

多変量データとパネルデータの 相関構造に関する注意と試み

統計数理研究所

リスク解析戦略研究センター

椿 広計

2011/06/22 LD2

ご朝食のお品書

- 食前酒
 - 多変量解析の家元争い
- 前菜
 - 相関構造の可視化に関する簡単な注意
- お料理
 - 相関行列の近似と逆相関行列の近似
 - 陸地、大陸棚、海溝の可視化
 - 潜在構造の測定と潜在構造からの影響
 - 飽和モデルと外的適合度
 - 探索的共分散構造モデリング
- デザート
 - 潜在成長曲線モデル
 - 個人的趣味の披露
- コーヒー
 - まとめ

0. 多変量解析におけるSEM派とGM派

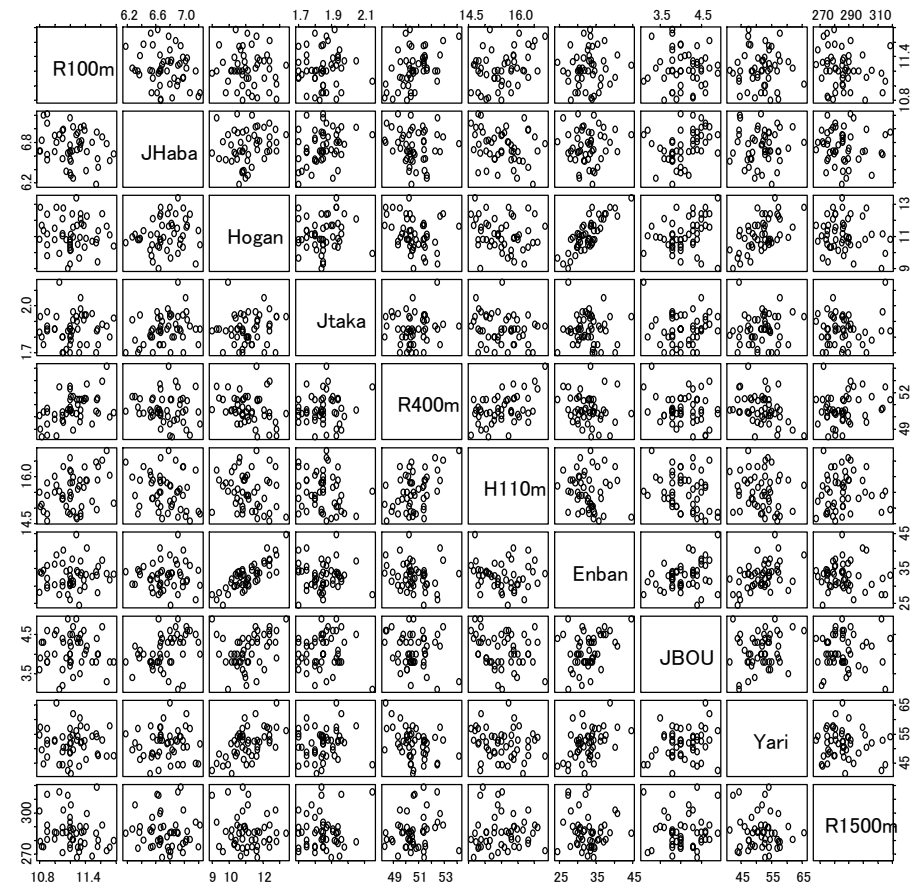
表千家と裏千家

- 相関観の違い
 - 親と子は似ている
 - 因果関係:共線性
 - Pearson, Yule
 - 兄弟は似ている→親の顔を見たい
 - 潜在因子による相関
 - Galton、Pearsonとその弟子

1. 表線形と裏線形の可視化

宮川雅巳(1997)の近代10種競技データ

- 多変量連関図で見る
連関性、すなわち相関
 - 今回の話は連関の
直線性は前提
 - 小生のノウハウ
RS学会誌連載講義
(98-00)
 - ノーマルスコアの相関
 - 順序相関の変換
- 相関係数は線形連関性の
表面的指標: 表線形
 - 裏線形→偏残差の世界

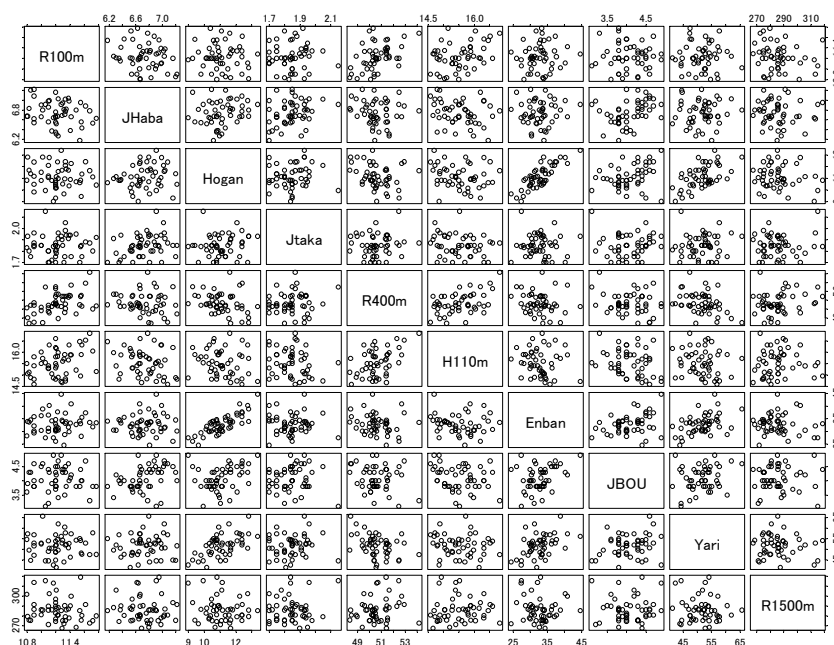


表と裏 $\Leftrightarrow$ 相関係数と偏相関係数

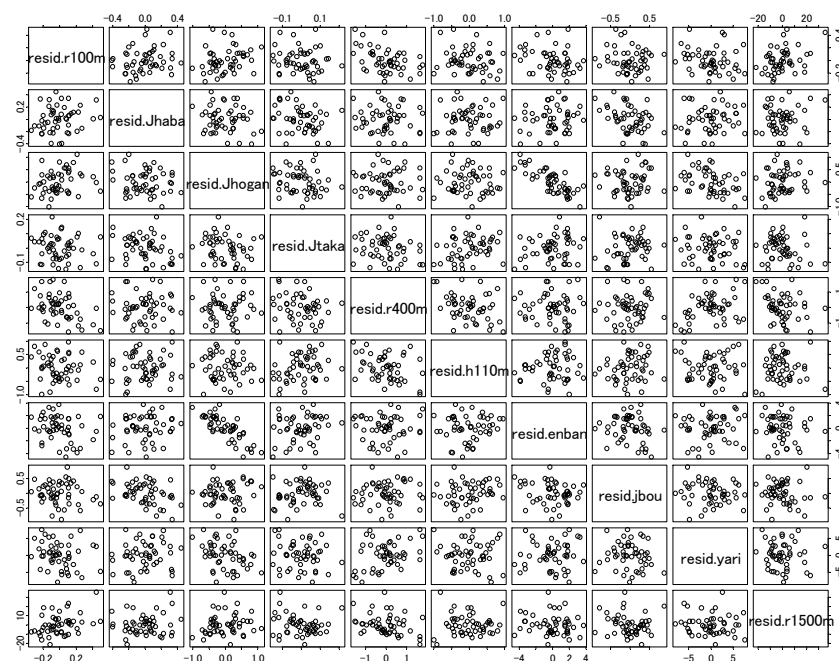
- 相関係数
 - 予測値ベクトルとして、各変量の平均値を全ての要素に代入
 - 残差ベクトルは、偏差ベクトル
- 偏相関係数
 - 説明変数ベクトル $\mathbf{A}$ の観測
 - $\mathbf{A}$ を与えたときの予測値 $E[X|\mathbf{A}]$, $E[Y|\mathbf{A}]$ を回帰分析などで求める
 - 残差ベクトルは、回帰分析の残差ベクトル
 - Y のあるべき値（予測値）の評価に、 X の情報も使うこともある： $E[Y|X, \mathbf{A}]$

続はじめの一步：裏の散布図行列

左：相関、表の線形構造、右：偏相関：裏の線形構造



	R100m	JHaba	Hogan	Jtaka	R400m	H110m	Enban	JBOU	Yari	R1500m
R100m	1.00000000	-0.2212658	-0.08048278	0.02374159	0.46714759	0.35349598	0.08178199	-0.024185659	0.002354373	-0.178333960
JHaba	-0.22126582	1.00000000	0.17530955	0.30633141	-0.11880247	-0.24715894	0.04206234	0.366095835	0.03206739	-0.052128480
Hogan	-0.080482784	0.17530955	1.00000000	0.11183797	-0.27943415	-0.18309876	0.82205357	0.286767544	0.568080842	-0.132397529
Jtaka	0.023741586	0.30633141	0.11183797	1.00000000	0.09540192	-0.19313381	-0.02306999	0.036756140	0.109432558	0.090328546
R400m	0.467147592	-0.11880247	-0.27943415	0.09540192	1.00000000	0.46484941	-0.17211757	-0.183989783	-0.378975351	0.205052015
H110m	0.353495981	-0.24715894	-0.18309876	-0.19313381	0.46484941	1.00000000	-0.12730684	-0.258411813	-0.157039448	0.030687538
Enban	0.081781985	0.04206234	0.82205357	-0.02306999	-0.17211757	-0.12730684	1.00000000	0.323853725	0.44508161	-0.080785475
JBOU	-0.024185659	0.36609583	0.28676754	0.03675614	-0.18398978	-0.25841181	0.32385372	1.00000000	0.230148636	0.007896979
Yari	0.002354373	0.03206739	0.56808084	0.10943255	-0.37897535	-0.15703945	0.4450816	0.230148636	1.00000000	-0.084705321
R1500m	-0.178333960	-0.05212848	-0.13239753	0.09032855	0.20505201	0.03068754	-0.08078548	0.007896979	-0.084705321	1.00000000

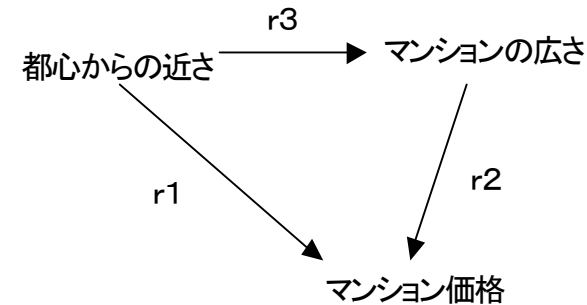


	resid.r100m	resid.JHaba	resid.JHogan	resid.Jtaka	resid.r400m	resid.h110m	resid.enban	resid.jbou	resid.yari	resid.r1500m
[1.]	1.0000000	0.18924452	0.244965184	-0.1333256	-0.444303915	-0.16947452	-0.244857290	-0.12585335	-0.20984347	0.36541182
[2.]	0.1892445	1.0000000	-0.15675722	-0.2862522	-0.044907759	0.04406201	0.139496918	-0.3803832	0.12294789	0.12994388
[3.]	0.2449652	-0.15675722	1.0000000	-0.1820716	-0.006798533	-0.06001828	-0.79908219	0.06440396	-0.36850430	0.16329777
[4.]	-0.1333256	-0.28625217	-0.18207164	1.0000000	-0.181916190	0.25553487	0.199982464	0.11425614	-0.12256078	-0.12885599
[5.]	-0.4443039	-0.04490776	-0.006798533	-0.1819162	1.0000000	-0.33637435	-0.005263093	0.06059999	0.35516086	-0.29885348
[6.]	-0.1694745	0.04406201	-0.060018276	0.2555350	-0.336374345	1.0000000	0.08899749	0.16048917	-0.05491616	-0.01180164
[7.]	-0.2448573	0.13949692	-0.79908219	0.1999825	-0.005263093	0.0889975	1.0000000	-0.18991455	0.08443561	-0.11278853
[8.]	-0.1258533	-0.3803832	0.064403968	0.1142561	0.005999993	0.16048917	-0.189914551	1.0000000	-0.09797739	-0.10951239
[9.]	-0.2098435	0.12294789	-0.368504296	-0.1225608	0.355160867	-0.05491616	0.084435606	-0.09797739	1.0000000	-0.09637360
[10.]	0.3654118	0.12994388	0.163297774	-0.1288560	-0.298853476	-0.01180164	-0.112788535	-0.10951239	-0.09637360	1.0000000

何故、偏相関構造に着目??

表の世界は夢のまた夢???

- 相関係数行列上で同様な操作可能! ?
 - － それでは実際的な解釈は出来ない! ?
 - 相関は分解して解釈
 - － 「直接効果」
 - － 「間接効果」
 - － 「擬似効果」
 - 分解した成分の合計が0になるのが、「無相関」
 - 偏相関係数が評価しているのは、根源的関連性としての「直接効果」
 - － 相関は、意味もなく足して0になる可能性
 - － 偏相関が0に近くなるのは、現象論的必然性の可能性



相関の分解の概念例：マンション価格と都心からの近さの相関 $=r_1+r_2\times r_3$
ただし、 r_1, r_2, r_3 は偏相関、 r_1 が直接効果、 $r_2\times r_3$ が間接効果

GMの成功要因：自然母数と裏線形

- 対数尤度関数

$$l(\mathbf{z}) = -\mathbf{z}^T \Sigma^{-1} \mathbf{z} / 2 + \log |\Sigma^{-1}|^{1/2} - \log(2\pi)^{1/2}$$

逆分散共分散行列（情報行列）の

i 行 j 列要素； σ^{ij}

– 自然母数は逆共分散

- 第 i 変数と第 j 変数の偏相関係数：

– $\sigma^{ij} / (\sigma^{ii} \sigma^{jj})^{1/2}$

- 多変量正規分布仮定の方便としての意義

– 一次、二次モーメントに関する最小情報量分布

2. 共分散行列を近似するのか、 逆共分散行列を近似するのか？

- 主成分分析・主因子分析の考え方
 - 乗法モデル(特異値分解)による共分散行列の最小二乗近似
 - 「主成分」や「因子」といった潜在変量ベクトル $\mathbf{Q}$ を観測変量の背後に導入
 - $\mathbf{Z}$ を $\mathbf{Q}$ に回帰し、残差ベクトルの共分散行列を最小化
 - » q 次元潜在ベクトル $\mathbf{Q}$ の共分散行列： Ω
 - » $\mathbf{Z}$ と $\mathbf{Q}$ の共分散行列： $\mathbf{B} \rightarrow$ 回帰残差の共分散行列： $\Sigma = \mathbf{B} \Omega^{-1} \mathbf{B}^T$
 - » Ω を単位行列と仮定し、通常 of 最小二乗基準で最小化
 $\rightarrow \Sigma$ は第 q 主成分までの固有空間を使って最良近似可能
 - 回転による単純構造 $\rightarrow$ パス係数の節約
- データの近似と共分散構造の近似との本質的な差異

伝統的主成分分析の問題 相関係数行列の近似？

- 相関行列の大きな固有値に対応する主成分(固有空間)への解釈の偏重
 - 相関行列のどのような近似に意味が有るか
 - GMの成功からの教訓
 - 逆相関(偏相関)行列の近似にも意味
 - 小さな固有値に対応する固有空間の重要性
 - 従来からも, 共線性と外れ値の発見(芳賀先生)
 - Mardia, Kent and Bibby(1979)主成分変数選択

Devianceと対数固有値プロットの効用

- 共分散行列、相関係数行列の近似の良さは？

$$\text{Dev} = \text{tr}(S\hat{\Sigma}^{-1}) - n \log |\hat{\Sigma}^{-1}| + \text{cons}$$

$$\text{Dev} = n \left\{ \text{tr}(\mathbf{R}\hat{\mathbf{P}}^{-1}) - \left(\log |\hat{\mathbf{P}}^{-1}| + \sum_{i=1}^p \log v_i \right) \right\} + \text{const} \quad ($$

固有値を定数倍する操作での

Devianceの増加は固有値の大きさに依存しない

対数固有値を定数だけずらす操作でも同様

共分散(相関)行列の第*i*固有ベクトル p_i , 固有値 λ_i

どの主成分を軽視すべきか？

- 主成分分析におけるケチなモデルとは？

- q個の主成分を解釈することの定義

- 真の共分散行列:

$$\Sigma_{true} = \sum_{i=1}^p \lambda_{(i)} \mathbf{p}_i \mathbf{p}_i^T$$

- 節約モデルとは: $\hat{\Sigma}_q = \sum_{i \in I} \lambda_{(i)} \mathbf{p}_i \mathbf{p}_i^T + \sigma^2 \sum_{i \notin I} \mathbf{p}_i \mathbf{p}_i^T,$

$$q = \#I$$

- λ : 固有値、 $\mathbf{p}$: 固有ベクトル
- Anderson(1984)参照

デヴィアンス最小化とは？

主成分分析の場合

- 固有値の変動係数あるいは、対数固有値の分散の小さいものを解釈しない主成分とする

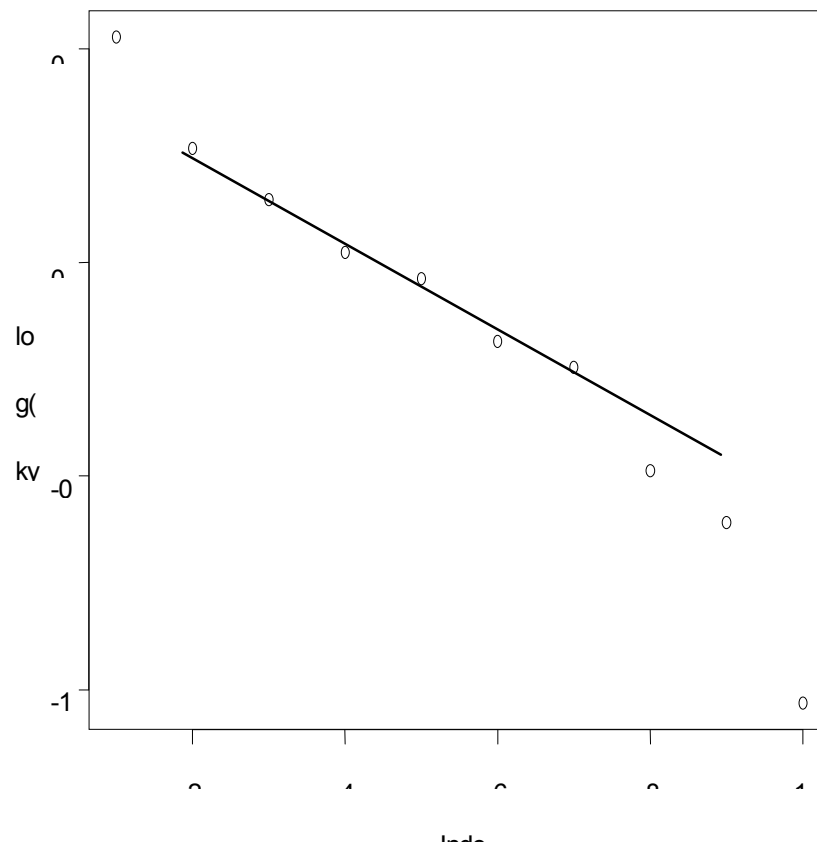
$$\log \left(\frac{\sum_{i=1}^n \lambda_i}{n} \right) - \frac{\sum_{i=1}^n \log \lambda_i}{n} = - \frac{\sum_{i=1}^n \log \lambda_i / \bar{\lambda}}{n}$$
$$\approx \frac{\sum_{i=1}^n (\lambda_i - \bar{\lambda})^2}{2\bar{\lambda}}$$

Devianceを最小にする主成分選択指針

- 固有値の大小は, 全く本質的でない!
- 固有値を大きさの順番に整列し,
対数固有値の変化の少ない部分を
解釈しないのが良い近似を与える
 - Relative Dispersionの最小化
 - 固有値の大きさを飛び越えて選択する必要はない
 - 小さい固有値を誤差空間→主成分分析
 - 大きい固有値を誤差空間→GM
 - 両者の統合が可能
 - 因子分析でも同様(計量行列がadaptiveとなる)

対数固有値プロット: 大陸, 大陸棚, 海溝

主成分分析における節約的モデルを求めて



- 第4固有値から第7固有値までは、ほぼ同じスピードで減少
- 3つの共線性
 - 第7-10固有値はそれより顕著なスピードで減少
- 3つの因子性
 - 第一固有値から第3固有値までの減少も若干スピードが速い
- 第4から第7までの固有値は、減少のスピードが緩やか
 → **これらの対数固有値が全て等しいと想定**しても、逸脱度増大は有意ではない
 - $n(4\log - \sum \log \lambda_i) = 4.70$ (自由度3)

途中固有空間節約モデルの意義 多変量解析におけるARMAモデル？

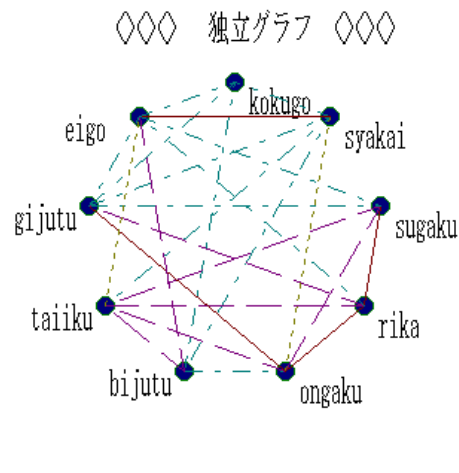
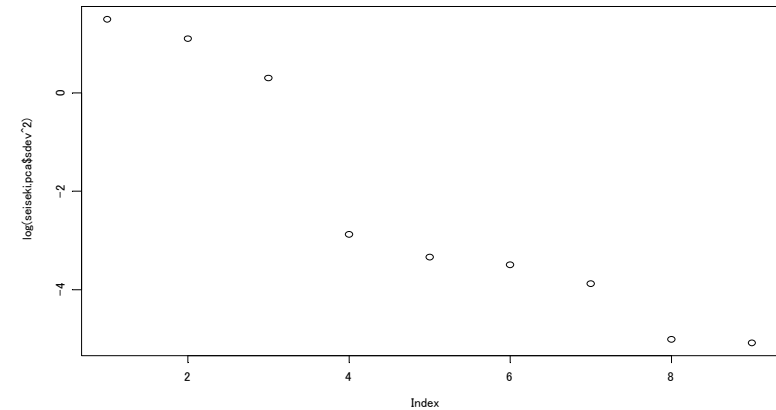
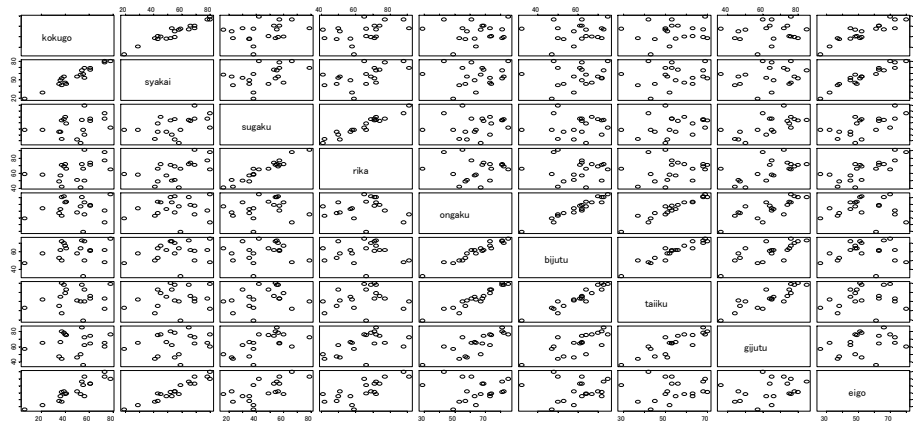
$$\mathbf{Z}^* = \mathbf{A}\mathbf{Z}^* + \mathbf{B}\mathbf{f} + \boldsymbol{\varepsilon}$$

多変量データの変動を3成分に分解

- 1) 共線性 (隠れた線形関係)
- 2) 因子で説明される相関構造
- 3) 個体差 (独自性) $\bar{\lambda}$

$$\left\{ \mathbf{I} + \sum_{i=8}^{10} \left(\sqrt{\frac{\bar{\lambda}}{\lambda_i}} - 1 \right) \mathbf{p}_i \mathbf{p}_i^T \right\} \mathbf{Z}^* = \sum_{i=1}^3 (\lambda_i - \bar{\lambda})^{1/2} \mathbf{p}_i f_i + \boldsymbol{\varepsilon}$$

グラフィカルモデルが不得意とするデータ 円川(1988)対数固有値プロットによる見極め



左下三角が偏相関係数の推定値, 右上三角が相関係数の差

n= 20 逸脱度= 6.925(df = 11) p=0.8051

GFI= 0.928 AGFI= 0.704 NFI= 0.983 SRMR= 0.007

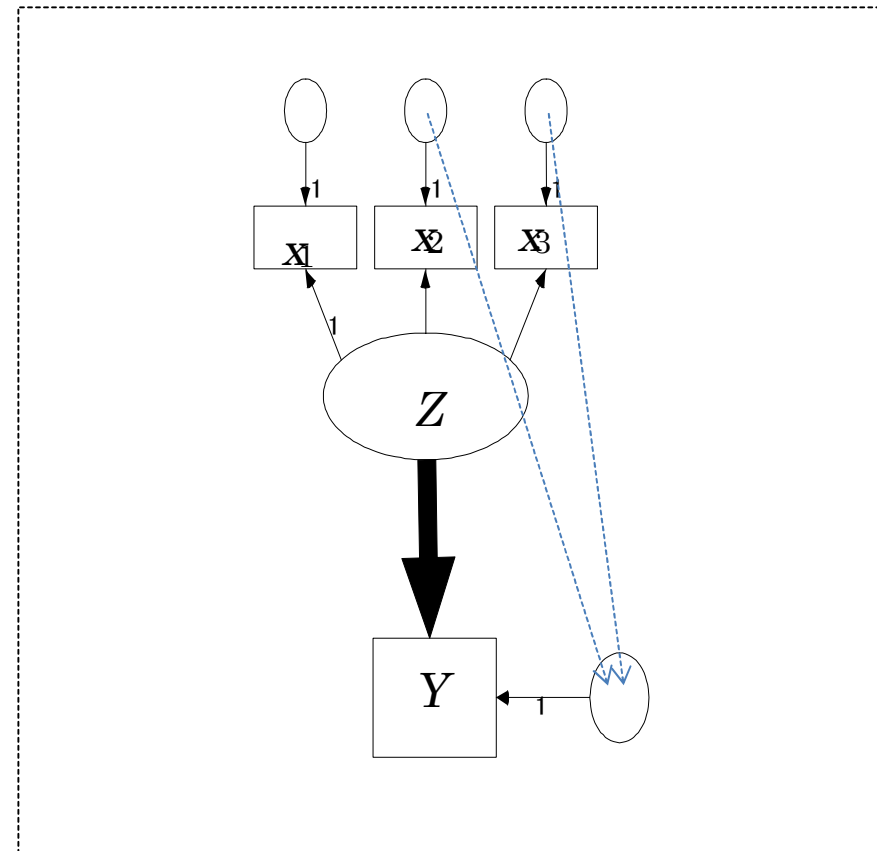
◇◇◇ 偏相関係数の推定 ◇◇◇

kokugo syakai sugaku rika ongaku bijutu taiiku gijutu eigo

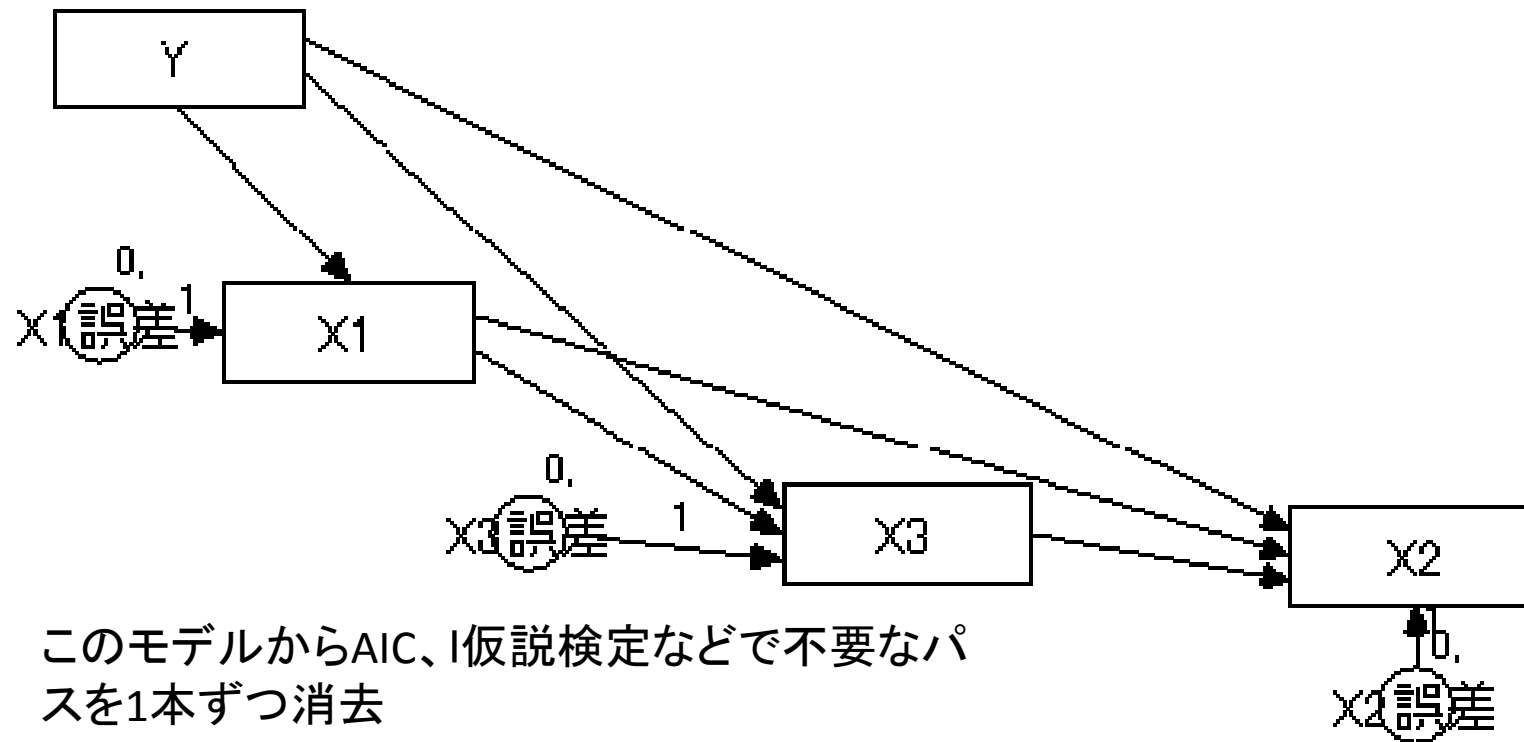
1kokugo	***		-0.027	-0.023	-0.000				-0.009
2syakai	0.362	***		-0.011	-0.011				
3sugaku	-0.001	-0.001	***						-0.015
4rika	-0.001	-0.001	0.895	***					-0.010
5ongaku	-0.001	0.170	0.558	-0.702	***				-0.002
6bijutu	0.213	0.332	-0.001	-0.001	0.331	***			0.011
7taiiku	-0.001	-0.280	-0.447	0.413	0.426	0.498	***		0.000
8gijutu	-0.258	-0.239	-0.243	0.549	0.634	-0.000	-0.000	***	
9eigo	0.221	0.775	-0.202	0.277	-0.000	-0.410	0.179	0.202	***

3. 潜在因子の測定と 潜在因子からの影響を分離する

- 検証的因子モデル当てはめにおける測定と因果との非分離
 - 結果変数を変えると潜在概念も変わる
 - 測定の不変性要請
 - 条件付き原理に違反
 - 因子得点に基づくパス解析を擁護
- 探索的共分散構造分析における飽和モデルの便利さ
 - パス解析におけるサイモンブレイロック手順

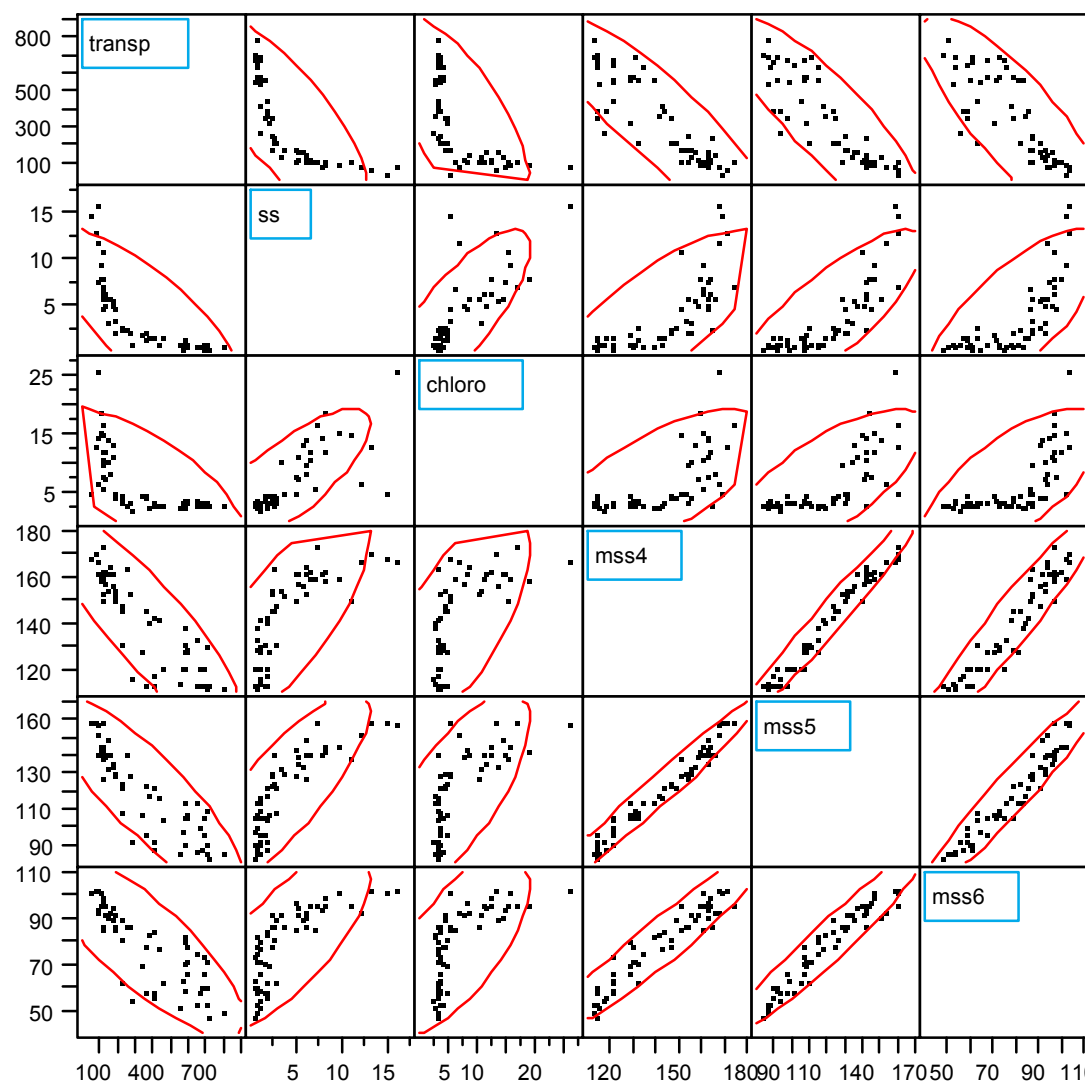


パス解析と飽和モデル： 完全逐次モデル



このモデルからAIC、t仮説検定などで不要なパスを1本ずつ消去
→サイモン・ブレイロック法
グラフィカルモデリングにおける共分散選択と同じ意義

散布図行列



多変量回帰モデルは飽和モデル

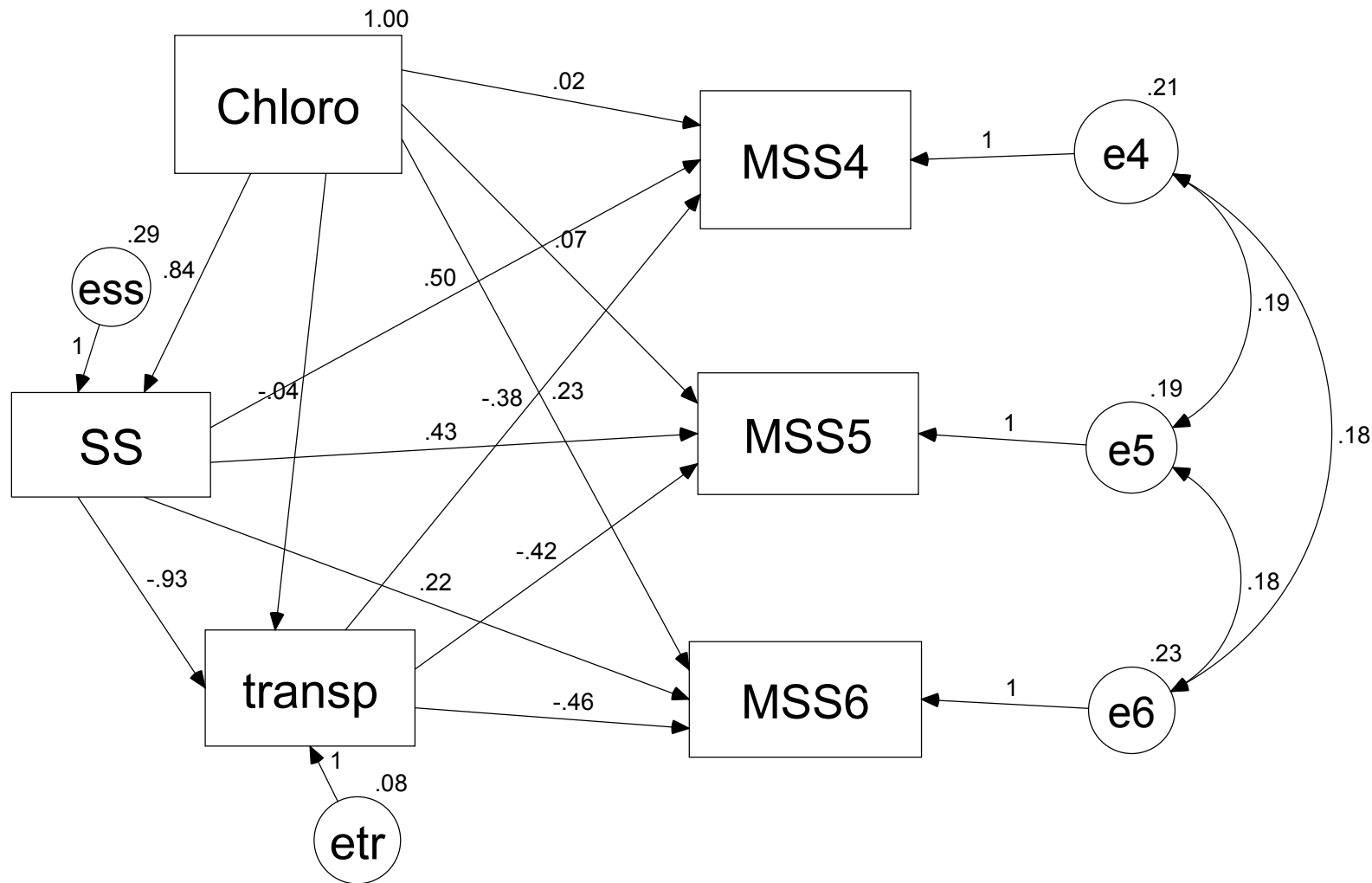


Fig.11 Estimated Causal Model for Remote Sensing Data

逐次パス消去後のモデル

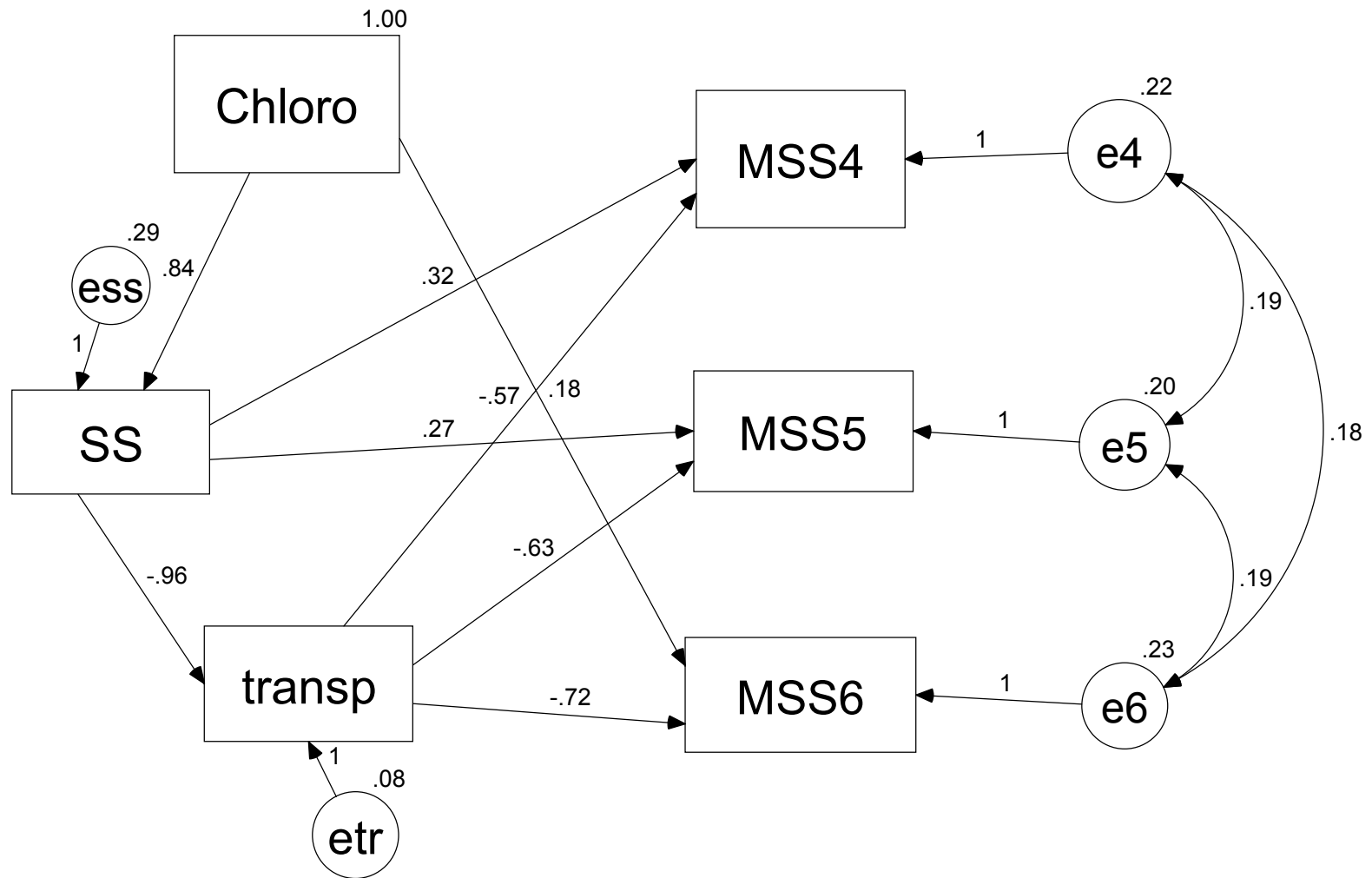
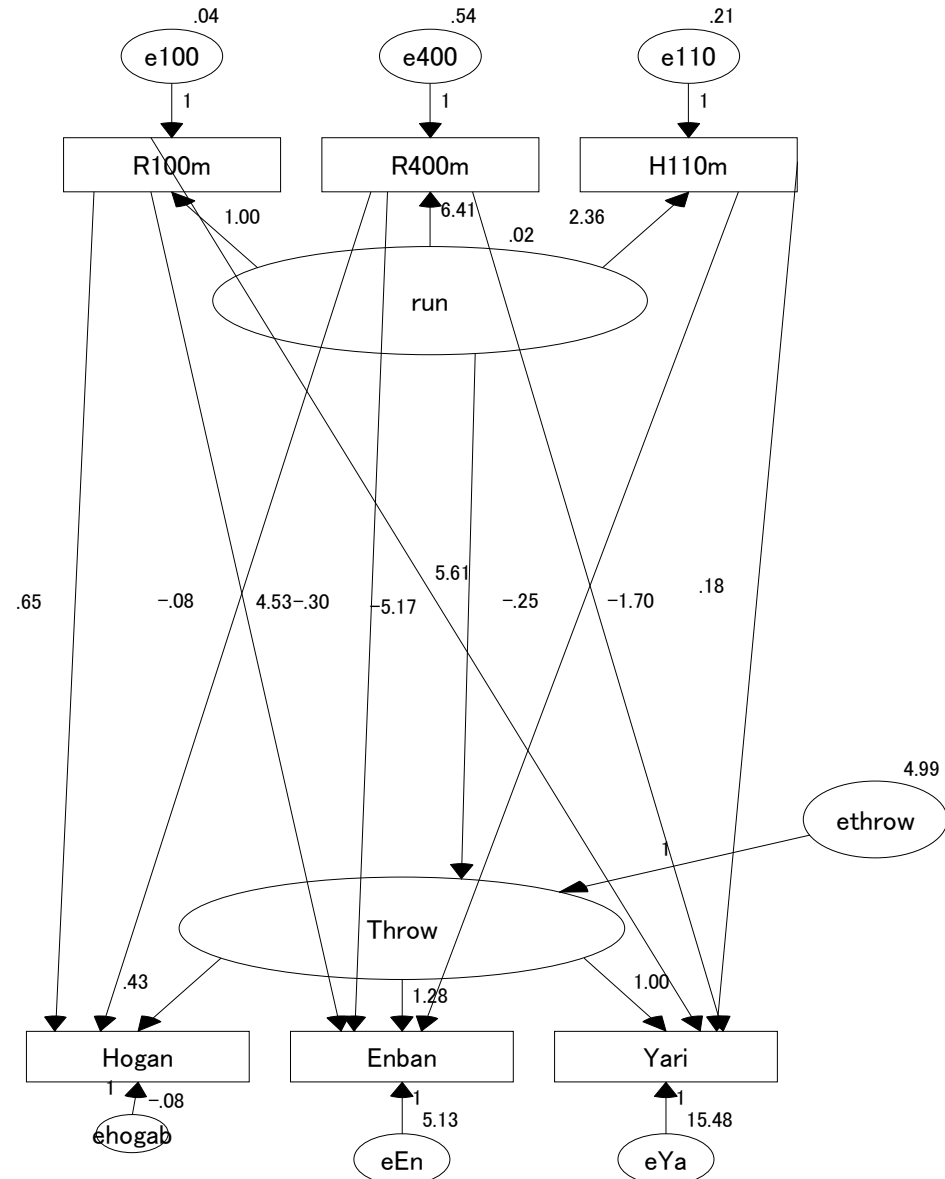


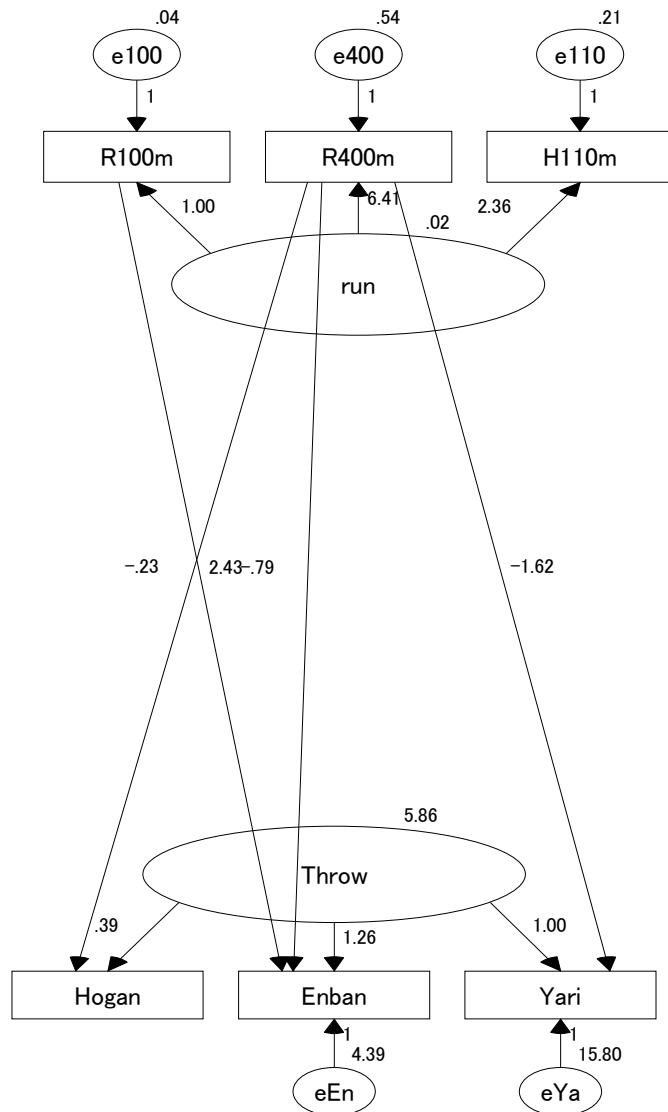
Fig.12 Causal Model for Remote Sensing Data after Model Selection

多重指標モデルと飽和化：試論

- 因子得点のパス解析は適合度上の問題が少ないのにSEMは問題多発
- 因子得点のパス解析とはどんなモデルに対応？
- 条件付推論の
代用手法としての**飽和化**
 - 飽和化＝測定モデルの結果
系変量に対する追加の不変
性確保
- 外的適合度問題
 - Cox and Wermuth



変数減少手順によるパスや分散の消去:サイモン・ブレイロック手順に類似



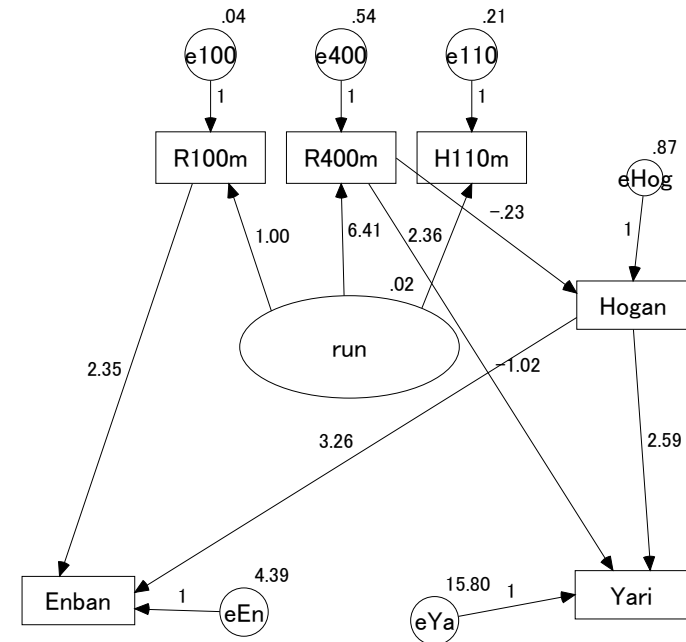
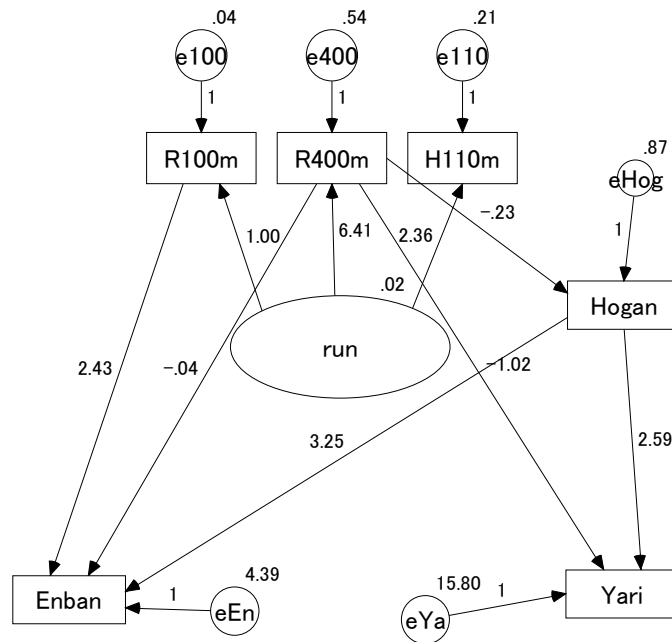
- 特徴

- 逸脱度3.32(自由度6)
- P値:0.77 AGFI:0.924

- 多重指標モデルは
外的適合性に問題あり？

- 走力から投擲力へのパスが消える.
- 400mから投擲力系3変数への
直接パスが消去されない
- 投擲力は砲丸投げと等価

等価なモデルから更にパスが消去可能！
逸脱度3.337(自由度7)AGFI:0.935



4. 動的因子構造

- 状態空間モデルの専門家を呼びなさい
 - 非線形・非ガウスベイズ型モデリング
 - 統計数理研究所のお家芸
 - 講演者の出番はない:
- せいぜい、Remarkとしての経験談
 - そこで潜在成長曲線モデル
 - アルツハイマー病の自然経過
 - パネルデータの場合
 - 同時潜在成長曲線モデル
 - 角埜氏との格闘

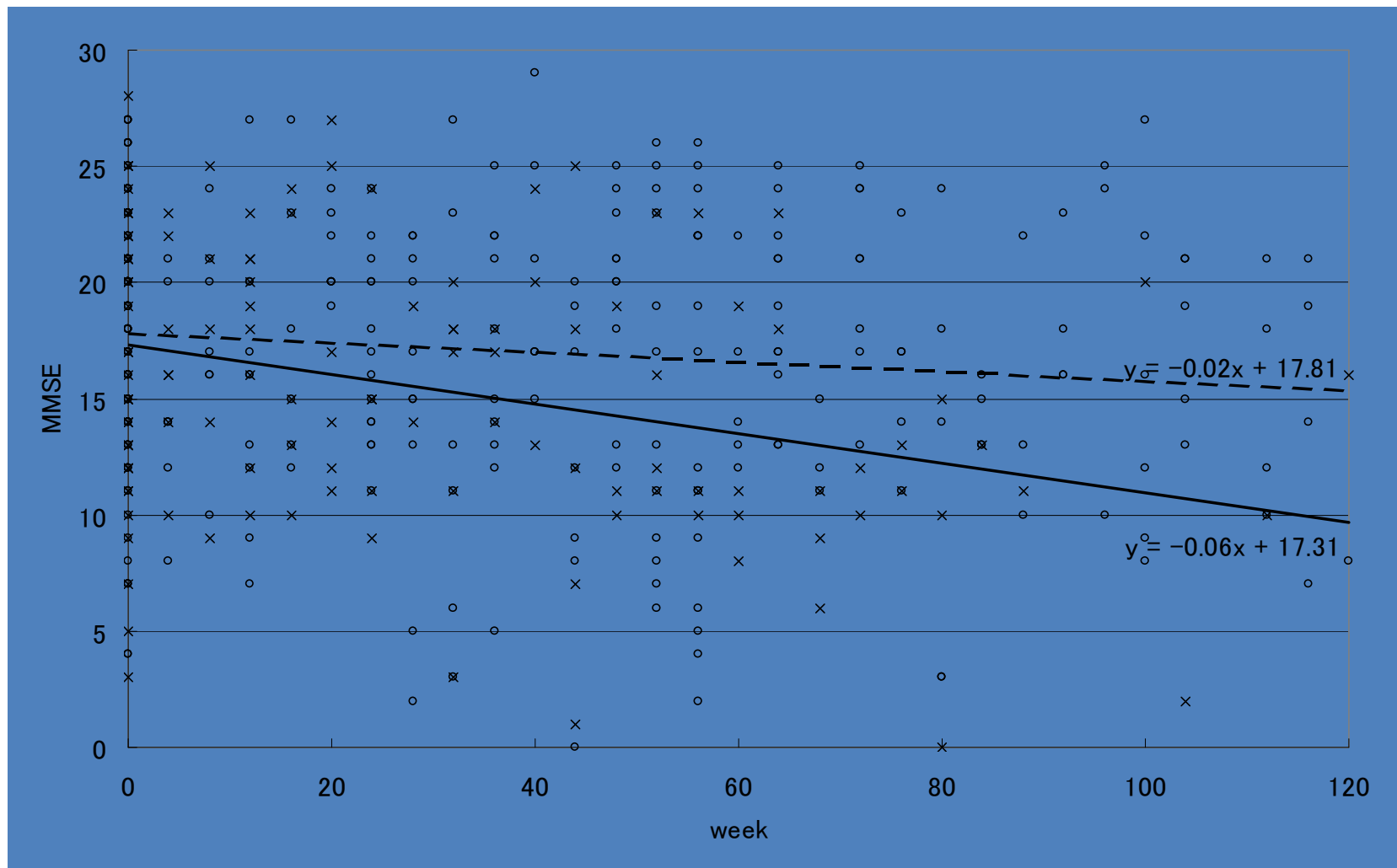
Virtual Repeated Measurement Data

Arai, Tsubaki et al.(2001)

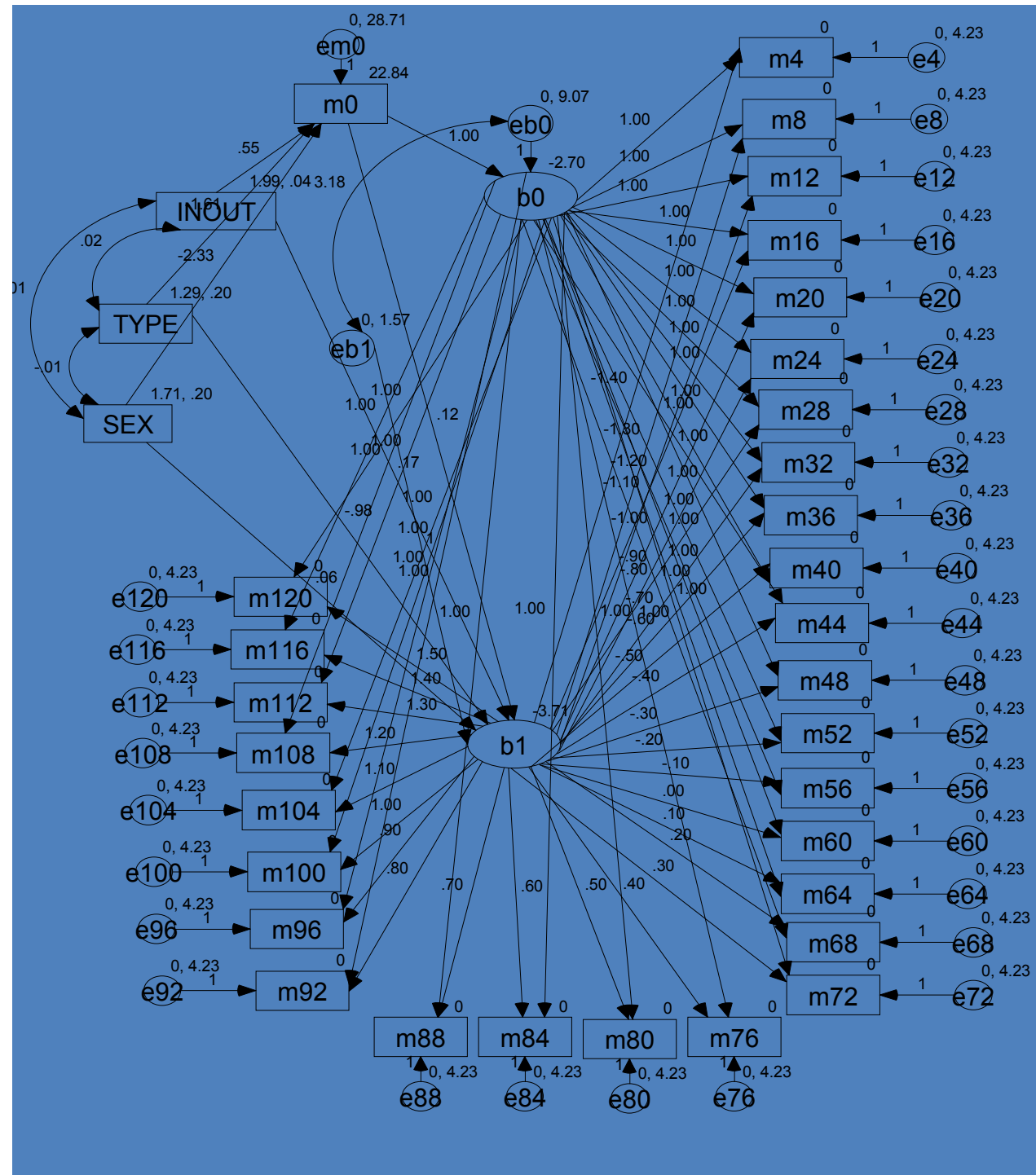
m0	m4	m8	m12	m16	m20	m24	m28	m32	m36
14						11			
7							5		
22					23				
17									
15				15				17	
9								6	
14							13		
18							15		
17			20						18
15	14	16		15		17			

Data

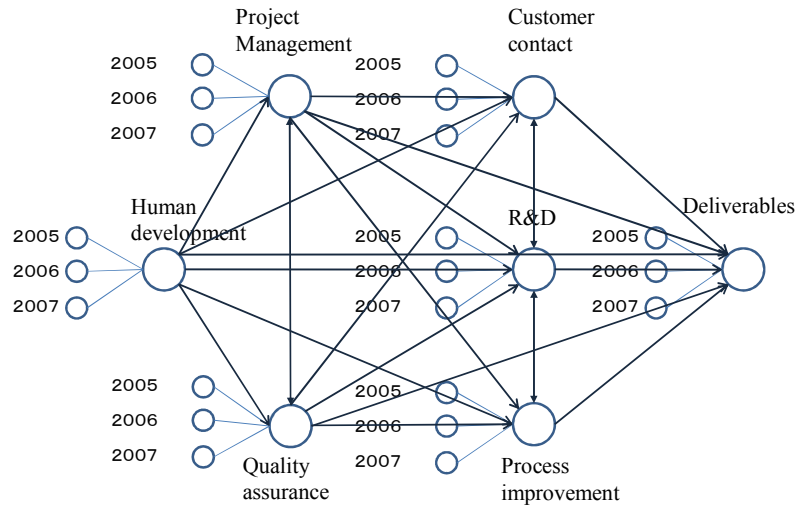
AD (cross, solid line) and SDAT (circle, dash line)



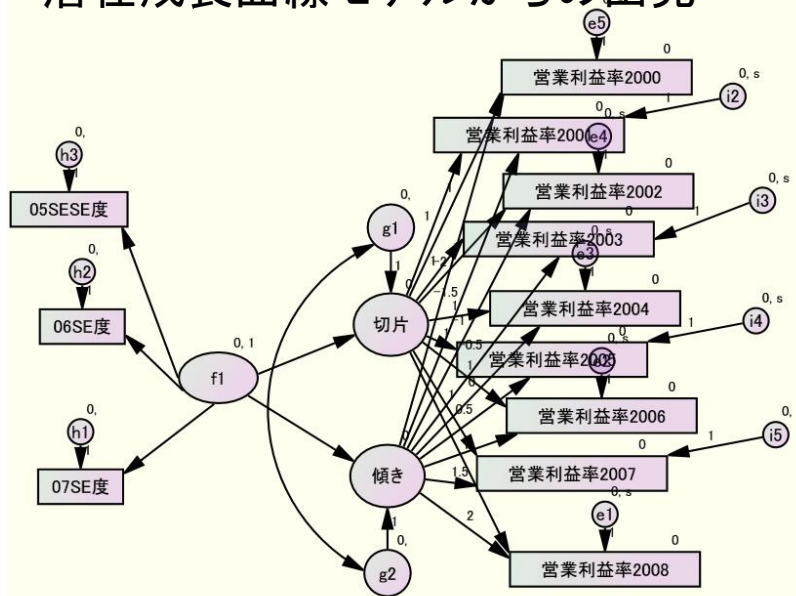
Path Diagram
and
the Results



基本仮説

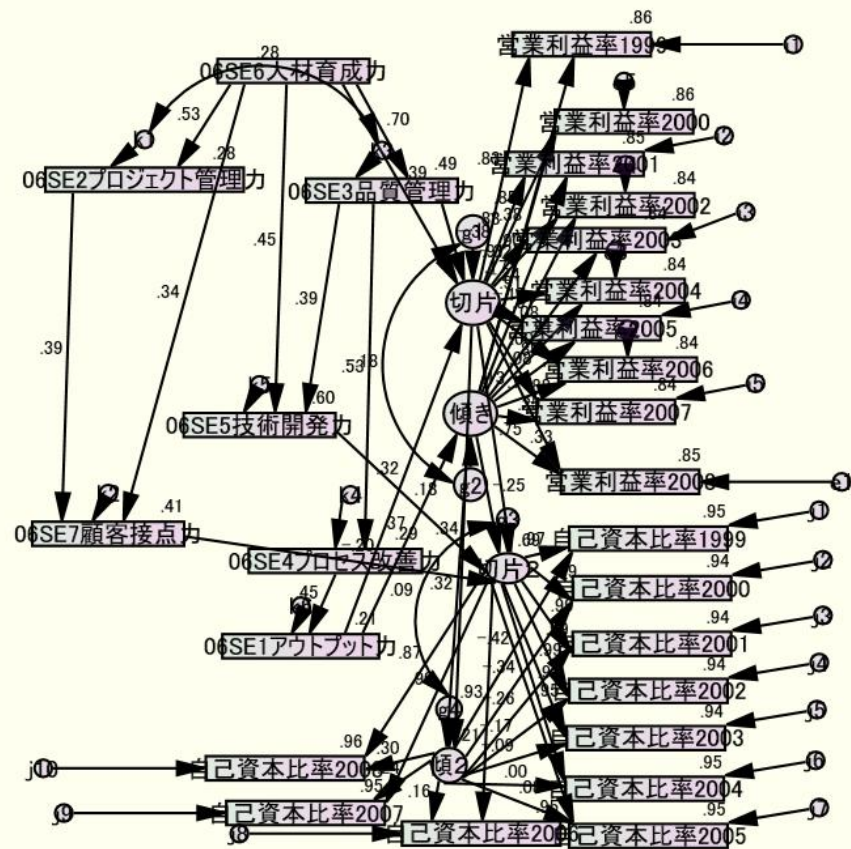


潜在成長曲線モデルからの出発

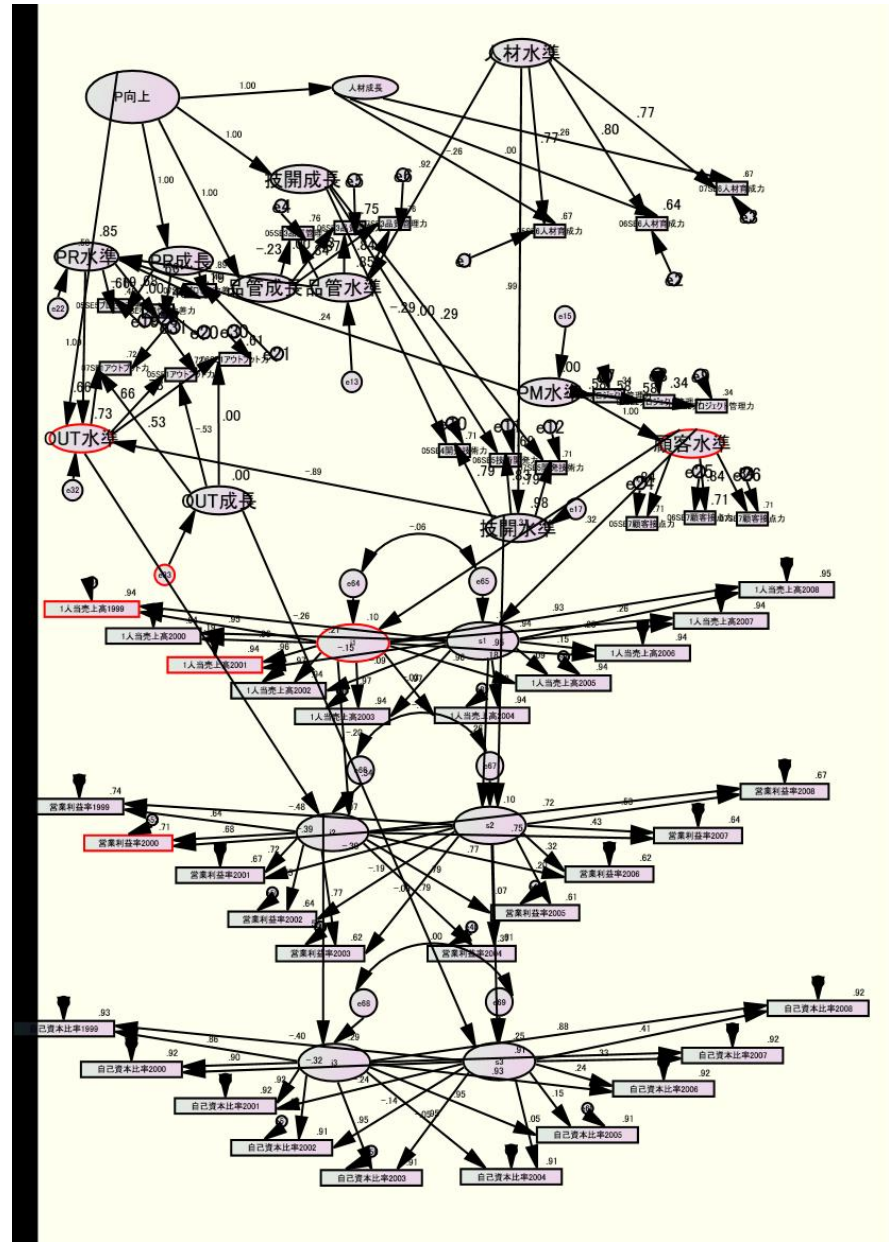
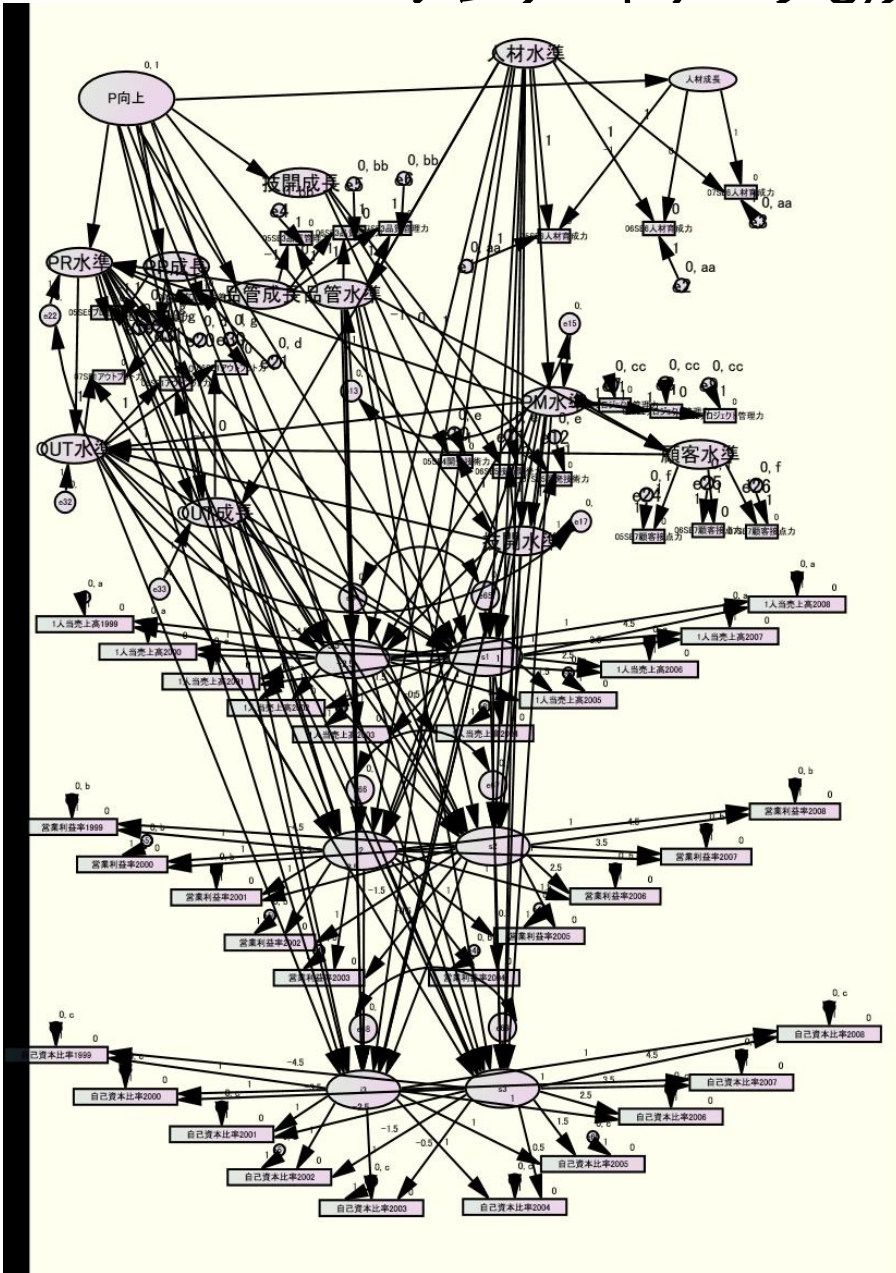


角埜恭央氏SE度調査2005-2007 ＋財務データとの格闘

成長曲線の潜在因子(切片、傾き)の 構造モデル導入



角埜恭央氏の研究で用いた統計モデル アンケートデータも成長曲線モデル一部導入



おわりに

- 探索的共分散構造分析への期待
 - 潜在因子か顕在因果構造か
 - 隠れた因子構造(大きな固有ベクトル空間)
 - 隠れた顕在因果構造(小さな固有ベクトル空間)
 - 因子構造(測定モデル)の不変性要求
 - 不適解多発の検証的因子モデリング
 - 因子モデルを含む場合の探索的モデリングの今後
- 動的因子モデリング、パネルデータのモデリング
 - 理論なき実証(AR型因果モデリング)
 - AICなどで優位性: 予測には良いのでしょう
 - Longitudinalな相関構造が、Cross Sectionalな相関構造をブロックする
 - 水準因子、成長因子間関係としての解釈(Cross Sectionの構造を維持)
 - 解釈としては美しいが適合度は(水準因子、成長因子)?
 - 成長の非線形性が適合を著しく劣化させる
- 非線形性潜在構造モデル
 - 非線形性については、スプラインモデルなどセミパラメトリックモデリング(GAM)で対応可能
 - この辺はState Space Modelの方々に聴いて下さい
- 質的データ、質量混合の多変量解析
 - 多変量プロビットモデリングは常に可能だが、筆者の好みではない
 - 多変量離散指数型分布族への拘り

ご清聴有難うございました

夏目(1901)に見る非線形潜在構造モデル

