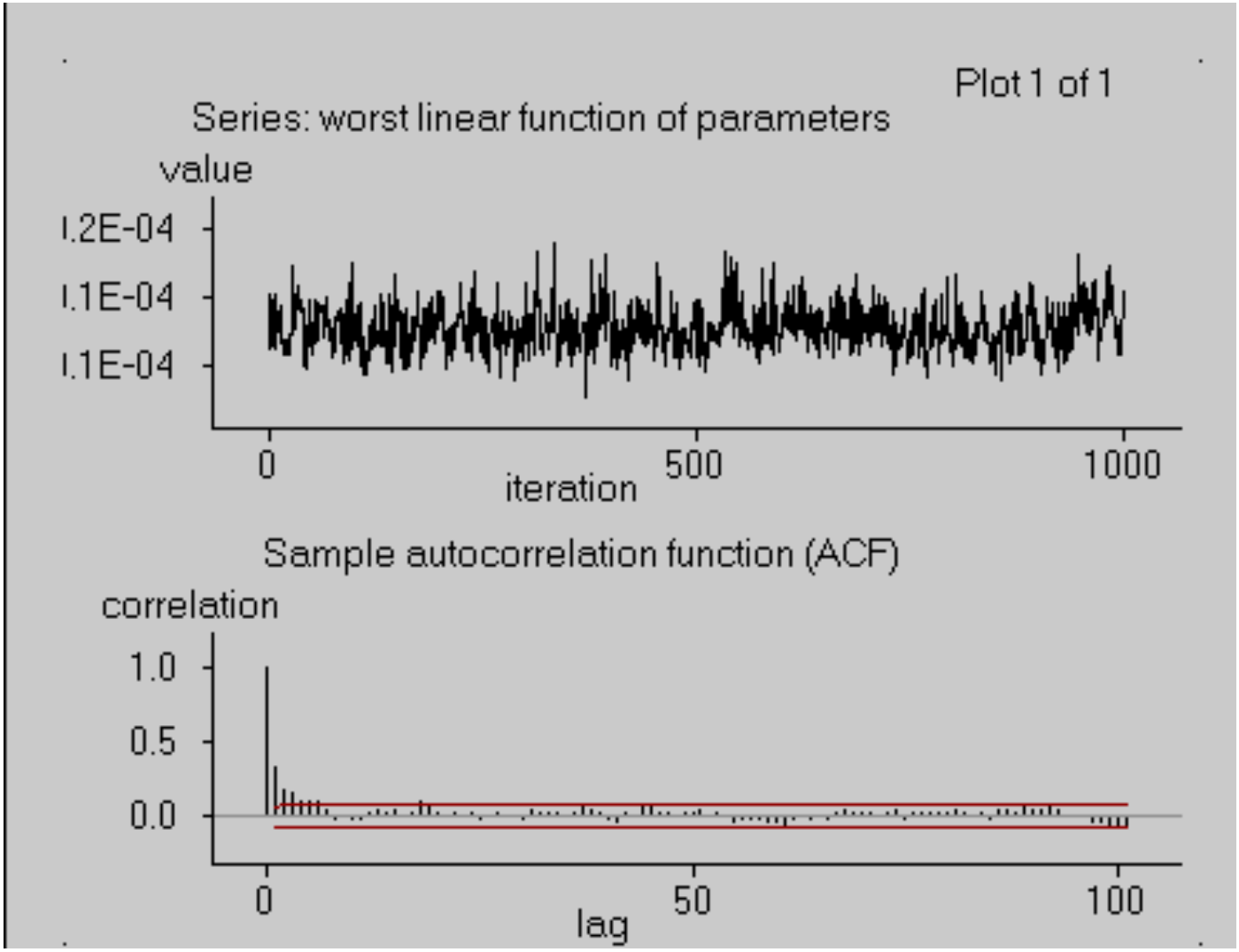
Bayesian Convergence

- Markov Chain Monte Carlo
- Random Walk Simulation
- Problem of autoregressive behavior
- Independent random draws produce the "posterior distribution" that imputations are sampled from.
- Bayesian convergence is in the eye of the beholder. No standard rules.

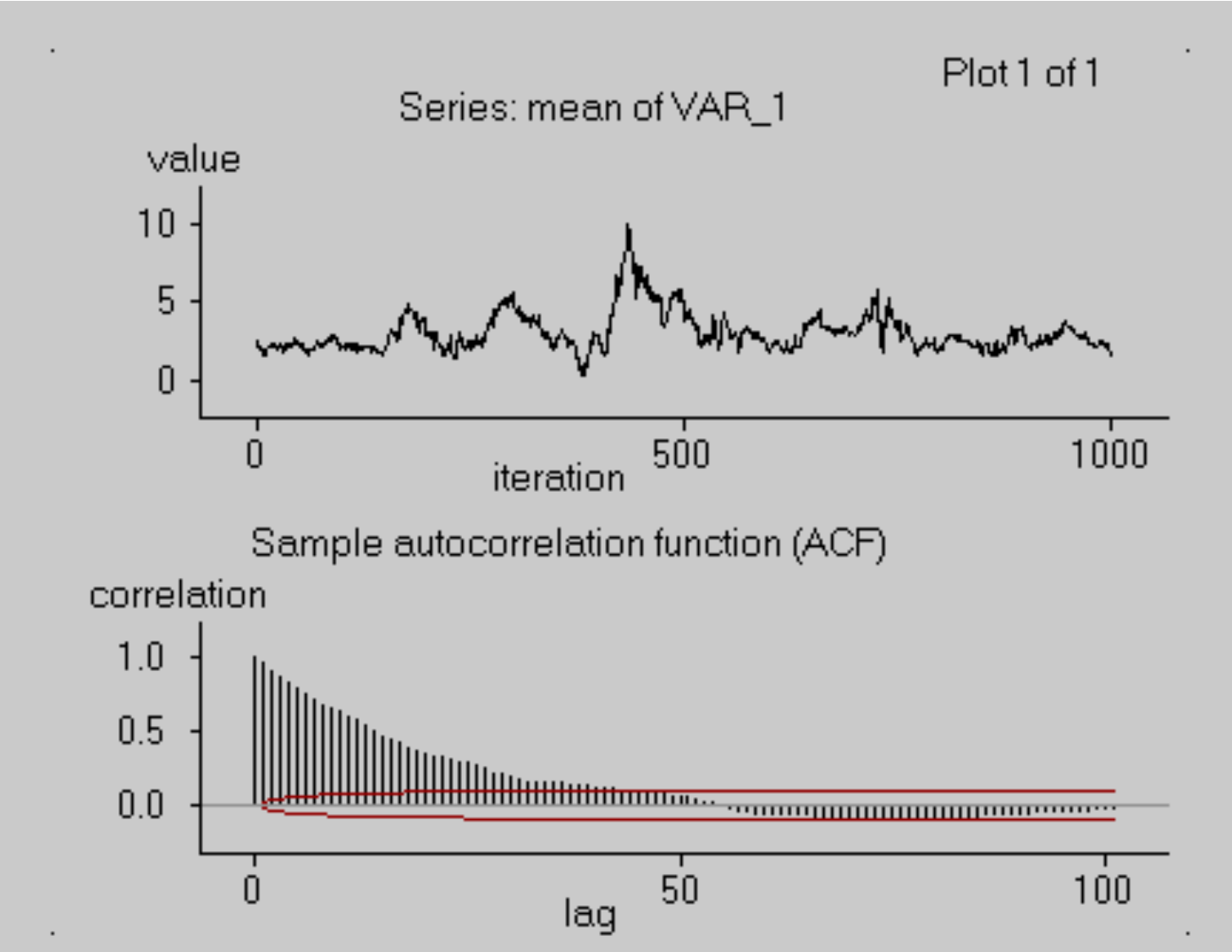
Ocular Shock Test of Convergence

- Well Implemented in MI software
- Has to be evaluated for all estimated parameters (this really sucks)
- Two Plots to Assess:
  - Parameter Value Plot
  - Autocorrelation Function Plot
- Be careful about the range of assessment
- Worst linear function - lucky if available

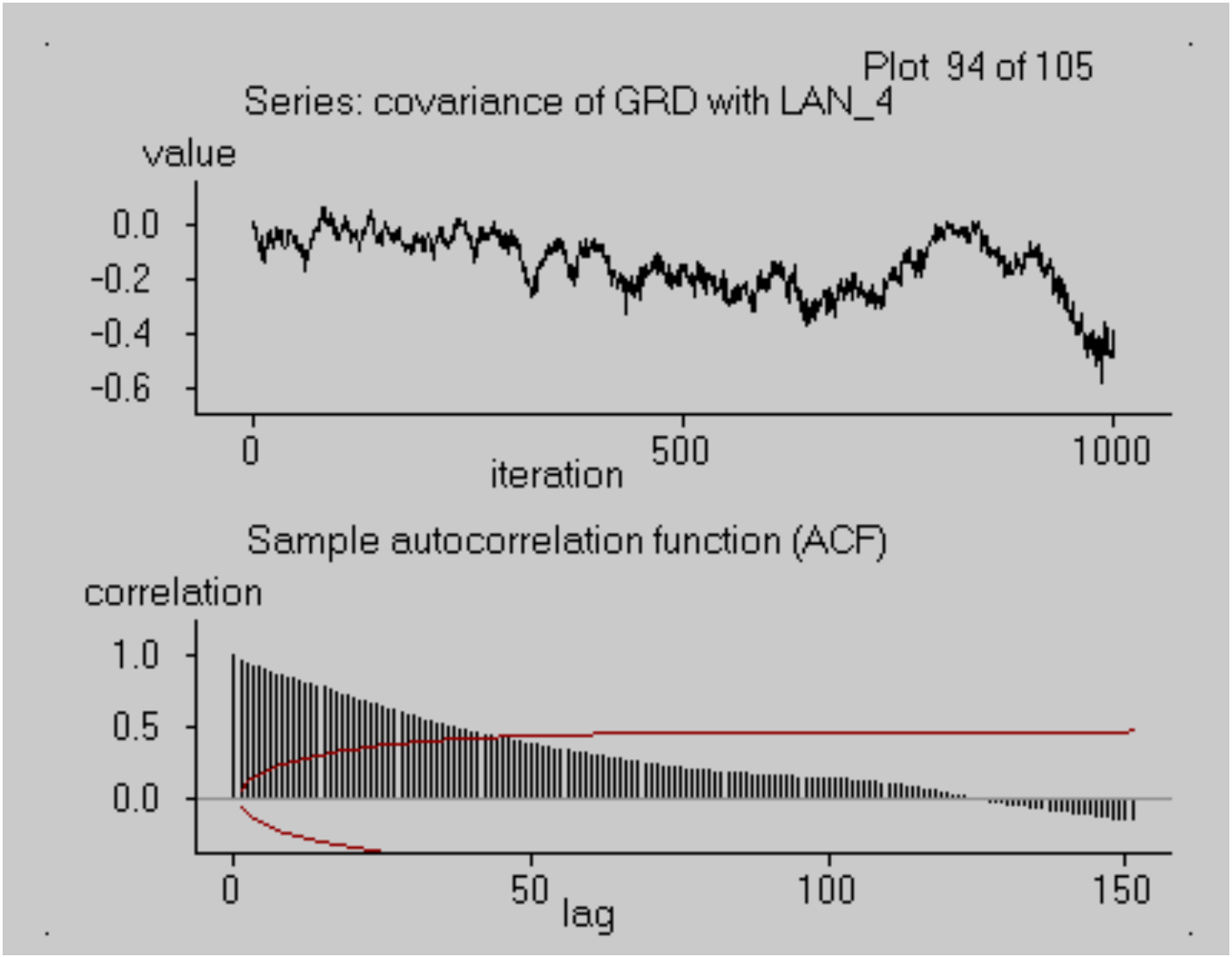
Quickly Converging Model



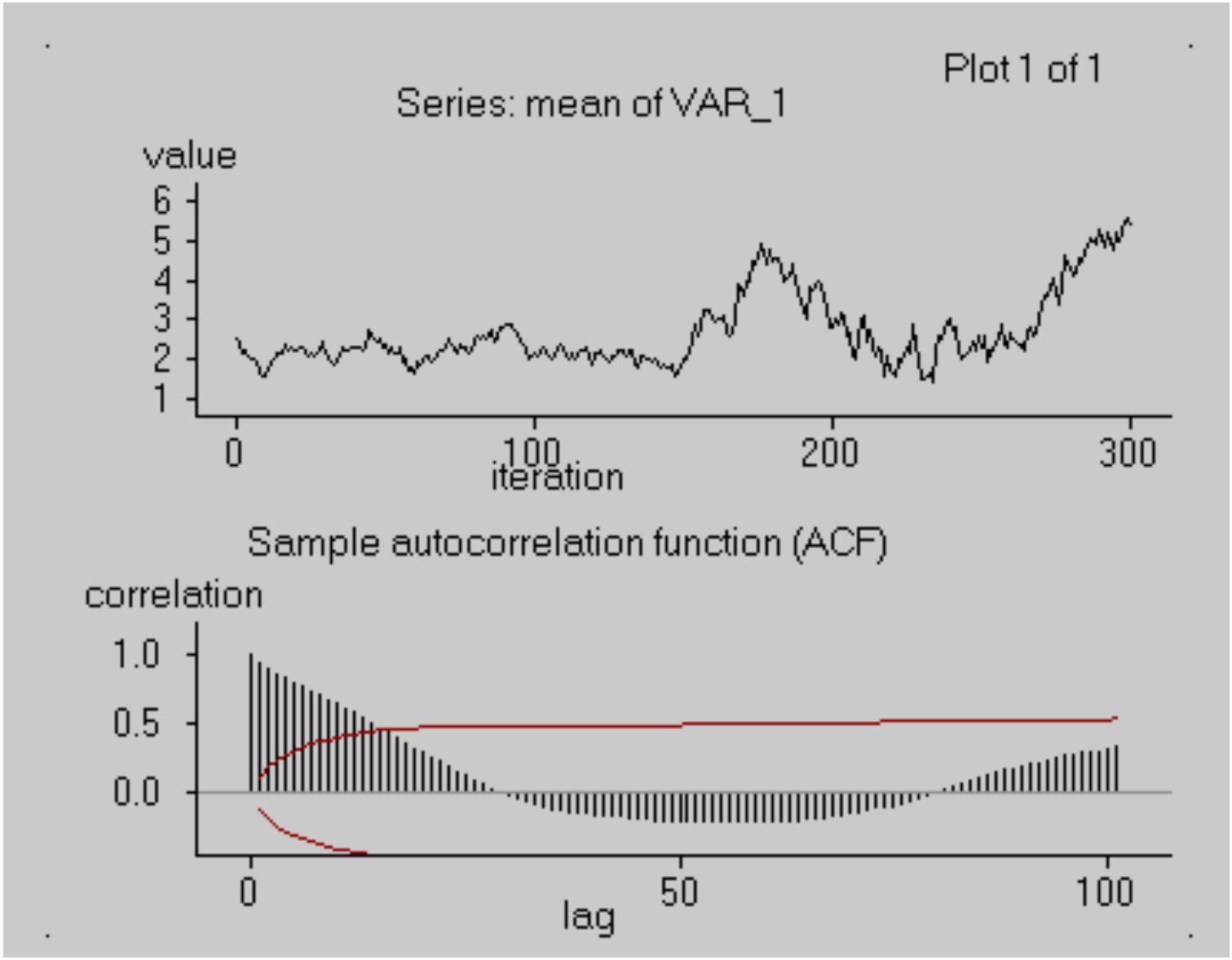
Slowly Converging Model



Pathological Situation
No Convergence



Did Not Yet Reach Convergence



Pseudo Multilevel Model

- Random Effect of the Intercept
  - Dummies for each level 1 unit (but one)
  - Pro: no distributional assumption of the variance of the intercept
  - Con: eats up degrees of freedom
- Random Effects of slopes
  - Interaction between the above dummy and the independent variable
  - Same pros and cons
- Same can be done with imputation model
  - Impact of ignoring random effects?

Level 2 missing (sucks)

- If you do Schafer suggests the following
  - Collapse your *level 1* variables by averaging across your *level 2* units
    This produces a single level dataset
  - Impute the single level dataset 10 times (Use a single level procedure)
  - Take the 10 *level 2* datasets remerge them with the *level 1* data (exclude?)
  - Impute level 1 missing once for each 10 using a multilevel imputation technique
- Assumptions of this approach (iterative?)

MI Support in Software

- HLM and Mplus
- Maybe Stata (clarify, micombine - ?,?)
- Maybe R (zelig - ?)
- MlWin can do imputation
  May also combine (possibly with hacking)

Rubin's Rules

- Combining results is still easy
- Use NORM like for single dataset
- One point of confusion is random effects
- But they also have parameter estimates and standard errors
- Combine like you combine coefficients and standard errors
- Don't forget about the error covariances

Direct Estimation of Multilevel Models

Direct Estimation of MLMs

- It is computationally intensive (requires numerical integration)
- Level 1 missing seems OK
- Missing IVs: make IVs into DVs
- Problem of auxiliary variables

Implementation

- In Mplus
  - Same as with SEM models
  - Multilevel SEM model
  - Downside: limited to unstructured error covariance matrix. (No AR1 band-diagonal)
- Mplus does level 2 missing with monte-carlo integration
  - Unstable
- MlWin's multilevel factor analysis (??)

Practical Considerations

- Getting good starting values
  - Really easy for most models
  - Run the model with all complete cases
  - Take results and use as starting values
  - Tedious, but worth it