Thinning Measurement Models and Questionnaire Design

Ricardo Silva
University College London
ricardo@stats.ucl.ac.uk

CSML PI Meeting, January 2013
UK Data Archive Study Number 6570 - National Health Service National Staff Survey, 2009
Not for Distribution

The Joy of Questionnaires

National NHS Staff Survey 2009

### YOUR JOB AND ORGANISATION

7. To what extent do you agree or disagree with the following statements about your immediate manager?

<table>
<thead>
<tr>
<th></th>
<th>Strongly disagree</th>
<th>Disagree</th>
<th>Neither agree nor disagree</th>
<th>Agree</th>
<th>Strongly agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A. encourages those who work for her/him to work as a team.
B. can be counted on to help me with a difficult task at work.
C. gives me clear feedback on my work.
D. asks for my opinion before making decisions that affect my work.
E. is supportive in a personal crisis.
Postulated Usage

- A measurement model for latent traits
- Information of interested postulated to be in the latents
Task

- Given a measurement model, select a subset of given size $K$ that preserves information about latent traits

> “Longer surveys take more time to complete, tend to have more missing data, and have higher refusal rates than short surveys. Arguably, then, techniques to reducing the length of scales while maintaining psychometric quality are worthwhile.” (Stanton, 2002)

- Criterion:

$$
\left\langle KL(\mathcal{P}_M(X \mid Y) \mid \mid \mathcal{P}_M(X \mid Y_z)) \right\rangle_{\mathcal{P}_M(Y)}$$

$X$ = latent variables, $Y$ = observations, $Y_z$ = a subset of size $K$
Optimization Problem

- Solve for

\[ z^* = \arg \max_z \left\{ \sum_{i=1}^{p} z_i \langle \log(\mathcal{P}_M(Y_i | X)) \rangle_{\mathcal{P}_M(X,Y_i)} + \mathcal{H}_M(Y_z) \right\} \]

\[ \equiv \arg \max_z \mathcal{F}_M(z) \]

subject to each \( z_i \in \{0, 1\} \) and

\[ \sum_{i=1}^{p} z_i = K \]

where \( H_M(Y_z) \) is marginal entropy

- Notice: equivalent to minimising conditional entropy
  - Other divergence functions could be used in principle
Issues

- Hard combinatorial optimization problem
- **Even objective function cannot be calculated exactly**

- Focus: models where small marginals of $Y$ can be calculated without major effort
- Working example: multivariate probit model

\[
X \sim \mathcal{N}(0, \Sigma), \quad Y_i^* \sim \mathcal{N}(\Lambda_i^T X + \lambda_i; 0, 1),
\]
\[
Y_i = 1, \text{ if } Y_i^* > 0, \text{ and } 0 \text{ otherwise}
\]
An Initial Approach

- Gaussian approximation

\[ F_{M;N}(z) \equiv \sum_{i=1}^{p} z_i \langle \log(P_M(Y_i \mid X)) \rangle_{P_M(X,Y_i)} + 0.5 \log |\Sigma_z| \]

maximize this by greedy search

- Covariance matrix: use “exact” covariance matrix
  - Doable in models such as the probit without an Expectation-Propagation scheme
Alternatives

- More complex approximations

- But how far can we intertwine the approximation with the optimisation procedure?
  - i.e., use a single global approximation that can be quickly calculated to any choice of $Z$
Alternative: Using Entropy Bounds

\[ \mathcal{H}(Y \mid S) \leq \mathcal{H}(Y \mid S') \text{ for } S' \subseteq S \]

- Exploit this using a given ordering \( e \) of variables (Globerson and Jaakkola, 2007) and a choice of neighborhood \( N(e, i) \)

\[ \mathcal{H}(Y_1, Y_2, \ldots, Y_p) = \sum_{i=1}^{n} \mathcal{H}(Y_{e(i)} \mid Y_{e(1:i-1)}) \leq \sum_{i=1}^{p} \mathcal{H}(Y_{e(i)} \mid Y_{N(e,i)}) \]

- Optimize ordering using a permutation-optimization method to get tightest bound
  - E.g., Teyssier and Koller (2005), Jaakkola et al., (2010), etc.
  - Notice role of previous assumption: each entropy term can be computed in a probit model for small \( N(e, i) \)
Method I: Bounded Neighborhood

- Plug in approximate entropy term into objective function

\[
F_{M;D}(\mathbf{z}) \equiv \sum_{i=1}^{p} z_i \langle \log(\mathcal{P}_M(Y_i \mid \mathbf{X})) \rangle_{\mathcal{P}_M(\mathbf{X},Y_i)} + \sum_{i=1}^{p} z_{e(i)} \mathcal{H}_L(\mathbf{z}, i)
\]

where

\[
\mathcal{H}_L(\mathbf{z}, i) \equiv \mathcal{H}_M(Y_{e(i)} \mid Y_z \cap N(e, i)) = \sum_{S \in \mathcal{P}(N(e, i))} \left[ \prod_{j \in S} z_j \right] \left[ \prod_{k \in N(e, i) \setminus S} (1 - z_k) \right] \mathcal{H}_M(Y_{e(i)} \mid S)
\]

- ILP formulation: first define \( q_M \equiv \prod_{m \in M} z_m \)

\[
q_M = 1 \Rightarrow \{ \forall m \in M, z_m = 1 \} \quad \Leftrightarrow \quad \forall m \in M, q_M - z_m \leq 0
\]
\[
q_M = 0 \Rightarrow \{ \exists m \in M \text{ s.t. } z_m = 0 \} \quad \Leftrightarrow \quad \sum_{m \in M} z_m - q_M \leq |M| - 1
\]

Glover and Woolsey (1974)
Method I: Interpretation

- Entropy for full model:
  - That of a Bayesian network (DAG model) approximation

- Entropy for submodels ($K$ items):
  - That of a crude marginal approximation to the DAG approximation

Starting model

DAG approximation with $|N| = 1$, $z = (1, 1, 1)$

Marginal with $|N| = 1$, $z = (1, 0, 1)$

Actual marginal used, $z = (1, 0, 1)$
Method II: Tree-structures

- Motivation: different selections for $z$ don’t wipe out marginal dependencies

\[
\mathcal{H}(Y_{e(i)} \mid Y_{e(1:i-1)}) \leq \min_{j \in e(1:i-1)} \mathcal{H}_M(Y_{e(i)} \mid Y_j),
\]

\[
\mathcal{F}_{M;\text{tree}}(z) \equiv \sum_{i=1}^{p} z_i \langle \log(\mathcal{P}_M(Y_i \mid X)) \rangle_{\mathcal{P}_M(X,Y_i)} + \sum_{i=1}^{p} z_{e(i)} \cdot \min_{\{Y_j \in Y_{e(1:i-1) \cap Y_z}\}} \mathcal{H}(Y_{e(i)} \mid Y_j)
\]

\[
\min_{\{Y_j \in Y_{e(1:i-1) \cap Y_z}\}} \mathcal{H}(Y_{e(i)} \mid Y_j) \equiv \sum_{j=1}^{i-1} q_i^{(j)} H_i^{(j)} + q_i^{(i)} \mathcal{H}_M(Y_{e(i)})
\]

\[
q_i^{(j)} \equiv z_i^{(j)} \prod_{k=1}^{j-1} z_i^{(k)}
\]
Method II: Interpretation

- For a fixed ordering, find a directed tree-structured submodel of size $K$

Tree approximation for $z = (1, 1, 1, 1)$

Tree approximation for $z = (1, 1, 0, 1)$ if $Y_4$ more strongly associated with $Y_2$ than $Y_1$
Evaluation

- Synthetic models, evaluation against a fairly good heuristic, the reliability score

\[ F_{M; \mathcal{R}}(z) = \sum_{i=1}^{p} w_i z_i, \text{ subject to } z_i \in \{0, 1\}, \sum_{i=1}^{p} z_i = K \quad w_i = \Lambda_i^T \Sigma \Lambda_i. \]

- Evaluation:
  - for different choices of \( K \), obtain selection \( Y_z \) for each of the four methods
  - given selection, evaluate how well latent posterior expectations are with respect to model using all \( Y \)
Evaluation

**Improvement ratio: high signal**

**Improvement ratio: low signal**

**Mean error: high signal**

**Mean error: low signal**
Evaluation: NHS data

- Model fit using 50,000 points, pre selection of 60+ items
  - Structure following structure of questionnaire

- Relative improvement over reliability score and absolute mean squared errors

<table>
<thead>
<tr>
<th></th>
<th>$s_F;D$</th>
<th>$s_F;tree$</th>
<th>$s_F;N$</th>
<th>$s_F;random$</th>
<th>$m_F;tree$</th>
<th>$m_F;R$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K = 32$</td>
<td>7.8%</td>
<td>6.3%</td>
<td>0.01%</td>
<td>-16.0%</td>
<td>0.238</td>
<td>0.255</td>
</tr>
<tr>
<td>$K = 50$</td>
<td>10.5%</td>
<td>11.9%</td>
<td>7.6%</td>
<td>-0.05%</td>
<td>0.123</td>
<td>0.140</td>
</tr>
</tbody>
</table>
Conclusion

- How to combine the latent information criterion with other desirable criterion?

- Integrating the order optimization step with the item selection step

- Incorporating uncertainty in the model

- Relation to adaptive methods in computer-based questionnaires