This paper is © IEEE and appears reformatted in IEEE Transactions on Visualization and Computer Graphics, 23(1):51-60, 2017.
DOI: 0.1109/TVCG.2016.2598479

A Visual Analytics Approach for Categorical Joint Distribution Reconstruction from Marginal Projections

Cong Xie, Wen Zhong, and Klaus Mueller, *Senior Member, IEEE*

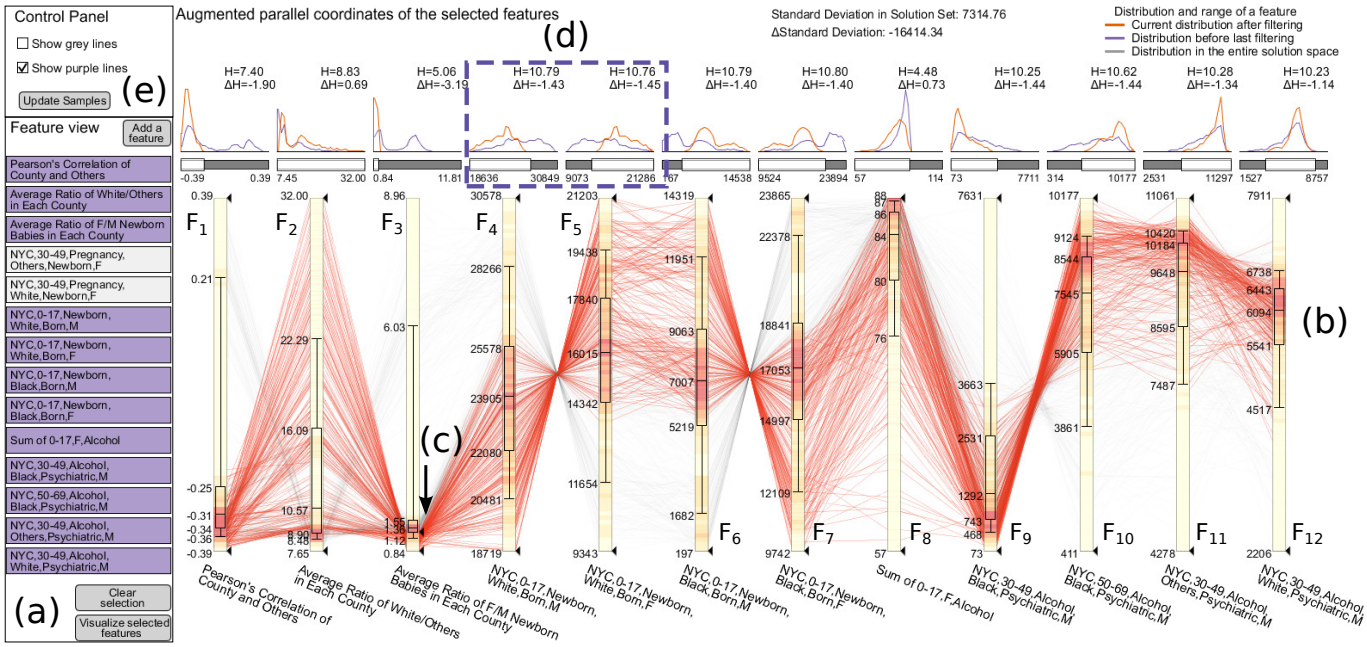

Fig. 1. The interface of our visual analytics framework for joint distribution reconstruction. (a) The features of the joint reconstruction solution space are defined. The user selects a subset of features (highlighted in purple) and visualizes it with (b) augmented parallel coordinates. Box plots and heat maps integrated into the axes bars show the distributions of the features. (c) Constraints can be added by filtering the range in each axis. (d) The probability density functions of the features before and after filtering are visualized as line charts. The bars below the line charts show the ranges of features after filtering. (e) The control panel for visualization.

Abstract—Oftentimes multivariate data are not available as sets of equally multivariate tuples, but only as sets of projections into subspaces spanned by subsets of these attributes. For example, one may find data with five attributes stored in six tables of two attributes each, instead of a single table of five attributes. This prohibits the visualization of these data with standard high-dimensional methods, such as parallel coordinates or MDS, and there is hence the need to reconstruct the full multivariate (joint) distribution from these marginal ones. Most of the existing methods designed for this purpose use an iterative procedure to estimate the joint distribution. With insufficient marginal distributions and domain knowledge, they lead to results whose joint errors can be large. Moreover, enforcing smoothness for regularizations in the joint space is not applicable if the attributes are not numerical but categorical. We propose a visual analytics approach that integrates both anecdotal data and human experts to iteratively narrow down a large set of plausible solutions. The solution space is populated using a Monte Carlo procedure which uniformly samples the solution space. A level-of-detail high dimensional visualization system helps the user understand the patterns and the uncertainties. Constraints that narrow the solution space can then be added by the user interactively during the iterative exploration, and eventually a subset of solutions with narrow uncertainty intervals emerges.

Index Terms—Joint Distribution Reconstruction, Solution Space, High-dimensional Data, Multivariate Data, Parallel Coordinates

1 Introduction

Fusing data from multiple sources or tables can often bring valuable insight. However, most of the time one cannot recover the full joint distributions that way. Instead one often faces a multi-set of marginal distributions which is only partially helpful in discerning the relationships within the fused domain. Take, for example, Andy, a researcher in public health who wishes to study a set of emerging diseases in the context of patient demographics. He contacts a few hospitals in his county and obtains two bivariate tables - diseases vs. age and diseases vs. gender. However, to Andy's great dismay knowing these two marginal distributions is insufficient to recover the tri-variate joint distribution: age vs. gender vs. disease. For example, while he now knows (1) how many females have disease A and (2) how many 20-30 years olds have disease A, he cannot determine how many 20-30 years old females have disease A. This is because the first group also contains other age groups, while the second group also contains males. Or taken another way, when visualized in a parallel coordinates plot, one

- *Cong Xie, Wen Zhong, and Klaus Mueller are with Computer Science Department, Stony Brook University. E-mail: {coxie, wezzhong, mueller}@cs.stonybrook.edu.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

could not create a set of polylines crossing all three axes: gender, age, and disease. One could only display disjoint polylines connecting the two axis pairs.

Andy's problem is one of reconstructing a 3D array from two 2D arrays. It is an under-constrained reconstruction problem. Assume we have 4 age groups, 4 diseases, and 2 genders. Then we have $4 \times 4 \times 2 = 32$ unknowns but only $4 \times 4 + 4 \times 2 = 24$ knowns.

This setting occurs in the reconstruction of 3D objects from their 2D projections, such as in computed tomography where we have a set of 2D projections acquired from different viewpoints. In this case often iterative reconstruction methods are applied. They seek to enforce an optimization criterion that can be written as follows:

$$\boldsymbol{x} = argmin_x ||\boldsymbol{W}\boldsymbol{x} - \boldsymbol{y}||_2 + \lambda || \bigtriangledown \boldsymbol{x}||$$

where $\boldsymbol{x}$ is the array to be reconstructed, $\boldsymbol{y}$ is the marginal data, $\boldsymbol{W}$ is the system matrix that relates $\boldsymbol{x}$ and $\boldsymbol{y}$, and λ is a constant. The first term is the data fidelity term, while the second is the regularization term that imposes some domain constraint on the reconstruction, in this case smoothness. Enforcing smoothness helps to steer the iterative reconstruction algorithm away from solutions with noise and streaks. While they would all fulfill the fidelity term, they are not plausible since the imaged object did not contain noise and streaks. This helps in trimming down the search space for the actual solution. As such, the smoothness constraint represents the domain knowledge. In Andy's setting, however, the smoothness term may not be applied since two of the three dimensions are categorical attributes. In this case the iterative approach will generate results with large joint distribution error because of the large and uncertain search space.

We propose interactive solution space analysis as a more accurate strategy for joint distribution reconstruction. It begins with a space of all plausible solutions constructed from the marginal distributions – we use a Monte Carlo based sampling procedure to generate a uniform distribution of solution space samples. We then refine it automatically by various forms of domain knowledge available as additional data. In Andys case, this additional data could be the marginal distributions of age and gender available per county from online census reports. Although these data may not exactly match the marginal distributions of the health data, they can still serve as additional constraints to refine the solution space.

Now, with this refined solution space in place, we put the human expert into the loop of the final reconstruction process. For example, the expert will know that females do not have prostate problems, or that more female than male babies suffer from fever for a certain disease. Applying these constraints will further reduce the number of possible solutions, which in many cases will also affect the probabilities of certain values in other attributes. All this will likely inspire the expert to apply further constraints, setting off an iterative refinement process.

To enable an effective discourse with the human expert an equally effective visual interface and knowledge assimilation algorithm is required. Our paper describes such a visual analytics framework. At its heart is a level-of-detail high-dimensional visualization technique based on an augmented parallel coordinates visualization interface. It can show a large amount of samples (e.g., 10 million) along with their uncertainties. Guided by the visualization, the user can then trim down the solution set by adding constraints interactively. We believe that the visual analysis of the solution space - the first of its kind to the best of our knowledge - can provide great value in the reconstruction of joint distributions in many types of settings and applications.

Our paper is structured as follows. Section 2 reviews related work. Section 3 defines the problem and gives an overview of our approach. Section 4 and Section 5 introduce the techniques of solution space construction and reduction. Solution sampling is presented in Section 6. Our visual exploration framework is described in Section 7, and two real world cases are used to validate our system in Section 8. Section 9 ends with conclusions and an outlook into future work.

2 Related Work

The reconstruction of numerical field data from their projections has been researched for decades and many algorithms have been proposed, such as the algebraic reconstruction technique (ART), simultaneous ART (SIRT) [6], multiplicative ART (MART) [19], and expectation maximization (EM) [16], all of which have applications in CT reconstruction. Most of these algorithms adopt an iterative framework and allow domain knowledge to be incorporated either using a related joint distribution as initial value or setting a regularization criterion.

However, these approaches cannot be applied to categorical data. Unlike medical 3D reconstruction where there are usually many 2D marginal projections, the marginal information of categorical data is usually insufficient. For example, a 5D joint distribution can have no more than 10 2D marginal distributions, which leaves the traditional iterative methods with an enormous solution space. By running an iterative method multiple times, we can obtain a result set with very small marginal error but extremely large standard deviation (as shown in Table 3). Even domain experts often cannot decide which result is better, and this reveals the impracticability of these approaches. In addition, because of the discreteness of categorical data, useful regularization such as smoothing cannot be employed.

Instead of finding one solution with large joint error, our visual analytics framework allows the user to explore the whole solution space to find a set of results that not only satisfy the given marginal information but also the applied domain knowledge requirements which render these solutions superior.

2.1 Categorical Data Reconstruction

Some of the iterative reconstruction methods are also used in categorical datasets, such as census data. Iterative proportional fitting (IPF) [15] revises a joint distribution which is close to the true solution with a set of marginal distributions. But IPF cannot always find a preferred solution with insufficient information as it uses a similar process as MART. Statistical methods use GMM [28] or Bayesian approaches [10] [26] to estimate the real joint distribution based on a prior joint distribution from smaller subsets of data. However, the prior distribution is unavailable in most cases. Some sampling methods [20] [17] [11] reconstruct contingency tables from the marginal frequencies by sampling the lattice points in high dimensional space. These techniques are not scalable to large datasets. Our method, on the other hand, reduces the solution space of reconstructions, making it capable of dealing with large datasets. Besides, with the assistance of iterative visual interactions, the domain expert is asked to focus on a specific part of the solution set in a large solution space.

To the best of our knowledge, most of the existing approaches are not able to show the features and the uncertainties of the solutions. In particular, no interactive approach has been described for exploring and filtering the solution space interactively with domain knowledge.

2.2 High Dimensional Uncertainty Visualization

Numerous methods [22] have been proposed for the interactive visual exploration of uncertainties in high dimensional space. Rados et al. [27] propose quantitative visual interaction methods which reduce the uncertainties. Torsney-Weir et al. [33] guide the user to find principal parameters for image segmentation algorithms. Wu et al. [35] visualize subjective opinions with degrees of uncertainty. Sanyal et al. [29] propose visual exploration tools for numerical weather model ensemble uncertainty. Some approaches are based on parallel coordinates. Chad et al. [32] allow users to explore hurricane trends with an interactive parallel coordinates plot. Wang et al. [34] create high dimensional data by sketching. Projection-based approaches [9] [30] [12] are also used for analyzing the uncertainties in high dimensional dataset. Berger et al. [8] use scatter plots to explore samples with respect to multi-target values. Correa et al. [13] incorporate uncertainty information into PCA projections and k-means clustering.

3 Problem Definition and Approach Overview

This paper seeks to address the following problem: Given k categorical attributes $A_1, A_2, ..., A_k$, and l marginal distributions $Y_1, Y_2, ..., Y_l$, estimate the joint probability $p(t)$ for each tuple $t = \{A_1 = a_1, A_2 = a_2, ..., A_k = a_k\}$. For instance, we wish to estimate the joint distribution of three attributes: A_1: "Age", A_2: "Gender", A_3: "Disease". Suppose Y_i is one marginal distribution over A_1 and

A_2. This can be represented as a vector $\boldsymbol{y_i}$ of $m_1 \times m_2$ probabilities: $\boldsymbol{y_i} = \{p(A_1 = 1, A_2 = 1), p(A_1 = 1, A_2 = 2), ..., p(A_1 = m_1, A_2 = m_2)\}$. Here $p(A_1 = a_1, A_2 = a_2)$ is a marginal probability. Before introducing the theory, Table 1 shows some related notations.

The reconstruction of the joint distribution is accomplished by our visual analytics approach with the following steps:

Step 1 Construction and reduction of the solution space: The solution space P of the joint distribution is constructed from a set of linear equalities and inequalities, which are defined by the marginal distributions. The dimension of P is reduced by a set of additional constraints - mainly the domain knowledge provided by the user.

Step 2 Sampling the reduced solution space: The Hit-and-Run sampler, which is a Markov Chain Monte Carlo method, is used to uniformly sample the solution space P.

Step 3 Visual exploration and filtering of the solution space: A set of features are extracted from the samples. These features are visualized in our augmented parallel coordinates display. The user can then explore and filter the solution space using a level-of-detail visual exploration loop.

Typically, the user iteratively loops over step 3 until a satisfactory solution subset is achieved. See Fig. 3 and Section 7 for a detailed illustration of this visual exploration loop.

4 Construction of the Solution Space

This section describes how the solution space of joint distribution is constructed from a set of linear equalities and inequalities, which are provided by the known marginal distributions.

4.1 Marginal Frequency Constraint

Each marginal probability in $\boldsymbol{y_i} (1 \leq i \leq l)$ is the sum of a set of joint probabilities. For $\boldsymbol{y_i}$, there is a system of linear equations:

$$\begin{array}{rcl}
\sum_{i=1}^{m_3} p(A_1 = 1, A_2 = 1, A_3 = i) & = & p(A_1 = 1, A_2 = 1) \\
\sum_{i=1}^{m_3} p(A_1 = 1, A_2 = 2, A_3 = i) & = & p(A_1 = 1, A_2 = 2) \\
\ldots & & \ldots \\
\sum_{i=1}^{m_3} p(A_1 = 1, A_2 = m_2, A_3 = i) & = & p(A_1 = 1, A_2 = m_2) \\
\sum_{i=1}^{m_3} p(A_1 = 2, A_2 = 1, A_3 = i) & = & p(A_1 = 2, A_2 = 1) \\
\ldots & & \ldots \\
\sum_{i=1}^{m_3} p(A_1 = m_1, A_2 = m_2, A_3 = i) & = & p(A_1 = m_1, A_2 = m_2)
\end{array} \quad (1)$$

Eq. 1 can be represented in linear matrix form:

$$\boldsymbol{W_i x} = \boldsymbol{y_i}, \quad (2)$$

where

$$\boldsymbol{x} = \{p(A_1 = 1, A_2 = 1, A_3 = 1), p(A_1 = 1, A_2 = 1, A_3 = 2), ...,$$

$$..., p(A_1 = m_1, A_2 = m_2, A_3 = m_3)\}$$

is the joint probability vector to be estimated. $\boldsymbol{W_i}$ is a matrix with $m_1 \cdot m_2$ rows and $m_1 \cdot m_2 \cdot m_3$ columns.

When there are multiple marginal distributions, Eq. 3 represents these as a set of linear matrix functions. Usually the number of rows in Eq. 3 is much smaller than the number of columns, rendering the joint distribution solution as non-unique.

$$\begin{array}{c}
\boldsymbol{W_1 x} = \boldsymbol{y_1} \\
\boldsymbol{W_2 x} = \boldsymbol{y_2} \\
\ldots \\
\boldsymbol{W_l x} = \boldsymbol{y_l}
\end{array} \quad (3)$$

Notation	Description
A_i	The ith attribute in the joint distribution
k	The number of attributes
a_i	The a_ith level of the ith attribute
m_i	The number of levels of the ith attribute
$p(A_1{=}a_1, A_2{=}a_2, ..., A_k{=}a_k)$	The probability of $\{A_1 = a_1, A_2 = a_2, ..., A_k = a_k\}$
$f(A_1{=}a_1, A_2{=}a_2, ..., A_k{=}a_k)$	The frequency of $\{A_1 = a_1, A_2 = a_2, ..., A_k = a_k\}$
N	The sum of joint frequencies
$\boldsymbol{x}$	A vector of all joint probabilities $\{p(1,1,..,1), p(1,1,..,2), ...p(m_1,m_2,..,m_k)\}$
$x^{(i)}$	The ith component of vector $\boldsymbol{x}$
n_0	The number of joint probabilities
$\boldsymbol{y_i}$	A vector of the marginal probabilities of the ith marginal distribution.
l	The number of known marginal distributions
n	The number of joint probabilities to be estimated after reduction
$P \subseteq \mathbb{R}^n$	The polytope of the solution space in $\mathbb{R}^n$
$p \in P$	A point in polytope P

Table 1. Mathematical notations

4.2 Uniting Multiple Marginal Distributions

Since the marginal distributions may originate from different data sources, inconsistent granularities of marginal information should be considered when unifying them into a single constraint framework. For example, if the "Age" attribute is at a finer granularity than the granularity we design for, its levels "0-5" and "5-10" could be merged into a coarser level "0-10"; and vice versa for the coarse granularity case. In this way, different granularities constraints are incorporated in Eq. 3.

Different sources may also cause marginal inconsistencies. For example, the marginal distribution of the general census data may be inconsistent with the patient data because the former reflects the county level population and the latter reflects the hospital level population. There are obviously some dependencies but the age-gender fractions are likely not identical. Even marginal distributions from the same sources can be slightly different due to some errors. Eq. 3 may have no solution if the equations of different marginal probabilities conflict.

The inconsistencies are resolved by adding an error tolerance vector $\boldsymbol{\varepsilon_i}$ $(1 \leq i \leq l)$ to each marginal distribution, as shown in Eq. 4.

$$\begin{array}{c}
-\boldsymbol{\varepsilon_1} \leq \boldsymbol{W_1 x} - \boldsymbol{y_1} \leq \boldsymbol{\varepsilon_1} \\
-\boldsymbol{\varepsilon_2} \leq \boldsymbol{W_2 x} - \boldsymbol{y_2} \leq \boldsymbol{\varepsilon_2} \\
\ldots \\
-\boldsymbol{\varepsilon_l} \leq \boldsymbol{W_l x} - \boldsymbol{y_l} \leq \boldsymbol{\varepsilon_l}
\end{array} \quad (4)$$

where $\boldsymbol{\varepsilon_i}$ is defined as a nonnegative constant vector, which is proportional to the marginal distributions $\boldsymbol{y_i}$ or set by the user. Particularly, when all marginal distributions match perfectly, $\boldsymbol{\varepsilon_i}$ can be set as a zero vector, and Eq. 4 will degenerate to Eq. 3.

5 Dimension Reduction of the Solution Space

In the following, we distinguish "variable" from "attribute". An attribute A_i is a singular characteristic of an entity (e.g., A_1: "Age", A_2: "Gender", and A_3: "Disease"). A variable $x^{(i)}$, on the other hand, represents a joint probability, such as $p(\{$ "0-17", "Male", "Fever"$\})$.

Eq. 4 defines a closed convex solution polytope in high dimensional space $\mathbb{R}^{n_0}$, where n_0 equals the number of variables to be estimated. This is a massive space where even a small number k of attributes can already generate a huge number of variables $n_0 = \prod_{i=1}^{k} m_i$. Likewise, the number of samples needed to probe it also grows exponentially with the dimensionality k. In the following, we propose several reduction techniques that can reduce the number of variables to be estimated.

5.1 Lower Bound and Upper Bound Constraints

Since each joint probability is nonnegative and it is less than or equal to any of its corresponding marginal probabilities, there exists a lower bound $b_l^{(i)} = 0$ and upper bound $b_u^{(i)}$ for $x^{(i)}$, where

$$0 \leq x^{(i)} \leq b_u^{(i)}.$$

Furthermore, a marginal probability of 0 renders all of the corresponding joint probabilities to 0. For instance, if the marginal probability of $\{A_2$ ="male", A_3 = "pregnancy"$\}$ is zero, then all of the joint probabilities of tuples with A_2 = "male", A_3 = "pregnancy" have to be zero as well. This can help to reduce a large number of independent variables because not all combinations are valid in some datasets.

5.2 Domain Knowledge Constraints

Domain knowledge adds additional constraints. In the reduction step only linear constraints are employed since they will keep the solution space convex. Non-linear constraints can be applied in form of features in the later interactive part (See Section 7.1).

5.2.1 Moments

Covariance and expectation are the two main forms of domain knowledge constraints. Covariance information describes the relations of attributes. For example, if an attribute A_1 is independent of the other attributes, then the covariance is 0 and the corresponding constraint $E(A_1,...,A_l) - E(A_1)E(A_2,...,A_l) = 0$ could be employed. Conditional expectation such as $E(A_1|A_2)$ can be considered a constraint as well, if the marginal distribution of A_1 and A_2 is not provided. These two kinds of constraints are represented in a linear matrix form (Eq. 5).

$$\boldsymbol{W_0 x = d_0} \tag{5}$$

5.2.2 Range of Variables

Besides the upper bound and lower bound found by the nonnegative property of distributions, a more restricted range could be provided with domain knowledge. For example, the user can set the probability of the tuple $t = \{$ "female", "0-2", "fever" $\}$ to be larger than 0.3. Together with the upper bound and lower bound described in Section 5.1, it can be summarized as:

$$\boldsymbol{b_l \leq x \leq b_u} \tag{6}$$

If $b_l^{(i)} = b_u^{(i)}$, the value of $x^{(i)}$ will be fixed and it can be removed.

5.3 Integer Constraints

For categorical datasets, marginal information is usually presented as frequencies, we can estimate the joint frequencies $f(t)$ and then calculate the joint probabilities following $p(t) = \frac{f(t)}{N}$, where N is the sum of joint frequencies. In this case, each element in $\boldsymbol{x}$ is a integer joint frequency, the solution sets will consist of all the integer points in the solution space. This will reduce the infinite solution set to a finite number of solutions. Moreover, since $x^{(i)}$ has to be integer, we can check the maximum value $x_{max}^{(i)}$ and minimum value $x_{min}^{(i)}$ of $x^{(i)}$. If there is only one integer value $x_{integer}^{(i)}$ between $x_{max}^{(i)}$ and $x_{min}^{(i)}$, $x^{(i)}$ can only be $x_{integer}^{(i)}$. Therefore, the solution space will be reduced further.

Eq. 4, Eq. 5, and Eq. 6 define a closed convex solution space polytope in $\mathbb{R}^{n_0}$ and the proposed dimension reduction techniques are adopted to reduce the dimensions (remove the variables with fixed value). Following this way, the number of variables in the joint frequencies is decreased to n, which can be much smaller than n_0 in some cases. The reduced solution space is then a polytope $P \subseteq \mathbb{R}^n$.

6 Sampling of High Dimensional Solution Space

For the solution space polytope $P \subseteq \mathbb{R}^n$, each point $p \in P$ is a reconstruction solution. Features such as the shape of the polytope can be learned by uniformly sampling points in it. We use the Markov Chain Monte Carlo based Hit-and-Run (HR) sampler [31] which is one of the most efficient methods for sampling in a closed convex polytope.

6.1 Hit-and-Run Sampler

The standard HR sampler algorithm is presented as follows:

Step 1: Let $p_i \in P$ be the point in the ith iteration (Fig. 2 (a)).

Step 2: Uniformly generate a random direction $d_i \in \mathbb{R}^n$, i.e. d_i is uniformly distributed in the n dimensional hypersphere (Fig. 2 (b)).

Step 3: Find the chord L inside P through p_i along the directions d_i and -d_i. So any point p on L can be represented as $P \cap \{p | p = p_i + \lambda d_i\}$ (Fig. 2 (c)).

Step 4: Uniformly sample a point p_{i+1} on L as the next point, which is used as the next starting point (Fig. 2 (d)).

Step 5: Repeat Step 1 - 4 for s steps until the sampling distribution becomes stationary.

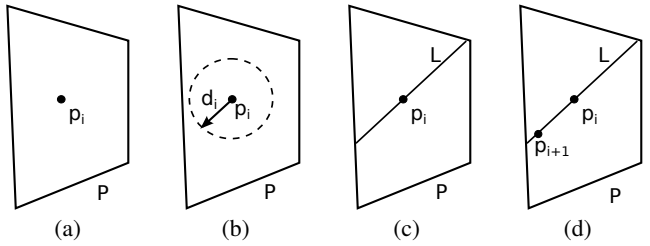

Fig. 2. We use an example in $\mathbb{R}^2$ to illustrate the Hit-and-Run (HR) algorithm. The polygon represents the solution space, and each point in the polygon is a solution for the reconstruction. (a) A point $p_i \in P$ in ith iteration. (b) A direction d_i is generated uniformly. (c) Following direction d_i, find the chord L through p_0 inside P. (d) Select p_{i+1} uniformly on L as the next point.

The HR sampler is a MCMC chain which reaches the stationary distribution (i.e. generate a uniformly distributed sample) in polynomial time [23]. To obtain c uniformly distributed samples, c chains of HR should be employed. Each of these chains iterates for at least s steps and the last point of each chain is kept as one sample. With the number of dimensions n increasing, s will become extremely large for the MCMC chain to reach a stationary distribution. Hence acceleration methods such as starting point generation are used to cut down s.

6.2 Acceleration Methods

HR converges rapidly from any interior point of a convex body [24], while it is slow when the initial point of iteration is on the border or in a corner of P. So we propose a strategy to generate an interior point as the starting point to avoid this situation.

Firstly, n_p random points on the boundary of P are generated and added to a point set S. Each point $p \in S$ is generated by a linear optimization (Eq. 7) with a random objective function (i.e., $\boldsymbol{c}$ and $\boldsymbol{d}$ are selected randomly).

$$\begin{aligned} min \quad & \boldsymbol{c}^T \boldsymbol{x} - \boldsymbol{d} \\ s.t. \quad & \boldsymbol{x} \in P,\ P \subseteq \mathbb{R}^n \end{aligned} \tag{7}$$

Then the center point of S is defined as $p_c = \frac{1}{n_p} \sum p$, which is used as the starting point. p_c has a very low probability on the border since points in S can be in different regions of P. In addition, since p_c does not have to be the exact center of P, a small number n_p of points is sufficient. Hence the cost of the start point generation process is low. Other acceleration methods such as artificial centering HR strategy [21] for direction selection can be used as well.

6.3 Lattice Point Generation

If the joint frequencies are estimated, the solution will be integer points, i.e. lattice points in P. Existing HR-based lattice sampling methods [7] are only applicable for hyper-rectangles. Other [20] methods need a larger convergence time than HR. Here we propose a simple and fast rounding strategy for generating a lattice point based on a random point $p \in P$ from the original HR sampler. The result will be a set of nearly uniformly distributed lattice points in P.

Suppose p is in a unit hypercube $C \subseteq \mathbb{R}^n$, whose vertices are lattice points. A straightforward way is finding the nearest vertex p_l of p by

rounding. However, p_l can be outside of P when p is near the border. And since a C has 2^n vertices, checking all its vertices is also impractical. Eq. 8 is used for finding a lattice point that locates between p and p_c. Since the solution space p is convex, Eq. 8 makes sure $p_l \in P$.

$$p_l^{(i)} = \begin{cases} \lfloor p^{(i)} \rfloor, & \text{if } p^{(i)} \geq p_c^{(i)} \\ \lceil p^{(i)} \rceil, & \text{otherwise} \end{cases} \quad (8)$$

Here $\lfloor \bullet \rfloor$ and $\lceil \bullet \rceil$ are floor and ceiling functions. $p_l^{(i)}$, $p^{(i)}$, $p_c^{(i)}$ are the ith component of p_l, p, and p_c, separately. The complexity of HR will not increase since the rounding process finishes in O(n) time.

6.4 Stopping Criterion

Although HR converges in polynomial time, the number of iterations is still high when n is large. Different convergence diagnostic methods [14] for testing whether the chains are converged can help to stop the iteration in a timely manner. We use Gelman and Rubin's convergence diagnostic [18] as the HR stopping criterion. A $\hat{R}$ value for all of the chains is computed to judge whether the sampling chains reach stationary distributions. $\hat{R}$ is mainly dominated by the ratio of $\hat{V}$ and W, where $\hat{V}$ is the variance of the current last points in the c parallel chains and W is the average of the within-chain variance. A chain's within-chain variance is the variance of all points it has generated so far. If $\hat{R}$ is around 1, the chains are considered converged. In this case each chain has sampled the space sufficiently and has not become stuck in a local area. In practice a value of 1.2 is satisfactory for most problems [18]. The $\hat{R}$ value is tracked during the sampling and once it is below a threshold (e.g., 1.2), we stop the sampling chains. Section 8.4 illustrates the sampling mechanism with two case studies.

7 Visual Exploration of Solution Space

The large solution space is visualized with high dimensional visualization techniques enabling the user to interactively explore interesting solution regions at different levels. Our first approach is shown with grey edges in Fig. 3. HR is used to sample the solution space online and update the samples when the user adds new a constraint (Fig. 3 (a)). The resulting sample set can be visualized via a parallel coordinates interface (Fig. 3 (b)), where the user can interactively trim down the solution space. After the filtering (Fig. 3 (c)), the system increases the sampling density in the reduced solution space to allow for exploration in details. The above process is repeated until the user identifies a satisfactory subset of solutions.

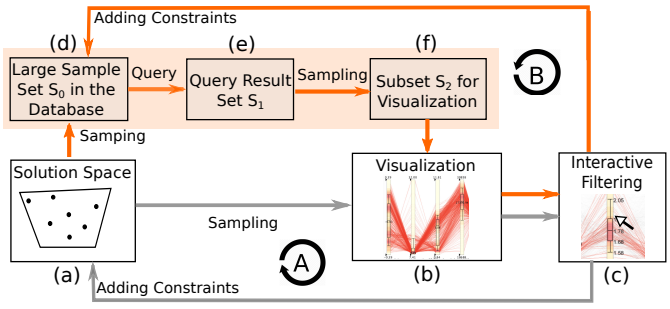

Fig. 3. Strategies for interactive solution space refinement Loop A (grey edges). Samples from (a) the solution space are (b) visualized and explored interactively. (c) When the user sets constraints, more samples are generated in the reduced space. By repeating this process, the solution space is trimmed down. Loop B (orange edges). The steps in the orange boxes are added to save the time of MCMC sampling. (d) Large amounts of samples S_0 are generated and stored in the database in advance. (e) Instead of sampling the solution space, we query the database to get a subset of the samples S_1 which meet the user's constraints. (f) A smaller set S_2 is sampled from S_1 for visual exploration since the capacity of the visualization is limited.

This approach works well with small solution spaces (e.g., $n = 50$). However, when the dimensionality of the solution space is large, the sampling process converges very slowly. For example, when $n = 4,505$, each chain needs to iterate for more than 52,000 steps to converge (as shown in Section 8.4). This requires the user to wait for about 10 minutes, which is prohibitively long.

To overcome this problem, we design Loop B (orange edges). Here a large set of samples S_0 (e.g., 10M samples) is generated and stored in a database beforehand (Fig. 3 (d)). Then, when the user sets constraints, a subset of $S_1 \in S_0$ which satisfies the user's constraints is retrieved from the database (Fig. 3 (e)). However, the number of samples of S_1 may still be too large (e.g., 10,000) for a visually manageable parallel coordinates plot (or any other such display). Instead, a smaller set S_2 (e.g., 2,000 samples) is generated by sampling from S_1 uniformly (Fig. 3 (f)) for visualization. At the same time, the uncertainties of the samples in S_1 (e.g., their ranges or distribution variances) are also shown to the user, providing more guidance for the exploration. S and S_1 are updated whenever a new constraint is added. After multiple iterations of this visual exploration activity, the result will be a smaller sample set S_1 whose uncertainty is also small.

7.1 Defining Features of Solution Space

When the dimensionality of the solution space n is large (e.g., $n = 4,000$), there will be too many axes in the parallel coordinates display. A way to solve this problem is to select or extract a small number (e.g., $n \leqslant 10$) of features and use these in place of the native attributes. However, the extracted features of most dimension reduction algorithms (e.g., PCA or t-SNE [25]) are too complex to be interpreted, making it hard for the user to set meaningful constraints on them.

A proper feature is expected to satisfy the following principles: 1). Explainability. Using a representative variable as a feature is preferred. 2). Relevance to the user's task. Some features may not be able to align with the user's existing questions. For example, a doctor who studies the distribution of patients in New York City may not care about other counties in the dataset. 3). Efficiency on reducing the solution space. Adding constraints to features with high uncertainties can filter the solution space to a large extent.

According to the above principles, different types of features are computed automatically or set by the user manually.

- Variables with high uncertainties. Variables whose ranges or entropies are larger than a given threshold are selected.
- Sum or expectation of a group of variables. These linear features allow users to set constraints interactively based on provided visual hints (e.g., whiskers of a boxplot).
- Ratio or correlation of variables. Users are typically more cognizant of the ratio of two variables than their exact difference. For example, the user may not know the absolute difference of the populations of females and males in New York, while he/she know that ratio is about 1.0. Many other relations can also be applied (e.g., spurious correlation).

Clusters of samples can also serve as features. However, since the samples are continuously and uniformly distributed in the polytope P, it is difficult to locate clusters in practice.

The predefined features are shown in the feature view (Fig. 1 (a)). The user is also allowed to add new features interactively via the feature definition view (See Fig. 6 (a) for the commuter dataset). If the amount of features exceeds the capacity of the visualization (e.g., limits on the number of display axes), the user can select a part of them from the feature view (purple features in Fig. 1 (a)) for visualization.

7.2 Augmented Parallel Coordinates Visualization

Initially, we use standard parallel coordinates to visualize samples in S_2. Each polyline represents a sample and each axis is a feature (Fig. 4 (a)). Although this configuration is able to visualize the general pattern of polylines crossing the axes, interpreting a feature's distribution from the density of the polylines can be misleading. Moreover, no information about the changes of a feature's distribution is provided after filtering. For example, when the user decreases the upper bound of F_8 in Fig. 4 (a), it is difficult to know whether there are changes in both

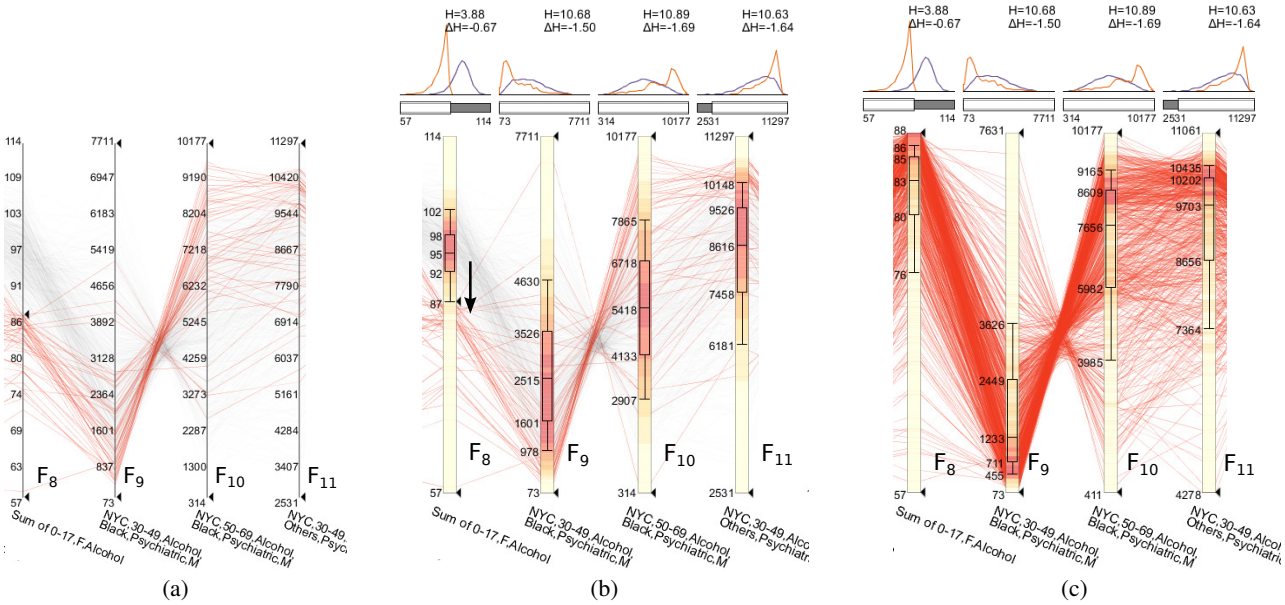

Fig. 4. (a) Standard parallel coordinates. (b) Our augmented parallel coordinates visualize the distributions and the changes of features. In the health data case, when the user reduces the range of the first axis (the sum of females under 17 years old who have "Alcohol/Drug Use"), the distributions of F_9, F_{10}, and F_{11} change accordingly, and their uncertainties decrease. (c) The axes of F_8 to F_{11} are zoomed in to fit the reduced ranges. More samples in the reduced solution space are added to show more details.

mean value and variance of F_{11}'s distribution, and how they change (4th axis in Fig. 4 (a)). Information of changes, however, is critical for understanding the effects of constraints on reducing the uncertainties of the features.

We hence modify the parallel coordinates to encode the feature distribution. One natural idea is to add visual representations to the axis [32], which can be done by changing each axis line to a bar (Fig. 5 (a)). A box plot is placed inside (Fig. 5 (b)) to show the quartiles of the feature's distribution in S_2. The band in the middle of the box is the median, and the bottom and top are the first and third quartiles. The ends of the whiskers can encode the bilateral standard deviation, or the 10th and 90th percentile, or others. Similar to scale markers on standard parallel coordinates (Fig. 4 (a)), the box plot provides helpful hints for filtering the axis range. The user can also optionally hide the labels of quartile values if their exact value is not important. If two labels overlap, one of them will be hidden to avoid visual clutter. Since comparing the box plot against the PDF may assist understanding, a 1D heat map is also provided to show the feature's PDF.

The user can set a constraint by filtering the feature's range using a cursor (Fig. 5 (c)). Polylines whose feature values are out of range will fade out. Since changes of each feature's distribution or range can also be critical to provide exploration guidance, a few auxiliary visualizations are available. A line chart shows the change of a feature's PDF in S_1 (Fig. 5 (d)). The purple and orange lines represent the PDFs before and after the user's last filtering. A grey line shows the PDF of a feature in the original entire solution space. The previous distributions can be hidden (i.e., purple and grey lines) (Fig. 1 (e)). A window on a bar (Fig. 5 (e)) indicates the current value range of a feature. Via this line chart and window range bar the user can obtain valuable visual feedback on the effect the applied constraint has with regards to the feature's distribution. For example, when the user decreases the upper bound of the first axis (F_8) in Fig. 4 (b), the distribution of F_9 (orange line, second line chart in Fig. 4 (b)) is skewed to the left compared to its previous distribution (purple line, second line chart in Fig. 4 (b)). Conversely, the line chart of F_{10} has the opposite skewness change. To indicate uncertainty, the entropy value of the current PDF and its change after filtering are also shown (Fig. 5 (f)).

7.3 Level-of-detail Visual Exploration

The user can zoom in and add more samples from the reduced solution space to explore the zoomed region in details. This is done by updating

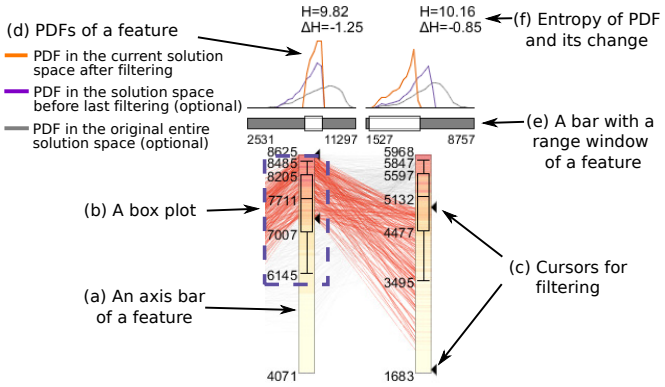

Fig. 5. Our augmented parallel coordinates for visualizing distribution and change of each axis. (a) Each axis bar incorporates a heat map which represents a feature. (b) The box plot of each bar shows the distribution of a feature. (c) The user can filter the range via cursors provided at each axis. (d) The current and previous probability density functions (PDFs) of features are visualized by orange, purple, and grey lines. (e) A window shows the value range of a feature after filtering. (f) The change of entropy and the current entropy are also indicated.

S_1 with the new constraints and sampling S_2 uniformly from S_1. Each parallel coordinate axis will be scaled appropriately to fit the reduced range. The box plots and heat maps are also changed accordingly. For example, after the user narrows the range of the first axis (F_8) in Fig. 4 (b), the axes scale to fit this new range and the covered range is enriched with samples from the reduced solution space(Fig. 4 (c)).

By repeatedly adding constraints, the solution space will be continually reduced. A solution set is satisfied if the standard deviation of S_1 is smaller than a threshold and/or the number of solutions remaining in S_1 is less than a given number (e.g., 10,000).

8 Case Studies

We conducted two case studies to show the usability of our system.

8.1 Participants and Procedure

We recruited two volunteers who were interested in reconstructing their data. The first volunteer was a graduate student from our university's public health department. He had been working with the public online New York Health Data query system [4] to find relations between sociodemographics and diseases in New York State. However, the query system only provides partial information, so our system presented him an opportunity to extract a more complete view. The other volunteer was a computer science student with a research focus on urban computing. He was interested in the relations of the departure times to work and the different modes of transportation. He wanted to learn it by reconstructing a joint distribution of commuter information in US from multiple tables of partial information. Both of our users were not experts in visualization and had basic statistical knowledge.

Before our study, we had several discussions with each user to learn about the attributes of the joint distribution they wanted to recover. Together, we searched online for additional data to supplement the primary sources. In discussions, we identified the constraints they wanted to set. After the solution space was constructed and sampled, we had another discussion to predefine a set of interesting features of the solution space. Each case study started with a 25-minute demo to introduce our approach. We then asked each person to use the system, followed by an interview to gather evaluations and subjective feedback.

Health Case	
Attributes of joint distribution reconstruction	Number of levels
A_1: "County (C)"	$m_1 = 9$
A_2: "Age Group (A)"	$m_2 = 5$
A_3: "Major Diagnostic Category (M)"	$m_3 = 26$
A_4: "Race (R)"	$m_4 = 4$
A_5: "Service Category (S)"	$m_5 = 6$
A_6: "Sex (G)"	$m_6 = 3$
Commuter Case	
Attributes of joint distribution reconstruction	Number of levels
A_1: "Age Group (A)"	$m_1 = 4$
A_2: "Gender (G)"	$m_2 = 2$
A_3: "Occupation Category (O)"	$m_3 = 6$
A_4: "Departure Time Period (D)"	$m_4 = 8$
A_5: "Mode of Transportation (T)"	$m_5 = 5$

Table 2. The attributes and the levels our users sought to reconstruct, for both the health and the commuter case.

8.2 Case Study 1: Public Health Data

The first user aimed to reconstruct the joint frequencies of six attributes (see Table 2) from the New York Health Data query system [4]. Ten 3D marginal frequencies were returned by the query system: C-A-M, C-A-R, C-A-S, C-A-G, C-M-R, C-M-S, C-M-G, C-R-S, C-R-G, C-S-G. A census dataset [5] was identified and served as supplementary data. It contained a 4D marginal distribution of C-A-S-G.

8.2.1 Construction and Sampling of the Solution Space

"Age Group" and "Race" in the census dataset had smaller granularities and we aggregated them to match the granularity of the query data. The marginal distributions from these two sources were not consistent since the census data provides general demographic information, which includes the patients but also other state residents. The tolerance for marginal error of the census data was set to 5% in Eq. 4.

The total number of variables to be estimated was $n_0 = \prod_{i=1}^{6} m_i = 84,240$. But not all of the combinations of attribute levels were meaningful. For example, no person matched the condition "MDC" = "Newborn", "Age Group" = "70+". Also, many variables could be removed by checking their lower and upper bounds (Section 5.1). The assumption that both "Race" and "County" were independent with other variables was added by the user as a constraint (Section 5.2.1). Since the patient number is an integer, the solution space could be further reduced using the method described in Section 5.3. After all constraints were set, $n = 4,505$ variables remained post-reduction – only about 5% of the original number. 10 million samples from the solution space $P \subseteq \mathbb{R}^n$ were generated and stored in the database.

Then we identified some features of the solution sets. Variables with large ranges were extracted and most of them were related to "Newborn" or "Alcohol/Drug Use" in "Major Diagnostic Category (M)", such as $F_4 = \{$"NYC", "0-17", "Newborn", "White", "Newborn", "Male"$\}$, $F_5 = \{$"NYC", "0-17", "Newborn", "White", "Newborn", "Female"$\}$, $F_9 = \{$"NYC", "30-49", "Alcohol/Drug Use", "White", "Psychiatric", "Male"$\}$, $F_{10} = \{$"NYC", "50-69", "Alcohol/Drug Use", "White", "Psychiatric", "Male"$\}$. The user then went on and defined several additional features related to this initial feature set. The new features included the sum of females under age 17 with "Alcohol/Drug Use" (F_3), and the average ratio of female-to-male newborn babies in different counties (F_8).

8.2.2 Visual Exploration

The user selected 12 features from the predefined features in the feature view (Fig. 1 (a)) and visualized them with the augmented parallel coordinates interface (Fig. 1 (b)). He found that the initial standard deviation of the solution space was about 2.1×10^5, which was similar to the results obtained by the traditional iterative approaches (see Table 3). The user commented that according to his knowledge, few females under "0-17" engage in "Alcohol/Drug Use". So he lowered the upper bound of F_8 (see first axis in Fig. 4 (b)) to the 10th percentile (bottom whisker of the box plot). After the filtering most of the polylines with high values in F_9 and low values in F_{10} faded out. The user then found that the original PDFs of F_9 and F_{10} were flat, indicating high uncertainty (purple lines, line charts of F_9 and F_{10} in Fig. 4 (b)). After filtering, the distribution of F_9 skewed to the left, while F_{10} had an opposite pattern in its distribution (orange lines, line charts of F_9 and F_{10} in Fig. 4 (b)). This was an indication that a preferred solution might have a low F_9 value and a high F_{10} value. The user then zoomed into the axes, and added more samples from the reduced solution space to the parallel coordinates (Fig. 4 (c)).

The user noticed that the ranges of the features which related to "Newborn" were still high, such as the number of white male and female babies (F_4 and F_5 in Fig. 1). From the box plots he observed that the 80% confidence interval (between the box plot's top/bottom whiskers) of F_4 was above 20,000, while the 80% confidence interval of F_5 was below 20,000. The user found it difficult to reduce the axes ranges directly since he did not know the exact number of newborn babies in NYC. He decided to use the gender ratio of newborn babies (F_3 in Fig. 1) instead. He said that the numbers of females and males were about the same, so he shrank the ratio range to $[0.9, 1.1]$ (Fig. 1 (c)). The value range and distribution of F_4 were skewed to low values, and those of F_5 changed in the opposite way (orange line, range window in Fig. 1 (d)). The user estimated from the changes in the visualization that their expectations of distributions had moved closer to each other. He concluded that with the set ratio, the samples from the preferred solution space had smaller differences in the number of male and female babies.

After the exploration, the solutions in the remaining solution set S_1 met all of the marginal frequencies and domain knowledge. The standard deviation of S_1 was about 7.3×10^3 (See top of Fig. 1), which was only 3% of the standard deviation of the original solution space. A final solution could then either be a random choice from this tight solution set or the mean of the solutions.

8.3 Case Study 2: Commuter Case

Our second user found the US Commuting Dataset [3] online. He was interested in the joint distribution of the five attributes in the dataset (Table 2). The dataset provides six 2D marginal distributions: A-D, G-D, O-D, A-T, G-T, and O-T. Another census data [1] was found to provide a demographic distribution of "Age Group (A)" and "Gender (G)". Especially, the user was interested in the distribution of "Departure Time Period (D)" and "Mode of Transportation (T)".

8.3.1 Construction and Sampling of the Solution Space

The marginal distribution errors given in the dataset were used as the tolerance of errors in Eq. 4. The number of unknown variables was

$n_0 = \prod_{i=1}^{5} m_i = 1920$. After checking the lower and upper bounds of the variables, $n = 1800$ variables remained (Section 5.1). 10 million samples were generated and stored.

Variables with high uncertainty were computed and used as features, such as the probabilities of carpooling in different time periods (F_5 to F_7). After learning the user's interests in "Departure Time Period (D)" and "Mode of Transportation (T)", we defined a set of related features, such as expectation of D for each level in T (F_1 to F_4), and Pearson's correlation of D and T (F_8).

8.3.2 Visual Exploration

At the beginning of the study, the user added the ratios of female to male in different occupations (F_9 to F_{12}) using the feature definition panel (Fig. 6 (a)). He explained that he was also interested in them but they were not provided in the marginal distributions. Then a subset of features was selected (Fig. 6 (b)) and visualized with the augmented parallel coordinates. Because farming and constructing jobs are mainly done by men, the user lowered the ratio of female to male in "Farming" and "Constructing". This operation also shrank the range of F_8 to $[-0.01, 0.0]$ (Fig. 6 (c)), which indicated there might be a subtle relationship between departure time and transportation modes.

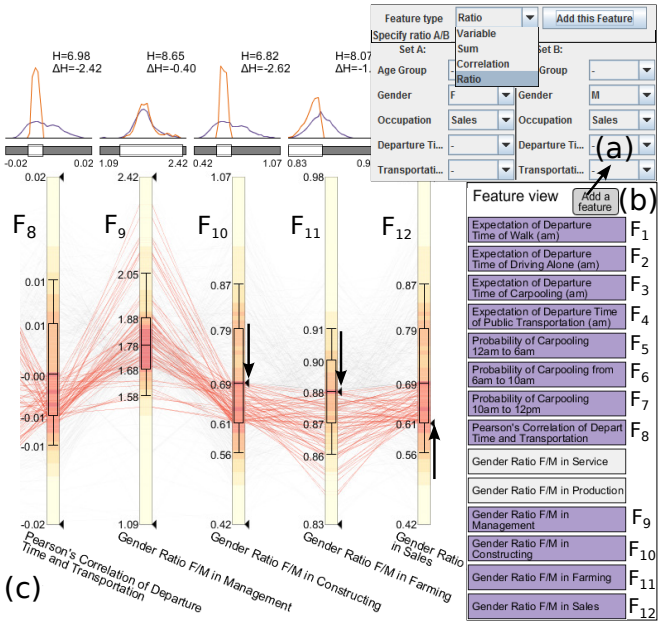

Fig. 6. (a) The user added new features with the feature definition view. (b) A subset of features (highlighted in purple) was selected, and visualized in the parallel coordinates. (c) The user adjusted the ratios of female to male in different occupations, such as "Constructing" (F_{10}), "Farming" (F_{11}), and "Sales" (F_{12}). The distribution of Pearson's correlations (F_8) between "Departure Time Period" and "Modes of Transportation" indicated subtle relationship of them in the reduced solution space.

Then the user checked the mean departure time of different transportation modes (F_1 to F_4 in Fig. 7). The average departure time of walk was later than other transportation modes. He observed that the distribution of F_4 was flat, which means the expected departure time for public transportation varied in a large range (purple line, line chart of F_4 in Fig. 7). The user thought the peak of public transportation usually happens in rush hours, so he reduced the range to about $[7:00, 9:00]$. He then zoomed in the axes and updated the samples in the parallel coordinates for a detailed exploration.

The expectations of departure time in different transportation modes changed accordingly. For example, the user checked the probabilities of carpooling in different time periods in the morning: 12am to 6am, 6am to 10am, and 10am to 12pm (F_5 to F_7 in Fig. 7). The line charts of F_5 to F_8 showed that their distributions were skewed into different directions (See dashed box in Fig. 7). From the boxplot of $F5$, the user observed that the 80% confidence interval of "probability of carpooling from 12am to 6am" moved from a lower region $[2.69\%, 5.78\%]$ to a higher region $[3.36\%, 5.87\%]$. This indicated the solutions in the preferred region have higher probabilities in earlier carpooling. Then user did a survey online and found a related report [2]. He concluded that this showed the carpooling pattern of people from the same household. For example, the married couples depart early to drop off a child at school or an old person at an older adult center on their way to work. This verified the user's result solutions as well as his constraints.

Most of the PDFs in Fig. 7 had spike shapes after filtering. This indicated less uncertainties of the features' distributions in the reduced solution set than those in the original set. The standard deviation was reduced from 2.28 to 0.89 after the exploration. With more domain knowledge, the user could trim down the solution space further.

8.4 Quantitative Analysis

For the solution space reduction method in Section 5, it is more efficient when the marginal distributions are sparse (i.e. there are a lot of zeros). In this case, the joint probabilities to be estimated are also sparse. About 90% variables are reduced in the health case, while only a small number of variables are removed in the commuter case. Our method is also effective in frequency reconstruction, since the integer reduction generates a much smaller and limited solution space.

For the sampling, the $\hat{R}$ values of the HR chains are tracked. Once the $\hat{R}$ of a iteration is below 1.2 (Fig. 8), the HR chains are regarded as converged. The number of convergence iterations are about $52,000$ and $14,000$ for the health case and the commuter case respectively (Fig. 8). Although the acceleration methods are applied, the converge iterations for both cases are still large. As a result, our strategy of generating and storing the samples in advance (Orange lines in Fig. 3) is necessary.

The scalability of our method is mainly decided by the performance of the HR sampler, whose complexity is polynomial [23]. However, the variables to be estimated grows exponentially with the number of attributes (e.g., 10 attributes may make $n = 10^{10}$), which is too large even for a polynomial sampler.

We test the traditional iterative methods and compare them with our approach in both cases (Table 3). For each iterative method, we test 10,000 trials with random initial values. All methods have small average marginal errors, which means their solutions are consistent with the given marginal distributions. Here the marginal error is the sum of the Euclidean distances between a result's marginal distributions and the given marginal distributions (i.e., NY Health Query System [4] in the health case and US Commuting Dataset [3] in the commuter case). The standard deviation of the solution set of each iterative method is large while that of our approach is small. This indicates that our approach reduces the uncertainty in the solution sets largely.

8.5 Feedback and Discussion

The user in the health case was satisfied with the result of our approach. He said, "The box plot provides markers of important positions on the axis, which gives me guidance for filtering." He told us that the line charts allowed him to compare different distributions. Especially, he thought that the comparison between the orange and purple lines (the distributions after and before each filtering operation) was critical for knowing the change of the solution space caused by the filtering operation. He also mentioned that the level-of-detail parallel coordinates allowed him to view more samples in the preferred region.

The user in the commuter case thought the exploration provided him insights into the relations of departure time and transportation modes. He commented, "The boxplot and distribution chart are frequently used in the statistics field, so they are easy to follow." He suggested that the system should give some filtering recommendations during the exploration. He agreed that returning a set of plausible solutions was better than finding only one result, whose uncertainty is unknown.

After filtering, the solution space may be narrow and the number of samples in S_1 may be small. In this case, loop A (Fig. 3) can be used if the following two conditions are satisfied: 1. The solution space

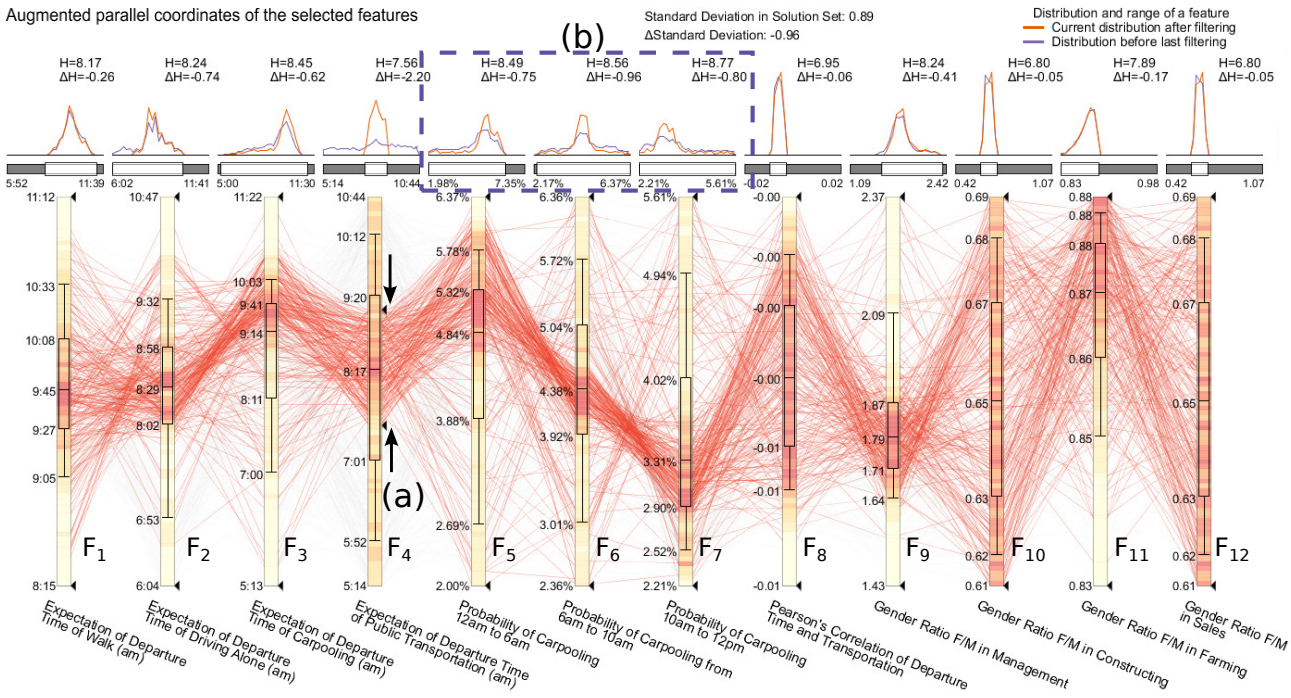

Fig. 7. (a) In the commuter case, the user set the expectation of departure time of public transportation (F_4) to around 7am to 9am (rush hours). This changed the pattern of the polylines across F_5 to F_7, which represented the probabilities of carpooling in different departure time periods. (b) As seen from the line charts in the dashed box, the distributions of F_5 to F_7 were skewed to different directions. The probability of carpooling in early mornings increased, and the chance of carpooling in late mornings decreased. So in the solution space, solutions with high probabilities of carpooling in earlier morning were more preferable. This case reflected the behavior of workers in the same family.

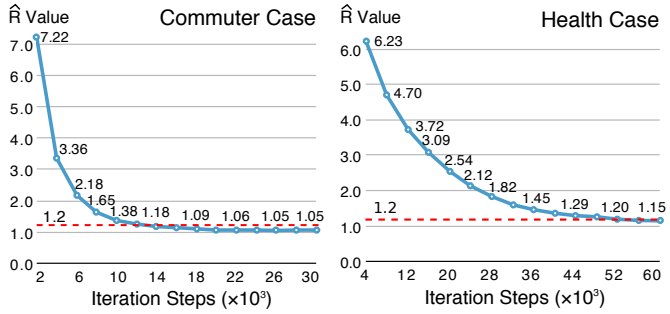

Fig. 8. $\hat{R}$ values of the HR chains at different iteration steps in the two cases. $\hat{R} < 1.2$ indicates convergence, which is around 52,000 iterations in the health case, and 14,000 in the commuter case, respectively.

Health Case			
Methods	Iterations	Standard Deviation	Marginal Error
ART	20	5.25×10^5	10^{-12}
MART	24	3.40×10^5	10^{-13}
SIRT	91	3.79×10^5	22.7
EM	90	3.70×10^5	27.4
This paper	-	7.30×10^3	10^{-13}
Commuter Case			
Methods	Iterations	Standard Deviation	Marginal Error
ART	12	3.15	10^{-17}
MART	14	2.85	10^{-17}
SIRT	39	2.60	6.31×10^{-4}
EM	40	2.20	2.02×10^{-4}
This paper	-	0.89	10^{-17}

Table 3. For our method and the traditional iterative methods, we test the average iterations, the average standard deviations in the result sets, and the average marginal error of the results. The high standard deviations of iterative methods indicate the uncertainties and joint errors of their solutions are high. (The sum of frequencies in the health case is 2,428,667, and the sum of probabilities in the commuter case is 1.0.)

remains convex (non-linear features may make the solution space non-convex). 2. The convergence iteration steps of HR sampling are less than a threshold (e.g., 1,000), which means sampling online is fast.

In some confidential datasets, such as patient records from a hospital, it is necessary to guarantee that marginal information would not reveal sensitive information. Given this situation, analysis of joint distribution reconstruction may also help to understand how to protect sensitive information. Our method may assist in testing whether the data can be reconstructed from marginal distributions.

9 Conclusion

We described a visual analytics approach to reconstruct joint distributions of data from multiple marginal distributions, allowing experts to inject domain knowledge to solve this otherwise ill-posed problem. Using a level-of-detail high dimensional visualization technique, the user can then trim down the solution set by adding constraints interactively to arrive a vastly reduced solution set.

While our approach is effective, there are a few aspects to be improved further. Firstly, other high dimensional visualization methods such as scatter plot or 2D projection could be added to visualize feature correlations and the preferred region in the solution space. Secondly, an effective method is needed to find if there is a model for a formal expression of the joint distribution. Finally, to test our method with actual data, we are applying for the complete NY health dataset.

Acknowledgments

This research was partially supported by NSF grant IIS 1527200 and the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the "IT Consilience Creative Program (ITCCP)" (NIPA-2013-

H0203-13-1001) supervised by NIPA.

REFERENCES

[1] Age and sex composition in the united states: 2009. http://www.census.gov/population/age/data/. 2009 (accessed March 30, 2016).

[2] Commuting in america 2013: the national report on commuting patterns and trends. http://traveltrends.transportation.org. 2013 (accessed March 30, 2016).

[3] Commuting (journey to work) dataset. http://www.census.gov/hhes/commuting/data/. 2009 (accessed March 30, 2016).

[4] Statewide planning and research cooperative system (sparcs). http://www.health.ny.gov/statistics/sparcs. 2016 (accessed March 30, 2016).

[5] United states census bureau. http://www.census.gov/. 2009 (accessed March 30, 2016).

[6] A. H. Andersen and A. C. Kak. Simultaneous algebraic reconstruction technique (sart): a superior implementation of the art algorithm. *Ultrasonic imaging*, 6(1):81–94, 1984.

[7] S. Baumert, S. Kiatsupaibul, A. Ghate, Y. Shen, R. L. Smith, and Z. B. Zabinsky. Discrete hit-and-run for sampling points from arbitrary distributions over subsets of integer hyper-rectangles. *Operations Research*.

[8] W. Berger, H. Piringer, P. Filzmoser, and E. Gröller. Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. In *Computer Graphics Forum*, volume 30, pages 911–920, 2011.

[9] Y.-H. Chan, C. D. Correa, and K.-L. Ma. The generalized sensitivity scatterplot. *Visualization and Computer Graphics, IEEE Transactions on*, 19(10):1768–1781, 2013.

[10] Y. P. Chaubey, F. Nebebe, and D. Sen. Estimation of joint distribution from marginal distributions. 2003.

[11] Y. Chen, I. Dinwoodie, and et al. Lattice points, contingency tables, and sampling. *Contemporary Mathematics*, 374:65–78, 2005.

[12] S. Cheng and K. Mueller. The data context map: Fusing data and attributes into a unified display. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):121–130, 2016.

[13] C. D. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. In *IEEE VAST*, pages 51–58, 2009.

[14] M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

[15] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[16] B.-y. Dong. Image reconstruction using em method in x-ray ct. In *Wavelet Analysis and Pattern Recognition, 2007. ICWAPR'07. International Conference on*, volume 1, pages 130–134. IEEE, 2007.

[17] M. Dyer, R. Kannan, and J. Mount. Sampling contingency tables. *Random Structures and Algorithms*, 10(4):487–506, 1997.

[18] A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical science*, pages 457–472, 1992.

[19] R. Gordon, R. Bender, and G. T. Herman. Algebraic reconstruction techniques (art) for three-dimensional electron microscopy and x-ray photography. *Journal of theoretical Biology*, 29(3):471–481, 1970.

[20] R. Kannan and S. Vempala. Sampling lattice points. In *Proceedings of the Twenty-ninth Annual ACM Symposium on Theory of Computing*, STOC '97, pages 696–700, New York, NY, USA, 1997. ACM.

[21] D. E. Kaufman and R. L. Smith. Direction choice for accelerated convergence in hit-and-run sampling. *Operations Research*, 46(1):84–95, 1998.

[22] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci. Visualizing high-dimensional data: Advances in the past decade. In *Proc. Eurographics Conf. Visualization*, pages 20151115–127, 2015.

[23] L. Lovász. Hit-and-run mixes fast. *Mathematical Programming*, 86(3):443–461, 1999.

[24] L. Lovász and S. Vempala. Hit-and-run from a corner. *SIAM J. Comput.*, 35(4):985–1005, Apr. 2006.

[25] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[26] D. S. Putler, K. Kalyanam, and J. S. Hodges. A bayesian approach for estimating target market potential with limited geodemographic information. *Journal of Marketing Research*, pages 134–149, 1996.

[27] S. Rado, R. Splechtna, K. Matkovic, M. Duras, E. Grller, and H. Hauser. Towards Quantitative Visual Analytics with Structured Brushing and Linked Statistics. *Computer Graphics Forum*, 2016.

[28] C. J. Romeo. Estimating discrete joint probability distributions for demographic characteristics at the store level given store level marginal distributions and a city-wide joint distribution. *Quantitative Marketing and Economics*, 3(1):71–93, 2005.

[29] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1421–1430, 2010.

[30] T. Schreck, T. Von Landesberger, and S. Bremm. Techniques for precision-based visual analysis of projected data. *Information Visualization*, 9(3):181–193, 2010.

[31] R. L. Smith. Efficient monte carlo procedures for generating points uniformly distributed over bounded regions. *Operations Research*, 32(6):1296–1308, 1984.

[32] C. A. Steed, J. E. Swan, T. Jankun-Kelly, P. J. Fitzpatrick, et al. Guided analysis of hurricane trends using statistical processes integrated with interactive parallel coordinates. In *IEEE VAST 2009*, pages 19–26, 2009.

[33] T. Torsney-Weir, A. Saad, and et al. Tuner: Principled parameter finding for image segmentation algorithms using visual response surface exploration. *IEEE TVCG*, 17(12):1892–1901, 2011.

[34] B. Wang, P. Ruchikachorn, and K. Mueller. Sketchpadn-d: Wydiwyg sculpting and editing in high-dimensional space. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2060–2069, 2013.

[35] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. Opinionseer: interactive visualization of hotel customer feedback. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1109–1118, 2010.