Joint Factor and Regression Analyses of Multivariate Ordinal Data – Application to Psychiatric Assessments

Guoqing Diao[1], Srikanth Gottipati[2] and Peter Zhang[2]

[1]Department of Statistics, Volgenau School of Engineering, George Mason University; [2]Otsuka Pharmaceutical Development and Commercialization, Inc.

INTRODUCTION

Efficacy assessments in CNS trials are multivariate ordinal data where each variable is an item in the assessment and is assigned a score.

- **The Positive and Negative Syndrome Scale (PANSS) is the most widely-used outcome instrument in randomized, controlled trials for assessing the effects of antipsychotic medication, and other treatments on the symptoms of schizophrenia.[1]**
- **It is well recognized that the number of factors used when analyzing the PANSS can impact the identification of subtypes of schizophrenia and/or the psychopathological processes underlying them, which may influence prognosis, therapeutic approaches, response to treatment, and prediction of related variables.**

However, as there are several subscales that contribute to the score encompassing the positive, negative and the psychopathology scores, it is difficult to find which of these may be indicative of pharmacotherapy or of relapse.

- **There have been several models ([2]-[5]) to reduce the dimensions -1) PCA and factor analyses models are used in literature to reduce the dimensionality, 2) Five-factor models for interpreting the PANSS are thought to be more representative of the syndromes of schizophrenia than the original 3 subscales: positive, negative, and psychopathology, especially in a chronic course. Five-factor models were used to determine the predictors associated with changes in psychosocial functioning, treatment adherence/ compliance, and treatment satisfaction.**
- **However, there has not been any literature where effects of covariates like treatment groups (placebo/drug), age, gender on latent factors were studied in a joint model.**
- **Current methods in literature that try to identify determinants of placebo/drug response largely rely on correlations of objective measures (demography, genetic, pharmacokinetic, and other covariates) with subjective efficacy measures like the total PANSS score.**
- **There are several issues with this approach - 1) it is not clear if sum of PANSS subscale scores have any relevance to patient disease/psychosis state or it is the latent factors which are representative of the disease state, 2) measurement of outcomes is highly subjective, but most current methods do not fully account for these biases.**

There is a rich literature on the analysis of multivariate ordinal data in the statistical community. Most existing methods assume that each ordinal outcome is determined by a latent variable which follows certain regression models.

- **Latent variable models [6], [7], [8], [9], [10] were developed for cross-sectional multivariate ordinal data of different types**
- **More recently, [11] describes a pairwise likelihood estimation for factor analysis models with cross-sectional multivariate ordinal data assuming multivariate normal latent variables.**
- **The aforementioned methods consider regression models for each ordinal outcome. When the dimension of the ordinal outcomes is high, the hypothesis testing problem typically involves a test with high degrees of freedom which consequently negatively impact the power of the test.**

To reduce the dimensionality of the testing problem in the regression modeling, a common practice is to allocate numerical scores to the levels of each ordinal variable. A summary score, usually the sum, is then used as a continuous outcome in a linear regression model.

- **However, this approach typically assume that the distances between two consecutive levels are the same across all ordinal variables.**
- **Furthermore, equal weights are assigned to each ordinal variable. In practice, different ordinal variables may belong to different domains.**
- **Additionally, some responses could be more important than others and therefore larger weights may be desired.**

An alternative approach to reduce the dimensionality is the factor analysis.

- **Standard factor analysis for continuous data may not be appropriate for ordinal data.**
- **To perform factor analysis with ordinal data, one approach is to first estimate the so-called polychoric correlation matrix and then perform standard factor analysis based on the estimated polychoric correlation matrix; for instance, see [11], [12] and [13].**
- **None of these papers consider the regression problem.**

OBJECTIVE

What's New?

- Here we propose a joint model for both factor analysis and regression analysis of multivariate ordinal data
- Our method allows one to assess the covariate effects on the (latent) factors.
- The new method is a key step in our efforts to build a valid and efficient toolbox for comparing the efficacy across different groups. In particular, we are interested in estimating the true effect of drug treatment in comparison to placebo while accounting for potential biases in the measurement of efficacy.
- The new method can potentially detect significant treatment effects missed by using the standard methods.
- It is possible that the drug may impact some factors but not all factors. The new method can provide insight on the domains (factors) that are affected or not affected by the treatment in a confirmatory factor model approach.

METHODS

Models

- Notation:
  - $\boldsymbol{Y}$: q-dimensional ordinal data
  - $\mathbf{X}$: p-dimensional covariates
  - $K_j, j = 1, \ldots, q$: number of levels for the jth ordinal variable
- Joint model for factor and regression analyses:

$$\boldsymbol{Y}^* = \boldsymbol{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}$$

and

$$\boldsymbol{\xi} = \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}$$

  - $\boldsymbol{Y}^*$: q-dimensional vector of underlying variables that determine the level of the observed ordinal variables $\boldsymbol{Y}$
  - $\boldsymbol{\xi}$: common factors (latent variables)
  - $\boldsymbol{\Lambda}$: the q by k matrix of factor loadings (particularly, Λ_{ij} is the factor loading of Y_i^*on factor j)
  - $\boldsymbol{\delta}$: the q-dimensional vector of residuals in the factor analysis model
  - $\boldsymbol{\beta}$: the k by p matrix of regression coefficients
  - $\boldsymbol{\varepsilon}$: a k-dimensional vector of residuals in the regression analysis
- The value of Y_j is determined by the latent variable Y_j^* such that

$$Y_j = \begin{cases} 1, & \alpha_{j0} < Y_j^* \leq \alpha_{j1} \\ & \cdots \\ K_j, & \alpha_{jK_j-1} < Y_j^* \leq \alpha_{jK_j} \end{cases}$$

where $-\infty = \alpha_{j0} < \alpha_{j1} < \cdots < \alpha_{jK_j-1} < \alpha_{jK_j} = \infty$.

- Assume
  - $\boldsymbol{\varepsilon} \sim \boldsymbol{N_k}(\mathbf{0}, \boldsymbol{\Phi})$, where diag($\boldsymbol{\Phi}$) = $\boldsymbol{I}$ and $\boldsymbol{I}$ is the identity matrix.
  - $\boldsymbol{\delta} \sim \boldsymbol{N_q}(\mathbf{0}, \boldsymbol{\Theta})$, where $\boldsymbol{\Theta} = \boldsymbol{I} - diag(\boldsymbol{\Lambda\Phi\Lambda^T})$.
  - Cov($\boldsymbol{\varepsilon}$, $\boldsymbol{\delta}$) = **0.**
- The joint factor and regression models can be written in one model as

$$\boldsymbol{Y}^* = \boldsymbol{\Lambda}\boldsymbol{\beta}\mathbf{X} + \boldsymbol{\Lambda}\boldsymbol{\varepsilon} + \boldsymbol{\delta}$$

- $\boldsymbol{Y}^*$ follows a multivariate normal distribution with mean $\boldsymbol{\mu} = \boldsymbol{\Lambda}\boldsymbol{\beta}\mathbf{X}$ and variance-covariance matrix

$$\boldsymbol{\Sigma} = \boldsymbol{\Theta} + \boldsymbol{\Lambda\Phi\Lambda^T} = \boldsymbol{I} + \boldsymbol{\Lambda\Phi\Lambda^T} \text{-} diag(\boldsymbol{\Lambda\Phi\Lambda^T}).$$

- Likelihood function based on n i.i.d. observations $\{(\boldsymbol{Y}_i, \boldsymbol{X}_i); i = 1, \ldots, n\}$, where $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iq})$, is

$$L(\theta) = \prod_{i=1}^{n} \int_{\alpha_{qY_{iq}-1}}^{\alpha_{qY_{iq}}} \cdots \int_{\alpha_{1Y_{i1}-1}}^{\alpha_{1Y_{i1}}} f(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}) d\boldsymbol{y},$$

where $\boldsymbol{\mu}_i = \boldsymbol{\Lambda\beta X}_i$ and $f(\boldsymbol{y}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ is the multivariate normal density function with mean $\boldsymbol{\mu}_i$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

- Pairwise likelihood

$$L_p(\theta) = \prod_{i=1}^{n} \sum_{1 \leq j < k \leq q} \int_{\alpha_{kY_{ik}-1}}^{\alpha_{kY_{ik}}} \int_{\alpha_{jY_{ij}-1}}^{\alpha_{jY_{ij}}} f(y_j, y_k; \mu_{ijk}, \boldsymbol{\Sigma}_{jk}) dy_j dy_k,$$

where $f(y_j, y_k; \mu_{ijk}, \boldsymbol{\Sigma}_{jk})$ is the bivariate normal density function with mean μ_{ijk} and variance-covariance matrix $\boldsymbol{\Sigma}_{jk}$.

- The maximum pairwise likelihood estimators (MPLEs) are consistent and asymptotically normal.
- Covariance of the MPLEs can be estimated by using the Sandwich estimators.

RESULTS

Simulation Studies

In the first set of simulations, we examine the performance of the MPLEs

- **Two covariates:** $X_1 \sim Bernoulli\left(\frac{1}{2}\right), X_2 \sim N(0,1)$
- $q = 6$
- $p = 2$
- $\boldsymbol{\beta} = \begin{pmatrix} 0.4 & 0.8 \\ -0.4 & 0.4 \end{pmatrix}$
- $\boldsymbol{\varepsilon} \sim N_2(0, I)$
- $\boldsymbol{\Lambda}' = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{10}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{10}} \end{pmatrix}$
- $n = 300$

Table 1. Summary statistics of estimates of β with $n = 300$ based on 500 replicates

Parameter	Bias	SE	SEE	CP
β_{11}	-0.009	0.188	0.185	0.940
β_{12}	0.032	0.136	0.133	0.964
β_{21}	-0.008	0.191	0.189	0.944
β_{22}	0.015	0.115	0.105	0.944

RESULTS CONTINUED

In the second set of simulations, we examine the type I error rates and powers of testing covariate effects on the latent factors

- proposed method (New)
- existing method by using the average of the ordinal scores as outcome for each factor (Average)
- linear regression model using the weighted sum based on true values of $\boldsymbol{\Lambda}$ (True), that is, $\boldsymbol{\Lambda}'\boldsymbol{Y}$

Two scenarios:

Scenario (a): $\boldsymbol{\Lambda}' = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{10}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{5}} & \frac{1}{\sqrt{10}} \end{pmatrix}$

Scenario (b): $\boldsymbol{\Lambda}' = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{5}} & \frac{1}{\sqrt{10}} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{5}} & \frac{1}{\sqrt{10}} \end{pmatrix}$

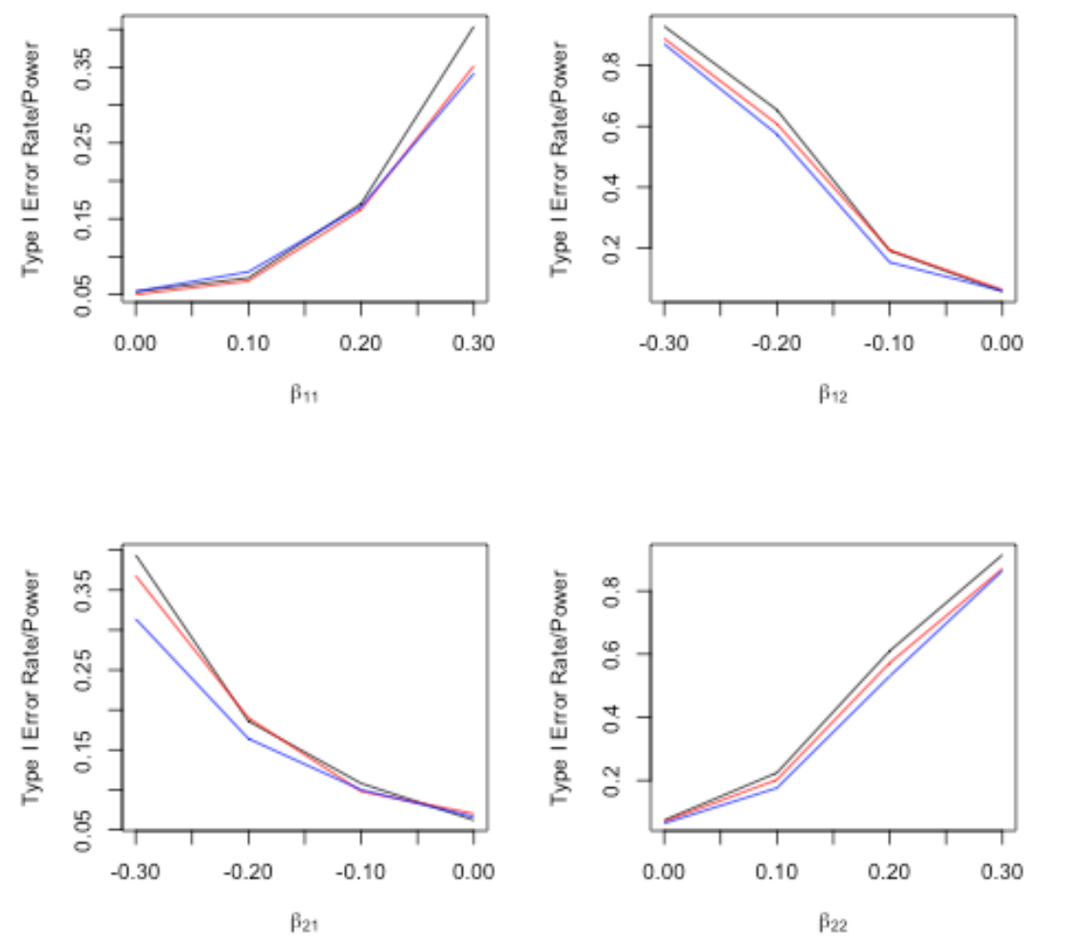

Figure 1. Type I error rates/powers for testing covariate effects under scenario (a). Black, red, and blue curves correspond to the "True", "Average", and "New" methods, respectively.

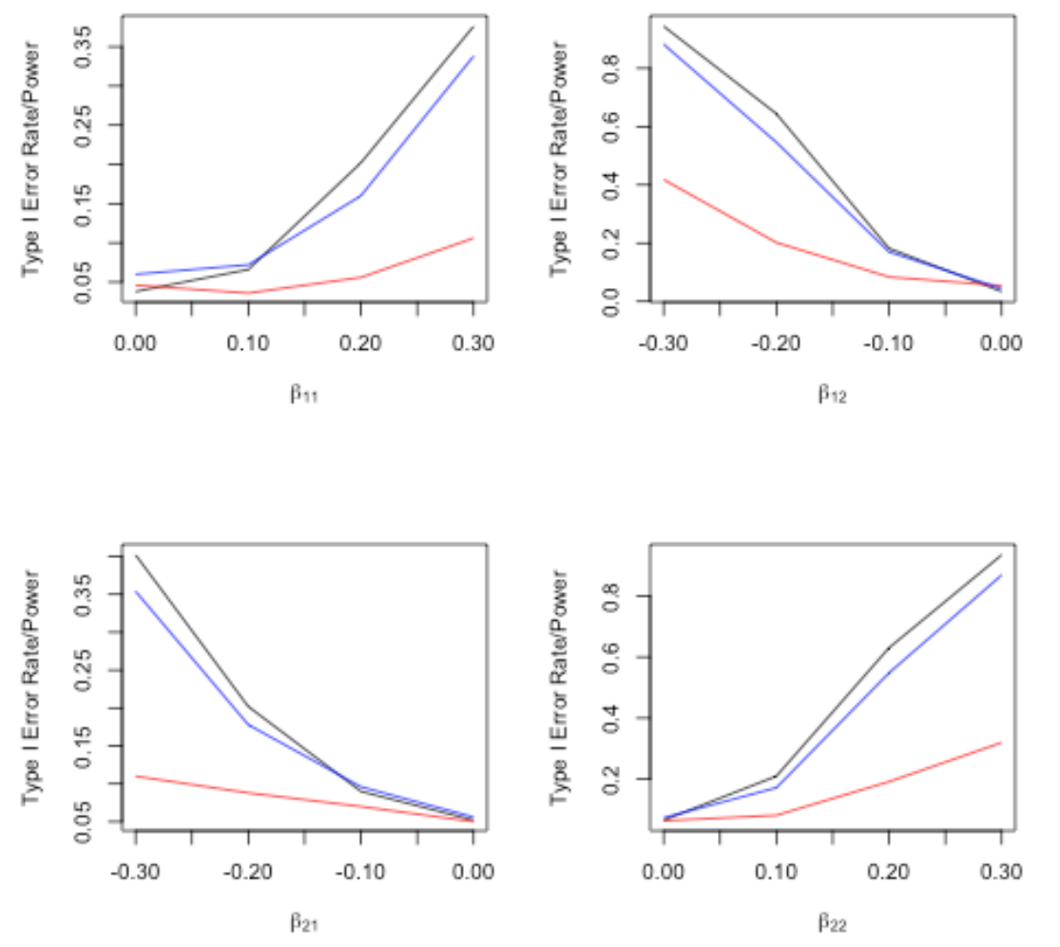

Figure 2. Type I error rates/powers for testing covariate effects under scenario (b). Black, red, and blue curves correspond to the "True", "Average", and "New" methods, respectively.

Summary of simulation results:

- Maximum pairwise likelihood estimators have little biases
- Standard error estimates (SEE) reflect the actual variation correctly
- 95% confidence intervals have correct coverage probabilities
- There is a slight loss of power of the proposed method for testing covariate effects compared to the "True" method assuming factor loadings are known
- The method using average of the ordinal scores may lead to substantial loss of power when some of the factor loadings are on different directions

APPLICATION TO PSYCHIATRIC ASSESSMENTS

Real Application

- We apply the proposed method on baseline PANSS scores measured in acutely relapsed schizophrenia inpatients enrolled in a placebo-controlled clinical trial to study efficacy of aripiprazole - NCT00080327. Sample size is 365.
- Positive and Negative Syndrome Scale (PANSS), consisting of 30 items, has been the most widely used measure of illness severity, and the PANSS total score is the gold standard primary efficacy measure in acute treatment studies of schizophrenia.
- Factor loadings of the PANSS was computed using the 5 factor model [2] - positive symptoms, negative symptoms, disorganized thinking and the associated symptom domains of hostility/excitement, and depression/anxiety.
- We evaluate the effects of covariates - (standardized) age, gender (male coded as 1 and female coded as 0), and race (White coded as 1 and others coded as 0) on the five latent factors.

Table 2. Effects of (standardized) age on the five factors

Factor	Estimate	SE	Test Statistic	p-value
1. Negative symptoms	0.025	0.063	0.389	0.697
2. Positive symptoms	-0.145	0.067	-2.150	0.032
3. Disorganized thought	0.010	0.061	0.168	0.867
4. Uncontrolled hostility/excitement	-0.159	0.063	-2.513	0.012
5. Anxiety/depression	-0.084	0.066	-1.261	0.207

Table 3. Effects of gender (male vs female) on the five factors

Factor	Estimate	SE	Test Statistic	p-value
1. Negative symptoms	0.241	0.146	1.653	0.098
2. Positive symptoms	0.033	0.221	0.147	0.883
3. Disorganized thought	-0.005	0.151	-0.036	0.971
4. Uncontrolled hostility/excitement	-0.221	0.150	-1.478	0.139
5. Anxiety/depression	-0.448	0.169	-2.65	0.008

Table 4. Effects of race (White vs others) on the five factors

Factor	Estimate	Standard Error	Test Statistic	p-value
1. Negative symptoms	0.070	0.121	0.579	0.562
2. Positive symptoms	0.484	0.138	3.500	<0.001
3. Disorganized thought	0.257	0.132	1.945	0.052
4. Uncontrolled hostility/excitement	0.343	0.119	2.883	0.004
5. Anxiety/depression	0.462	0.143	3.231	0.001

APPLICATION TO PSYCHIATRIC ASSESSMENTS

Table 5. Factor loading estimates

Factors and Items	Estimate	SE	Test Statistic	p-value
1. Negative symptoms				
Blunted affect	0.746	0.071	10.567	<1.0e-8
Emotional withdrawal	0.703	0.079	8.845	<1.0e-8
Poor rapport	0.727	0.061	11.947	<1.0e-8
Passive social withdrawal	0.717	0.070	10.294	<1.0e-8
Lack of spontaneity	0.770	0.082	9.399	<1.0e-8
Motor retardation	0.543	0.098	5.554	2.8e-8
Active social avoidance	0.562	0.099	5.655	1.6e-8
2. Positive symptoms				
Delusions	0.786	0.093	8.478	<1.0e-8
Hallucinatory behavior	0.382	0.106	3.611	3.0e-4
Grandiosity	0.262	0.116	2.258	0.024
Suspiciousness	0.335	0.103	3.256	1.1e-3
Stereotyped thinking	0.632	0.085	7.443	<1.0e-8
Somatic concern	0.218	0.181	1.209	0.227
Unusual thought content	0.644	0.072	8.950	<1.0e-8
Lack of judgment and insight	0.307	0.104	2.942	3.3e-3
3. Disorganized thought				
Conceptual disorganization	0.670	0.064	10.508	<1.0e-8
Difficulty in abstract thinking	0.489	0.095	5.171	2.3e-7
Mannerisms and posturing	0.519	0.085	6.085	<1.0e-8
Disorientation	0.487	0.069	7.082	<1.0e-8
Poor attention	0.701	0.060	11.578	<1.0e-8
Disturbance of volition	0.400	0.092	4.326	1.5e-5
Preoccupation	0.686	0.064	10.706	<1.0e-8
4. Uncontrolled hostility/excitement				
Excitement	1.004	0.115	8.721	<1.0e-8
Hostility	0.531	0.081	6.535	<1.0e-8
Uncooperativeness	0.567	0.094	6.053	<1.0e-8
Poor impulse control	0.589	0.089	6.611	<1.0e-8
5. Anxiety/depression				
Anxiety	0.908	0.128	7.100	<1.0e-8
Guilt	0.347	0.128	2.701	6.9e-3
Tension	0.673	0.124	5.431	5.6e-8
Depression	0.335	0.117	2.868	4.1e-3

Table 6. Correlation matrix estimate of the five factors

(*= significant at level 0.005; **= significant at level 0.0005)

Factors	1. Negative symptoms	2. Positive symptoms	3.Disorganized thought	4. Uncontrolled hostility/excitement	5. Anxiety/ depression
1. Negative symptoms	1	0.246*	0.560**	-0.011	-0.102
2. Positive symptoms	0.246**	1	0.740**	0.390**	0.107
3. Disorganized thought	0.560**	0.740**	1	0.317**	0.038
4. Uncontrolled hostility/excitement	-0.011	0.390**	0.317**	1	0.372**
5. Anxiety/depression	-0.102	0.107	0.038	0.372**	1

Summary of Application Results

- The proposed maximum composite likelihood estimators have little biases, the standard error estimates accurately, and the 95% confidence intervals have correct coverage probability.
- In the real data analysis, we consider the 5-factor model of the PANSS score with the following five factors: 1) Negative symptoms; 2) Positive symptoms; (3) Disorganized thought; (4) Uncontrolled hostility/excitement; and (5) Anxiety/Depression.
- We include age, gender, and race in the regression model and evaluate their effects on the five latent factors. The ranges of the factor loadings for the five factors are: (1) 0.543 – 0.770; (2) 0.218 – 0.786; (3) 0.400 – 0.701; (4) 0.531 –1.004; and (5) 0.335 – 0.908. Males appear to have lower values in factor "Anxiety/Depression" with a parameter estimate of -0.448 (p-value=0.008).
- Compared to other race groups, "White" patients have higher values in latent factors "Positive symptoms", "Uncontrolled hostility/excitement", and "Anxiety/Depression" with parameter estimates of 0.484 (p-value<0.001), 0.343 (p-value=0.004), and 0.462 (p-value=0.001).
- Older patients have lower values in latent factors "Positive symptoms" and "Uncontrolled hostility/excitement" with p-values of 0.032 and 0.012, respectively.

CONCLUSIONS

- Improved outcomes and disease states measured in psychiatric patients to further CNS drug development.
- Improved power for testing covariate effects on the latent factors
  - outperforms the existing methods in which each item is equally weighted in defining the factor scores while the factor loadings are different across different items.
- Other subjective factors unaddressed in this work but will be addressed in future
  - rater to rater variability,
  - rater bias stemming from the order in which he/she rates patients
- Longitudinal modeling of data will be in future work to compare efficacy between treatment arms

REFERENCES

1. Kay, S.R., Fiszbein, A., and Opler, L.A. The Positive and Negative Syndrome Scale (PANSS) for Schizophrenia. Schizophr Bull, (1987); 13 (2): 261-276.
2. Marder SR, Davis JM, Chouinard G. The effects of risperidone on the five dimensions of schizophrenia derived by factor analysis: combined results of the North American trials. J Clin Psychiatry. 1997;58:538–546.
3. Wallwork RS, Fortgang R, Hashimoto R, Weinberger DR, Dickinson D. Searching for a consensus five-factor model of the Positive and Negative Syndrome Scale for schizophrenia. Schizophrenia Res. 2012;137:246–250.
4. Langeveld, J., Andreassen, O. A., et. al. Is there an optimal factor structure of the Positive and Negative Syndrome Scale in patients with first-episode psychosis? Scandinavian Journal of Psychology, (2013);54(2), 160-165.
5. Reininghaus, U., Priebe, S. and Bentall, R.P. Testing the Psychopathology of Psychosis: Evidence for a General Psychosis Dimension. Schizophr Bull, (2013); 39 (4): 884-895.
6. Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997), "Latent variable models for mixed discrete and continuous outcomes," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59, 667–678.
7. Shi, J.-Q. and Lee, S.-Y. (2000), "Latent variable models with mixed continuous and poly- tomous data," Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62, 77–87.
8. Lee, S.-Y. and Song, X.-Y. (2004), "Maximum likelihood analysis of a general latent variable model with hierarchically mixed data," Biometrics, 60, 624–636.
9. Teixeira-Pinto, A. and Normand, S.-L. T. (2009), "Correlated bivariate continuous and bi- nary outcomes: issues and applications," Statistics in medicine, 28, 1753–1773.
10. Qaqish, B. F. and Ivanova, A. (2006), "Multivariate logistic models," Biometrika, 93, 1011– 1017.
11. Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., and Jöreskog, K. G. (2012), "Pairwise likelihood estimation for factor analysis models with ordinal data," Computational Statistics & Data Analysis, 56, 4243–4258.
12. Kolenikov, S., Angeles, G., et al. (2004), "The use of discrete data in PCA: theory, simulations, and applications to socioeconomic indices," Chapel Hill: Carolina Population Center, University of North Carolina, 1–59.
13. Baglin, J. (2014), "Improving your exploratory factor analysis for ordinal data: a demon- stration using FACTOR," Practical Assessment, Research & Evaluation, 19, 2.

ACKNOWLEDGMENTS/ DISCLOSURES

This research was supported by Otsuka Pharmaceutical Development & Commercialization, Inc. Srikanth Gottipati and Peter Zhang are employees at Otsuka Pharmaceutical Development & Commercialization, Inc

One or more authors report potential conflicts which are described in the program.

International Society for CNS Clinical Trials and Methodology - Autumn Conference, October 15th, 2018