Unfolding-Model-Based Visualization: Theory, Method and Applications

Yunxiao Chen[*], Zhiliang Ying[†], Haoran Zhang[‡]

Abstract

Multidimensional unfolding methods are widely used for visualizing item response data. Such methods project respondents and items simultaneously onto a low-dimensional Euclidian space, in which respondents and items are represented by ideal points, with person-person, item-item, and person-item similarities being captured by the Euclidian distances between the points. In this paper, we study the visualization of multidimensional unfolding from a statistical perspective. We cast multidimensional unfolding into an estimation problem, where the respondent and item ideal points are treated as parameters to be estimated. An estimator is then proposed for the simultaneous estimation of these parameters. Asymptotic theory is provided for the recovery of the ideal points, shedding lights on the validity of model-based visualization. An alternating projected gradient descent algorithm is proposed for the parameter estimation. We provide two illustrative examples, one on users' movie rating and the other on senate roll call voting.

KEY WORDS: Multidimensional Unfolding; Data Visualization; Distance Matrix Completion; Item Response Data; Embedding.

[*]Department of Statistics, London School of Economics and Political Science. Address: Columbia House, Room 5.16, Houghton Street, London, WC2A 2AE. Email: y.chen186@lse.ac.uk

[†]Department of Statistics, Columbia University

[‡]Shanghai Center for Mathematical Sciences, Fudan University

1 Introduction

Multidimensional unfolding (MDU) methods are widely used as an important data visualization tool in social and behavioral sciences such as psychology (Van Deun et al., 2007; Papesh and Goldinger, 2010), political science (Poole, 2000, 2005; Clinton et al., 2004a; Bakker and Poole, 2013), and marketing (DeSarbo and Hoffman, 1987; DeSarbo et al., 1997; Ho et al., 2010). It is regarded as the dominant method in the scaling of both preferential choice and attitude (Mair et al., 2015). The basic idea of MDU is to place both respondents and items in a joint Euclidean space based on data, with the understanding that respondents tend to prefer items that are close to them in the space. This joint visualization may lead to better understanding and interpretations of both the respondents and the items, as compared with separately visualizing the respondents and the items by themselves. MDU has its origin in psychology (Bennett, 1956; Bennett and Hays, 1960; Hays and Bennett, 1961; Coombs, 1964). It is closely related to multidimensional scaling (MDS) methods (Kruskal, 1964; Kruskal and Wish, 1978; Borg and Groenen, 2005) and several other recent approaches to nonlinear dimension reduction and manifold learning (Tenenbaum et al., 2000; Lu et al., 2005; Chen and Buja, 2009; Zhang et al., 2016).

MDU methods can be categorized into two types, algorithm-based and model-based. Algorithm-based methods (e.g., Takane et al., 1977; Greenacre and Browne, 1986; De Leeuw and Mair, 2009) estimate the ideal points by minimizing a certain objective function, also known as the stress function in the literature of MDU. The classical algorithm-based methods have been implemented in the R package *smacof* (De Leeuw and Mair, 2009) that is widely used for MDU and MDS analysis. Model-based methods (e.g., DeSarbo and Hoffman, 1987; Hinich, 2005; Bakker and Poole, 2013), however, infer the locations of the ideal points by making use of a probabilistic model. Such a model typically assumes that, up to some measurement error, the similarity between a person and an item is a decreasing function of some defined

distance between the corresponding ideal points. The specification of MDU models is closely related to item response theory models in psychometrics (see e.g., Embretson and Reise, 2000; Rabe-Hesketh and Skrondal, 2004; Bartholomew et al., 2011).

The MDU problem is closely related to MDS. The key difference is that data for the former do not contain direct measurement of within-set (i.e., person-person and item-item) similarities, while data for MDS typically have such information. Largely due to the missing information contained in the within-set similarities, the MDU problem tends to be more challenging. As a result, degenerate solutions are often encountered in the applications of MDU methods, in which case the visualization and the corresponding interpretations convey no information (e.g., Busing et al., 2005; Borg and Groenen, 2005), while MDS results tend to be more stable. These empirical observations suggest that it is of importance to study the validity of MDU solutions, which motivates the research in this paper.

This paper studies the visualization of MDU from the statistical perspective. First, for binary choice data, we formulate the MDU problem into a parameter estimation problem under a general family of probabilistic MDU models, where the respondent and item ideal points are treated as parameters to be estimated. Second, an estimator is proposed for the ideal points and an asymptotic theory is provided for this estimator, shedding lights on the validity of model-based visualization. Finally, an efficient alternating projected gradient algorithm is proposed for the computation which is scalable to large-scale problems.

We illustrate the proposed method through two applications, one on movie rating and the other on senate roll call voting. The movie dataset is a subset from the famous MovieLens dataset (Harper and Konstan, 2016). We unfold the 943 users and 338 movies in the dataset. Specifically, we study the users' movie watching decisions. Based on the ideal points of movies in a two-dimensional space, it is found that one dimension of the space corresponds to the popularity of the movies and the other

dimension corresponds to the release date of the movies. Good understanding of the user ideal points is further obtained based on their distances to the movie ideal points. The senate voting dataset is based on the senate roll call voting records from the 108th congress in 2003-2004. Based on the unfolding of the senators and roll calls, it is found that most of the ideal points lie around a one-dimensional line, with the two extremes of the line representing the most liberal and the most conservative political standings.

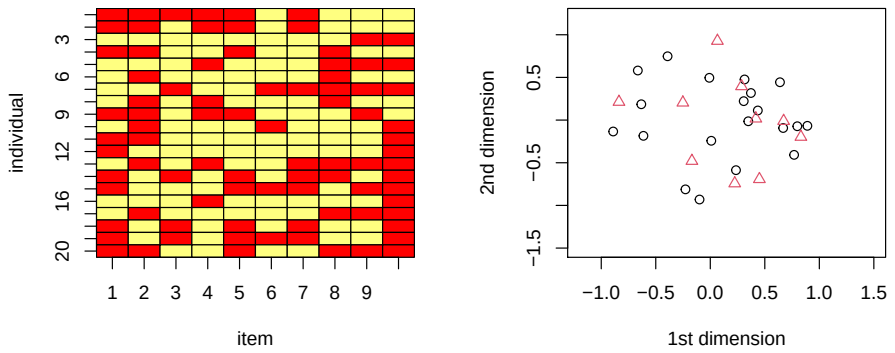
The rest of the paper is organized as follows. In Section 2, we introduce a family of MDU models and formulate the problem of joint configuration recovery into an estimation problem. In Section 3, we propose an estimator, for which statistical theory is established that guarantees the consistency of configuration recovery under reasonable conditions. Simulation studies and real data examples are presented in Sections 4 and 5, respectively. We end with discussions on future directions in Section 6. An application to cluster analysis, proofs of the theoretical results, and numerical comparison with classical MDU methods are provided as supplementary materials.

2 Distance-based MDU

2.1 Distance-based Unfolding Model for Binary Data

Consider N respondents making choice on J binary items (e.g., "agree/disagree"). Let Y_{ij} be a random variable, denoting the response from respondent i to item j, taking value 0 or 1, and let y_{ij} be its realization. For example, such data can come from senate roll call voting, where the respondents are senators and the items correspond to roll calls. Response $Y_{ij} = 1$ means that senator i supports roll call j and $Y_{ij} = 0$ otherwise.

We provide a simulated example in Figure 1 to illustrate MDU analysis. Panel (a)



(a) (b)

Figure 1: An illustrative example. Panel (a): The heatmap of a response matrix, where 0 and 1 responses are represented by red and yellow colors, respectively. Panel (b): The respondent and item ideal points where black circles represent respondents and red triangles represent items.

shows the heat map of an observed response matrix which consists of 20 respondents and 10 items, where 0 and 1 responses are represented by red and yellow colors, respectively. Given choice data in panel (a), an MDU method aims at representing respondents and items by ideal points in the same low-dimensional Euclidian space $\mathbb{R}^K$ as in panel (b) of Figure 1 that can be easily visualized, where the respondent-respondent, respondent-item, and item-item relationships are captured by the between-points distance. The dimension K of the Euclidian space is often set to be 2 or 3 for the purpose of visualization.

One way to conduct MDU is via a statistical model. An MDU model typically assumes that each respondent/item is associated with a true ideal point in $\mathbb{R}^K$ that is represented by a K-dimensional parameter vector. Let $\boldsymbol{\theta}_i = (\theta_{i1}, ..., \theta_{iK})^\top$ and $\mathbf{a}_j = (a_{j1}, ..., a_{jK})^\top$ denote the parameter vectors of respondent i and item j, respectively. It is assumed that response Y_{ij} is determined by the Euclidian distance between $\boldsymbol{\theta}_i$ and $\mathbf{a}_j$ in $\mathbb{R}^K$. Finally, we use $\Theta^N = (\theta_{ik})_{N\times K}$ and $A^J = (a_{jk})_{J\times K}$ to denote the matrices containing all the person and the item ideal points, respectively. Under

such a statistical model, the goal of MDU becomes to estimate the person and item parameters based on data.

In this paper, we focus on MDU models taking the form

$$P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i, \mathbf{a}_j) = f(\|\boldsymbol{\theta}_i - \mathbf{a}_j\|^2), \tag{1}$$

where $\|\cdot\|$ denotes the standard L_2 norm and $f : [0, \infty) \to [0, 1]$ is a pre-specified link function. It is assumed that the responses Y_{ij} are conditionally independent, given the ideal points $\boldsymbol{\theta}_i$ and $\mathbf{a}_j$, $i = 1, ..., N, j = 1, ..., J$. This model falls under the general framework of the MDU threshold model for binary choice data (see DeSarbo and Hoffman, 1987). According to the form of (1), the distribution of data only depends on the squared distance between every pair of person and item ideal points, $d_{ij} = \|\boldsymbol{\theta}_i - \mathbf{a}_j\|^2, i = 1, ..., N, j = 1, ..., J$. The matrix $D_{N,J} = (d_{ij})_{N \times J}$ is known as the corresponding *partial distance matrix*, where the subscripts of $D_{N,J}$ emphasize the dependence of this matrix on the numbers of respondents and items.

In addition, the link function f is often assumed to be a monotone decreasing function, so that a larger distance implies a lower probability of $Y_{ij} = 1$. An example of such a link function is $f(x) = 2/(1 + \exp(x))$. When $f(x)$ takes this form, $P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i, \mathbf{a}_j) = 1$ when the distance between $\boldsymbol{\theta}_i$ and $\mathbf{a}_j$ is 0, i.e., the two points are identical, and the probability $P(Y_{ij} = 1 \mid \boldsymbol{\theta}_i, \mathbf{a}_j)$ decays towards 0 when the distance increases.

In what follows, we provide two remarks on this modeling framework.

Remark 1 *We remark on the link function f which plays a similar role as the dissimilarity transformation function in the classical MDS and MDU methods (e.g., Chapter 9, Borg and Groenen, 2005). Assuming a pre-specified f is similar to assuming an identity transformation in classical MDU.*

In classical MDS and MDU, the dissimilarity transformation function can be un-

known and estimated from data parametrically or non-parametrically. Similar treatment can be applied to the link function f. For example, one may assume

$$f(\|\boldsymbol{\theta}_i - \mathbf{a}_j\|^2) = g(\beta_0 + \beta_1\|\boldsymbol{\theta}_i - \mathbf{a}_j\|^2),$$

where $g : \mathbb{R} \to [0, 1]$ is a given monotone decreasing function and β_0 and β_1 are additional parameters to be estimated from data together with the person- and item-specific parameters. This form is similar in spirit to the interval transformation in classical MDU. When no constraint is imposed on the scales of $\boldsymbol{\theta}_i$s and $\mathbf{a}_j$s, β_1 needs to be fixed to be a constant (e.g., $\beta_1 = 1$) for model identifiability. One may also estimate f non-parametrically, for example, by using monotone splines.

Under suitable regularity conditions, our theoretical development in Section 3 can be extended to the case when f also needs to be estimated from data.

Remark 2 *Although we focus on binary data, the introduced modeling framework can be easily extended to other types of preference data, such as rating and ranking data. For example, consider rating data $Y_{ij} \in \{0, ..., T\}$, where 0, 1, ..., T are $T+1$ ordered response categories. A higher category implies a higher level of agreement between the respondent and the item. Then one can assume the following unfolding model*

$$P\left(Y_{ij} \geq t \mid \boldsymbol{\theta}_i, \mathbf{a}_j\right) = g\left(d_t + \|\boldsymbol{\theta}_i - \mathbf{a}_j\|^2\right), \tag{2}$$

for $t \in \{1, ..., T\}$, where $g : \mathbb{R} \to [0, 1]$ is a given monotone decreasing function and d_1, ..., d_T are additional model parameters. It implies that the larger the distance, the smaller the probability for Y_{ij} to take a large value. This model is closely related to the graded response model (Samejima, 1997) in item response theory. For another example, consider ranking data consisting of pair-wise comparisons, where each response is a comparison between two items j and j'. Following the same idea as above, one may model the probability that item j is preferred over j' to take the

form $g(\|\boldsymbol{\theta}_i - \mathbf{a}_j\|^2 - \|\boldsymbol{\theta}_i - \mathbf{a}_{j'}\|^2)$. *That is, the probability decreases with the difference of their squared distances to person* i. *Our theoretical results and computational algorithm given below can be adapted to these situations.*

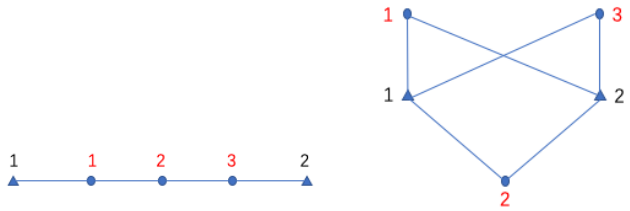
2.2 Recovery of Configuration

Our main goal is the simultaneous recovery of the ideal points $\boldsymbol{\theta}_i$ and $\mathbf{a}_j$, based on the observed binary responses $y_{ij}, i = 1, ..., N, j = 1, ..., J$. Since the model only relies on the Euclidian distance between the ideal points, two sets of points lead to the same model if they have the same configuration, i.e., one set of points can be obtained by applying an isometry mapping to the other. This is because, the distance between points is invariant under an isometry mapping. An isometry mapping F in $\mathbb{R}^K$ takes the form

$$F(\mathbf{x}) = O\mathbf{x} + \mathbf{b}, \quad \forall \mathbf{x} \in \mathbb{R}^K,$$

where O is a $K \times K$ orthogonal matrix and $\mathbf{b}$ is a vector in $\mathbb{R}^K$ (see, e.g., Olver, 1999). We further denote $\mathcal{A}_K$ as the set of all isometry mappings on $\mathbb{R}^K$. Without additional information, the best possible result one can expect is recovering the ideal points up to an isometry mapping. We refer to this problem as the recovery of ideal point configuration.

It is worth noting that regularity conditions are needed to ensure the recovery of the configuration. That is, it is possible that there exist multiple sets of ideal points with different configurations that lead to the same distribution of Y_{ij}s. In other words, the configuration of $\{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N, \mathbf{a}_1, ..., \mathbf{a}_J\}$ may not be unique only given the partial distance matrix. This is known as the situation of degeneration, in which case the visualization does not convey information or can even be misleading. A simple example is given in Figure 2, where the two different configurations in the two panels have the same partial distance matrix.



(a) (b)

Figure 2: An example of degenerate situation: The triangles represent item points and circles represent person points. The two configurations in $\mathbb{R}^2$ share the same partial distance matrix, where $d_{11} = 1^2$, $d_{12} = 3^2$, $d_{21} = 2^2$, $d_{22} = 2^2$, $d_{31} = 3^2$, and $d_{32} = 1^2$.

Following the above discussion, the validity of unfolding-model-based visualization relies on the accuracy of configuration recovery, a problem to be discussed. Specifically, we consider the following loss function for configuration recovery,

$$\min_{F \in \mathcal{A}_K} \frac{\sum_{i=1}^{N} \|\boldsymbol{\theta}_i^* - F(\hat{\boldsymbol{\theta}}_i)\|^2}{N} + \frac{\sum_{j=1}^{J} \|\mathbf{a}_j^* - F(\hat{\mathbf{a}}_j)\|^2}{J}, \tag{3}$$

where $\boldsymbol{\theta}_i^*$ and $\mathbf{a}_j^*$ denote the true ideal points and $\hat{\boldsymbol{\theta}}_i$ and $\hat{\mathbf{a}}_j$ denote the estimates from data $(y_{ij})_{N \times J}$. Note that (3) quantifies the accuracy of configuration recovery in an average sense, where isometry indeterminacy is bypassed by the minimization in (3) with respect to all isometry mappings in $\mathcal{A}_K$. We call (3) the *average loss* for the recovery of ideal point configuration. Error bounds will be established for (3) under reasonable conditions, which ensures the accurate recovery of the loss function when both N and J are large.

2.3 Connection with Other Scaling Methods

MDU is closely related to MDS, a class of methods for visualizing the similarity pattern between data points (Borg and Groenen, 2005). More precisely, MDS maps a set of variables onto a low dimensional space, based on data measuring the similarity between variables. As pointed out in Chapter 14, Borg and Groenen (2005), MDU

can be viewed as a special case of MDS, where the set of variables in MDS composes of both the respondents and items and the item response data $(y_{ij})_{N\times J}$ are regarded as measures of similarity between the respondents and the items, while the similarities within the two sets (i.e., respondents and items) are structurally missing; see Figure 3 for an illustration that is a reproduction of Figure 14.1 of Borg and Groenen (2005).

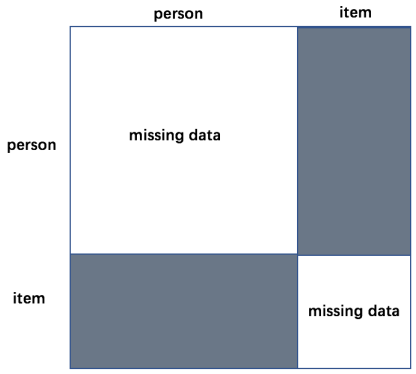


Figure 3: In MDU, the diagonal blocks are missing. All we observe are the off-diagonal blocks.

Little statistical theory has been developed for the recovery of configuration based on MDS models. The most relevant work is Zhang et al. (2016), in which an error bound is developed for the recovery of the complete distance matrix, under a linear MDS model without structurally missing data. However, little discussion is provided on the recovery of ideal point configuration, under an MDU setting.

The recovery of configuration is relatively easier under the setting of MDS with no structurally missing data. This is because, the complete data matrix of similarities will provide sufficient information on the complete distance matrix. The accurate recovery of the complete distance matrix further implies the accurate recovery of configuration under weak conditions, due to the one-to-one relationship between the complete distance matrix and the ideal point configuration as described in Proposition 1. Under the MDU setting, the recovery of configuration requires additional regularity conditions, due to the lack of direct measurement of within-set distances.

Proposition 1 *For* $\{\mathbf{x}_1, ..., \mathbf{x}_n\} \subset \mathbb{R}^K$, $\{\mathbf{y}_1, ..., \mathbf{y}_n\} \subset \mathbb{R}^K$, *if* $\|\mathbf{x}_i - \mathbf{x}_j\| = \|\mathbf{y}_i - \mathbf{y}_j\|$ *for all* i *and* j, *then there exists an isometry mapping* $F \in \mathcal{A}_K$ *such that* $F(\mathbf{x}_i) = \mathbf{y}_i$ *for all* $i = 1, ..., n$.

MDU is also related to other scaling methods for binary data such as item response theory (IRT; Embretson and Reise, 2000; Reckase, 2009) and multiple correspondence analysis (Gifi, 1990; Le Roux and Rouanet, 2010). Specifically, probabilistic models are available from IRT for multivariate binary data. An IRT model also represents respondents and items by low-dimensional parameter vectors, say $\boldsymbol{\theta}_i$ and $\mathbf{a}_j$. It also assumes that the probability of $Y_{ij} = 1$ is a function of $\boldsymbol{\theta}_i$ and $\mathbf{a}_j$. In this sense, the model introduced above can be viewed as a special IRT model, in which the probability of $Y_{ij} = 1$ is assumed to be a monotone decreasing function of $\|\boldsymbol{\theta}_i - \mathbf{a}_j\|$. However, the classical IRT models (see e.g., Embretson and Reise, 2000; Reckase, 2009) are not specified in this way. Consequently, it does not make sense to visualize the person and item parameter vectors jointly.

Multiple correspondence analysis is an algorithm-based approach that can be applied to binary data and produce low-dimensional scores for both respondents and items. These score vectors can be plotted jointly in the same space. However, as a common issue with algorithm-based approaches, the meaning of the distance between the score vectors is not clear and the uncertainty associated with the visualization is hard to quantify.

3 Theoretical Results

3.1 Configuration Recovery based on Perturbed Partial Distances

We first study the recovery of configuration from a perturbed partial distance matrix, when both N and J grow to infinity. Let $\boldsymbol{\theta}_i^*$, $i = 1, ..., N$, and $\mathbf{a}_j^*$, $j = 1, ..., J$ be the true person and item ideal points in $\mathbb{R}^K$, respectively, and let $D_{N,J}^*$ be the corresponding partial distance matrix. In addition, let $\tilde{\boldsymbol{\theta}}_i \in \mathbb{R}^{K_+}$ and $\tilde{\mathbf{a}}_j \in \mathbb{R}^{K_+}$ correspond to a perturbed version of the true configuration, satisfying

$$\|\tilde{D}_{N,J} - D_{N,J}^*\|_F^2 = o(NJ) \tag{4}$$

and $K_+ \geq K$, where $\tilde{D}_{N,J}$ denotes the partial distance matrix given by the perturbed configuration and $\|\cdot\|_F$ denotes the matrix Frobenius norm. One can think of K_+ as the latent dimension of the MDU model being applied to data, and $\tilde{\boldsymbol{\theta}}_i$ and $\tilde{\mathbf{a}}_j$ as some estimates of the person and item ideal points. For the time being, we treat K_+, $\tilde{\boldsymbol{\theta}}_i$, and $\tilde{\mathbf{a}}_j$ as given.

Based on the definition of matrix Frobenius norm, the left side of (4) has NJ terms, each of which is a squared distance between a true person-item distance and its perturbed value. Equation (4) implies that the perturbed partial distance matrix converges to the true one in an average sense, when both N and J grow to infinity.

We denote $\boldsymbol{\theta}_i^+ = ((\boldsymbol{\theta}_i^*)^\top, \mathbf{0}^\top)^\top$ and $\mathbf{a}_j^+ = ((\mathbf{a}_j^*)^\top, \mathbf{0}^\top)^\top$ in $\mathbb{R}^{K_+}$ as the embedding of the true ideal points in $\mathbb{R}^{K_+}$, where $\mathbf{0}$ denotes a zero vector. In what follows, we show that

$$\min_{F \in \mathcal{A}_{K_+}} \frac{\sum_{i=1}^N \|\boldsymbol{\theta}_i^+ - F(\tilde{\boldsymbol{\theta}}_i)\|^2}{N} + \frac{\sum_{j=1}^J \|\mathbf{a}_j^+ - F(\tilde{\mathbf{a}}_j)\|^2}{J} \to 0$$

as N and J grow to infinity, under reasonable conditions on the true ideal points.

Throughout this paper, we assume that ideal points are constrained in a compact set in $\mathbb{R}^K$.

A0. There exists a constant M such that $\|\boldsymbol{\theta}_i^*\| \leq M$ and $\|\mathbf{a}_j^*\| \leq M$ for all i and j.

To impose regularity conditions on the true configuration of the $N + J$ ideal points, which can vary with N and J, we introduce the notion of anchor points, two finite sets of points in $\mathbb{R}^K$ satisfying certain regularities that are independent of N and J.

Definition 1 *Two sets of points, $\{\mathbf{b}_1^*, ..., \mathbf{b}_{k_1}^*\}, \{\mathbf{c}_1^*, ..., \mathbf{c}_{k_2}^*\} \subset B_{\mathbf{0}}^K(M)$, are called a collection of anchor points of $\mathbb{R}^K$, if they satisfy conditions A1 and A2 below, where $B_{\mathbf{0}}^K(M)$ denotes a closed ball in $\mathbb{R}^K$ centered at $\mathbf{0}$ with radius M.*

Let $D^* = (\|\mathbf{b}_i^* - \mathbf{c}_j^*\|^2)_{k_1 \times k_2}$ be the partial distance matrix based on the anchor points, whose entries are assumed to be all positive (i.e., there is no identical points).

A1. There exists $\eta > 0$ such that for any partial distance matrix $D \in \mathbb{R}^{k_1 \times k_2}$ satisfying $\|D - D^*\|_F < \eta$, D has a unique configuration.

A2. Both $\{\mathbf{b}_1^*, ..., \mathbf{b}_{k_1}^*\}$ and $\{\mathbf{c}_1^*, ..., \mathbf{c}_{k_2}^*\}$ can affine span $\mathbb{R}^K$.

Remark 3 *According to Definition 1, we still get a collection of anchor points when slightly perturbing the points in a given anchor point collection in $\mathbb{R}^K$.*

According to condition A1, the anchor points are well-behaved points whose configuration can be uniquely determined by the partial distance matrix, even after a small perturbation. In addition, thanks to A2, the anchor points will help to anchor the rest of the points in $\mathbb{R}^K$, i.e., determining the configuration of a larger set of respondent and item ideal points.

Following the above concept of anchor points, it is intuitive that if there exist anchor points $\{\mathbf{b}_1^*, ..., \mathbf{b}_{k_1}^*\}$ and $\{\mathbf{c}_1^*, ..., \mathbf{c}_{k_2}^*\}$, satisfying that each $\mathbf{b}_i^*$ is surrounded

by sufficiently many respondent ideal points and each $\mathbf{c}_j^*$ is surrounded by sufficiently many item ideal points; that is, there exist a sufficient number of anchor points. Then it is relatively easy to recover the configuration of the ideal points from a perturbed partial distance matrix. This intuition is formalized by condition A3 below.

A3. There exists a collection of anchor points $\{\mathbf{b}_1^*, ..., \mathbf{b}_{k_1}^*\}$ and $\{\mathbf{c}_1^*, ..., \mathbf{c}_{k_2}^*\} \subset B_{\mathbf{0}}^K(M) \subset \mathbb{R}^K$ and $0 < \epsilon < M/10$ such that the closed balls with $\mathbf{b}_1^*, ..., \mathbf{b}_{k_1}^*$ and $\mathbf{c}_1^*, ..., \mathbf{c}_{k_2}^*$ as centers and radius ϵ, denoted by $B_{\mathbf{b}_1^*}(\epsilon), ..., B_{\mathbf{b}_{k_1}^*}(\epsilon)$ and $B_{\mathbf{c}_1^*}(\epsilon), ..., B_{\mathbf{c}_{k_2}^*}(\epsilon)$, do not overlap. The following two conditions are required to hold.

(1) For any $\mathbf{b}_1 \in B_{\mathbf{b}_1^*}(\epsilon), ..., \mathbf{b}_{k_1} \in B_{\mathbf{b}_{k_1}^*}(\epsilon)$ and $\mathbf{c}_1 \in B_{\mathbf{c}_1^*}(\epsilon), ..., \mathbf{c}_{k_2} \in B_{\mathbf{c}_{k_2}^*}(\epsilon)$, $\{\mathbf{b}_1, ..., \mathbf{b}_{k_1}\}$ and $\{\mathbf{c}_1, ..., \mathbf{c}_{k_2}\}$ are also a collection of anchor points.

(2) When N and J grow to infinity,

$$p_i = \liminf_{N\to\infty} \frac{\sum_{l=1}^N 1_{\{\|\boldsymbol{\theta}_l^* - \mathbf{b}_i^*\| < \epsilon\}}}{N} > 0, \quad i = 1, ..., k_1,$$
$$q_j = \liminf_{J\to\infty} \frac{\sum_{l=1}^J 1_{\{\|\mathbf{a}_l^* - \mathbf{c}_j^*\| < \epsilon\}}}{J} > 0, \quad j = 1, ..., k_2.$$

Theorem 1 *Suppose that A0 and A3 are satisfied for the true ideal points $\boldsymbol{\theta}_i^*$ and $\mathbf{a}_j^*, i = 1, ..., N, j = 1, ..., J$. Let $\tilde{\boldsymbol{\theta}}_i, \tilde{\mathbf{a}}_j \in B_{\mathbf{0}}^{K_+}(M)$ correspond to a perturbed version of the true configuration, for some $K_+ \geq K$. Further let $\tilde{D}_{N,J}$ be the corresponding partial distance matrix. Suppose that $\|\tilde{D}_{N,J} - D_{N,J}^*\|_F^2 = o(NJ)$, when N and J grow to infinity. Then*

$$\limsup_{N,J\to\infty} \left(\min_{F\in\mathcal{A}_K} \frac{\sum_{i=1}^N \|\boldsymbol{\theta}_i^+ - F(\tilde{\boldsymbol{\theta}}_i)\|^2}{N} + \frac{\sum_{j=1}^J \|\mathbf{a}_j^+ - F(\tilde{\mathbf{a}}_j)\|^2}{J} \right) \leq C\epsilon^2. \tag{5}$$

where C is a constant that does not depend on N and J. If there exists a fixed collection of anchor points, for which A3 is satisfied for any sufficiently small $\epsilon > 0$, then we

have

$$\limsup_{N,J\to\infty}\left(\min_{F\in\mathcal{A}_K}\frac{\sum_{i=1}^{N}\|\boldsymbol{\theta}_i^{+}-F(\tilde{\boldsymbol{\theta}}_i)\|^2}{N}+\frac{\sum_{j=1}^{J}\|\mathbf{a}_j^{+}-F(\tilde{\mathbf{a}}_j)\|^2}{J}\right)=0. \tag{6}$$

Remark 4 *Theorem 1 shows that the configuration can be recovered asymptotically when both N and J grow to infinity and suitable conditions hold. The conditions required by Theorem 1 are quite mild. It first requires all the true and perturbed ideal points to be located in a compact set. Second, as will be shown in Proposition 2 below, condition A3 is satisfied with high probability when the true person and item points are i.i.d. samples from two distributions satisfying mild conditions, respectively. Finally, it requires that the perturbation of the partial distance matrix is not too large, i.e., $\|\tilde{D}_{N,J}-D^{*}_{N,J}\|_F^2=o(NJ)$. As will be shown in Proposition 3, this condition holds with high probability when $\tilde{D}_{N,J}$ is given by a likelihood-based estimator.*

Proposition 2 *Suppose that $\boldsymbol{\theta}_1^{*},...,\boldsymbol{\theta}_N^{*}$ and $\mathbf{a}_1^{*},...,\mathbf{a}_J^{*}$ are independent and identically distributed samples from distributions P_1 and P_2, where P_1 and P_2 have positive and continuous density functions within a ball $G\subset B_{\mathbf{0}}^{K}(M)$. Then A3 holds almost surely for any sufficiently small $\epsilon>0$.*

Remark 5 *We remark that constant C is determined and only determined by the configuration of the anchor points in A3, according to our proof in the supplementary material. Roughly, the more regular the set of anchor points is (in terms of affine spanning $\mathbb{R}^K$), the smaller the value of C.*

Remark 6 *As discussed in Section 2.2, we can only recover the ideal points up to an isometry mapping. This isometry mapping may be fixed if one is willing to make further assumptions such as non-negativity (Donoho and Stodden, 2004; Hoyer, 2004) and sparsity (Chen et al., 2019b). In that case, one may further interpret each coordinate of the latent space. We leave this problem for future investigation.*

3.2 Likelihood-based Estimation

In what follows, we propose a constrained maximum likelihood estimator and show its properties. Given the assumptions of the MDU model, our likelihood function takes the form

$$L(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N, \mathbf{a}_1, ..., \mathbf{a}_J) = \prod_{i=1}^{N} \prod_{j=1}^{J} f(\|\boldsymbol{\theta}_i - \mathbf{a}_j\|^2)^{y_{ij}} (1 - f(\|\boldsymbol{\theta}_i - \mathbf{a}_j\|^2)^{1-y_{ij}}.$$

Based on this likelihood function, we consider the following estimator

$$\begin{aligned} (\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_N, \hat{\mathbf{a}}_1, ..., \hat{\mathbf{a}}_J) = & \underset{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N, \mathbf{a}_1, ..., \mathbf{a}_J \in \mathbb{R}^{K_+}}{\arg\min} - \log L(\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N, \mathbf{a}_1, ..., \mathbf{a}_J) \\ & s.t. \quad \|\boldsymbol{\theta}_i\| \leq M, \ \|\mathbf{a}_j\| \leq M, \quad i = 1, ..., N, \ j = 1, ..., J. \end{aligned} \tag{7}$$

where K_+ and M are pre-specified. We denote $\hat{D}_{N,J}$ as the partial distance matrix based on $\hat{\boldsymbol{\theta}}_i$s and $\hat{\mathbf{a}}_j$s from (7).

We impose the following regularity condition on the link function f, which requires f to be neither too steep nor too flat in the feasible domain. Similar conditions are assumed in Davenport et al. (2014) for solving a 1-bit matrix completion problem.

A4. The link function $f : \mathbb{R} \to (0, 1)$ is a smooth and monotone decreasing function, satisfying $L_{4M^2} < \infty$ and $\beta_{4M^2} < \infty$, where

$$L_\alpha = \sup_{|x| \leq \alpha} \frac{|f'(x)|}{f(x)(1 - f(x))}, \quad \text{and} \quad \beta_\alpha = \sup_{|x| \leq \alpha} \frac{f(x)(1 - f(x))}{|f'(x)|^2}.$$

Proposition 3 *Suppose that A0 and A4 are satisfied and $K_+ \geq K$. Then there exist C_1 and C_2 independent of N and J, such that*

$$\frac{1}{NJ} \|\hat{D}_{N,J} - D^*_{N,J}\|_F^2 \leq C_1 M^2 L_{4M^2} \beta_{4M^2} \sqrt{\frac{N + J}{NJ}} \sqrt{1 + \frac{\log(NJ)}{N + J}},$$

with probability at least $1 - C_2/(N + J)$.

Proposition 3 implies that $\|\hat{D}_{N,J} - D^*_{N,J}\|_F^2 = o_p(NJ)$, which, combined with Theorem 1, leads to Theorem 2 below.

Theorem 2 *Suppose that A0, A3 and A4 are satisfied and $K_+ \geq K$. Then*

$$\lim_{N,J\to\infty} P\left(\min_{F\in\mathcal{A}_{K_+}} \frac{\sum_{i=1}^N \|\boldsymbol{\theta}_i^+ - F(\hat{\boldsymbol{\theta}}_i)\|^2}{N} + \frac{\sum_{j=1}^J \|\mathbf{a}_j^+ - F(\hat{\mathbf{a}}_j)\|^2}{J} \leq C\epsilon^2\right) = 1, \quad (8)$$

where $\hat{\boldsymbol{\theta}}_i$ and $\hat{\mathbf{a}}_j$, $i = 1, ..., N$, and $j = 1, ..., J$, are given by (7), ϵ is from condition A3, and C is a constant independent of ϵ, N, and J.

Remark 7 *We remark that if A3 holds for any sufficiently small ϵ, then (8) implies that the loss*

$$\min_{F\in\mathcal{A}_{K_+}} \frac{\sum_{i=1}^N \|\boldsymbol{\theta}_i^+ - F(\hat{\boldsymbol{\theta}}_i)\|^2}{N} + \frac{\sum_{j=1}^J \|\mathbf{a}_j^+ - F(\hat{\mathbf{a}}_j)\|^2}{J}$$

converges to zero in probability. Further note that according to Proposition 2, A3 holds with high probability for any sufficiently small ϵ, under a random design for the true ideal points. Therefore, the loss can be shown to converge to zero in probability, under this random design. This result is summarized in Theorem 3 below.

Theorem 3 *Suppose that A0 and A4 are satisfied and $K_+ \geq K$. Further suppose that $\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_N^*$ and $\mathbf{a}_1^*, ..., \mathbf{a}_J^*$ are independent and identically distributed samples from distributions P_1 and P_2, respectively, where P_1 and P_2 have positive and continuous density functions within a ball $G \subset B_{\mathbf{0}}^K(M)$. Then for $\hat{\boldsymbol{\theta}}_i$ and $\hat{\mathbf{a}}_j$, $i = 1, ..., N$, and $j = 1, ..., J$, given by (7), the loss function*

$$\min_{F\in\mathcal{A}_{K_+}} \frac{\sum_{i=1}^N \|\boldsymbol{\theta}_i^+ - F(\hat{\boldsymbol{\theta}}_i)\|^2}{N} + \frac{\sum_{j=1}^J \|\mathbf{a}_j^+ - F(\hat{\mathbf{a}}_j)\|^2}{J}$$

goes to 0 in probability as N and J grow to infinity.

Remark 8 *We remark that the probability measures in Theorems 2 and 3 are slightly different. The probability in Theorem 2 is based on the conditional distribution of Y_{ij}s*

given $\boldsymbol{\theta}_i^*$ *and* $\mathbf{a}_j^*$*, while that for Theorem 3 is based on the joint distribution of* Y_{ij}*,* $\boldsymbol{\theta}_i^*$ *and* $\mathbf{a}_j^*, i = 1, ..., N, j = 1, ..., J$*.*

Remark 9 *A stress function is a squared error loss function that plays an important role in the classical MDS/MDU algorithms. It serves not only as the objective function in the search for the MDS/MDU solution, but also as the basis for assessing the goodness-of-fit of the solution (Mair et al., 2016). In the proposed framework, the negative joint log-likelihood function plays a similar role as the stress function. It replaces the squared loss in the stress function by a loss function based on the Kullback–Leibler divergence. Similar goodness-of-fit measures in classical MDU can be developed under the proposed framework, based on the negative joint log-likelihood.*

Remark 10 *We remark on the choice of latent dimension. Theorems 2 and 3 suggest that as long as we choose* K_+ *to be no less than the true dimension* K*, then the unfolding result is asymptotically valid. When there is no such prior knowledge about an upper bound of* K*, one can estimate the latent dimension* K *using data. Several methods from factor analysis and network data analysis may be adapted to the current problem, such as trace-norm regularization (Bach, 2008), cross-validation (Chen and Lei, 2018; Li et al., 2020), and information criteria (Bai and Ng, 2002). We believe that consistency results on the selection of* K *can be established.*

Remark 11 *We point out that the result of Proposition 3 can be easily extended to other MDU models, such as models with additional parameters in the link function and models for rating and ranking data. Then, by making use of Theorem 1, the results of Theorems 2 and 3 can also be extended to these models.*

We propose an alternating minimization algorithm for solving (7). To handle the constraints in (7), a projected gradient descent update is used in each iteration. For

$\mathbf{x} \in \mathbb{R}^{K_+}$, we define the following projection operator:

$$\mathrm{Proc}_M(\mathbf{x}) = \underset{\|\mathbf{y}\| \le M}{\arg\min} \|\mathbf{y} - \mathbf{x}\| = \begin{cases} \mathbf{x} & \text{if } \|\mathbf{x}\| \le M, \\ M\mathbf{x}/\|\mathbf{x}\| & \text{if } \|\mathbf{x}\| > M. \end{cases}$$

Algorithm 1 (Alternating minimization algorithm)

Input*: Data $(y_{ij})_{N\times J}$, pre-specified dimension K_+, constraint M, iteration number $m = 1$, and the initial values $\boldsymbol{\theta}_1^{(0)}, ..., \boldsymbol{\theta}_N^{(0)}$ and $\mathbf{a}_1^{(0)}, ..., \mathbf{a}_J^{(0)}$ in $\mathbb{R}^{K_+}$.*

Alternating minimization*: at the mth iteration, perform*

(a) For each respondent i, update

$$\boldsymbol{\theta}_i^{(m)} = Proc_M\left(\boldsymbol{\theta}_i^{(m-1)} + \varrho \mathbf{s}_i^{(m-1)}(\boldsymbol{\theta}_i^{(m-1)})\right),$$

where

$$\begin{aligned} &\mathbf{s}_i^{(m-1)}(\boldsymbol{\theta}) \\ =&\frac{\partial}{\partial \boldsymbol{\theta}}\left(\sum_{j=1}^{J} y_{ij} \log f(\|\boldsymbol{\theta} - \mathbf{a}_j^{(m-1)}\|^2) + (1 - y_{ij}) \log\left(1 - f(\|\boldsymbol{\theta} - \mathbf{a}_j^{(m-1)}\|^2)\right)\right). \end{aligned}$$

The step size $\varrho > 0$ is chosen by line search.

(b) For each item j, update

$$\mathbf{a}_j^{(m)} = Proc_M\left(\mathbf{a}_j^{(m-1)} + \varrho \tilde{\mathbf{s}}_j^{(m-1)}(\mathbf{a}_j^{(m-1)})\right),$$

where

$$\begin{aligned} &\tilde{\mathbf{s}}_j^{(m-1)}(\mathbf{a}) \\ =&\frac{\partial}{\partial \mathbf{a}}\left(\sum_{i=1}^{N} y_{ij} \log f(\|\boldsymbol{\theta}_i^{(m)} - \mathbf{a}\|^2) + (1 - y_{ij}) \log\left(1 - f(\|\boldsymbol{\theta}_i^{(m)} - \mathbf{a}\|^2)\right)\right). \end{aligned}$$

The step size $\varrho > 0$ is chosen by line search.

Iteratively perform steps (a) and (b) until convergence. Let m^* be the last iteration number upon convergence.

Output: $\hat{\boldsymbol{\theta}}_1 = \boldsymbol{\theta}_1^{(m^*)}, ..., \hat{\boldsymbol{\theta}}_N = \boldsymbol{\theta}_N^{(m^*)}$ *and* $\hat{\mathbf{a}}_1 = \mathbf{a}_1^{(m^*)}, ..., \hat{\mathbf{a}}_J = \mathbf{a}_J^{(m^*)}$.

Remark 12 *Since* (7) *is not a convex optimization problem, there is no guarantee that Algorithm 1 finds the global optimal solution. However, we point out that the previous theoretical results hold even when* $\{\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_N, \hat{\mathbf{a}}_1, ..., \hat{\mathbf{a}}_J\}$ *is not a global optimal point. Specifically, Proposition 3 and Theorems 2 and 3 hold for any* $\{\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_N, \hat{\mathbf{a}}_1, ..., \hat{\mathbf{a}}_J\}$ *satisfying the constraints in* (7) *and*

$$L(\hat{\boldsymbol{\theta}}_1, ..., \hat{\boldsymbol{\theta}}_N, \hat{\mathbf{a}}_1, ..., \hat{\mathbf{a}}_J) \geq L(\boldsymbol{\theta}_1^*, ..., \boldsymbol{\theta}_N^*, \mathbf{a}_1^*, ..., \mathbf{a}_J^*). \tag{9}$$

According to our simulation study, estimates given by Algorithm 1 are likely to satisfy (9).

3.3 Analyzing Missing Data

We further discuss the configuration recovery problem when data have many missing values, which is commonly encountered in practice. Denote matrix $\Omega = (\omega_{ij})_{N \times J}$, where $\omega_{ij} = 1$ indicates that response y_{ij} is observed and $\omega_{ij} = 0$ indicates y_{ij} is missing. We consider the simple case of uniformly missing, as described in condition A5. We point out that this assumption can be relaxed to analyzing data that have non-uniformly missing entries, following the developments in Cai and Zhou (2013) for solving a 1-bit matrix completion problem.

A5. Entries of Ω, ω_{ij}, are independent and identically distributed Bernoulli random variables with

$$P(\omega_{ij} = 1) = \frac{n}{NJ}.$$

Under this condition, there are on average n entries of the data matrix $(y_{ij})_{N\times J}$ that are observable. Thanks to the ignorable missingness, given Ω and the observed data, the likelihood becomes

$$L^{\Omega}(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_N,\mathbf{a}_1,...,\mathbf{a}_J)=\prod_{\omega_{ij}=1} f(\|\boldsymbol{\theta}_i-\mathbf{a}_j\|^2)^{y_{ij}}(1-f(\|\boldsymbol{\theta}_i-\mathbf{a}_j\|^2)^{1-y_{ij}}.$$

We still consider a constrained maximum likelihood estimator

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_1^{\Omega},...,\hat{\boldsymbol{\theta}}_N^{\Omega},\hat{\mathbf{a}}_1^{\Omega},...,\hat{\mathbf{a}}_J^{\Omega}) = &\underset{\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_N,\mathbf{a}_1,...,\mathbf{a}_J\in\mathbb{R}^{K_+}}{\arg\min} -\log L^{\Omega}(\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_N,\mathbf{a}_1,...,\mathbf{a}_J)\\ s.t. \quad &\|\boldsymbol{\theta}_i\|\le M,\ \|\mathbf{a}_j\|\le M, \quad i=1,...,N,\ j=1,...,J.\end{aligned} \tag{10}$$

Let $\hat{D}^{\Omega}_{N,J}$ denote the partial distance matrix for $\hat{\boldsymbol{\theta}}_1^{\Omega},...,\hat{\boldsymbol{\theta}}_N^{\Omega},\hat{\mathbf{a}}_1^{\Omega},...,\hat{\mathbf{a}}_J^{\Omega}$. Proposition 4 presents a missing-data version of Proposition 3. It implies that we can still recover the partial distance matrix if n is large enough.

Proposition 4 *Suppose that A0, A4 and A5 are satisfied and $K_+\ge K$. Then there exist C_1 and C_2 independent of N and J, such that*

$$\frac{1}{NJ}\|\hat{D}^{\Omega}_{N,J}-D^{*}_{N,J}\|_F^2\le C_1M^2L_{4M^2}\beta_{4M^2}\sqrt{\frac{N+J}{n}}\sqrt{1+\frac{NJ\log(NJ)}{n(N+J)}} \tag{11}$$

with probability at least $1-C_2/(N+J)$.

Remark 13 *If $n>(N+J)\log(NJ)$, then the right side of (11) goes to 0 as N and J grow to infinity, which means $\|\hat{D}^{\Omega}_{N,J}-D^{*}_{N,J}\|_F^2=o_p(NJ)$. Following the discussion in Section 3, $\{\hat{\boldsymbol{\theta}}_1^{\Omega},...,\hat{\boldsymbol{\theta}}_N^{\Omega},\hat{\mathbf{a}}_1^{\Omega},...,\hat{\mathbf{a}}_J^{\Omega}\}$ provides a consistent estimate of the ideal point configuration. This consistency result is summarized in Proposition 5, which is a missing-data version of Theorem 3 under the random design.*

Proposition 5 *Suppose that A0, A4 and A5 are satisfied, and $K_+\ge K$, and $n>(N+J)\log(NJ)$. Further suppose that $\boldsymbol{\theta}_1^{*},...,\boldsymbol{\theta}_N^{*}$ and $\mathbf{a}_1^{*},...,\mathbf{a}_J^{*}$ are independent and*

identically distributed samples from distributions P_1 and P_2, where P_1 and P_2 have positive and continuous density functions within a ball $G \subset B_{\mathbf{0}}^K(M)$. Then the loss function

$$\min_{F \in \mathcal{A}_{K_+}} \frac{\sum_{i=1}^N \|\boldsymbol{\theta}_i^+ - F(\hat{\boldsymbol{\theta}}_i^\Omega)\|^2}{N} + \frac{\sum_{j=1}^J \|\mathbf{a}_j^+ - F(\hat{\mathbf{a}}_j^\Omega)\|^2}{J}$$

goes to 0 in probability as N and J grow to infinity.

4 Simulation Studies

In what follows, simulation studies are conducted to verify our theoretical results. Specifically, we consider a random design where the true ideal points are generated from distributions. All the analyses in this section, as well as those in Section 5, are based on our implementation of Algorithm 1 in statistical software R[1].

4.1 Study I

Setting. We first consider a setting where K_+ is chosen to be exactly K. We consider MDU in a two-dimensional latent space, i.e., $K = K_+ = 2$. Diverging sequences of J and N are considered, by letting $J = 200, 400, ..., 1000$ and $N = 20J$. For given N and J, 100 independent datasets are generated. For each dataset, we first sample $\boldsymbol{\theta}_i^*$s and $\mathbf{a}_j^*$s uniformly from $B_{\mathbf{0}}^2(1)$, a ball in $\mathbb{R}^2$ with center $\mathbf{0}$ and radius 1. Then given the ideal points, response data Y_{ij} are generated under the link function $f(x) = 2/(1 + \exp(x + 0.1))$. It can be easily verified that condition A4 is satisfied for this link function.

For each dataset, we obtain an estimate of the ideal points, by applying Algorithm 1 ten times with random starting points and then choosing the result that gives the largest likelihood function value. The use of multiple starting points sub-

[1]An R package has been developed and can be downloaded from `https://github.com/hrzhang16/mmdu`.

	$J = 200$	$J = 400$	$J = 600$	$J = 800$	$J = 1000$
25%	0.0630	0.0323	0.0218	0.0164	0.0131
median	0.0647	0.0328	0.0222	0.0167	0.0134
75%	0.0666	0.0336	0.0227	0.0170	0.0135

Table 1: Simulation Study I: The average squared Frobenius loss of partial distance when J increases from 200 to 1000. For each J, the table shows the 25%, 50% and 75% quantiles of the loss based on 100 independent experiments.

stantially reduces the risk of the algorithm converging to bad local minima. In the application of Algorithm 1, the constraint M is set to 1.5.

Results. We first check the obtained likelihood function values for the 100 datasets. As we point out in Remark 12, Proposition 3 and Theorems 2 and 3 still hold as long as the estimate satisfies (9), even if the global solution to the optimization (7) is not obtained. It is found that by using ten random starting points, the likelihood function at the estimated parameters is always larger than that at the true parameters for all the 100 datasets.

We then present the average squared Frobenius loss for the recovery of the partial distance matrix, $\|\hat{D}_{N,J} - D^*_{N,J}\|_F^2/(NJ)$. These results are given in Table 1 which presents the 25%, 50%, and 75% quantiles of the loss based on the 100 datasets. From this table, we see that the loss tends to decrease as the sample size increases, supporting the result of Proposition 3.

Table 2 presents the results on loss (3) for configuration recovery, where the best isometry mapping F in (3) is obtained by solving an optimization problem given the true and estimated ideal points. Similar to the results on partial distance matrix recovery, the loss (3) also decreases towards 0 as J grows large, which is consistent with the result of Theorem 3.

Finally, the computation time on a standard desktop machine[2] for solving (7) is shown in Table 3. It is worth pointing out that since the update of person and item

[2]All the computation is conducted on a single Intel®Gold 6130 core.

	$J = 200$	$J = 400$	$J = 600$	$J = 800$	$J = 1000$
25%	0.0158	0.0079	0.0053	0.0040	0.0032
median	0.0160	0.0080	0.0053	0.0040	0.0032
75%	0.0162	0.0080	0.0054	0.0040	0.0032

Table 2: Simulation Study I: The average loss for configuration recovery when J increases from 200 to 1000. For each J, the table shows the 25%, 50% and 75% quantiles of the loss based on 100 independent experiments.

	$J = 200$	$J = 400$	$J = 600$	$J = 800$	$J = 1000$
25%	98.2	109.6	144.1	191.5	254.6
median	113.8	120.1	156.0	201.5	272.6
75%	128.9	138.0	176.9	213.9	286.7

Table 3: Simulation Study I: The computation time of optimization (7) when J increases from 200 to 1000. For each J, 25%, 50% and 75% quantiles of the computation time from 100 independent experiments are shown.

parameters in each iteration of Algorithm 1 can be run in parallel, the computation can be further speeded up substantially by parallel computing.

4.2 Study II

Setting. We now consider a setting where $K_+ > K$. We take the same setting as in Study I, except that we set $K_+ = 3$ when fitting the MDU model. The same as Study I, for each pair of N and J, 100 independent datasets are generated. For each dataset, Algorithm 1 is applied similarly, using 10 random starting points and constraint parameter $M = 1.5$.

Results. The results are given in Tables 4 through 6. Similar to Tables 1–3, these three tables also show the results on partial distance matrix recovery, configuration recovery, and computation time, respectively. Comparing with the results of Study I, we see that both losses for the recovery of partial distance matrix and configuration tend to be larger. This is due to the overfitting brought by adding unnecessary parameters in the model. The computation time also increases compared with that

	$J = 200$	$J = 400$	$J = 600$	$J = 800$	$J = 1000$
25%	0.0734	0.0384	0.0261	0.0198	0.0159
median	0.0758	0.0390	0.0265	0.0200	0.0161
75%	0.0780	0.0398	0.0269	0.0204	0.0163

Table 4: Simulation Study II: The average squared Frobenius loss of partial distance when J increases from 200 to 1000. For each J, the table shows the 25%, 50% and 75% quantiles of the loss based on 100 independent experiments.

	$J = 200$	$J = 400$	$J = 600$	$J = 800$	$J = 1000$
25%	0.0853	0.0568	0.0452	0.0386	0.0343
median	0.0862	0.0573	0.0455	0.0390	0.0345
75%	0.0877	0.0580	0.0459	0.0392	0.0346

Table 5: Simulation Study II: The average loss for configuration recovery when J increases from 200 to 1000. For each J, the table shows the 25%, 50% and 75% quantiles of the loss based on 100 independent experiments.

of Study I.

5 Real Examples

5.1 Example I: Movie Data

Background. We apply MDU to a movie rating dataset from the famous MovieLens project (see e.g., Harper and Konstan, 2016). The dataset analyzed in this paper is a subset of a benchmark MovieLens dataset collected during a seven-month period

	$J = 200$	$J = 400$	$J = 600$	$J = 800$	$J = 1000$
25%	106.3	264.0	639.3	1294.0	2302.5
median	110.0	286.5	698.8	1407.7	2480.8
75%	112.9	308.9	793.8	1551.0	2841.1

Table 6: Simulation Study II: The computation time of optimization (7) when J increases from 200 to 1000. For each J, 25%, 50% and 75% quantiles of the computation time from 100 independent experiments are shown.

from September, 1997 through April, 1998[3]. This subset contains 943 users and 338 movies, obtained by selecting movies that have been rated by at least 100 users. Unlike many analyses of MovieLens data that focus on the rating scores, we consider to unfold the rating behavior itself (i.e., rated/not rated) which may also reveal the users' preference patterns. More precisely, we let $Y_{ij} = 1$ if movie j has been rated by user i and $Y_{ij} = 0$ otherwise.

Analysis. For visualization purpose, we unfold the data onto a two-dimensional space. To apply the MDU model introduced in this paper, we need to specify the link function f. We assume f to take the logistic form $f(x) = 2/(1 + \exp(x + \delta))$, where δ is a pre-specified small positive constant. For any $\delta > 0$, it is easy to check that the regularity condition A4 is satisfied. The results presented below are based on the choice $\delta = 0.1$, but we point out that other choices of δ ($\delta = 0.05, 0.15, 0.2$) have also been tried which all lead to very similar results. The constraint constant M is set to 3.5 when applying Algorithm 1. After obtaining the estimate, we transform the estimated ideal points by an isometry mapping, so that the x-axis corresponds to the dimension along which the estimated movie ideal points have the highest variance. As will be described in the sequel, under this isometry mapping of the estimated ideal points, both the x- and y-axes receive good interpretations.

Results. The results from the MDU analysis are presented in Figures 4 through 6. Figure 4 jointly visualizes the estimated movie and user points. As we can see, the movies and the users tend to form two giant clusters that only slightly overlap.

We investigate the movie points. First, the y-axis of the space largely indicates, if not perfectly, the popularity of the movies. The movies with a smaller $\hat{a}_{j2}$ value tends to be rated more frequently. Roughly speaking, the shorter the average distance from a movie to the user points, the more often the movie is rated. In fact, the Kendall's

[3]The dataset can be downloaded from `https://grouplens.org/datasets/movielens/100k/`

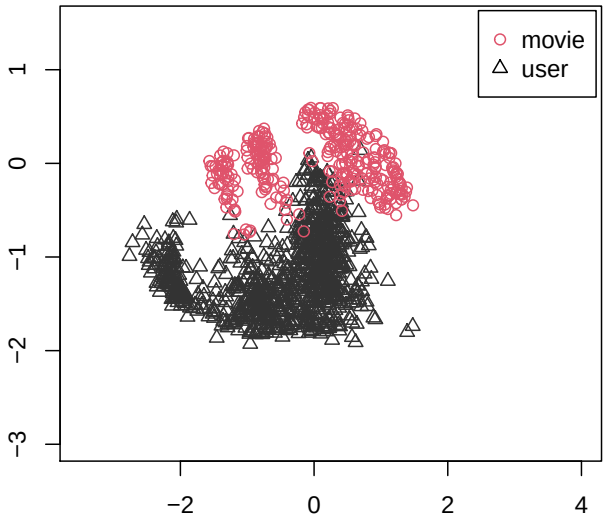
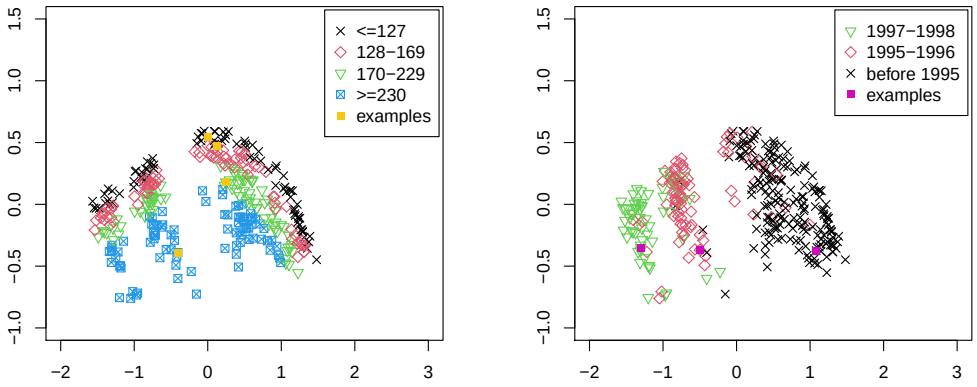


Figure 4: Analysis of movie rating data: Simultaneous visualization of the estimated movie and user points.

tau rank correlation between $\hat{a}_{j2}$s and the numbers of ratings received by the movies is -0.66. This phenomenon is further reflected by panel (a) of Figure 5, where movies are stratified by the numbers of ratings they received into four categories. These four categories tend to be ordered along the y-axis. We list four movies as examples, as indicated in panel (a) of Figure 5. From the top to the bottom, they are Batman Forever (1995), Golden Eye (1995), Get Shorty (1995) and The Godfather (1972), respectively. Based on our interpretation of the y-axis, these four movies are ordered from the least popular to the most popular.

Second, the x-axis of the space seems to indicate the release time of the movies. The Kendall's tau correlation between $\hat{a}_{j1}$s and the release dates of the movies is -0.70. As shown in panel (b) of Figure 5, where the movies are stratified into three categories, namely "before 1995", "1995-1996", and "1997-1998". According to this figure, the clustering patten of the movies can be largely explained by the three



(a) (b)

Figure 5: Analysis of movie rating data. Panel (a): Visualization of movie points, with movies stratified into four equal-size categories based on the numbers of rating. Movies with numbers of rating less than 127, 128-169, 170-229 and more than 230 are indicated by black, red, green and blue points, respectively. Yellow points represent example movies. Panel (b): Visualization of movie points, with movies stratified into three categories based on their release time. Movies released in 1997-1998, 1995-1996, and before 1995 are indicated by green, red and black points, respectively. Purple points represent example movies.

categories based on the movie release dates. From the right to the left of the space, the points correspond to movies from the relatively older ones to the relatively more recent ones. For example, the three movies indicated in panel (b) of Figure 5 are, from left to right, Citizen Kane (1941), Twelve Monkeys (1995) and The Devil's Own (1997), respectively.

The interpretation of the latent space based on movies facilitates the interpretation of the user points. First, the y-axis corresponds to the users' activeness. Roughly speaking, the shorter the average distance from a user point to the movies points, the more active the user is. The Kendall's tau rank correlation between $\hat{\theta}_{i2}$s and the numbers of ratings given by the users is 0.73. This is further shown via Figure 6, where users are classified into four equal-size groups depending on the number of movies they rated. These groups of users, from the most active one to the least active

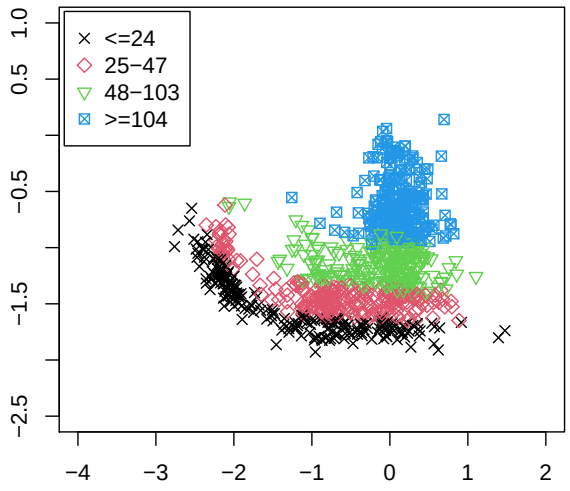


Figure 6: Analysis of movie rating data: Visualization of user points, with users classified into four equal-size categories based on the numbers of rating. Users who rated less than 24, , 25-47, 48-103 and more than 104 movies are indicated by black, red, green and blue points, respectively.

one, lie from the top to the bottom. Second, based on the alignment of movies along the x-axis, the user points from right to left may be interpreted as the ones who tend to more frequently rate relatively older movies to the ones who tend to more frequently rate relatively more recent ones.

5.2 Example II: Senate Roll Call Voting Data

Background. We now analyze a senate roll call voting dataset from the 108th congress. This dataset contains the voting records from 100 senators to 675 roll calls in years 2003 and 2004[4]. Among the 100 senators, there are 48 from the Democratic party, 51 from the Republican party, and one independent politician. For each roll call j, the vote of senator i is recorded in three ways, "Yea", "Nay" and "Not Voting", treated as $Y_{ij} = 1, 0$, and missing, respectively.

[4]The dataset can be downloaded from `https://legacy.voteview.com/dwnl.htm`.

Analysis. Similar analysis as the previous one is conducted. Specifically, we unfold the data into a two-dimensional space. The same link function f and constraint constant M are adopted as in the analysis of movie data. After getting the estimate, we transform the estimated ideal points by an isometry mapping, so that the x-axis corresponds to the dimension along which the estimated senate ideal points have the highest variance.

Results. The results are presented in Figures 7 through 9. In Figure 7, the ideal points of both roll calls and senators are visualized simultaneously. As we can see, most of the roll calls and all the senators tend to lie around a one-dimensional line. This visualization is still valid, in the sense that even when the true latent dimension is one, according to Theorem 3, unfolding the data in a two-dimensional space is still consistent.

This phenomenon of degeneration is quite consistent with the overall unidimensional patten in the congress voting data throughout the history. It has been well recognized in the political science literature (Poole et al., 1991; Poole and Rosenthal, 1991) that senate voting behavior is essentially unidimensional, though slightly different latent space models are used in that literature. For example, Poole et al. (1991) concluded that "to the extent that congressional voting can be described by a spatial model, a unidimensional model is largely (albeit not entirely) sufficient."

We first interpret the senators. In Figure 8, all the senator points are visualized with their party membership indicated by different point types. In Table 7, we rank the senators based on their value of $\hat{\theta}_{i1}$, which is presented along the x-axis. According to this table, the Democrats tend to lie on the left and the Republicans tend to be on the right. In fact, this ranking is largely consistent with *National Journal*'s liberalness ranking of the senators in 2003. *National Journal*'s ranking result, which is replicated in Clinton et al. (2004b), is obtained by unfolding the senators' votes on 62 key roll

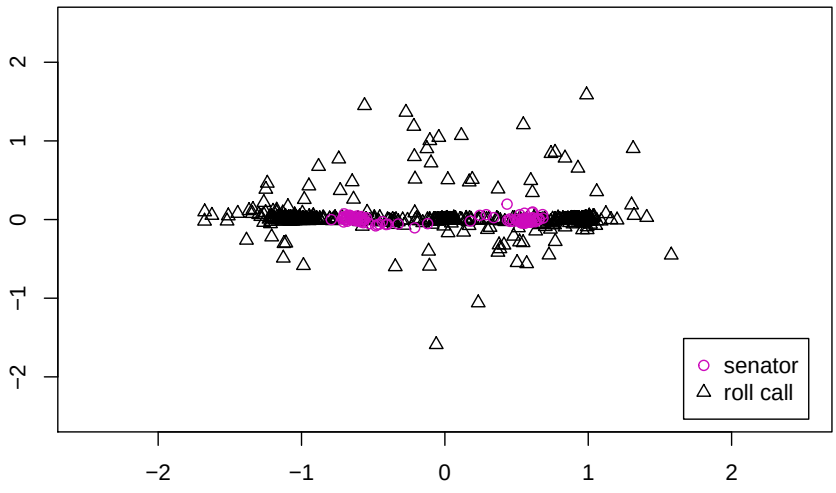


Figure 7: Analysis of senator roll call data: Simultaneous visualization of the estimated senator and roll call ideal points.

calls using a model given in Clinton et al. (2004a). The Kendall's tau rank correlation between the result in Table 7 and that given by the *National Journal* is 0.79. In fact, Senator John Kerry is ranked the most liberal by both our model and by *National Journal* and Senator Craig L. Thomas, who is the most conservative senator according to the ranking of *National Journal*, is the third most conservative senator given by our model.

From Figure 8 and Table 7, it is also worth noting that there is a Democrat whose estimated ideal point is mixed together with those of the Republicans. This senator is Zell Miller from the state of Georgia. He is a conservative Democrat and in fact, he supported Republican President George W. Bush against the Democratic nominee John Kerry in the presidential election in 2004.

In this congress, there is an independent senator, Jim Jeffords from the state of Vermont, who does not belong to either of the two major parties. As we can see from both Figure 8 and Table 7, his ideal point lies on the left, mixed with many ideal

	Name	State		Name	State		Name	State
1	Kerry	D-MA	35	Johnson	D-SD	69	Grassley	R-IO
2	Sarbanes	D-MD	36	Lieberman	D-CT	70	Bond	R-MO
3	Reed	D-RH	37	Bingaman	D-NM	71	Roberts	R-KA
4	Harkin	D-IO	38	Nelson	D-FL	72	Gregg	R-NH
5	Graham	D-FL	39	Dorgan	D-ND	73	Allen	R-VI
6	Lautenberg	D-NJ	40	Conrad	D-ND	74	Domenici	R-NM
7	Edwards	D-NC	41	Carper	D-DE	75	Bennett	R-UT
8	Kennedy	D-MA	42	Pryor	D-AR	76	Dole	R-NC
9	Durbin	D-IL	43	Bayh	D-IN	77	Frist	R-TN
10	Levin	D-MI	44	Lincoln	D-AR	78	Brownback	R-KA
11	Akaka	D-HA	45	Landrieu	D-LO	79	Hatch	R-UT
12	Byrd	D-WE	46	Baucus	D-MT	80	Cochran	R-MS
13	Boxer	D-CA	47	Breaux	D-LO	81	Graham	R-SC
14	Corzine	D-NJ	48	Nelson	D-NE	82	Alexander	R-TN
15	Clinton	D-NY	49	Chafee	R-RH	83	Lott	R-MS
16	Leahy	D-VE	50	Snowe	R-ME	84	Chambliss	R-GE
17	Dodd	D-CT	51	Collins	R-ME	85	Burns	R-MT
18	Stabenow	D-MI	52	Specter	R-PE	86	Bunning	R-KE
19	Mikulski	D-MD	53	Mccain	R-AZ	87	Crapo	R-ID
20	Feingold	D-WI	54	Dewine	R-OH	88	Mcconnell	R-KE
21	Rockefeller	D-WE	55	Campbell	R-CO	89	Ensign	R-NV
22	Hollings	D-SC	56	Smith	R-OR	90	Cornyn	R-TX
23	Kohl	D-WI	57	Coleman	R-MN	91	Sununu	R-NH
24	Inouye	D-HA	58	Warner	R-VI	92	Santorum	R-PE
25	Schumer	D-NY	59	Murkowski	R-AK	93	Craig	R-ID
26	Cantwell	D-WA	60	Voinovich	R-OH	94	Inhofe	R-OK
27	Dayton	D-MN	61	Hutchison	R-TX	95	Allard	R-CO
28	Murray	D-WA	62	Lugar	R-IN	96	Enzi	R-WY
29	Wyden	D-OR	63	Miller	D-GE	97	Sessions	R-AL
30	Daschle	D-SD	64	Fitzgerald	R-IL	98	Thomas	R-WY
31	Biden	D-DE	65	Talent	R-MO	99	Kyl	R-AZ
32	Feinstein	D-CA	66	Hagel	R-NE	100	Nickles	R-OK
33	Jeffords	I-VE	67	Stevens	R-AK			
34	Reid	D-NV	68	Shelby	R-AL			

Table 7: Analysis of senator roll call data: Ranking of senators based on $\hat{\theta}_{i1}$.

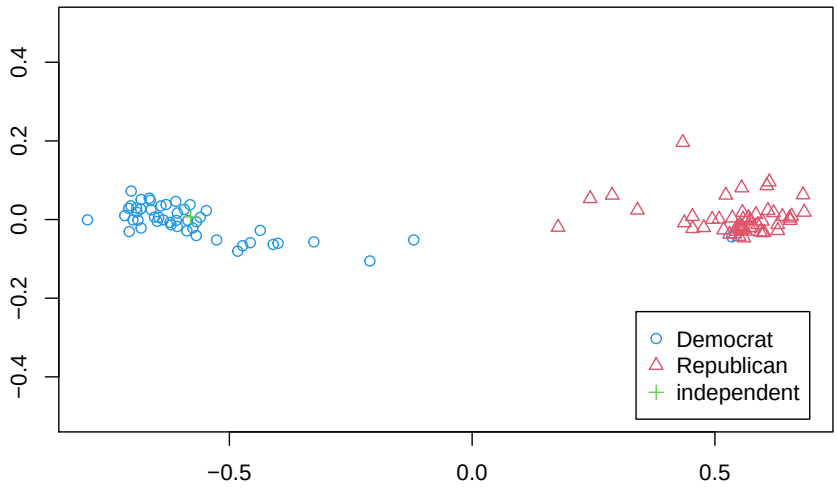


Figure 8: Analysis of senator roll call data: Visualization of senator points, where senators are classified by their party membership. Specifically, The Democrats, Republicans and an independent politician are indicated by blue, red, and green, respectively.

points of the Democrats. This is also consistent with Senator Jim Jeffords' political standing. In fact, he left Republican party to become an independent and began caucusing with the Democrats since 2001.

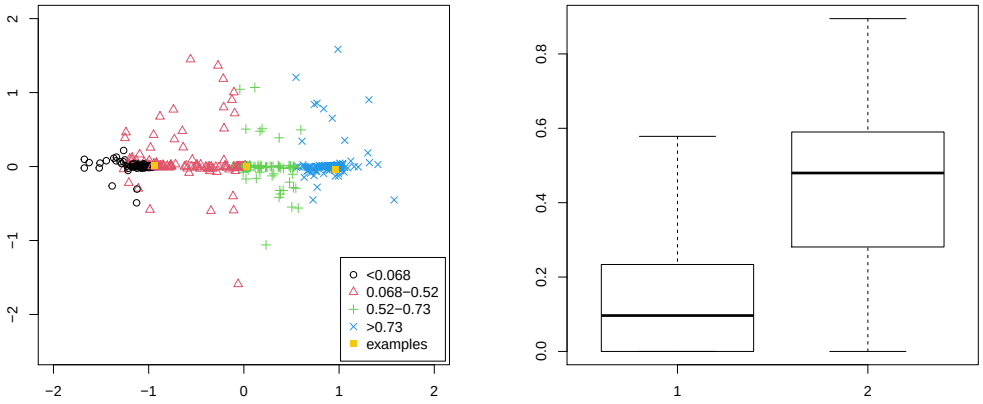
We now investigate the roll calls. The value of $\hat{a}_{j1}$, i.e., the roll calls' coordinate on the x-axis, seems to represent the roll calls' liberalness-conservativeness. The more liberal roll calls lie on the left and the more conservative ones lie on the right. This interpretation is further confirmed by the voting records for the roll calls. In particular, for each roll call, we calculate the proportion of Republicans among the senators who voted "Yea". A larger value of this proportion indicates that the roll call is more conservative. As we can see from panel (a) of Figure 9, for roll calls from the left to the right, this proportion increases. In fact, the Kendall's tau rank correlation between $\hat{a}_{j1}$s and the proportions of "Yea" from Republicans is as high as 0.88. We present the content of three roll calls as representative examples. As indicated in panel

(a) of Figure 9, these roll calls have substantially different coordinates along the x-axis. From left to right, they are (1) "To improve the availability of contraceptives for women", (2)"Confirmation Thomas J. Ridge, of Pennsylvania, to be Secretary of Homeland Security", and (3)"To provide financial security to family farm and small business owners by ending the unfair practice of taxing someone at death".

Although most of the roll calls lie near the x-axis (i.e., $\hat{a}_{j2} \approx 0$), there are still quite a few roll calls which spread out on the y-axis. It seems that the voting on such roll calls is heterogeneous within both parties. Specifically, we measure heterogeneity of voting within each party by a cross entropy measure, defined as

$$\mathrm{CE}_j^{(i)} = -p_{jy}^{(i)} \log p_{jy}^{(i)} - p_{jn}^{(i)} \log p_{jn}^{(i)} - p_{jm}^{(i)} \log p_{jm}^{(i)},$$

where $i = 1, 2$ indicate Democrat and Republican, respectively, and $p_{jy}^{(i)}$, $p_{jn}^{(i)}$, and $p_{jm}^{(i)}$ denote the proportions of "Yea", "Nay", and "Not voting" within the party for the jth roll call. Cross entropy is a commonly used measure of heterogeneity (Chapter 9, Friedman et al., 2001). The larger the cross entropy, the more heterogeneous voting behavior within a party. In panel (b) of Figure 9, we present the box plots of $\min\{\mathrm{CE}_j^{(1)}, \mathrm{CE}_j^{(2)}\}$, for roll calls lying near the x-axis ($|\hat{a}_{j2}| \leq 0.05$) and for those spreading out along the y-axis ($|\hat{a}_{j2}| > 0.05$). According to panel (b) of Figure 9, the roll calls in the latter group ($|\hat{a}_{j2}| > 0.05$) tend to have a larger value of $\min\{\mathrm{CE}_j^{(1)}, \mathrm{CE}_j^{(2)}\}$, implying that the voting tends to be more heterogeneous within both parties for these roll calls. The latter group contains roll calls, such as "To provide for the distribution of funds under the infrastructure performance and maintenance program", "To enhance the role of Congress in the oversight of the intelligence and intelligence-related activities of the United States Government", and "To strike provisions relating to energy tax incentives". Many of such roll calls may be explained by constituency specific factors.



(a) (b)

Figure 9: Analysis of senator roll call data. Panel (a): Visualization of roll call points, where roll calls are classified by the proportion of "Yea" from Republicans. Specifically, roll calls who have the proportions less than 0.068, 0.068-0.52,0.52-0.73 and larger than 0.73 are indicated by black, red, green and blue points, respectively. The yellow solid points are example roll calls to be discussed. Panel (b): Box plots of $\min\{\mathrm{CE}_j^{(1)}, \mathrm{CE}_j^{(2)}\}$, for roll calls lying near the x-axis ($|\hat{a}_{j2}| \leq 0.05$) one the left and for those spreading out along the y-axis ($|\hat{a}_{j2}| > 0.05$) on the right.

6 Concluding Remarks

In this paper, we provide a statistical framework for studying unfolding-model-based visualization. An estimator, together with an algorithm for its computation, is proposed, whose performance is examined by simulation studies. Under reasonable conditions, we provide asymptotic results for the recovery of ideal-point configuration. The proposed method is applied to two datasets, one on movie rating and the other on senator voting, for which interpretable results are obtained.

The ideal points obtained from the proposed method can be used in further analysis. For example, one can use the estimated person points as covariates in regression analysis. For another example, one may further conduct cluster analysis on the respondents and items, for example, by applying the K-means algorithm (MacQueen, 1967). In fact, as discussed in the supplementary material, there is a connection between our unfolding model and the stochastic co-blockmodel (Choi and Wolfe, 2014; Rohe et al., 2016) for bi-cluster analysis. When data follow a stochastic co-blockmodel, then our consistency result for the unfolding model further guarantees the consistency of bi-cluster analysis.

The current analysis may be extended along multiple directions. First, the current analysis keeps the latent dimension K fixed. In fact, the theoretical results established in this paper can be generalized to a setting where K also diverges, a more appropriate setting for data of a very large scale. Second, it is possible to make statistical inference about the person and item ideal points, such as testing whether a person point is closer to one item point than another. Making statistical inference under our model is closely related to statistical inference for low-rank matrix completion (see e.g., Chen et al., 2019a; Xia and Yuan, 2019), but the non-linear link function in our model brings more challenges and thus methods and theory remain to be developed. Third, although we focus on binary data in this paper, the proposed modeling framework, theory and computational algorithm can be extended to other types of data, such as ratings

and rankings. Finally, it may also be of interest to extend the current framework to the modeling and analysis of large-scale preferential choice data with informatively missing data entries.

Acknowledgement

We thank the editor, associate editor and two anonymous reviewers for their valuable comments. We also thank Prof. Ming Yuan for helpful discussions at an early stage of this work, and Prof. Abdo Alfakih for suggesting references on the uniqueness of Euclidian distance matrix completion.

References

Bach, F. R. (2008). Consistency of trace norm minimization. *Journal of Machine Learning Research*, 9:1019–1048.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221.

Bakker, R. and Poole, K. T. (2013). Bayesian metric multidimensional scaling. *Political Analysis*, 21:125–140.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons, West Sussex, UK.

Bennett, J. F. (1956). Determination of the number of independent parameters of a score matrix from the examination of rank orders. *Psychometrika*, 21:383–393.

Bennett, J. F. and Hays, W. L. (1960). Multidimensional unfolding: Determining the dimensionality of ranked preference data. *Psychometrika*, 25:27–43.

Borg, I. and Groenen, P. J. (2005). *Modern multidimensional scaling: Theory and applications*. Springer, New York, NY.

Busing, F. M., Groenen, P. J., and Heiser, W. J. (2005). Avoiding degeneracy in multidimensional unfolding by penalizing on the coefficient of variation. *Psychometrika*, 70:71–98.

Cai, T. and Zhou, W.-X. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, 14:3619–3647.

Chen, K. and Lei, J. (2018). Network cross-validation for determining the number of communities in network data. *Journal of the American Statistical Association*, 113:241–251.

Chen, L. and Buja, A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, 104:209–219.

Chen, Y., Fan, J., Ma, C., and Yan, Y. (2019a). Inference and uncertainty quantification for noisy matrix completion. *Proceedings of the National Academy of Sciences*, 116:22931–22937.

Chen, Y., Li, X., and Zhang, S. (2019b). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*. (in press).

Choi, D. and Wolfe, P. J. (2014). Co-clustering separately exchangeable network data. *The Annals of Statistics*, 42:29–63.

Clinton, J., Jackman, S., and Rivers, D. (2004a). The statistical analysis of roll call data. *American Political Science Review*, 98:355–370.

Clinton, J. D., Jackman, S., and Rivers, D. (2004b). "The most liberal senator"? Analyzing and interpreting congressional roll calls. *PS: Political Science & Politics*, 37:805–811.

Coombs, C. H. (1964). *A theory of data*. Wiley, New York, NY.

Davenport, M. A., Plan, Y., van den Berg, E., and Wootters, M. (2014). 1-bit matrix completion. *Information and Inference: A Journal of the IMA*, 3:189–223.

De Leeuw, J. and Mair, P. (2009). Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software*, 31:1–30.

DeSarbo, W. S. and Hoffman, D. L. (1987). Constructing MDS joint spaces from binary choice data: A multidimensional unfolding threshold model for marketing research. *Journal of Marketing Research*, 24:40–54.

DeSarbo, W. S., Young, M. R., and Rangaswamy, A. (1997). A parametric multidimensional unfolding procedure for incomplete nonmetric preference/choice set data in marketing research. *Journal of Marketing Research*, 34:499–516.

Donoho, D. and Stodden, V. (2004). When does non-negative matrix factorization give a correct decomposition into parts? In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in neural information processing systems*, pages 1141–1148. MIT Press, Cambridge, MA.

Embretson, S. E. and Reise, S. P. (2000). *Item response theory*. Psychology Press, Hove, UK.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer, New York, NY.

Gifi, A. (1990). *Nonlinear multivariate analysis*. Wiley, New York, NY.

Greenacre, M. J. and Browne, M. W. (1986). An efficient alternating least-squares algorithm to perform multidimensional unfolding. *Psychometrika*, 51:241–250.

Harper, F. M. and Konstan, J. A. (2016). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5:1–19.

Hays, W. L. and Bennett, J. F. (1961). Multidimensional unfolding: Determining configuration from complete rank order preference data. *Psychometrika*, 26:221–238.

Hinich, M. J. (2005). A new method for statistical multidimensional unfolding. *Communications in Statistics—Theory and Methods*, 34:2299–2310.

Ho, Y., Chung, Y., and Lau, K. (2010). Unfolding large-scale marketing data. *International Journal of Research in Marketing*, 27:119–132.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5:1457–1469.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional scaling*. Sage, Beverly Hills, CA.

Le Roux, B. and Rouanet, H. (2010). *Multiple correspondence analysis*. Sage, Newbury Park, CA.

Li, T., Levina, E., and Zhu, J. (2020). Network cross-validation by edge sampling. *Biometrika*, 107:257–276.

Lu, F., Keleş, S., Wright, S. J., and Wahba, G. (2005). Framework for kernel regularization with application to protein clustering. *Proceedings of the National Academy of Sciences*, 102:12332–12337.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Cam, L. M. L. and Neyman, J., editors, *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297. University of California Press, Berkeley, CA.

Mair, P., Borg, I., and Rusch, T. (2016). Goodness-of-fit assessment in multidimensional scaling and unfolding. *Multivariate Behavioral Research*, 51:772–789.

Mair, P., De Leeuw, J., and Wurzer, M. (2015). Multidimensional unfolding. In *Wiley StatsRef: Statistics Reference Online*. NewYork: Wiley.

Olver, P. J. (1999). *Classical invariant theory*. Cambridge University Press, Cambridge, UK.

Papesh, M. H. and Goldinger, S. D. (2010). A multidimensional scaling analysis of own- and cross-race face spaces. *Cognition*, 116:283–288.

Poole, K. T. (2000). Nonparametric unfolding of binary choice data. *Political Analysis*, 8:211–237.

Poole, K. T. (2005). *Spatial models of parliamentary voting*. Cambridge University Press, Cambridge, UK.

Poole, K. T. and Rosenthal, H. (1991). Patterns of congressional voting. *American Journal of Political Science*, 35:228–278.

Poole, K. T., Rosenthal, H., and Koford, K. (1991). On dimensionalizing roll call votes in the US congress. *American Political Science Review*, 85:955–976.

Rabe-Hesketh, S. and Skrondal, A. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Chapman and Hall/CRC, New York, NY.

Reckase, M. (2009). *Multidimensional Item Response Theory*. Springer, New York, NY.

Rohe, K., Qin, T., and Yu, B. (2016). Co-clustering directed graphs to discover asymmetries and directional communities. *Proceedings of the National Academy of Sciences*, 113:12679–12684.

Samejima, F. (1997). Graded response model. In van der Linden, W. J. and Hambleton, R. K., editors, *Handbook of modern item response theory*, pages 85–100. Springer, New York, NY.

Takane, Y., Young, F. W., and De Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42:7–67.

Tenenbaum, J. B., De Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.

Van Deun, K., Heiser, W. J., and Delbeke, L. (2007). Multidimensional unfolding by nonmetric multidimensional scaling of Spearman distances in the extended permutation polytope. *Multivariate Behavioral Research*, 42:103–132.

Xia, D. and Yuan, M. (2019). Statistical inferences of linear forms for noisy matrix completion. *arXiv preprint arXiv:1909.00116*.

Zhang, L., Wahba, G., and Yuan, M. (2016). Distance shrinkage and Euclidean embedding via regularized kernel estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:849–867.