Latent Factor Decomposition Model: Applications for Questionnaire Data*

Connor J. McLaughlin[1], Efi G. Kokkotou[2], Jean A. King[3], Lisa A. Conboy[4], Ali Yousefi[1]

Abstract— **The analysis of clinical questionnaire data comes with many inherent challenges. These challenges include the handling of data with missing fields, as well as the overall interpretation of a dataset with many fields of different scales and forms. While numerous methods have been developed to address these challenges, they are often not robust, statistically sound, or easily interpretable. Here, we propose a latent factor modeling framework that extends the principal component analysis for both categorical and quantitative data with missing elements. The model simultaneously provides the principal components (basis) and each patients' projections on these bases in a latent space. We show an application of our modeling framework through Irritable Bowel Syndrome (IBS) symptoms, where we find correlations between these projections and other standardized patient symptom scales. This latent factor model can be easily applied to different clinical questionnaire datasets for clustering analysis and interpretable inference.**

I. INTRODUCTION

This application proposes a solution to issues faced in clinical datasets built from patient questionnaires. The first issue is in the prominence of missing data, especially in longer questionnaires. Additionally, these questions often result in a dataset with many fields of both categorical and continuous data, which renders them difficult to interpret. The question of how best to combine individual fields to summarize the state of a patient's condition is the focus of this research. The efficacy of our method is demonstrated through a dataset on Irritable Bowel Syndrome (IBS) [1].

We can approach inference of questionnaire data as a problem of dimensionality reduction, i.e. how best to transform the many clinical measurements into fewer, representative variables. One common method for this transformation is to take a summation of patient responses. These aggregate measures can be effective for expressing the magnitude of patient condition severity, as done in the Whorwell Severity Scale for IBS [2]. This method, however, concedes information about the specific condition of each patient; two patients with dissimilar symptoms may be summarized with the same aggregate score. To capture this underlying structure, we must utilize more robust dimensionality reduction techniques.

The most prominent method for dimensionality reduction is Principal Component Analysis (PCA) [3], which reduces a dataset to a set of uncorrelated variables which explain the most variance present in the data. While this technique is useful for dimensionality reduction without losing information, this approach is not equipped to handle categorical data, as found in questionnaires. Multiple Correspondence Analysis (MCA) [4] is equivalent to PCA for nominal data, and may also be extended to handle both categorical and continuous measures. MCA may also be extended to handle mixed data types (combinations of categorical and continuous measures), known as Factor Analysis of Mixed Data (FAMD). The drawback to these methods is that the handling of missing data is not embedded in the method and it requires a secondary process. Furthermore, these methods are developed with pre-set assumptions on the categorical and quantitative data distributions, which may not hold for certain datasets. The outcome of these measures is a point estimate of data points' coordinates in the principal component space, which can mask the statistical significance of these projections. More recently, flexible techniques such as t-SNE [5], UMAP [6], or VAEs [7] may also be used to reduce the dimensionality of mixed data, but the interpretability of the learned transformation is complex.

To answer challenges in the analysis of questionnaire data, we propose a probabilistic model which may simultaneously handle both missing data and mixed data types in the search of principle components and new data points' coordinates. This model reduces questionnaire responses into a low dimensional approximately-orthogonal basis (principal components), which allows for the separation of different underlying conditions. This mapping allows clinicians to better capture the condition affecting each patient. Our principle components and corresponding point coordinates provide interpretable pictures of how different symptoms are grouped, and how much of each symptom group is present for each patient. Furthermore, the latent variables for each patient may be used as a predictor of treatment outcome, or as an alternative measure for monitoring disease progression.

The structure of our paper is as follows: Section II introduces our modeling framework formulation. Section III shows the interpretation and results of our model for IBS. Section IV provides an analysis of the significance of our model.

II. MATERIALS AND METHODS

A. Proposed Model Framework

Our method finds the first d principal components in patient questionnaire data with n patients and m symptoms. Vector $\mathbf{y_n}$ has m elements representing the nth patient's responses to the questionnaire. Each element of the vector represents a symptom, which may be present, not present, or not reported. Vector $\mathbf{x_n}$ with d elements, is the latent weight vector for the nth patient, where the principal components are defined by a matrix $\mathbf{A_{mxn}}$ and intercept vector $\mathbf{b_{mx1}}$. Using $\mathbf{A}$ and $\mathbf{x_n}$, the probability of observing a $\mathbf{y_n}$ is defined by:

$$P(y_{nj}|x_n) \sim \text{Bernoulli}(P_{nj})_{j=1...m} \quad (1a)$$
$$\text{logit } P_{nj} = A_j x_n + b_j \quad (1b)$$

Where y_{nj} is the jth element of the vector $\mathbf{y_n}$. $\mathbf{A_j}$ is the jth row of the basis $\mathbf{A}$, and vector $\mathbf{b}$ is the expected outcome of the jth symptom, in logit scale, across all patients. Note that $\mathbf{x_n}$ is not observed, and it is a latent variable in our model.

Here, we assume the latent variables follow a multivariate normal with common prior μ_0 and covariance of Σ_0: $x_n \sim N(\mu_0, \Sigma_0)$.

The likelihood of observing the symptoms $\mathbf{Y_{nj}}$ given $\mathbf{x_n}$ is defined by:

$$L(y_{nj}|x_n) = \begin{cases} P_{nj} & \text{if } Y_{nj} = 1 \text{ (present)} \\ 1 - P_{nj} & \text{if } Y_{nj} = 0 \text{ (not present)} \\ 1 & \text{not reported} \end{cases} \quad (2)$$

and the full likelihood of observed symptoms across patients is defined by:

$$L(y_{1..N}; x_{1..N}, \mu_0, \Sigma_0, A, b) = \prod_{n=1}^{N} \prod_{j=1}^{M} L(y_{nj}|x_n) \quad (3)$$

We assume the symptoms are independent given $\mathbf{x_n}$, and that each participants responses are independent from others. The posterior distribution on our model parameters and our latent variable is defined by:

$$P(x_{1...N}, A, b, y_{1...N}; \mu_0, \Sigma_0)$$
$$\propto L(y_{1...N}; x_{1...N}, \mu_0, \Sigma_0, A, b) \prod_{n=1}^{N} P(x_n; \mu_0, \Sigma_0) \quad (4)$$

where $P(x_n; \mu_0, \Sigma_0)$ is the likelihood of $\mathbf{x_n}$ given (μ_0, Σ_0) using the multivariate normal prior.

We can also include continuous measurements in our model. Let vector $\mathbf{z_n}$, with k elements, represent the nth patient's continuous measurements. Each element, as with our categorical data, may be measured or not reported. The likelihood of observing measurement $\mathbf{z_{nk}}$ given $\mathbf{x_n}$ is defined by:

$$L(z_{nk}|x_n) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma_k} \exp \frac{-(z_{nk} - A_k x_n - b_k)^2}{2\sigma_k^2} & \text{if measured} \\ 1 & \text{not reported} \end{cases} \quad (5)$$

where we assume $\mathbf{A}$ and $\mathbf{b}$ to have $m+k$ rows. We assume $\mathbf{z_{nk}}$ follows a normal distribution with mean $\mathbf{A_k x_n} + \mathbf{b_k}$ and unknown variance σ_k. Other continuous distributions, such as log-normal or exponential, could also be used to model $\mathbf{z_{nk}}$, depending on the type of measurements. For cases of mixed data, our model likelihood is defined by:

$$L(z_{1...N}, y_{1...N}; x_{1...N}, \mu_0, \Sigma_0, A, b) = L(y_{1...N}; x_{1...N}, \mu_0, \Sigma_0, A, b)$$
$$* \prod_{n=1}^{N} \prod_{k=1}^{K} L(z_{nk}|x_n) \quad (6)$$

where we assume categorical and continuous measurements are statistically independent given $\mathbf{x_n}$. We can similarly define the posterior of the model parameters like in equation 4, where we only have categorical data. Note that we now have a new set of parameters σ_k^2 which need to be estimated along other parameters in the model.

The latent factor model defined in equation 6 can represented as a graphical model as shown in Figure 1. Note that $\theta = \{\mu_0, \Sigma_0\}$ is the prior on our latent variable $\mathbf{X}$.
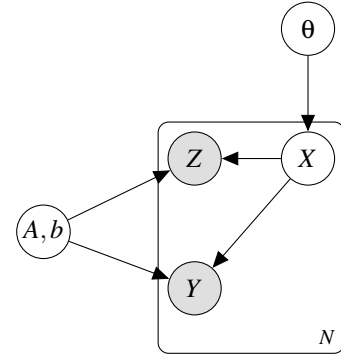

Fig. 1. Graphical model for latent factor model. θ is the set of parameters characterizing prior distributions of the latent variable $\mathbf{X}$; $\mathbf{Z}$ and $\mathbf{Y}$ are continuous and categorical observed variables. $\mathbf{A}$ and $\mathbf{b}$ are the observation model free parameters.

B. Parameter Optimization

Our objective is to find $\mathbf{A}$ and $\mathbf{b}$ along with $\mathbf{x_{1...N}}$ which maximize the likelihood defined in equation 4 or 6. This objective can be expanded to mixed data by including $\sigma_{1...k}$ in the optimization step. The optimization function is defined by:

$$\underset{x_{1...N}, A, b}{\text{argmax}} \ P(x_n, A, b, y_n; \sigma_0, \Sigma_0)$$
$$s.t. ||A^T A - I||_F^2 < \gamma \quad (7)$$

where $||A^T A - I||_F^2$ is the Frobenius norm and γ is a small positive number. This norm will imply matrix $\mathbf{A}$ to be approximately orthonormal.

When $\mathbf{x_n}$ are known, we can use out-of-box numerical methods to find $\mathbf{A}$ and $\mathbf{b}$ elements. Here, we must simultaneously estimate $\mathbf{x_{1..N}}$ and $\mathbf{A}$ and $\mathbf{b}$ elements; thus we use expectation-maximization (EM) techniques to recursively find posterior estimation of $\mathbf{x_{1..N}}$ and $\mathbf{A}$ and $\mathbf{b}$ elements that maximize the posterior, as in [8] [9]. The EM technique is

an established solution for maximum likelihood estimation in the presence of latent and/or missing data points. We use sampling techniques, e.g. Monte-Carlo, to find the posterior of $x_{1..N}$ and use those samples to find what $\mathbf{A}$ and $\mathbf{b}$ will maximize the expected log of the posterior. Our implementation can be found at the github link: `https://github.com/YousefiLab/LatentFactorQuestionnaire`.

In the following section, we show an application of our methodology in the IBS dataset. We focus on reported IBS symptom data, so our model contains only categorical data.

III. RESULTS

A. IBS Dataset

We studied the dataset from the clinical trial initially described in [1]. This includes samples from 230 IBS patients, with measurements taken at three time points throughout the study. Measurements consist primarily of questionnaire data based on IBS pain and reported symptoms. The primary outcome measures analyzed in previous studies are the IBS Global Improvement Scale [10], IBS Symptom Severity Scale [2], and the IBS Quality of Life (IBS-QOL) measure [11]. Each of these scales are formed through questionnaire data, either in single questions or through the aggregation of many questions (e.g. 34 questions for IBS-QOL). In our analysis, we sought to create a more robust representation of patients' overall symptoms, based on a series of 19 binary symptom variables. We focus our analysis on the group of patients who received treatment (non-waitlist group), in the form of sham acupuncture and patient-practitioner interaction. We excluded data points where the patient dropped out of the study after the baseline measurements (n=126). Our proposed model is especially useful for the analysis of conditions such as IBS, which have high inter-patient variability [12].

B. Basis and Latent Variables

We applied our modeling framework on the IBS dataset under the assumption that two principal components may capture the structure of reported symptoms. Figure 2 shows each of the questionnaire symptoms plotted in the transformed principle components coordinates. We see clear groupings of symptoms along each of these components. Along the first component, we note the importance of symptoms associated with the lack of psychological or sleep-related issues, such as insomnia, dry mouth, and difficulty concentrating (lower left of figure). Along the second component, we note the importance of symptoms generally associated with physical issues, such as diarrhea, headache, muscle/joint pain, and loss of appetite (upper right of figure). This offers a clear interpretation for physicians, as we can now understand how each latent component describes the patients' overall disease burden.

Our centering term $\mathbf{b}$ provides the expected log-odds of each symptom across all patients. In figure 2 we can see that headache has the largest intercept term, indicating it may be a more common symptom across patients, whereas palpitations
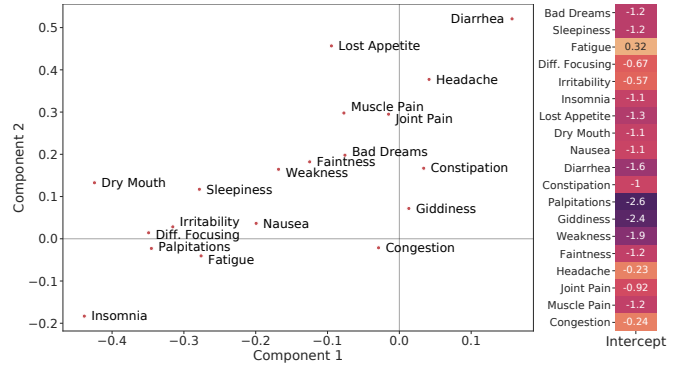

Fig. 2. Fitted parameters $\mathbf{A}$, $\mathbf{b}$ on IBS symptom data. (Left) Basis $\mathbf{A}$ shown with each component in logit space, meaning symptoms with higher values in basis n have higher probability given positive x_n. (Right) The intercept term $\mathbf{b}$, which represents the expected log-odds of each symptom across patients.

has a strong negative log-odds, indicating that it may be a rarer and more severe symptom.

Given this context, latent variable estimates for each patient provide insight into the observed symptoms for each patient. For example, a patient with a higher value of x_1 and a lower value of x_2 may have more physical pain symptoms and few psychological or sleep symptoms. These variables provide effective summary measures which may result in distinctive clusters, showing commonalities across patient groups.

In the next section, we discuss how these latent variables relate to other known measures of IBS.

C. Correlation at Baseline

We subsequently examined the potential correlation of our summary statistics (x_1 and x_2) with other patient metrics measured at baseline using established scales. In table I, we show Pearson correlation coefficients of x_1 and x_2 with IBS-specific and non-specific measures of physical pain and psychological distress. Recall that x_1 indicates a lack of psychological symptoms, and x_2 indicates the presence of physical pain symptoms. We note that there is a consistent negative correlation with x_1 across all metrics, suggesting perhaps that alleviation of the psychological cluster of symptoms might have a positive impact on the overall patient symptomatology. For x_2, there is a positive correlation with anxiety and both general and IBS-specific pain, which indicates that treatment of diarrhea, lost appetite, and headache may result in reduced anxiety and pain.

In a broader sense, these correlation results imply that x_1 and x_2 bring new and meaningful summary metrics for each patient which can be used for finding associations with symptom clusters and established outcome measures.

D. Comparison of Outcome Measures

Given the result demonstrated in the previous section, we additionally explore the relationship between the change in latent variables and treatment impact.

We fit linear regression models using the change in each latent variable to predict the change in each of the metrics described above. We find statistically significant results for both

Metric	X1	X2
Carroll Depression Scale	-.321**	.155
Beck Anxiety Inventory	-.315***	.389***
McGill Pain Scale	-.315**	.284**
IBS-QOL	-.279**	.151
IBS-SSS	-.235*	.298**

TABLE I. Correlation between latent variables and other metrics taken at study baseline. $^*p<0.05$; $^{**}p<0.01$, $^{***}p<0.001$.
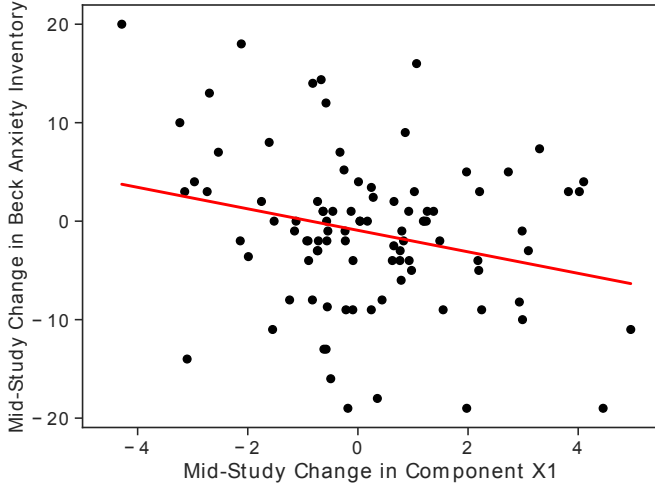

Fig. 3. Linear Regression analysis of delta x_1 vs. delta Beck Anxiety Inventory. Mid-study change corresponds to values reported 3 weeks after treatment compared to baseline.
Δ Anxiety = -1.092 (Δx_1) - 0.931
$P > |t| = 0.0223$; $R^2 = 0.064$

Beck Anxiety Inventory and IBS-SSS. Change in $\mathbf{x_1}$ could predict change in Beck Anxiety Inventory, indicating that measures to reduce symptoms such as insomnia, dry mouth, and fatigue may reduce patient anxiety. Additionally, change in $\mathbf{x_2}$ could predict change in IBS-SSS, which indicates that limiting diarrhea, loss of appetite, and headache, might have a positive impact in IBS-specific symptoms, including abdominal pain and distortion.

IV. CONCLUSION

We approached the challenges presented by questionnaire data through a probabilistic model framework. The output of our model is an approximately orthonormal basis, which we show is effective in visualizing groupings of responses as reported in questionnaire data, as well as latent data points which we express as an interpretable treatment outcome measure.

Through our study in IBS, we found interpretable clusters of patient symptoms, with one grouping of physical pain symptoms and another of psychological and sleep-related symptoms. Using the latent space projection for each patient resulted in statistically significant findings for both correlation with other baseline clinometrics as well as Linear Regression modeling of metric deltas.

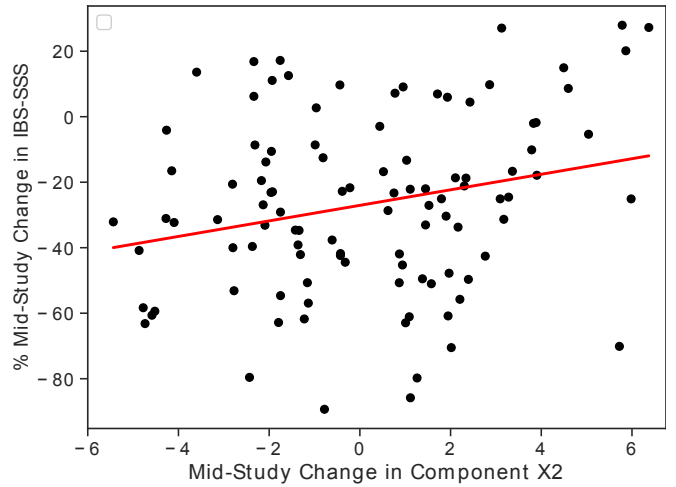This framework can be applied to other clinical questionnaire datasets with missing and mixed data types. In

Fig. 4. Linear Regression analysis of delta x_2 vs. delta (%) IBS Symptom Severity Scale. Mid-study change corresponds to values reported 3 weeks after treatment compared to baseline.
Δ IBS-SSS = 2.373 (Δx_2) - 27.069
$P > |t| = 0.0148$; $R^2 = 0.060$

addition to the model outputs used in the application of IBS, our model may be used to obtain probability estimates for each missing data point, and the distribution of each patient data projection can also be studied to further understand the significance of these summary statistics. With each path of analysis, our solution offers interpretable results, a key element for clinical analysis which is missing from other widely used techniques such as t-SNE and UMAP.

REFERENCES

[1] A. J. Lembo, L. Conboy, J. M. Kelley, R. S. Schnyer, C. A. McManus, M. T. Quilty, C. E. Kerr, D. Drossman, E. E. Jacobson, R. B. Davis, and T. J. Kaptchuk, "A treatment trial of acupuncture in IBS patients," *The American Journal of Gastroenterology*, vol. 104, pp. 1489–1497, May 2009.

[2] C. Y. FRANCIS, J. MORRIS, and P. J. WHORWELL, "The irritable bowel severity scoring system: a simple method of monitoring irritable bowel syndrome and its progress," *Alimentary Pharmacology & Therapeutics*, vol. 11, pp. 395–402, Apr. 1997.

[3] K. P. F.R.S., "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[4] M. Tenenhaus and F. Young, "An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data," *Psychometrika*, vol. 50, pp. 91–119, 02 1985.

[5] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[6] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," 2020.

[7] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling incomplete heterogeneous data using vaes," 2020.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[9] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[10] S. Gordon, V. Ameen, B. Bagby, B. Shahan, P. Jhingran, and E. Carter *Digestive Diseases and Sciences*, vol. 48, no. 7, pp. 1317–1323, 2003.

[11] D. A. Drossman, D. L. Patrick, W. E. Whitehead, B. B. Toner, N. E. Diamant, Y. Hu, H. Jia, and S. I. Bangdiwala, "Further validation of the ibs-qol: a disease-specific quality-of-life questionnaire," *The American Journal of Gastroenterology*, vol. 95, no. 4, pp. 999–1007, 2000.

[12] F. Mearin, E. Baró, M. Roset, X. Badía, N. Zárate, and I. Pérez, "Clinical patterns over time in irritable bowel syndrome: Symptom instability and severity variability," *American Journal of Gastroenterology*, vol. 99, pp. 113–121, Jan. 2004.