Latent Factor Regression Models for Grouped Outcomes

Dawn Woodard
Operations Research and Information Engineering
Cornell University

with T. M. T. Love, S. W. Thurston, D. Ruppert
S. Sathyanarayana and S. H. Swan

Joint Stastistical Meetings 2012, San Diego

Outline

Overview

- Multiple-outcome regression models pool information across outcomes, for higher power to detect a covariate effect
- Two traditional approaches:
  - **Random effect models:** random effects induce correlations between outcomes (cf. Sammel, Lin, & Ryan 1999; Roy, Lin, & Ryan 2003)
  - **Latent factor models:** continuous latent variables manifested by outcomes (cf. Muthén 2002; Sanchez et al. 2005)
- Our focus: **Multiple outcomes nested in domains** such as motor function, intelligence, and attention
- Motivating example: Relating sexually dimorphic traits in male infants to factors including exposure to phthalates (industrial chemicals)

Overview

- We show **the random effect models are a special case of the continuous latent factor framework, even when the multiple outcomes are nested in domains**
- Not surprising since the continuous latent factor framework is extremely general, and non-identifiable in unrestricted case.
- Allows us to view modeling options as a spectrum between parsimonious random effect models and flexible latent factor models
- We **introduce a set of models along this spectrum, and show that they are identifiable**

Overview

- **Gives a set of general-purpose tools for modeling and sensitivity analysis in the case of outcomes nested in domains**
- Yields estimates of the exposure/covariate effect both at the domain and outcome levels
- Shares information across outcomes, in part by **shrinkage of the estimated effect across outcomes**.
- Models **differ in degree of shrinkage**

Modeling Grouped Outcomes

Random effect regression model for outcomes nested in domains

(Thurston, Ruppert, Davidson 2009). Subjects $i \in \{1, \ldots, n\}$, (standardized) outcomes $j \in \{1, \ldots, p\}$:

$$Y_{ij} = (\mathbf{b}_{o,j} + \mathbf{b}_{\mathcal{D},d(j)} + \mathbf{b})\mathbf{X}_i + \phi_i + \psi_{i,d(j)} + \epsilon_{ij}$$

$d(j) \in \{1, \ldots, D\}$: domain of outcome j
$\mathbf{b}$: common covariate effect
$\mathbf{b}_{\mathcal{D},d(j)}$: random domain-level covariate effect
$\mathbf{b}_{o,j}$: random outcome-level covariate effect
ϕ_i: subject random effect
$\psi_{i,d(j)}$: subject-domain random effect

Modeling Grouped Outcomes

$$Y_{ij} = (\mathbf{b}_{o,j} + \mathbf{b}_{\mathcal{D},d(j)} + \mathbf{b})\mathbf{X}_i + \phi_i + \psi_{i,d(j)} + \epsilon_{ij}$$

- Captures positive correlation between outcomes even after accounting for covariates
- Shrinks covariate effect across domains and across outcomes within a domain

Modeling Grouped Outcomes

Continuous latent factor framework (non-identifiable)
(cf. Sammel & Ryan 1996; Muthén 2002; Sanchez et al. 2005)

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta}_o \mathbf{X}_i + \boldsymbol{\Lambda}\boldsymbol{\xi}_i + \boldsymbol{\epsilon}_i$$
$$\boldsymbol{\xi}_i = \boldsymbol{\beta}_{\mathcal{D}} \mathbf{X}_i + \mathbf{B}\boldsymbol{\xi}_i + \boldsymbol{\zeta}_i.$$

$\mathbf{Y}_i$: outcomes vector
$\boldsymbol{\beta}_o$, $\boldsymbol{\beta}_{\mathcal{D}}$: regression coefficients matrices
$\boldsymbol{\Lambda}$: factor loadings matrix
$\boldsymbol{\xi}_i$: latent factors vector
$\mathbf{B}$: matrix relating latent factors
$\boldsymbol{\epsilon}_i$, $\boldsymbol{\zeta}_i$: independent residuals

Modeling Grouped Outcomes

- Take the # latent factors = D
- Take nonzero elements of $\mathbf{\Lambda}$ to be the $(j, d(j))$ elements. E.g., with $D = 2$ and $p = 4$ where $d(1) = d(2) = 1$ and $d(3) = d(4) = 2$ then

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_3 & \lambda_4 \end{pmatrix}^T.$$

- Induce **B** by setting $\boldsymbol{\xi}_i = \boldsymbol{\beta}_{\mathcal{D}}\mathbf{X}_i + \phi_i + \boldsymbol{\psi}_i$

Yields

$$Y_{ij} = \alpha_j + \boldsymbol{\beta}_{o,j}\mathbf{X}_i + \lambda_j \xi_{i,d(j)} + \epsilon_{ij}$$
$$\xi_{ik} = \boldsymbol{\beta}_{\mathcal{D},k}\mathbf{X}_i + \phi_i + \psi_{ik}$$

Modeling Grouped Outcomes

$$
\begin{aligned}
Y_{ij} &= \alpha_j + \boldsymbol{\beta}_{o,j}\mathbf{X}_i + \lambda_j \xi_{i,d(j)} + \epsilon_{ij} \\
\xi_{ik} &= \boldsymbol{\beta}_{\mathcal{D},k}\mathbf{X}_i + \phi_i + \psi_{ik}
\end{aligned} \tag{1}
$$

- This model is identifiable so long as $\boldsymbol{\beta}_{o,j}$ are random effects and $\lambda_j = 1$ for the first outcome in each domain
- Additionally taking $\lambda_j = 1$ for all j and making $\boldsymbol{\beta}_{\mathcal{D},k}$ random effects yields the random effect model of Thurston et al. (2009)
- We investigate models of the form (1)

Modeling Grouped Outcomes

$$Y_{ij} = \alpha_j + \boldsymbol{\beta}_{o,j}\mathbf{X}_i + \lambda_j \xi_{i,d(j)} + \epsilon_{ij}$$
$$\xi_{ik} = \boldsymbol{\beta}_{\mathcal{D},k}\mathbf{X}_i + \phi_i + \psi_{ik}$$

- We obtained this from the latent factor framework by making assumptions about $\mathbf{\Lambda}$ and $\mathbf{B}$
- Contrast our choice of $\mathbf{B}$ (induce via subject random effect ϕ_i) with more standard choice: assign latent factors particular interpretations, like "motor function" and "intelligence", then manually select a small number of nonzero elements in $\mathbf{B}$ corresponding to hypothesized associations among the factors

Modeling Grouped Outcomes

$$Y_{ij} = \alpha_j + \boldsymbol{\beta}_{o,j}\mathbf{X}_i + \lambda_j \xi_{i,d(j)} + \epsilon_{ij}$$
$$\xi_{ik} = \boldsymbol{\beta}_{\mathcal{D},k}\mathbf{X}_i + \phi_i + \psi_{ik}$$

Our models (identifiable):

A. As above

B. Make $\boldsymbol{\beta}_{\mathcal{D},k}$ random effects

C. Like **A** but sets $\boldsymbol{\beta}_{o,j} = 0$

D. Like **C** but sets $\lambda_j = 1$ for all j

Use Bayesian inference and MCMC computation

Simulation Study

Vary:

- Models used to simulate & fit data
- sample size n
- # outcomes per domain
- factor loading (λ_j) values
- ...

Evaluate bias and RMSE for:

- domain-specific covariate effects $\beta_{\mathcal{D},k}$
- outcome-specific covariate effects $\text{OS}_j \triangleq \beta_{o,j} + \lambda_j \beta_{\mathcal{D},d(j)}$

Simulation Study

λ_j	Some $\lambda_j = 0.5$		All $\lambda_j = 1$						
Simulated	Model A		Model A			Model B		Model D	
Estimation	A	D	A	B	D	A	B	A	D
bias $\beta_{\mathcal{D},k}$	-.011	-.041	-.014	-.017	-.001	-.002	-.005	-.008	.000
bias OS_j	-.001	.002	-.005	-.005	-.003	.001	.000	-.004	-.001
RMSE $\beta_{\mathcal{D},k}$	.054	.063	.039	.041	.036	.039	.039	.033	.026
RMSE OS_j	.041	.061	.040	.040	.050	.040	.041	.035	.025

Even when the model is misspecified:

RMSE is an order of magnitude smaller than the parameter range (.5) for OS_j, and $\approx$5$\times$ smaller than the range (.2) for $\beta_{\mathcal{D},k}$

Bias is 2 orders of magnitude smaller than the range for OS_j and 5-25$\times$ smaller for $\beta_{\mathcal{D},k}$.

Simulation Study

λ_j	Some $\lambda_j = 0.5$		All $\lambda_j = 1$						
Simulated	Model A		Model A			Model B		Model D	
Estimation	A	D	A	B	D	A	B	A	D
bias $\beta_{\mathcal{D},k}$	-.011	-.041	-.014	-.017	-.001	-.002	-.005	-.008	.000
bias OS_j	-.001	.002	-.005	-.005	-.003	.001	.000	-.004	-.001
RMSE $\beta_{\mathcal{D},k}$	.054	.063	.039	.041	.036	.039	.039	.033	.026
RMSE OS_j	.041	.061	.040	.040	.050	.040	.041	.035	.025

The highest error occurs if we fit too simple of a model (D) when the data are simulated from a more complex model (A)

The lowest error occurs if the model is simple (D) and we fit the simple model (D)

Fitting Model B does not yield lower error than fitting Model A, even if the data are from Model B.

Study for Future Families

Study for Future Families:

Relate sexually dimorphic traits in male infants to factors including phthalate exposure

7 outcomes, in 3 domains:

- 4 skinfold thickness metrics
- body mass index and weight percentile-for-age
- head circumference percentile-for-age

Predictors including:

- phthalate score
- infant age
- gestational age
- mother's age & race

Study for Future Families

- Applied Models A, C, and D (Model B not appropriate because phthalate effect not hypothesized to be similar, or even have the same sign, in all domains)
- Find no evidence of a phthalate effect in any model
- Estimated covariate effects show differences between models

Study for Future Families

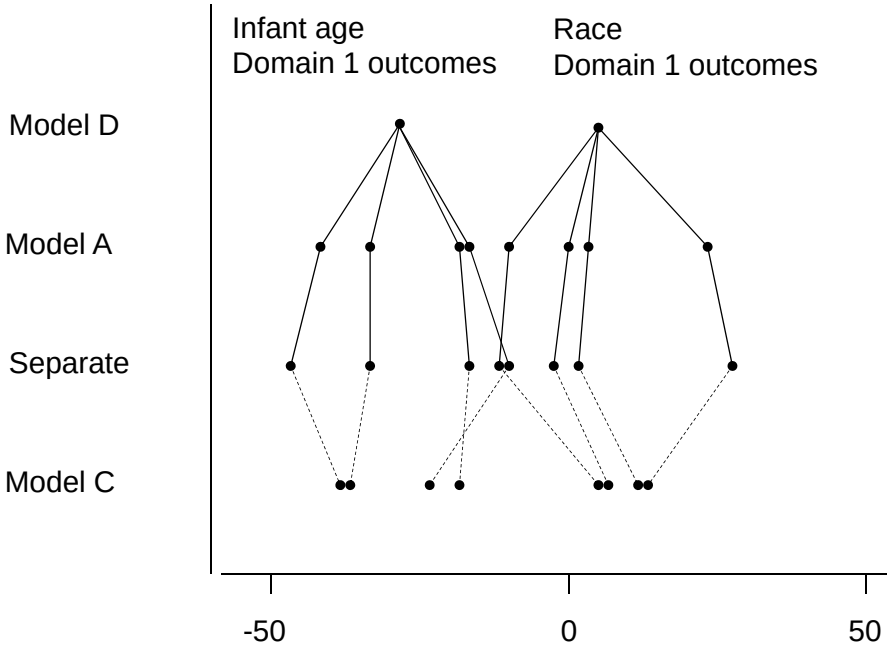
Model D coefficient estimates (common within a domain):

Covariate	Domain 1	Domain 2	Domain 3
Infant Age	$\mathbf{-28.0}$ $(\mathbf{-41.5}, \mathbf{-14.9})$	$\mathbf{-19.5}$ $(\mathbf{-36.0}, \mathbf{-3.3})$	16.1 $(-2.7, 34.9)$
Mother's Age	-11.3 $(-24.9, 1.9)$	0.13 $(-13.2, 13.4)$	9.97 $(-9.4, 29.5)$
Gestational Age	8.36 $(-4.9, 21)$	$\mathbf{18.1}$ $(\mathbf{2.0}, \mathbf{34.1})$	10.4 $(-8.3, 29.1)$
Race (Cauc./Non)	5.06 $(-7.9, 18.0)$	4.56 $(-8.1, 17.3)$	-12.3 $(-31.1, 6.4)$

- positive relationship between gest. age and BMI/weight
- negative relationship between infant age and both skinfold thickness and BMI/weight

Study for Future Families

Some coefficient point estimates ($\times$ 100) for Models A, C, D and separate regressions:



Model A estimates are shrunk relative to separate regressions, and Model D estimates are shrunk relative to Model A. Model C estimates are very different.

Study for Future Families

Although interval estimates get wider Model D $\rightarrow$ C $\rightarrow$ A, there are as many significant covariate effects in Models C & A as in D.

Slightly different set in Model A, which allows covariate effects to have different significance within a domain

Conclusions

- Showed that random-effect models for multiple outcomes nested in domains are a special case of the continuous latent factor framework
- Introduced identifiable models extending the random-effect models
- Simulations show excellent accuracy in point estimation of outcome-specific and domain-specific covariate effects, even in misspecified cases
- Applied to the Study for Future Families to understand factors relating to sexually dimorphic traits in male infants
- We recommend the use of Models A & D, which showed complementary strengths in the simulation study and different degrees of shrinkage in the data analysis

Paper available at: http://people.orie.cornell.edu/woodard/