

# Visualizing Flow of Uncertainty through Analytical Processes

Yingcai Wu, *Member, IEEE*, Guo-Xun Yuan, and Kwan-Liu Ma, *Fellow, IEEE*

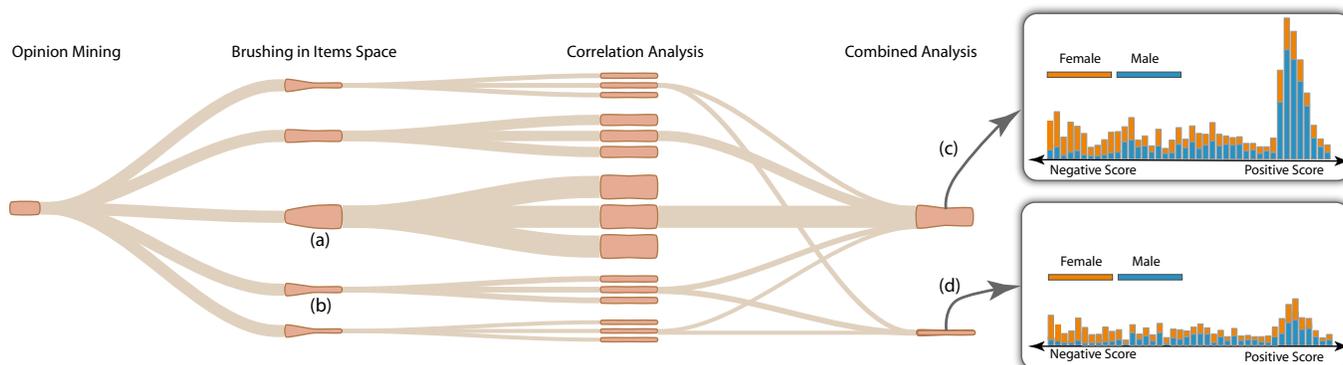


Fig. 1. Uncertainty flow showing variations of uncertainty along different analysis processes with four major steps: opinion mining, brushing in items space, correlation analysis, and combined analysis. Uncertainty arises during opinion mining, and increases, decreases, splits, or merges in subsequent steps. (a) and (b) show that uncertainty increases and decreases, respectively. (c) and (d) show two different results obtained by merging the results of correlation analysis. (c) reveals that female customers complain more while male customers complain less, whereas (d) only reveals that female customers complain more. Compared with (c), (d) is more reliable as it excludes a highly uncertain result of correlation analysis for the combined analysis.

**Abstract**—Uncertainty can arise in any stage of a visual analytics process, especially in data-intensive applications with a sequence of data transformations. Additionally, throughout the process of multidimensional, multivariate data analysis, uncertainty due to data transformation and integration may split, merge, increase, or decrease. This dynamic characteristic along with other features of uncertainty pose a great challenge to effective uncertainty-aware visualization. This paper presents a new framework for modeling uncertainty and characterizing the evolution of the uncertainty information through analytical processes. Based on the framework, we have designed a visual metaphor called uncertainty flow to visually and intuitively summarize how uncertainty information propagates over the whole analysis pipeline. Our system allows analysts to interact with and analyze the uncertainty information at different levels of detail. Three experiments were conducted to demonstrate the effectiveness and intuitiveness of our design.

**Index Terms**—Uncertainty visualization, uncertainty quantification, uncertainty propagation, error ellipsoids, uncertainty fusion.

## 1 INTRODUCTION

Uncertainty can arise in any stage of an analytical process, for applications as diverse as financial analysis, business intelligence, network monitoring, and geo-spatial data analysis [27, 35, 42]. For instance, the input data for analysis is often uncertain because of ineffective data sampling, inaccurate measurements, or corrupted data [28]. Data processing or transformation, which is usually required in data analysis to account for the growing scale and complexity of the data, is another source of uncertainty because of the lack of knowledge of the parameters used (i.e., the uncertainty of parameters) [21]. Even the analytical reasoning process has been reported to result in a certain degree of uncertainty information [46].

When uncertainty arises, properly informing analysts is very important. Otherwise, the reliability of the conclusions drawn based on the data will be questionable, which can lead to erroneous decisions and undesired consequences [21, 27]. The amount of uncertainty should also be faithfully revealed to analysts because the manner in which the uncertainty is expressed can significantly affect decision making [11]. Underestimated uncertainty can mislead analysts, making them over-

confident of the results. Overestimated uncertainty, on the other hand, may cause analysts to abandon valuable findings that are actually trustworthy. Thus, effectively and faithfully showing uncertainty is demanded in data analysis processes whenever uncertainty is present.

In visual analytics where data is usually analyzed in an explorative and iterative manner [23], uncertainty becomes extremely complicated, thus hindering the fulfillment of the effectiveness and faithfulness requirement. Given that data is transformed, uncertainty does not remain the same, but is rather dynamic and can change through the entire analysis process [9], in which the uncertainty may increase, decrease, split, or merge. For example, uncertainty in a data set may change along with the subsequent data transformation and processing steps, where the amount of uncertainty may increase or decrease. For high-dimensional data analysis, analyzing the data a few dimensions at a time and then switching to other dimensions is commonly practiced. Uncertainty in this scenario splits with the dimensions. Analysts may combine some of the findings discovered using different data dimensions to make decisions, where the uncertainty associated with the findings are merged. The complex and dynamic characteristics of the uncertainty hereby pose a great challenge in tracking and managing uncertainty effectively to make better decisions.

Researchers in visual analytics have developed various methods [12, 27, 28] to visually represent uncertainty information. Nevertheless, existing methods present uncertainty to analysts for only a particular stage of the data analysis process. Users may not be able to obtain a clear overview of how the uncertainty originates and spreads through the process, especially for complicated analysis with a num-

• Yingcai Wu, Guo-Xun Yuan, and Kwan-Liu Ma are with the University of California, Davis. E-mail: {ycwu, gxyuan, klma}@ucdavis.edu.

Manuscript received 31 March 2012; accepted 1 August 2012; posted online 14 October 2012; mailed on 5 October 2012.

For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

ber of transformations. The ability to visualize the flow of uncertainty through the entire analysis is still absent from the field. In addition, current methods often focus on visualizing one-dimensional uncertainty information. In a practical multivariate data analysis problem, however, uncertainty is largely multidimensional as data variables are not always independent of one another. For instance, sampling multidimensional data can produce multidimensional uncertainty, which is difficult to visualize using existing methods. Therefore, with the growing popularity of multivariate data analysis, an increasing need for developing new techniques for visualizing the flow of multidimensional uncertainty information through the visual analytics process exists.

This work focuses on the characterization and visualization of the flow of uncertainty information through visual analytics processes. Specifically, we propose a new framework to model uncertainty using error ellipsoids and to characterize the dynamic changes of the uncertainty information including uncertainty integration, transformation, and propagation through the processes. The error ellipsoids can be regarded as a multidimensional generalization of standard deviations, which allows for better understanding and more intuitive characterization of uncertainty information. This work enables multi-level uncertainty visualization accomplished through a set of uncertainty analysis techniques. A new flow-style visual metaphor is designed at the overview level to intuitively visualize the flow of uncertainty in complex analytical processes. Guided by the overview, users are allowed to interact with the uncertainty flow to analyze multidimensional uncertainty information in detail using a matrix visualization of the projections of the ellipsoids. Users can even drill down to individual data items to see more detailed uncertainty information.

The contributions of this work are as follows:

- A framework for describing uncertainty transformation and propagation based on error ellipsoids is created.
- A visual metaphor to visually summarize the flow of uncertainty over an entire visual analytics process is introduced.
- A visualization system to show the uncertainty information depicted by error ellipsoids at different levels of detail is built.

## 2 RELATED WORK

The growing scale and complexity of data pose a great challenge for analysts in discovering interesting patterns directly from its raw form [9]. Thus, the transformation or simplification of data is usually needed before such patterns can be shown [23]. Card et al. [7] presented a well-accepted model suggesting that information visualization is essentially a series of transformations. Other models [3, 23] have also been proposed to include automated data analysis algorithms. These transformations may create uncertainty information [21, 28] that may spread throughout the entire visualization process [9].

Effectively showing uncertainty in data visualization can improve trustworthiness and avoid misleading analysis [11, 27]. Pang et al. [28] presented a comprehensive survey of traditional uncertainty visualization techniques. A typology for visualizing uncertainty in intelligence analysis was proposed by Thomson et al. [38]. Zuk and Carpendale [46] extended this typology by including the uncertainty of reasoning. Researchers have proposed a variety of new uncertainty visualization methods such as ambiguation [27] and summary plots [29]. The evaluation of different techniques has received considerable attention recently [4, 11, 32, 45]. Sanyai et al [32] compared four methods, such as traditional errorbars, and found that the efficiency of these methods was highly dependent on the tasks performed. Deitrick and Edsall [11] conducted an empirical evaluation of the influence of uncertainty visualization on decision making. Their results reveal that the degree of the influence is affected by the manner in which the uncertainty is expressed. Zuk and Carpendale [45] presented an interesting analysis of uncertainty visualization based on three well-established theories. Bisantz et al. [4] compared the effects of displaying uncertainty using numeric and graphical representations of uncertainty.

Recent studies mostly focus on visual representations of uncertainty. Wu et al. [42] designed a circular wheel representation to show the uncertainty of extracted customer opinions. Slingsby et al. [35] used simple interactive graphics such as bar charts to visualize uncer-

tainty in geodemographics data. In contrast, our work aims to characterize and visualize the flow of uncertainty through the analytics pipeline. To our knowledge, this area has not been previously studied. Feng et al. [12] used density plots to visualize multidimensional uncertainty that is uncorrelated. Correa et al. [9] used covariance matrices to represent uncertainty. However, the process for visualizing these covariance matrices intuitively and effectively remains unclear. This work introduces a new framework to characterize multidimensional uncertainty information using error ellipsoids.

A simple definition of uncertainty does not exist. Skeels et al. [34] presented a comprehensive classification of uncertainty for information visualization. Hunter and Goodchild define uncertainty as the degree of the lack of knowledge about the amount of error [19]. Uncertainty can also be characterized by accuracy, reliability, precision, and consistency in literature [28]. In metrology, uncertainty is a parameter that depicts the dispersion of the measured values [22, 37]. Meteorologists generally classify uncertainty sources as either random or systematic uncertainty based on the method used for evaluation. Both types of uncertainty are represented by an estimated standard deviation, termed *standard uncertainty*. This uncertainty definition has been well accepted in many different disciplines. This work uses error ellipsoids, a generalization of standard deviation in multidimensional space to model uncertainty information.

Uncertainty quantification is the quantitative characterization of the effect of uncertainty information on system outcome. Roy and Oberkampf [30] described an uncertainty framework for uncertainty modeling and quantification in scientific computing. Perturbation methods [9], moment analysis [25], and operator-based methods [44] are three widely used non-sampling techniques. However, these techniques can only be applied to systems with small uncertainties [43]. Monte Carlo Sampling (MCS) is a sampling-based technique for uncertainty quantification [13, 31]. Compared with other non-sampling methods, MCS is more general and can be used to estimate the uncertainty of complex systems where other methods fail. Traditional MCS has been improved to achieve faster convergence using a variety of techniques such as Latin supercube sampling [17] and quasi Monte-Carlo [26] at the cost of additional restrictions posed by these methods [43]. This work employs traditional MCS technique to quantify the uncertainty information in a visual analytics pipeline.

## 3 UNCERTAINTY IN VISUAL ANALYTICS

This section describes a framework for characterizing the flow of uncertainty information in a visual analytics process.

### 3.1 Uncertainty Framework

In a non-trivial analytical process the development of uncertainty is complex. For instance, new uncertainty from data transformations could be integrated into the process, and the amount of uncertainty could increase or decrease through the process. Additionally, the uncertainty could be split into different pieces as the data is divided into different parts or could be merged when the separated parts are combined for an overall analysis. Moreover, multivariate data analysis can yield complicated multidimensional uncertainty information. To facilitate uncertainty analysis and visualization, we have developed a new framework to depict the dynamic change of the uncertainty. This framework was inspired by Correa et al.'s uncertainty framework [9], which shows how uncertainty is modeled, propagated, and aggregated in the analysis process. The previous framework has three elements:

- **Uncertainty Modeling** models uncertainty using a set of covariance matrices. However, the abstract, theoretical, covariance matrices lack an intuitive visual representation that conveys the meaning of uncertainty. Additionally, these matrices do not support the description of the complicated dynamic variation of the uncertainty information in analytical processes.
- **Uncertainty Propagation** suggests that the uncertainty is propagated as data is transformed. A Taylor series method is used to evaluate the amount of uncertainty propagated by the transformation. The method, however, can only be used for a data transformation with a small amount of uncertainty. Furthermore,

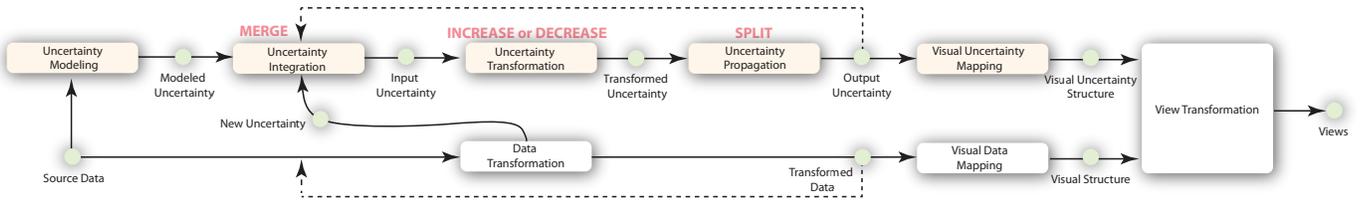


Fig. 2. Uncertainty framework for characterizing the variation of uncertainty information in a visual analytics process. The bottom shows a general information visualization process. The top part illustrates the uncertainty framework parallel to the bottom process, thus depicting how the uncertainty of the process is modeled, integrated, transformed, propagated, and visualized along with the visualization.

although the process suggests that the visual analytics process is often a complex network of transformations, the associated uncertainty network has not been characterized by the framework.

- **Uncertainty Aggregation** shows that data transformation itself can introduce a certain amount of uncertainty information, such as loss of information, to the analytical process. This uncertainty is aggregated via its addition to the uncertainty propagated by the same transformation. The linear aggregation method described in [9], however, fails to aggregate multidimensional uncertainty information because covariance matrices cannot be linearly combined, especially when their sizes differ.

We extend this framework to better characterize variations of multidimensional uncertainty in analysis processes. Figure 2 shows that our new framework (top) parallels a general visualization pipeline (bottom). The new framework includes five main elements.

### 3.1.1 Uncertainty Modeling

Uncertainty modeling mathematically defines uncertainty information (Section 4.1). This method serves as the basis for all other elements of the framework because such elements primarily address the modeled uncertainty. Uncertainty should be effectively and intuitively modeled to facilitate the characterization and visualization of the variation of the uncertainty. For the sake of simplicity, our framework uses the widely adopted Gaussian model, where uncertainty is modeled by a covariance matrix. The covariance matrix can be further characterized and represented using an error ellipsoid, which allows for intuitive, multi-level visualization of the uncertainty. More importantly, the error ellipsoid can better capture the dynamic variation of uncertainty.

### 3.1.2 Uncertainty Integration

Uncertainty integration characterizes how the modeled uncertainty can be merged together. This process occurs whenever different pieces of uncertainty information are combined to account for all uncertainty information. For instance, an analyst may apply different transformations to the same data and then combine the results for further analysis. The uncertainty introduced or propagated by different transformations must be merged to determine the uncertainty of the current system. Unlike the uncertainty aggregation of the previous framework, uncertainty integration allows for the true fusing of multidimensional uncertainty information of different data dimensions or different data portions. Details can be found in Section 4.4.

### 3.1.3 Uncertainty Transformation

Uncertainty transformation is responsible for the quantification of the transformed uncertainty information, given the input uncertainty from uncertainty integration. This method is similar to the uncertainty propagation of the previous framework. However, uncertainty transformation employs MCS to quantify the extent of uncertainty to account for a large amount of uncertainty. Depending on the transformation used, the quantified uncertainty information may either increase or decrease. The transformed uncertainty may be visualized via uncertainty visualization or may be propagated via uncertainty propagation to subsequent data transformations (See Figure 2). Section 4.2 describes the uncertainty transformation in greater detail.

### 3.1.4 Uncertainty Propagation

Uncertainty propagation in this work is different from that of the previous framework. This method determines how the transformed uncertainty is passed on to the subsequent data transformations, where the transformed uncertainty may either be separated into pieces or may remain unchanged. For example, when different transformations are applied to different data portions, the transformed uncertainty will be split accordingly. The propagated uncertainty (output uncertainty) may be further merged by the uncertainty integration in the subsequent data transformation with other propagated uncertainty information or with newly introduced uncertainty from data transformations. Section 4.4 presents how the propagated uncertainty information is derived.

### 3.1.5 Visual Uncertainty Mapping

Visual uncertainty mapping visualizes the flow of uncertainty through the analytics processes, where uncertainty may split, merge, shrink, or expand. We design a new flow-style visual metaphor, uncertainty flow, to visually represent the overall evolution of uncertainty information. Section 5 presents the detailed visual design.

## 3.2 System Overview

Figure 3 shows the overview of our approach. It runs parallel to a visual analytics system and maintains a history graph to record the history of all data transformations. When an uncertainty visualization is requested by users, the approach quantifies the uncertainty (modeled as a multidimensional error ellipsoid) of every data item using an MCS technique. All three types of uncertainty (including integrated, transformed, and propagated uncertainty) are measured. The method subsequently computes an overall error ellipsoid of every type of uncertainty for the data set by fusing the related error ellipsoids of all data items. The overall uncertainty magnitude is finally derived from the overall error ellipsoid. Based on this information, we can draw a visualization of the flow of uncertainty in the entire visual analytics. With the framework, users are allowed to interact with the uncertainty flow to see the uncertainty at different levels of detail.

## 4 UNCERTAINTY ANALYSIS

This section presents a set of techniques supporting uncertainty modeling, transformation, propagation, and integration.

### 4.1 Uncertainty Modeling

For multivariate analysis tasks, uncertainty information introduced by a typical data transformation is often multidimensional. Using standard deviation to model the uncertainty for each dimension is ineffective as the data dimensions are often correlated. In statistics, the uncertainty of multidimensional information is usually represented by a covariance matrix, which is a generalization of standard deviation [36].

Assume we have a  $k$ -dimensional data item of interest  $X$ . We consider the quantity as a normally-distributed random vector

$$X \equiv [X_1, X_2, \dots, X_k] \sim \mathcal{N}(\mu, \mathbf{C}) \quad (1)$$

where  $\mathcal{N}(\cdot, \cdot)$  denotes an  $n$ -dimensional normal distribution,  $\mu \in \mathbf{R}^k$  is the mean vector, and  $\mathbf{C} \in \mathbf{R}^{k \times k}$  is the covariance matrix of  $X$  which is regarded as the uncertainty of  $\mu$ .

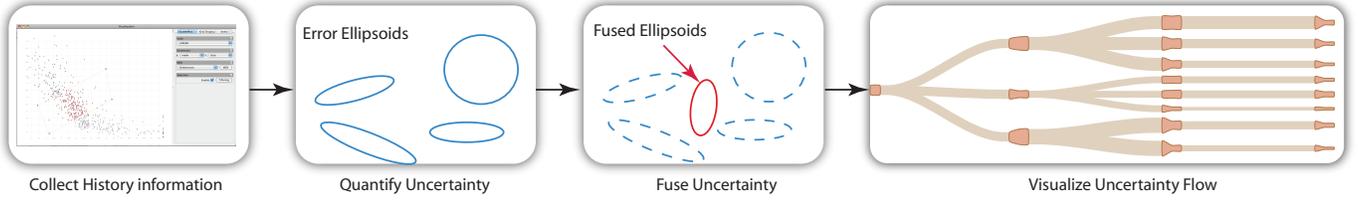


Fig. 3. System overview: the system automatically collects history information and estimates the error ellipsoids of every data item, which are subsequently combined to obtain an overall uncertainty level for drawing the uncertainty flow visualization.

Various methods, such as Corrgrams [14], have been proposed to directly visualize the covariance matrix  $\mathbf{C}$ . However, these direct visual representations cannot be easily understood because they do not intuitively show the dispersion of the values of  $X$ . Similarly difficult is the direct derivation of an overall uncertainty degree of  $X$  from  $\mathbf{C}$ , which is requested to provide an overview. Thus, we use an ellipsoid to depict the covariance matrix  $\mathbf{C}$  to address these issues [5], such that the uncertainty of  $X$  can be intuitively represented by the extent of the ellipsoid. Centered at  $\mu$ , the ellipsoid is defined by the equation

$$(x - \mu)^T \mathbf{C}^{-1} (x - \mu) \leq 1 \quad (2)$$

The eigenvectors and eigenvalues of  $\mathbf{C}$  define the principal directions and inverse squares of the semi-axes of ellipsoid  $E$ , respectively. A *standard error ellipsoid* is obtained, which can describe the uncertainty of  $X$  in a geometric manner. However, providing an overall picture is difficult. Instead, the ellipsoid volume can represent the overall uncertainty level of  $X$ . Volume  $U(E)$  can be computed by the equation

$$U(E) \equiv \frac{\pi^{\frac{k}{2}}}{\Gamma(\frac{k}{2} + 1)} \prod_{i=1}^k \sqrt{\xi_i}, \quad (3)$$

where  $E$  is an ellipsoid defined by (4),  $\xi_i \forall i$  denotes the eigenvalues of the covariance matrix  $\mathbf{C} \in \mathbf{R}^{k \times k}$ , and  $\Gamma(\cdot)$  is the Gamma function.

Our work is based on the error ellipsoid for quantifying the variation of uncertainty information in visual analytics. The intuitive geometrical features, as well as the capacity of the error ellipsoid to be fused and split, allow us to design an effective and interactive level-of-detail visualization of uncertainty in visual analytics.

## 4.2 Uncertainty Quantification

Uncertainty quantification refers to the quantitative estimation of the uncertainty of each data item  $X$ , which is the error ellipsoid of  $X$ . We employ a technique based on traditional MCS to evaluate the uncertainty that results from data transformation. First, the technique produces a set of independent samples of system inputs, then repeatedly runs the system for each sample, and finally collects all the system outputs. Based on this collection, the statistical uncertainty information (i.e., covariance matrices) of the system output can be estimated. With the covariance matrices, we can obtain the associated error ellipsoids.

Although the standard error ellipsoid is more powerful for characterizing uncertainty variation, it is also more complex and, without appropriate modification, can sometimes mislead analysts. A standard error ellipsoid of  $n$  dimensions derived from Equation (2) implies a totally different confidence level from that of  $m$  dimensions when  $m \neq n$ . For example, a 2D standard error ellipsoid indicates a mere 40% confidence level, whereas a 1D standard deviation represents a confidence level of 60%. Thus, we propose that a user be allowed to specify a desired confidence level and then create the error ellipsoids accordingly. More specifically, we scale the standard error ellipsoids based on the specified confidence to ensure the consistent meaning of the error ellipsoids of different dimensions.

Through a sampling process, we have a sample mean  $\bar{X} \equiv \frac{1}{n} \sum_{i=1}^n X_i$ , which follows a normal distribution. We define  $Z \equiv n(\bar{X} - \mu)^T \mathbf{C}^{-1} (\bar{X} - \mu)$  to remove the dependence of the distribution  $\bar{X}$  on the population parameters  $\mu$  and  $\mathbf{C}$ . In statistics,  $Z$  is a  $\chi^2$  random variable

with  $n$  degrees of freedom [1]. A value of  $Z$  corresponds to the contour of an ellipsoid with respect to a given confidence level  $CL \in [0, 1]$ . The equation of an ellipsoid has the form [5]

$$(x - \check{\mu})^T \check{\mathbf{C}}^{-1} (x - \check{\mu}) = \frac{z_{CL}}{n}, \quad (4)$$

where  $z_{CL}$  is an observation of  $Z$  with respect to the confidence level  $CL$ ,  $\check{\mu}$  is a *sample* mean vector, and  $\check{\mathbf{C}}$  is a *sample* covariance matrix obtained by a sampling process of  $X$ . The ellipsoid center is the vector  $\check{\mu}$ . Moreover, other geometric parameters of the ellipsoid, such as the orientation and magnitude of its semi-axes, can be determined by the eigenvectors and eigenvalues of the matrix  $\check{\mathbf{C}}$ . For example, we have a 3D ellipsoid defined by (4). Suppose that  $\xi_1$ ,  $\xi_2$ , and  $\xi_3$  are the eigenvalues of a sample covariance  $\check{\mathbf{C}}$  and  $a$ ,  $b$ , and  $c$  are the magnitude of its three axes. Then, they have the relationship as follows:

$$a = \sqrt{\frac{z_{CL} \xi_1}{n}}, \quad b = \sqrt{\frac{z_{CL} \xi_2}{n}}, \quad \text{and} \quad c = \sqrt{\frac{z_{CL} \xi_3}{n}}. \quad (5)$$

Given a confidence level  $CL$  specified by a user, we compute the quantile of the  $f_{\chi^2}^2(n)$  density function to determine the axis scaling factor  $z_{CL}$ . The computation involves solving an integral equation.

$$CL = \int_0^{z_{CL}} \frac{1}{2^{\frac{n}{2}} + \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} dx. \quad (6)$$

No analytical form exists. Thus, a numerical approximation is necessary. We illustrate a two-dimensional example in Figure 4. Given the estimated mean  $\check{\mu} = [0, 0]^T$  and covariance matrix  $\check{\mathbf{C}} = [v_1, v_2]$  where  $v_1 = [1, 0.2]^T$  and  $v_2 = [0.2, 1]^T$ , the corresponding error ellipse  $E$  centered at  $\check{\mu}$  has two semi-axes of length  $\sqrt{\xi_X} = 0.9$  and  $\sqrt{\xi_Y} = 1.1$ , where  $\xi_X$  and  $\xi_Y$  are the eigenvalues of  $\check{\mathbf{C}}$ . We then seek an error ellipse with a 95% confidence level. As described in Eq. (6), we compute the quantile  $z_{CL}$  of the  $\chi^2$  density function with 2 degrees of freedom, such that the area under the curve is 0.95 (as shown in Figure 4(a)). Finally, by scaling the two axis length of  $E$  with  $\sqrt{z_{CL}}$  as in Eq. (5), we can obtain the scaled error ellipse shown in Figure 4(b).

## 4.3 Uncertainty Combination

With the uncertainty quantification method, we can determine the error ellipsoid for every data item as the data is transformed. However, a means for deriving an error ellipsoid characterizing the overall uncertainty of the entire data set still needs to be determined. The overall uncertainty is needed to visualize the overall trend of uncertainty. However, determining the overall uncertainty is difficult considering the complexity of the uncertainty information. The linear combination used in [9] fails to combine multidimensional uncertainty information, which has dimensions that are not always independent of one another.

Estimating the overall uncertainty can be regarded as a process of finding an improved estimate based on previous multiple estimates of the same quantity  $\mu$  of interest, which is a well-studied problem in the field of information fusion [10]. Therefore, combining uncertainty to obtain its overall value by applying an optimal weighting scheme to achieve a minimum total error is reasonable. In particular, suppose that we conduct  $r$  independent sampling. We can then form  $r$  error

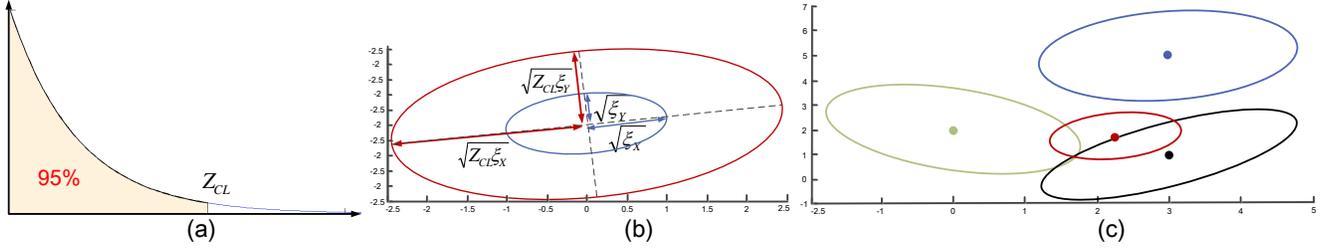


Fig. 4. (a) shows the quantile of the cumulative probability 0.95 in the  $\chi^2$  distribution with degree 2; (b) shows the scaled error ellipse corresponding to 0.95 confidence level; (c) describes the combined result (bold red) from the given three error ellipses (black, blue, and green).

ellipsoids using the procedure described in Section 4.2. By defining an improved estimate as the weighted sum of previous  $r$  estimates  $\bar{X}_W = \sum_{t=1}^r W_t \bar{X}_t$ . We can obtain the covariance of the weighted estimate  $\bar{X}_W$

$$C_{W,ij} = \sum_{t=1}^r (W_t C_t W_t^T)_{ij},$$

where  $W_t \in \mathbf{R}^{k \times k} \forall t$  are the weight matrices in the scheme. As previously described, the ellipsoid volume can be used as an uncertainty measure. We cast our goal as an optimization problem.

$$\begin{aligned} \min_{W_a \forall a} \quad & \det(C_W) = \det\left(\sum_{t=1}^r (W_t C_t W_t^T)_{ij}\right), \\ \text{st.} \quad & \delta_{ij} = \sum_{t=1}^r W_{t,ij}. \end{aligned} \quad (7)$$

with the solution of

$$W_a = \left(\sum_{t=1}^r C_t^{-1}\right) C_a^{-1}. \quad (8)$$

Finally, we can show the combined estimate  $\bar{X} = C_W \sum_{t=1}^r C_t^{-1} \bar{X}_t$  and its corresponding covariance matrix  $C_W = \left(\sum_{t=1}^r C_t^{-1}\right)^{-1}$ .

Figure 4(c) shows an example of combining three 2D estimated error ellipses in blue, green, and black. By solving problem (7), we can determine the optimal weights in Eq. (8) and finally we obtain the resulting error ellipse with minimum uncertainty (shown in red).

#### 4.4 Methods for Uncertainty Integration and Propagation

In visual analytics, data may be divided into a number of portions that are then transformed separately. The propagated uncertainty information of the data is similarly split into pieces. Two scenarios can occur depending on the data variables (dimensions).

- The data items include the original variables. In this case, the error ellipsoid of each data point remains the same.
- The data items include only a selected number of variables. The error ellipsoid of each data item in a data portion changes. The new ellipsoid can be obtained by simply projecting the original one on the space of the selected dimensions.

Notably, the overall ellipsoid for each data portion should be approximated by the uncertainty combination method (Section 4.3).

Uncertainty may also be merged whenever uncertainty integration occurs, such as when the propagated uncertainty merges with the newly added uncertainty of the data transformation. Another example is when the divided data portions are combined for further analysis. Two scenarios can occur based on whether the data items contain the same or different sets of variables.

- The data items include the same set of variables. The error ellipsoids of each data item are combined by the uncertainty combination method (Section 4.3).
- The data items include different sets of variables. We use the sampling-based uncertainty quantification method (Section 4.2) to re-approximate the error ellipsoids of each data item.

Notably, the overall ellipsoid for the merged data should also be computed using the uncertainty combination method (Section 4.3).

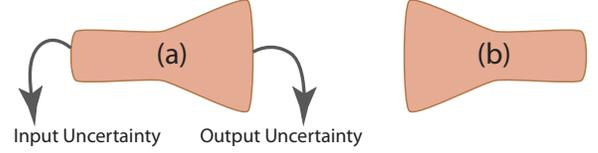


Fig. 5. (a) and (b) visually represent the analysis steps where uncertainty decreases and increases, respectively.

## 5 MULTI-LEVEL UNCERTAINTY VISUALIZATION

This section describes an interactive, multi-level visualization of the multidimensional uncertainty flow through a visual analytics process. The process follows the well-known visual information seeking mantra [33] and allows for level-of-detail exploration. At the overview level, we design a flow-style visual metaphor, uncertainty flow, to visually depict the overall uncertainty evolution in the analytical process. A user can interact with the uncertainty flow and choose to examine the uncertainty of a certain transformation. A matrix visualization shows the overall error ellipsoid of the transformation. The user is also allowed to drill down and see the single ellipsoid of a data point, which can be further reduced to a scalar value indicating the overall uncertainty level of the data point. This allows the use of existing visualization techniques to show the uncertainty on a visualization. Figure 6 illustrates the overall picture of the multi-level visualization.

### 5.1 Visualization of Uncertainty Flow

Visual analytics is often an explorative and iterative process with multiple cycles of analytical reasoning [40]. Uncertainty may arise, change, and spread over the complex visual analytics process. We must be able to show the uncertainty in the context of the visual analytics process to better understand, track, and visualize the flow of uncertainty. Thus, we should capture and model the analytics process. This section first briefly describes a tool called *history graph* to record and track the history of data transformations, followed by the detailed description of visual uncertainty flow.

#### 5.1.1 History Graph of Data Transformations

We utilize a tree-like graph to record the full history of the data transformations. Compared with the history tree, a widely-used provenance tool based on a tree data structure [6, 16, 24], our history graph allows a node to have multiple parents to account for uncertainty integration. We use roots to represent the input data sets under analysis. Other nodes represent the transformation operations, such as clustering performed in the analysis. An edge indicates the order of two transformations of the analysis process. The root-to-leaf paths of the graph represent all alternative analysis paths that have so far been explored. The system naturally supports branching historical timelines of transformations. A new branch is added to a node when a user steps back to the previous stage. In contrast to other tools, the history graph used in this system also records the temporal variation of uncertainty. As we described in our uncertainty framework, uncertainty changes as data is transformed. Storing these changes in the corresponding graph

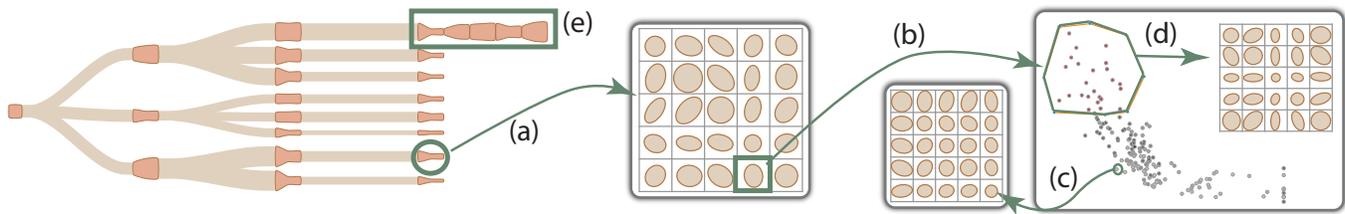


Fig. 6. (a) Selecting a node shows a matrix visualization of the overall error ellipsoid; (b) Selecting a cell of the matrix shows a scatterplot of the related variables to visualize the uncertainty degree of each data point; (c) Selecting a data point shows a matrix visualization of the error ellipsoid; (d) Selecting a group of data points shows a matrix visualization of their overall error ellipsoid; and (e) Combined nodes.

nodes is reasonable. Whenever any data transformation occurs, a parallel thread automatically records the transformation in the graph. The error ellipsoids of data points corresponding to input and output uncertainty are also computed and stored in the node.

### 5.1.2 Visual Metaphor for Uncertainty Flow

We design a new visual metaphor, *uncertainty flow*, to visualize the overall uncertainty evolution through the visual analytics process. Uncertainty flow is the core component of the visualization system, providing an intuitive visual summary of the provenance of uncertainty information in the process, which helps analysts develop a growing understanding of how and from where the uncertainty was introduced, integrated, transformed, and propagated. Furthermore, the recorded provenance of uncertainty may allow analysts to determine how and where to optimize the analysis process to reduce the uncertainty of the insight and conclusions. Uncertainty flow can also be used as a usual history mechanism for insight provenance and supports backtracking and alternative analysis paths.

**Preliminary Layout** The uncertainty flow is a graph drawn upon the history graph. We first generate a preliminary node-link layout of the graph and then decorate the layout using the uncertainty information. The layout is oriented from left to right (other orientations, such as top-down, are also allowed). The root representing the first data transformation is placed at the current view origin. All the other nodes representing data transformations are added to the view, starting from the roots, based on their hierarchical relations. The horizontal distance between a parent and its children, as well as that between two sibling nodes, are fixed but can be interactively assigned by users. The left-to-right ordering (edges) of the nodes represents the sequential order of the associated data transformations. The vertical ordering of the sibling branches is dynamically adjusted, such that the central branch is always the current exploration path, and all other branches are alternatively placed above or below this branch. The distance of a branch to the central branch encode the time when the branch was created. The closer a branch lies to the center, the more recently it was created. This process can ensure that the most recently explored paths always remain in the center focus region.

**Visual Uncertainty Encoding** We next draw the uncertainty on the preliminary layout. The uncertainty framework essentially defines two types of uncertainty information associated with a typical data transformation: input and output uncertainty. Effective visualization of uncertainty flow requires visual encoding in a faithful, concise, and intuitive manner on the preliminary layout.

Figure 5 shows our visual mapping of the uncertainty of a data transformation on the related node. We use a visual metaphor of a loudspeaker to intuitively encode the amplification of uncertainty information (Figure 5(a)). The uncertainty reduction, on the other hand, can also be intuitively encoded by an inverse loudspeaker (see Figure 5(b)). We used this visual encoding for the following reasons. First, the shape of a loudspeaker can intuitively convey the concept of “amplification” or “reduction” of the overall uncertainty. Second, the visual metaphor can facilitate a side-by-side visual comparison of input and output uncertainty information. Finally, using this visual metaphor allows us to easily combine a sequence of nodes,

creating a more compact layout (see Figure 6(e)).

The edge between two nodes represents the sequential order of the uncertainty distribution. We use the width of the edge to visually encode the overall uncertainty propagated from one node to another. The two nodes can be seamlessly connected by the edge, given that the width of the edge is identical to the length of the output base of one node as well as that of the input base of the other node.

### 5.1.3 User Interactions

Apart from general interactions (such as panning and zooming), uncertainty flow supports a variety of other interactions, such as the following:

- **Hover** When a user places the mouse over a node, a pop-up view displays the screenshot of the related visualization near the node.
- **Select** Double-click selection restores the visual analytics process to a certain step represented by the node. A new branch will be added to the node when the user continues to process the data. Single-click selection shows the matrix visualization of the uncertainty (see Figure 6(b)). Depending on the position clicked within the node, the matrix visualization of the input or output uncertainty is shown.
- **Oriente** The default orientation of the uncertainty flow is left-to-right. A user can change the orientation to top-down, bottom-up, or right-to-left. An animation will be created to show the smooth transition when the orientation changes.
- **Focus+Context** A selected node is viewed as the focus node. We consider the neighbors around the selected node in the graph as focus nodes. All ancestors of the selected node are also regarded as foci. We use the fisheye function [15] to compute the degree of interest (DOI) of the nodes. Based on the DOI of the nodes, we can combine the nodes with low DOI values (see Figure 6(e)), thus allowing for focus+context visualization.

## 5.2 Matrix Visualization of Error Ellipsoids

With the uncertainty flow, a user can quickly determine the overall degrees of the input, transformed, and output uncertainty information of a node. However, users may also want to see more details of the overall error ellipsoids from which the overall uncertainty degrees are derived. The error ellipsoids are often high-dimensional geometric objects and there is no method for drawing them directly. Thus, we design a matrix visualization showing the 2D projections of the ellipsoid (i.e., 2D ellipse) in the space of each pair of data variables to help the user visualize an ellipsoid. The user can then interactively select one or multiple cells on the matrix. Scatterplots or a parallel coordinate plot of the variables (represented by the selected cells) will then be presented to the user to reveal the uncertainty information of each data item (see Figure 6(b)).

## 5.3 Uncertainty Encoding on Visualizations

Our system can visualize not only the overall uncertainty information of the whole data set (by the uncertainty flow and matrix visualization), but also the uncertainty information of individual data items. Every data item has an error ellipsoid. Thus, we can similarly compute their

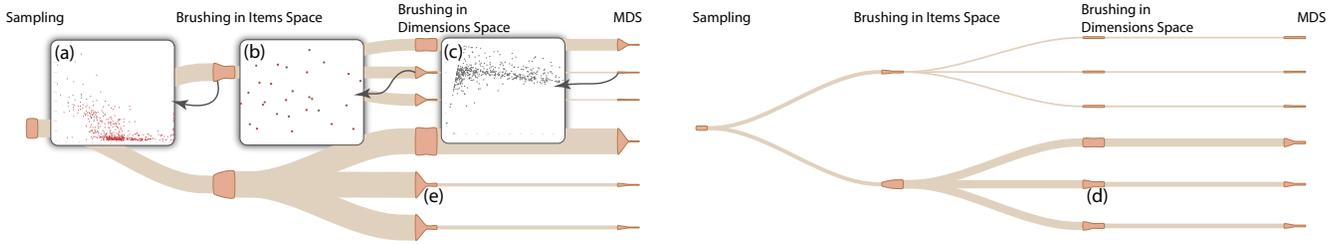


Fig. 7. Uncertainty flows (left and right) based on the scaled (with a 95% confidence level) and unscaled standard error ellipsoids, respectively. (a)-(c) show three dual analysis steps: brushing in items space, brushing in dimensions space, and reducing dimensions using MDS. The visualization (right) shows much less uncertainty than (left), which could mislead an analyst to become overconfident. Additionally, the unscaled uncertainty with different dimensions actually imply different confidence levels, which may confuse analysts. The input uncertainty with 14 dimensions and the output uncertainty with 6 dimensions in (d) imply 0.0001% and 1.4385% confidence levels, respectively. In contrast, they consistently represent the same confidence level (95%) in (d).

overall uncertainty degree, which can be visually encoded using existing uncertainty visualization methods. For example, we can use the opacity values of the glyphs (representing the data items) for a scatterplot to visually encode the overall uncertainty levels (see Figure 6(b)). When the user clicks a glyph on the scatterplot, the system can also create a pop-up view showing the matrix visualization of the related error ellipsoid (see Figure 6(c)). More importantly, the user can select a group of data items and combine their error ellipsoids to obtain an overall error ellipsoid for the group of data items, which can likewise be visualized using the matrix visualization (see 6(d)).

## 6 EXPERIMENTS AND DISCUSSION

In this section, we demonstrate the effectiveness and usefulness of our uncertainty framework and visualization techniques based on three experiments. Our techniques were implemented in Java and were tested on an Apple Macbook Pro with Intel Core i7 2.66GHz CPU and 4GB RAM. It took minutes to quantify the uncertainty information. Visual uncertainty flow can be created immediately when the uncertainty analysis is completed. We set the confidence level 95% for all results except the right uncertainty flow in Figure 7.

### 6.1 Multivariate Data Analysis

The first experiment is used to demonstrate the use of uncertainty flow in a multivariate data analysis process called dual analysis [39]. Dual analysis allows the analysts to explore data iteratively in two linked spaces, namely, items space and dimensions space. We use the “Boston Neighborhood Housing Prices” dataset with 506 data items and 14 dimensions for analysis. The purpose of the analysis is to detect correlations among the data dimensions.

In many practical applications, data is usually simplified to conserve computer resources, which may introduce uncertainty [9, 27]. Following the practice, an analyst starts the dual analysis process by randomly sampling the data. A sample contains only 20% of the original data items, thus introducing uncertainty. The uncertainty of each sampled data item is modeled as an error ellipsoid with 14 dimensions and is quantified by the method described in Section 4.2. Next, the data sample is visualized in the items space using a scatterplot. The analyst interactively brushes the data items in the scatterplot (see Figure 7(a)), and then observes the changes of the means and standard deviations of different dimensions in dimensions space using another scatterplot (see Figure 7(b)). The dimensions that do not deviate much due to the brushing are considered “stable”. The analyst intentionally brushes the most stable dimensions in the dimensions space. Multidimensional scaling (MDS) is applied using the selected dimensions.

Dual analysis is an explorative and iterative process, in which uncertainty from the sampling may change. Figure 7 presents two visual uncertainty flows for a dual analysis process. They are created using the standard error ellipsoids (defined in Equation (2)) and the error ellipsoids (defined in Equation (4)) scaled by a confidence level  $CL = 95\%$ , respectively. Comparing these two uncertainty flows, we can see that the right flow shows significantly less uncertainty than

the left flow. Using the unscaled uncertainty may mislead the analyst to become overconfident or too cautious about the analysis results, resulting in increased risk of erroneous decisions. Additionally, unscaled uncertainty has inconsistent implications for different numbers of data dimensions (see Section 4.2 for detailed discussion). Figure 7(d) shows such an example where the input uncertainty with 14 dimensions from the previous items space brushing implies only 0.0001% confidence, while the output uncertainty from the dimensions space brushing with 6 dimensions implies 1.4385% confidence. In contrast, they consistently represent 95% confidence in Figure 7(e). The comparison in Figure 7 demonstrates the usefulness of our uncertainty quantification to faithfully and consistently reveal uncertainty.

Our multi-level uncertainty visualization system can help analysts intuitively validate the analysis results from the perspective of uncertainty and find out how to reduce it. Assume that in a dual analysis process an analyst discovers an interesting split of the data items into two groups (see Figure 8(a)). Figure 8 shows the uncertainty flow which indicates that the finding contains a high degree of uncertainty, and so the finding is questionable. The uncertainty flow allows the analyst to quickly figure out that the uncertainty mainly increases in the previous brushing steps. By inspecting the overall error ellipsoid in the dimensions space brushing step using the matrix visualization (see Figure 8(b)), she immediately knows that two dimensions of the data (“*crim*” indicating “per capita crime rate by town” and “*medv*” indicating “the median value of owner-occupied homes” highlighted by the green rectangles) are the major factors that cause the uncertainty amplification. Thus, she goes back to the items space brushing step and checks the uncertainty in the items space. Figure 8(c) shows the uncertainty at the item level, where uncertainty is encoded as the opacity values of the data items. We can see that a region in the bottom right section of the scatterplot (in the green rectangle) contains mostly uncertain data items. The analyst filters out these uncertain data items (see Figure 8(d)), which leads to a lower degree of uncertainty in the related node. She performs the dual analysis again. Figure 8(e) presents the analysis result which is almost the same as the previous result. However, as we remove uncertain data items, the uncertainty degree of the result is reduced, leading to a more reliable finding.

### 6.2 Customer Opinion Analysis

Our uncertainty framework based on error ellipsoids is primarily used for multidimensional uncertainty information. However, as error ellipsoids are a generalization of one-dimensional standard deviation, the framework also works for the uncertainty represented by standard deviation. In the experiment, we visualize the uncertainty flow of a customer opinion analysis process. The opinion data set used was obtained from [www.TripAdvisor.com](http://www.TripAdvisor.com). It consists of 8 dimensions including user information, hotel information, and opinion scores with 3891 data items (opinions). The opinion scores were extracted by an opinion mining technique [18]. We modeled the uncertainty associated with the extracted opinion scores using one-dimensional error ellipsoids (standard deviations) to reflect the lexical and structural am-

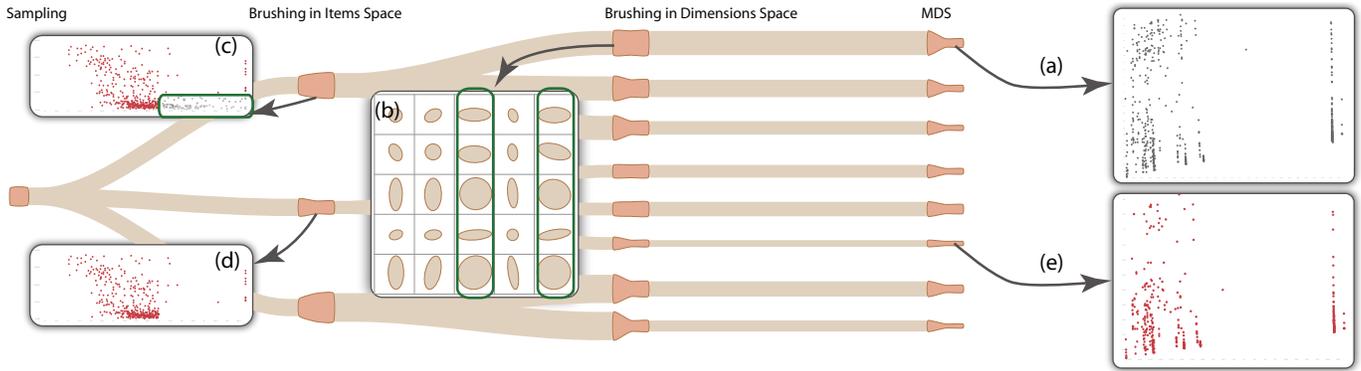


Fig. 8. Uncertainty flow of a dual analysis process. (a) A scatterplot showing an interesting split into two groups but with a high degree of uncertainty; (b) A matrix visualization of the uncertainty revealing that two dimensions amplify the uncertainty; (c) and (d) Scatterplots showing the uncertainty in items space; and (e) A scatterplot showing a revised analysis result by removing highly uncertain data items from (d).

biguity of human languages, as suggested by [42].

Finding opinion patterns regarding categorical information is a fundamental task of opinion analysis in hospitality research [2]. Figure 1 shows a typical opinion analysis process and the related visual uncertainty flow of the analysis process. It starts with opinion mining where the uncertainty arises, followed by user searching or filtering (brushing) in items space and correlation analysis for finding opinion patterns. For instance, the analyst selects opinions about high-class hotels in the items space brushing step (Figure 1(a)). In subsequent correlation analysis, she analyzes the opinion scores against customers’ ages, genders, and locations iteratively using a histogram/bar chart. This process is repeated for different classes of hotels. Finally, the analyst finds that there seems to be an opinion pattern regarding the genders of the customers. She combines the histograms in the previous correlation analysis steps for further analysis (Figure 1(c) and (d)).

In the analysis process, the uncertainty can be split into pieces as the analyst brushes the data, or the uncertainty is merged together as the analyst combines the analysis results. Moreover, uncertainty can increase or decrease as the data is analyzed. Our system can generate a visual uncertainty flow (see Figure 1) for the opinion analysis process. From the uncertainty flow, we can clearly see the variation of the uncertainty information through the process, in which the uncertainty information splits, merges, decreases (Figure 1(a)), or increases (Figure 1(b)). Note that the uncertainty in the correlation analysis steps remains the same because the opinion scores are not changed and the data items are just aggregated with respect to different opinion scores.

Underestimated uncertainty may result in questionable opinion patterns. Figure 1(c) presents an analysis result by combining the previous analysis results. The result reveals an interesting opinion pattern: female customers complain more than male customers, while male customers tend to provide more positive feedback. However, from the uncertainty flow, we can see that the pattern is not very trustworthy because two correlation analysis results used for the combined analysis contain a high degree of uncertainty. The uncertainty flow allows us to track the uncertainty information, which reveals that the uncertainty of the correlation analysis results mainly arises from the items space brushing steps, where high-class hotels are selected. We remove the uncertain correlation analysis results from our combined analysis result (see Figure 1(d)). The uncertainty of the combined result is re-approximated by fusing the uncertainty of the remaining three correlation analysis results. A more reliable opinion pattern is identified in Figure 1(d): female customers of economic hotels tend to complain more. This experiment clearly proves that our uncertainty visualization system can help the analyst quickly evaluate the results from the perspective of uncertainty, and find out how to reduce the uncertainty to improve the analysis result.

### 6.3 Combustion Simulation Analysis

The third experiment demonstrates that our approach can also facilitate scientific data analysis. The data set for analysis is a time-varying volume data set generated by a scientific combustion simulation. Combustion simulations trace a large number of particles and generate a huge amount of data [8]. The data contains more than one million moving particle records. Each record has two time-varying physical dimensions (temperature and mixture fraction). Recent advanced visualization systems can help users interactively explore particle behaviors in combustion simulations [41]. Sampling is often used to simplify the data, which introduces uncertainty. Other transformations, such as filtering and clustering, may also produce uncertainty.

We consider the analysis pipeline described in [41], which helps users classify trajectories of particles. The system pipeline has four main steps (see Figure 9 from the left to right). First of all, particles are sampled from the source data. Trajectories of the sampled particles are initially clustered into a number of groups. Next, according to the users’ domain knowledge, groups of particles are selected, followed by a semi-supervised classification. Finally, the centroids are obtained to represent their corresponding groups. It is an explorative analysis process and each step can repeat multiple times. In the analysis, uncertainty information arises from data sampling and clustering.

Figure 9 shows an uncertainty flow. Initially, the data is randomly sampled over particles and time space, which introduces two-dimensional uncertainty information. The trajectories of the sampled particles are then divided into  $k$  clusters by a model-based clustering method. Figure 9(a) and (b) show the clustering results with  $k = 10$  and  $k = 6$  in the upper and lower branches of the initial clustering step, respectively. The clustering generates a probability vector  $v$  of size  $k$  for each particle, where  $v_i$  indicates the probability of the particle being inside group  $i$ . This actually introduces a new source of uncertainty with  $k$  dimensions. In this step, the number of dimensions of every data item increases from 2 dimensions to  $2 + k$  dimensions, where  $k$  dimensions are from the related probability vector, and we merge the original sampling uncertainty and the new clustering uncertainty into an overall uncertainty.

In Figure 9, we see that the uncertainty variation induced by the clustering with  $k = 10$  is smaller. As we go back to the data with  $2 + k$  dimensions, we find that the  $k$ -dimensional sub-vectors have less deviation in the case of clustering  $k = 10$ . The uncertainty merge process stabilizes the results in this case and hereby reduces the amount of uncertainty according to Equation (3). For each clustering result, we choose two clusters and then try two different brushing operations. Note that in the brushing step, arbitrary brush-out may also cause uncertainty. For example, the uncertainty degree in the third branch is larger than the fourth one after the brushing operations. The first branch results in a small degree of uncertainty, which reflects that we have properly filtered out outliers from the selected clusters. In the final clustering step, according to the brushing information, parti-

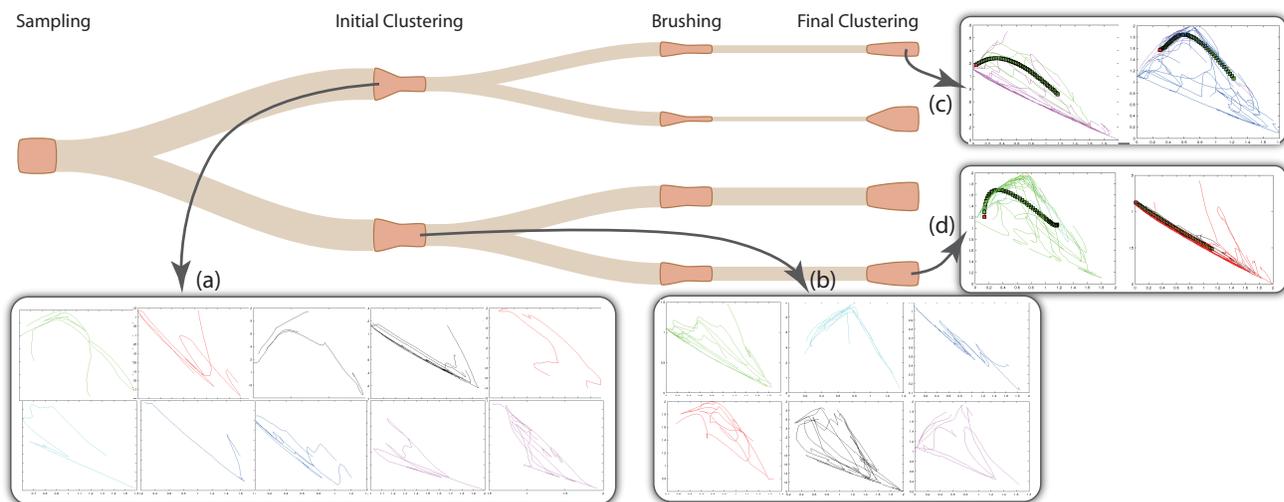


Fig. 9. Uncertainty flow of a combustion data analysis process. In the initial clustering step, (a) shows the resulting 10 trajectory clusters, and (b) shows 6 clusters. The uncertainty flow reveals that splitting more clusters leads to more certain results. In the brushing step, inconsistent curve brushing may magnify uncertainty. In the last step, we present two final clustering results with smaller uncertainty values in (c) and (d).

cles are re-clustered into two groups, where the probability vector of each data item is two-dimensional. We incorporate the obtained two-dimensional probability vectors again to merge all uncertainty error ellipsoids. Finally, we can see that the uncertainty value in the first branch is the smallest among the four cases. From this experiment, we can see that our uncertainty framework can effectively combine two different sources of uncertainty information from the data sampling and data clustering steps, respectively.

#### 6.4 Discussion

Effective visualization of uncertainty information is demanded in many data analysis applications. In the early stages of our research, we came across difficulties in modeling and quantifying the uncertainty information. We found that uncertainty information can arise in any stage of an analysis process and may increase, decrease, split, or merge through the entire process. With the increasing complexity of modern analytical processes, tracking and managing uncertainty information through the entire processes has become even more challenging. However, existing studies focus on a certain stage of an analytics process and fail to characterize the complex variations of the uncertainty along the entire process. This poses a great obstacle to developing an effective visualization system. To tackle this problem, we came up with a new uncertainty framework to model and characterize the variations of uncertainty information through analytical processes.

Another lesson we learned is the importance of preserving the consistency of estimated uncertainty information. Although our approach can quantitatively measure uncertainty information, the quantified uncertainty can be overestimated or underestimated, which may cause analysts to make erroneous decisions, as demonstrated in Figure 7. We propose scaling the quantified uncertainty information according to a confidence level specified by a user, such that the derived uncertainty information is consistent along the entire analysis.

Our system runs as a background process in parallel with the general visual analytics system to automatically collect the necessary history information of the analysis. The time-consuming uncertainty analysis using MCS is only performed when users request an uncertainty visualization from the system. It may take minutes to quantify the uncertainty information for each data item and obtain the overall uncertainty for every recorded analysis step. Thus, the system can be regarded as a post-processing step of a visual analytics process. Although it cannot run in real time, it is still very useful for analysts to revisit their analysis processes, validate their analysis results, and figure out a way to reduce the uncertainty. We plan to accelerate the uncertainty analysis by improving the Monte-Carlo sampling process

and using more advanced parallel computing techniques.

The framework is adapted from previous work [9] to better characterize the uncertainty variations along analysis processes. Although our experiments prove that the framework can be successfully applied to practical problems, there is still room for improvement. For instance, visual analytical reasoning is a rather complex iterative process. It is challenging to model and capture the analytical processes effectively [20]. In some complicated scenarios, the history tree-like graph structure used by the framework may fail to capture the analytical processes. The framework quantifies the uncertainty information using statistical techniques. It may not always work in some applications where it is more effective to model the uncertainty using other techniques such as fuzzy sets and evidence sets. We will investigate this issue and study how to adapt the framework to these applications.

#### 7 CONCLUSIONS

We present a new approach to characterizing, tracking, and visualizing complex uncertainty variations through analytical processes. One major benefit of this work is that it allows users to effectively manage uncertainty of the entire analysis, such that they can validate their results, locate the analysis steps that result in increasing uncertainty, and improve the results by reducing the uncertainty. Our approach can be used by different types of visual analysis applications, as demonstrated in our experiments. The generalizability of the approach is guaranteed by its three characteristics. First, it fully supports multidimensional uncertainty information, which usually arises in multivariate data analysis. Second, it can capture the uncertainty variations, such as uncertainty integration, transformation, and propagation, commonly found in visual analytics processes. Third, the techniques used by the approach, such as the uncertainty modeling based on error ellipsoids and the uncertainty evaluation based on MCS, are well-established techniques that have been used extensively in uncertainty quantification and other related fields. Therefore, we believe that our approach can be applied to many different practical applications.

#### ACKNOWLEDGMENTS

This research is sponsored in part by the National Science Foundation through grants CCF 0811422, CCF 1025269, OCI 0850566, and CCF 0808896, and the Department of Energy through grants DE-FC02-06ER25777 and DE-FC02-12ER26072. The authors would like to thank Jishang Wei and Hongfeng Yu for providing data and source code for the experiment of combustion simulation analysis.

## REFERENCES

- [1] A. Afifi, S. May, and V. A. Clark. *Computer-aided Multivariate Analysis*. CRC Press, 2004.
- [2] N. Au, R. Law, and D. Buhalis. The impact of culture on ecomplaints: Evidence from the chinese consumers in hospitality organization. In U. Gretzel, R. Law, and M. Fuchs, editors, *Information and Communication Technologies in Tourism 2010*, pages 285–296. Springer-Verlag Wien, 2010.
- [3] E. Bertini and D. Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *ACM SIGKDD Explorations Newsletter*, 11(2):9–18, 2010.
- [4] A. M. Bisantz, S. S. Marsiglio, and J. Munch. Displaying uncertainty: Investigating the effects of display format and specificity. *Human Factors*, 47(4):777–796, 2005.
- [5] S. Brandt. *Data Analysis: Statistical and Computational Methods for Scientists and Engineers*. Springer Verlag, 1999.
- [6] K. Brodlić, A. Poon, H. Wright, L. Brankin, G. Banecski, and A. Gay. GRASPARC: a problem solving environment integrating computation and visualization. In *Proceedings of the 4th conference on Visualization*, pages 102–109, 1993.
- [7] S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. *Readings in information visualization: using vision to think*. Morgan Kaufmann Publishers, 1st edition, 1999.
- [8] J. Chen, A. Choudhary, B. De Supinski, M. DeVries, E. Hawkes, S. Klasky, W. Liao, K. Ma, J. Mellor-Crummey, N. Podhorszki, et al. Terascale direct numerical simulations of turbulent combustion using s3d. *Computational Science & Discovery*, 2:015001, 2009.
- [9] C. D. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology*, pages 51–58, 2009.
- [10] J. E. Davis. Combining error ellipses. Technical report, Massachusetts Institute of Technology, 2007.
- [11] S. Deitrick and R. Edsall. The influence of uncertainty visualization on decision making: An empirical evaluation. In *Progress in Spatial Data Handling*, pages 719–738. Springer Berlin Heidelberg, 2006.
- [12] D. Feng, L. Kwock, Y. Lee, and R. Taylor. Matching visual saliency to confidence in plots of uncertain data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):980–989, 2010.
- [13] G. S. Fishman. *Monte Carlo Concepts, Algorithms, And Applications*. Springer-Verlag, 1996.
- [14] M. Friendly. Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, 56(11):316–324, 2002.
- [15] G. Furnas. Generalized fisheye views. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 16–23, 1986.
- [16] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1189–1196, 2009.
- [17] J. Heltona and F. Davis. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. *Reliability Engineering System Safety*, 81(1):23–69, 2003.
- [18] M. Hu and B. Liu. Mining and summarizing customer reviews. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.
- [19] G. Hunter and M. Goodchild. Managing uncertainty in spatial databases: Putting theory into practice. *Journal of Urban and Regional Information Systems Association*, 5(2):55–62, 1993.
- [20] T. J. Jankun-Kelly, K.-L. Ma, and M. Gertz. A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics*, 13(2):357–369, 2007.
- [21] C. R. Johnson and A. R. Sanderson. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications*, 23(5):6–10, 2003.
- [22] Joint Committee for Guides in Metrology (JCGM). *Guide to the Expression of Uncertainty in Measurement (GUM)*. International Organization for Standardization (ISO), 1998.
- [23] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In S. J. Simoff, M. H. Böhlen, and A. Mazeika, editors, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. Springer, 1st edition, 2008.
- [24] M. Kreuseler, T. Nocke, and H. Schumann. A history mechanism for visual data mining. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 49–56, 2004.
- [25] W. K. Liu, T. Belytschko, and A. Mani. Random field finite elements. *International Journal for Numerical Methods in Engineering*, 23(10):1831–1845, 1986.
- [26] H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*. Society for Industrial Mathematics, 1992.
- [27] C. Olston and J. D. Mackinlay. Visualizing data with bounded uncertainty. In *Proceedings of IEEE Information Visualization*, pages 37–40, 2002.
- [28] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1996.
- [29] K. Potter, J. Kniss, R. Riesenfeld, and C. Johnson. Visualizing summary statistics and uncertainty. *Computer Graphics Forum*, 29(3):823–832, 2010.
- [30] C. J. Roy and W. L. Oberkampf. A comprehensive framework for verification, validation and uncertainty quantification in scientific computing. *Computer Methods in Applied Mechanics and Engineering*, 200(25-28):2131–2144, 2011.
- [31] R. Y. Rubinstein. *Simulation and the Monte Carlo Method*. Wiley-Interscience, 2nd edition, 2007.
- [32] J. Sanyal, S. Zhang, G. Bhattacharya, P. Amburn, and R. Moorhead. A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1209–1218, 2009.
- [33] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, 1996.
- [34] M. Skeels, B. Lee, G. Smith, and G. Robertson. Revealing uncertainty for information visualizations. *Information Visualization*, 9(1):70–81, 2010.
- [35] A. Slingsby, J. Dykes, and J. Wood. Exploring uncertainty in geodemographics with interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2545–2554, 2011.
- [36] H. Smith and N. Richard. *Applied Regression Analysis*. Wiley-Interscience, 2003.
- [37] B. N. Taylor and C. E. Kuyatt. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. The National Institute of Standards and Technology (NIST), 1994.
- [38] J. Thomson, B. Hetzler, A. MacEachrenb, M. Gaheganb, and M. Pavel. A typology for visualizing uncertainty. In *Proceedings of Visualization and Data Analysis*, pages 146–157, 2005.
- [39] C. Turkay, P. Filzmoser, and H. Hauser. Brushing dimensions: A dual visual analysis model for high-dimensional datasets. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2591–2599, 2011.
- [40] Y. B. S. J. van Wijk. Supporting exploration awareness in information visualization. *IEEE Computer Graphics and Applications*, 29(5):34–43, 2009.
- [41] J. Wei, H. Yu, R. Grout, J. Chen, and K. Ma. Visual analysis of particle behaviors to understand combustion simulations. *Computer Graphics and Applications, IEEE*, 32(1):22–33, 2012.
- [42] Y. Wu, F. Wei, S. Liu, N. Au, W. Cui, H. Zhou, and H. Qu. Opinionseer: Interactive visualization of hotel customer feedback. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1109–1118, 2010.
- [43] D. Xiu. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.
- [44] F. Yamazaki, M. Shinozuka, and G. Dasgupta. Neumann expansion for stochastic finite-element analysis. *Journal of Engineering Mechanics*, 114(8):1335–1354, 1988.
- [45] T. Zuk and S. Carpendale. Theoretical analysis of uncertainty visualizations. In *Proceedings of Visualization and Data Analysis*, pages 66–79, 2006.
- [46] T. Zuk and S. Carpendale. Visualization of uncertainty and reasoning. In *Proceedings of the international symposium on Smart Graphics*, pages 164–177, 2007.