DOI: 10.1111/cgf.12389
Eurographics Conference on Visualization (EuroVis) 2014
H. Carr, P. Rheingans, and H. Schumann
(Guest Editors)

Volume 33 (2014), Number 3

LoVis: Local Pattern Visualization for Model Refinement

Kaiyu Zhao[1], Matthew O. Ward[1], Elke A. Rundensteiner[1], Huong N. Higgins[1]

[1]Worcester Polytechnic Institute, Worcester, MA 01760

Abstract

Linear models are commonly used to identify trends in data. While it is an easy task to build linear models using pre-selected variables, it is challenging to select the best variables from a large number of alternatives. Most metrics for selecting variables are global in nature, and thus not useful for identifying local patterns. In this work, we present an integrated framework with visual representations that allows the user to incrementally build and verify models in three model spaces that support local pattern discovery and summarization: model complementarity, model diversity, and model representivity. Visual representations are designed and implemented for each of the model spaces. Our visualizations enable the discovery of complementary variables, i.e., those that perform well in modeling different subsets of data points. They also support the isolation of local models based on a diversity measure. Furthermore, the system integrates a hierarchical representation to identify the outlier local trends and the local trends that share similar directions in the model space. A case study on financial risk analysis is discussed, followed by a user study.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical user interfaces (GUI)

1. Introduction

It is never a trivial task to select an appropriate subset of data variables for data analytical processes, such as data mining (*classification, regression, and clustering* [GE03]), and visual exploration [YPH*04, JJ09]. Various pipelines and metrics have been implemented for the different modeling processes in mining packages such as Weka [HFH*09] and R [R C12] for selecting *variables of interest*. However, the algorithm-centric packages usually lack the ability to incorporate domain knowledge [MP13]; furthermore, these methods lack of the flexibility to reveal *local patterns* [MP13, GWRR11]. In some cases, the *local patterns* might comply with the *global pattern* of the data which indicates the global pattern explains the data well; however, in other cases, the local patterns may behave rather differently from the global pattern and may even be opposite of the global pattern [BHO*75], which is known as Simpson's Paradox.

The task of selecting data variables of interest may become more challenging when considering the local subtleties in the data. Example 1 (Figure 1) shows two global models with bias towards opposite directions for part of the data space; Example 2 (Figure 2) shows different ways of defining multiple local models for the same data. Regarding the first example, we want to learn how the models complement each other locally, namely, *(a) on which parts of the data does one model have smaller errors than the other?* and *(b) on which parts of the data does one model overestimate the dependent variable while the other underestimates it?* Regarding the second example, we want to understand *(a) are there any local models that significantly overperforms the global model in terms of model fitness? (b) how many distinguishable local models are appropriate to describe the multiple trends in the data? (c) what are the best cutting values for isolating the local models?* Two example solutions are: 1) to build local models on every single data point; 2) to build one model for all the data points. However, the first case is overly complicated while the second case is not capable of capturing local patterns. In our approach, we are more interested in finding solutions inbetween the two examples. Regarding the isolated local patterns in example 2, a user may further ask, *(a) how different are these local models w.r.t. their direction (e.g., slope and intercept)? (b) do these local models comply with the direction of a representative trend? (c) are there any outlier trends to oppose the majority?* In this paper, we seek to answer the 3 sets of questions above by investigating three model spaces: **model complementarity** (Section 3.1), **model diversity** (Section 3.2), and **model representivity** (Section 3.3).

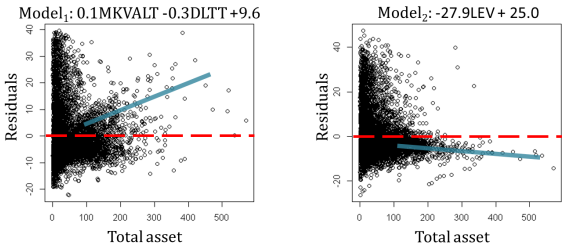

Figure 1: *The two plots show that the two models oppose each other in terms of bias. $Model_1$ has the tendency to underestimate and $Model_2$ tends to overestimate when the total asset grows. The y-axis shows the goodness of fit (residuals). The x-axis is the value of total assets (one of the independent variables). DLTT: Total long-term debt; LEV: Leverage; MKVALT: Market value*

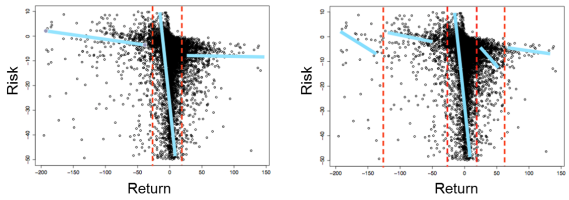

Figure 2: *The plots represent the linear relationship between two variables can be different when considering different partitions of data points. From a domain expert point of view, both high return and low return companies have relatively high risk; intermediate return (fluctuate around 0) companies tend to follow a trend that the risk is reversely proportional to the return.*

Our contributions are summarized as:

- ***A novel model selection environment:*** LoVis allows the user to interactively build and evaluate models at both global and local scales. The interactive exploration is guided by the visual designs in three model spaces.
- ***A novel approach for identifying complementary models:*** LoVis utilizes a pairwise comparison strategy for the model refining. Models that complement the *TBR (to-be-refined)* model are identified and combined *(union of variables)* to the TBR model.
- ***A novel way to examine goodness-of-fit:*** LoVis integrates a novel partitioning strategy for isolating local linear patterns. Strong and weak trends (in terms of goodness of fit) are visualized distinctly in a pattern space. The trend of interest is marked by a data partition (range query).
- ***A hierarchical representation for model summarization:*** We present a hierarchical view for presenting groups of local models, where each group can be interactively divided into smaller ones based on a similarity measure. During the dividing and merging process, the user may investigate the relationship between the size of a group and the divergence within it.

2. Related Work

Many methods for identifying local patterns exist. Guo et al. [GWR09] proposed a system to isolate linear trends by only including the data points within a user specified distance to a trend. Their idea of isolating multiple trends is similar to ours, except that our methods use partition-driven methods to describe the meaning of isolated linear trends. The local patterns in paper [GWRR11] are defined around a focal point; the relative positions of neighbouring points of it are visualized. In LoVis, however, we are instead interested in the local pattern of *a group* of data points and the comparisons between groups.

A partition based framework [MP13] compares the linear models in both 1-D and 2-D partitions of independent variables to facilitate variable selection. In LoVis, we are more interested in how the variables locally complement each other, how the performance of local models vary in different data partitions, and how to identify representativeness of local patterns. A maximal information coefficient (MIC) metric [RRF*11] was defined for identifying multiple types of pair-wise relationships via local analysis. In LoVis, we focus on one type of local relationship and investigate the local pattern of models formed by *multiple variables*.

Data partitioning is perhaps the most important step for identifying local patterns; an interactive framework [MBD*11] was implemented to guide the user to identify local relevance and aggregated global correlation. We do not intend to solve the problem of searching locally correlated feature sets and the corresponding subset of data points, which leads to an expensive optimization problem [GFVS12]. In our work, we use an overlapped partitioning strategy to capture the trends that otherwise might be lost due to less optimally chosen partition boundary.

The Rank-by-Feature Framework [SS04] is similar to our work; it provides quality metrics to measure the interestingness of lower projections (1-D and 2D) to facilitate the visual exploration process in high dimensional data. It has inspired our work in the sense of ranking views by importance. Models with diverse goodness of fit are believed to have more prediction power [BWHY05] and they may indicate the existence of a "lurking explanatory variable" [BHO*75]. Other techniques that focus on the application of quality measures are not specifically designed for local pattern discovery, though they indeed inspired us from various aspects. Scagnostics [WAG05] supplies metrics for identifying interesting structures (e.g., clumpy and stringy). The user-centric approach [JJ09] utilizes several quality metrics that could be combined and adjusted by the user. Peng et al. [PWR04] proposed a metric for reducing clutters in the visual rep-

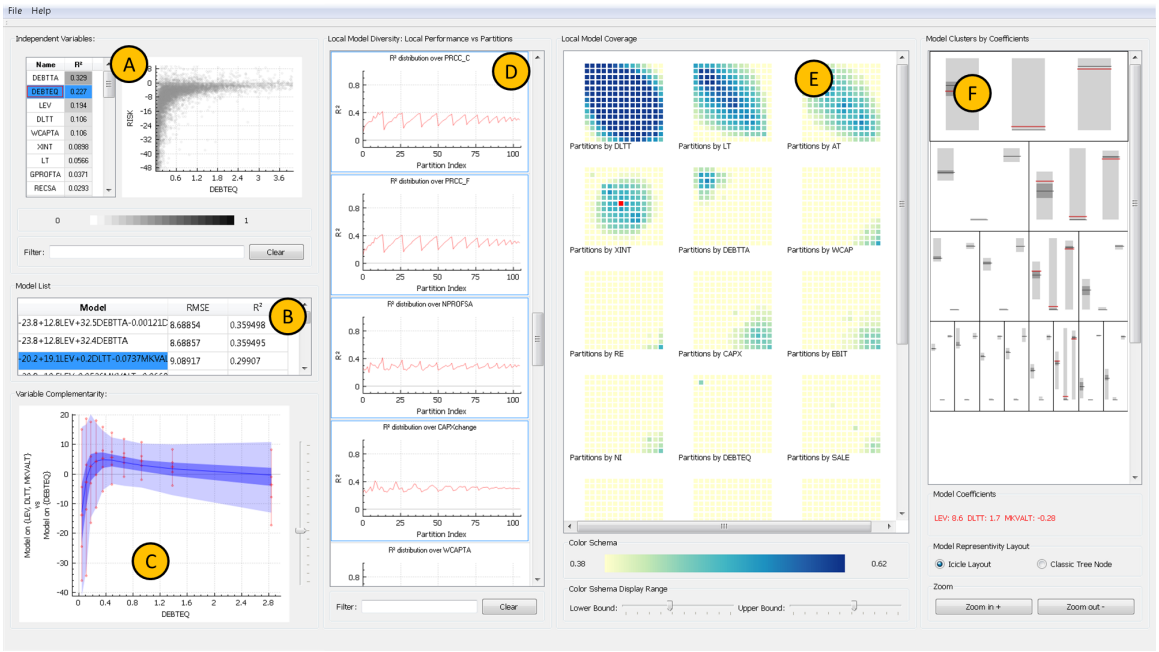

Figure 3: *Integrated analysis framework with 3 stages.* ***1)*** *Variables are ranked by the relevance to the dependent variable and the scatterplot (a) shows the relationship between a selected independent variable and the dependent variable. The global models built by the user are listed in (b). Model complementarity are presented in (c) for refining a user built model in (b).* ***2)*** *Local models can be derived from a selected global model and are presented in (d,e).* ***3)*** *The local models are grouped and summarized in a hierarchy (f).*

resentations. Peringer et al. [PBH08] suggested a quality measure integrated with data space brushing and linking. Tatu et al. [TMF*12] implemented a system that ranks data variables based on subspace cluster structures. The EnsembleMatrix [TLKT09] combines multiple model analysis with visual representations. It allows the user to visually examine the contrast of multiple classifiers and interactively combine them. This strategy motivated us to build a framework to investigate the relations between multiple models. Additionally, we allow the user to incrementally examine the model comparisons in terms of model complementarity and determine the best candidate models for combining.

3. Model Spaces for Visualization

We first categorize the model spaces according to the measurement (local measure or global measure) of models and amount of data the models describe (local model or global model). In the first space, for example, linear models are built on *all data points* and the performance (goodness of fit) of the models are measured on *all data points* using Coefficient of Determination (R^2) and Root Mean Squared Error (*RMSE*). This space together with 3 other spaces are shown in Table 1, where the *local measure* means the models are evaluated in a local data space that only involves a subset

	Global Measure	**Local Measure**
Global Model	R^2, *RMSE*	Model Complementarity
Local Model	Model Representivity	Model Diversity

Table 1: *Model spaces for visualization*

of data points. For example, companies with asset value below 1 million (small companies) and companies with asset value over 10 billions (large companies) can be two local data spaces in a financial dataset. The local models are the models specifically built in a local data space, such as a risk prediction model for small companies and another for large companies. Since the first space has already been commonly used by many other tools, the model spaces we primarily focused on in this paper are the other three:

- *Model Complementarity:* In this space (Section 3.1), we discuss how the model comparisons (Figure 3c) are performed to identify complement models. We also discuss how to characterize the degree of complementarity.
- *Model Diversity:* In this space (Section 3.2), we discuss how the local data spaces are generated via a partitioning

method. We also discuss how *reference variables* (variables used for cutting the data space into partitions) are ranked. We lastly discuss how the diversity is measured, ranked and visualized (Figure 3d,e).

- *Model Representivity:* In this space (Section 3.3), we discuss how the representivity of a group of local models is measured, which helps to determine how well a group of local models is represented by a single trend. We also discuss how the view (Figure 3f) is designed to seek balance between coverage of a group of local models and the divergence within the group.

3.1. Model Complementarity Visualization

This section introduces: 1) how we measure goodness of fit of a model locally; 2) how we compare models based on their local measures; and 3) how we visualize the model complementarity based on the model comparison.

Consider the following scenario: ***A financial analyst found that a risk model she built is dominated by large companies. This means that the fitness (measured by residuals) are smaller for large companies. She wants to find out what additional variables can help the model to perform better on smaller companies.***

To make the scenario more specific, the dependent variable she uses is the bankruptcy risk of companies labeled by financial analysts [WGG10]; the independent variables are financial attributes, such as working capital (WCAPTA), liability (DEBTTA and DEBTEQ), and total assets (AT); the residual is defined as $Y - \hat{Y}$, where Y is the dependent variable and $\hat{Y}$ is the predicted value. The analyst wants to learn on which portions of the data the model performs poorly, and on which portions of the data the model overestimates or underestimates. Hence, we need to investigate the model local performance in local data spaces using additional independent variables such as *total assets*. The relationship between residuals of a linear model and the additional independent variable can illustrate where the model performs poorly (the small companies in this scenario).

Now, we do a point-wise model comparison. In Figure 1, the residuals of two linear models are plotted against an additional independent variable, *total assets*. Both models predict rather poorly (large absolute values of residuals) for the smaller companies; and $model_1$ tends to underestimate (positive residuals) the risk of larger companies while $model_2$ tends to over-estimate (negative residuals). In practice, the two conditions for complementarity are: 1) *error complement*; 2) *bias complement*. For a list of local partitions $p_1, p_2, \ldots, p_n$, let the local errors of a model A be $e_1^a, e_2^a, \ldots, e_n^a$. The above two conditions for complementarity between model A and model B are defined as:

$$\begin{aligned} &\exists i : (|e_i^a| >> 0 \Rightarrow |e_i^b| \to 0) \\ &\quad \vee (|e_i^b| >> 0 \Rightarrow |e_i^a| \to 0) \quad (i \in \mathbb{N}, i \leq n) \end{aligned} \tag{1}$$

$$\exists i : (e_i^a \approx \varepsilon \Rightarrow e_i^b \approx -\varepsilon) \quad (\varepsilon \in \mathbb{R}) \tag{2}$$

In plain language, the two equations can be interpreted as: 1) the large errors of one model align with the small errors of another; 2) the over-estimation portion of one model aligns with the under-estimation portion of another.

A point-wise comparison becomes impractical as the number of data points gets larger. We were inspired by the visualizations for model local performance in [MP13], where the residuals of two models are compared in a 2-D space-filling display using $|Y - \hat{Y}_1| - |Y - \hat{Y}_2|$. Rather than showing the model differences we are instead interested in determining whether the combination of the two models is *cost-effective*. Adding each variable to a TBR model increases the model complexity. Hence we want to know which variable adds more performance to the TBR model. We believe the models that complement each other form a better *combined model* (union of variables). The performance of the combined models can be examined in the table presented in Figure 3b. In order to compare the local performance of two models, we use Tukey's 5-number summary [Tuk77] to measure the distribution of residuals. Two distinguishable forms of boxplot are used to differentiate the local measures of two models (Figure 4). Figure 4 opposes to Figure 3c, as the two models in Figure 3c share a common trend rather than complementarity. This particular design decision is made after experimenting with parallel bar charts and parallel box plots. The parallel bar charts only show the number of data points that fall into a particular partition, which is quite limited in determining the complementarity relationship. The parallel box plots provide more information but takes a lot of screen space. Finally, we chose vertical lines as alternative representations of box plots and added horizontal line connections and space filling to differentiate the two models.

Now we discuss how to define the local measures. We want to translate the local data spaces into a meaningful form. In our case, a data partition (or range query). To define the data partitions, we use a reference variable driven partitioning method [MBD*11], where the authors describe two decomposition strategies. We chose the decomposition strategy that allows comparisons across other variables because we need to compare models that are formed by multiple variables over the data partitions.

Next, we discuss variable rankings in our system. Variable ranking is utilized to support model refinement (Figure 3a) by showing the user the most promising variables first. The ranking score between an independent variable and the dependent variable are measured based on local partitions of the independent variable. Specifically, R^2 is computed over each partition of the independent variable and the final score is the maximum R^2 over the partitions. With the views de-

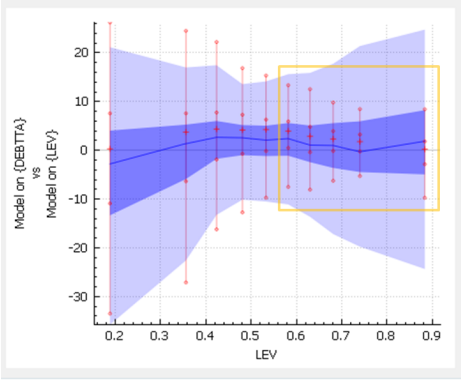

Figure 4: *A candidate model LEV complement the TBR model DEBTTA (in the yellow box). The y-axis represents the error spread of two models. Positive (Negative) values suggest bias towards underestimate (overestimate). The x-axis represents local partitions where the errors are estimated. The theme river design [HHN00] represents the residuals of the TBR model; and the red vertical lines represent the residuals of a candidate model (usually a uni-variate model).*

signed in this space, the tasks a user can perform are listed as follows:

- *Identify relevant variables:* The users may freely choose a variable according to either its relevance to the dependent variable, or their previous domain knowledge.
- *Identify model weaknesses:* The visualization of model local measures reveals the distribution of residuals in local data spaces. By examining the local measures, a user may learn which parts of the data are not described effectively.
- *Identify complementary variables:* The visualization of local measures and local comparisons helps the user to identify whether adding variables to an existing model is cost-effective. The effectiveness of this strategy is evaluated in Section 4.2.

3.2. Model Diversity Visualization

This section discusses the problem when simply adding variables does not significantly improve the model fitness. According to previous work, the reasons may be: 1) the trend is not linear, thus the refining process must consider the possible non-linear polynomials [MP13]; 2) there are multiple linear trends [GWR09]. In this work, we mainly focus on a domain-driven model coverage problem: seeking a way for isolating the multiple models and label the trends with range queries. A query for example can be "*companies with income above 1 million*".

After an interactive selection process, the financial analyst is not satisfied with the model. She suspects there are multiple local trends in the dataset; therefore she wants to break the dataset into a few partitions based on the size of the companies (total assets). Then, she builds local models in the partitions.

This task raises several interesting questions: 1) *how do we retain the domain meaning of each partition while we search for the local trends, and why is it important?* 2) *how do we define the partitions?* 3) *how do we illustrate the relationship between the possible ways of partitioning and the local trends each partition may have?*

For the first question, the analyst wants to isolate local trends into different data partitions, and she wants to know which companies (e.g., large companies or small companies) are associated with a local trend (Figure 2). To accomplish this task, we define a space $\mathcal{P} = \{p_1^1, p_2^1, \ldots; p_1^2, p_2^2, \ldots; \ldots; p_1^v, p_2^v, \ldots\}$ that contains partitions for v variables. Once we have the partitions ready the next steps are to identify a linear trend in each partition using *Robust Regression* (as implemented in R [Hub11]), and visualize the model goodness (Figure 5a). The variables used in the local models are selected using the process discussed in Section 3.1. In order to investigate the reasons why the trends are isolated into several data partitions, the very first step is to annotate the partitions with domain range queries. By linking a local trend to a domain related query, the analysts are able to target the subset of data and further investigate the local properties of the subset.

The discussions above lead to the second question. Specifically, *How do we assign the partition boundaries so that a trend is not divided into different partitions and irrelevant data points are minimized in a partition?* The question is also motivated by the representation of the piecewise linear ranking model [MP13]: 1) when using very coarse piece-sizes, partitions are large and may contain irrelevant data points; 2) when using very fine segments, a trend may be assigned into several partitions. To address that, we use a enumerated partitioning strategy considering all interesting reference variables for partitioning and all interesting sub-intervals of partitions. For example, $total\ assets : [0/100, 30/100]$ represents a 0th and 30th percentile interval on reference variable *total assets*. Each partition in space $\mathcal{P}$ thus can be defined as $p_k^R = R : [l,h]$ where R denotes the chosen reference variable; k represents the index of the partition; and l and h ($0 \leq l,h \leq 1$) represent lower and upper boundaries on the reference variable. The space $\mathcal{P}$ is populated by partitions of varying boundaries, which is discussed next together with the layout strategy.

We answer the third question by introducing the layout strategy of the diversity view (Figure 5a). In an n by n grid view (Figure 5a), the position (i, j) of a cell (Figure 5b) represents the boundaries $[i/n, j/n]$ of a data partition. The factor $1/n$ is a *minimum step size threshold* to avoid infinite number of partitions. Due to the symmetricity of the n by n grid and the trivial information on the diagonal we first remove the diagonal and the entries below the diagonal; and then fill the lower half of the grid according to the sym-

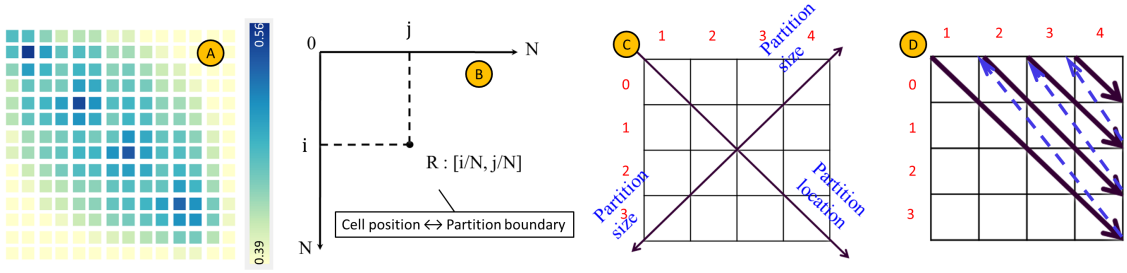

Figure 5: *The x-y position of any cell in the grid view (a) is determined by the lower (x) and upper (y) percentile threshold of a data partition. The relationship between x-y position and the partition boundary is shown in (b) and is indexed as in (c,d). Each cell is colored by the fitness of a local model in it. The diagonal and the orthogonal direction in (c) indicates two ways a data partition may change to another: expanding (add more data points) and shifting (add data points at one end and remove at the other). An alternative display of (a) (Figure 6) is transformed from (a) by the sequence in (d) where the main diagonal is walked from top left first followed by the second diagonal above it. The walk continues till the right top corner.*

metricity. We fill the grid because several test subjects felt the symmetric view is more pleasing to read while others have no preferences. In some cases a partition $R : [i/n, j/n]$ may not well cover a linear trend due to missing relevant data points or containing irrelevant data points. An alternative partition $R : [(i+\varepsilon)/n, (j+\omega)/n]$ $(\varepsilon, \omega \in \mathbb{Z})$ need to be compared to $R : [i/n, j/n]$ for getting better boundary positions. A vicinity relationship between the compared partitions are demonstrated in Figure 5c in two directions to help the comparisons. The diagonal direction corresponds to partition shifting (ε and ω change towards the same direction). The orthogonal direction represents the expanding or shrinking of a partition. The color of each cell in Figure 5a represents the goodness of fit of the trend in that partition. We use relative measure R^2 to measure the goodness of fit because the absolute fitness measure, such as RMSE, is often driven by the value of the independent variables which will cause unfair comparisons between data partitions. The absolute errors can be studentized [CW82] before the comparisons but it is beyond the scope of this paper.

To support the ranking and filtering of diversity views, we design a linear layout of the partitions (Figure 3d) which are ranked by the degree of fluctuations (Figure 6b,d). We use standard deviation of the local goodness of fit to quantify the fluctuations. The data partitions in a line chart (x-axis) is ordered by the diagonal walking sequence illustrated in Figure 5d. The more fluctuating line in Figure 6b indicates higher diversity. It suggests that the reference variable is effective in isolating multiple local trends. The smoother line in Figure 6d suggests the performance of isolated local models is similar to that of the global model. The diversity view is ordered and filtered using the same standard deviation measure. A user can perform the following tasks, using the views designed in this space:

- *Identify reference variables:* With the local model diversity measure, a reference variable is ranked based on the fluctuation local model. With the ranking metric, the user may identify variables that better isolate local models.
- *Identify multiple trends:* With the diversity representations, the user may identify multiple trends by reading the color spread in the diversity view.
- *Identify the size, location and strength of a local trend:* The user may identify the corresponding range query for a trend in the diversity view by reading the x-y position of the cells. The size and strength of the trend can also be identified by the color spread the cells.

3.3. Model Representivity Visualization

Let us continue our case scenario from Section 3.2: ***The financial analyst discovered that the local models perform rather well in some partitions*** ($profit : [0.3, 0.5]$, $assets : [0.4, 0.7]$, $sales : [0, 0.4]$)***. She wonders if it suggests the existence of a single model that can cover these local models. Furthermore, she also wants to know if that single model is robust, namely, are the local models it covers significantly diverging? Additionally, which data partitions contain trends that disagree with the majority of trends?***

To help her, we designed an interactive hierarchical visualization that represents the similarities between the isolated models. We measure the similarities using coefficient vectors of the models (e.g., slope and intercept in a 2-D case). We want to answer: 1) *do the isolated local trends point to a similar direction, and thus can be covered by a representative trend?* 2) *if yes, how much confidence can be assigned to such local trends?* 3) *if not, how different are the trends in terms of their directions in the hyperspace?*

A representative model in $\mathcal{S}$ is expected to be central and cover as many partitions in $\mathcal{P}$ as possible, while the divergence in $\mathcal{S}$ below a certain threshold ξ. We define $\mathcal{S}$ as:

$$\min_{\forall \mathcal{S} \subset \mathcal{P}} (|\mathcal{P}| - |\mathcal{S}|) \text{ subject to } Div(\mathcal{S}) < \xi$$

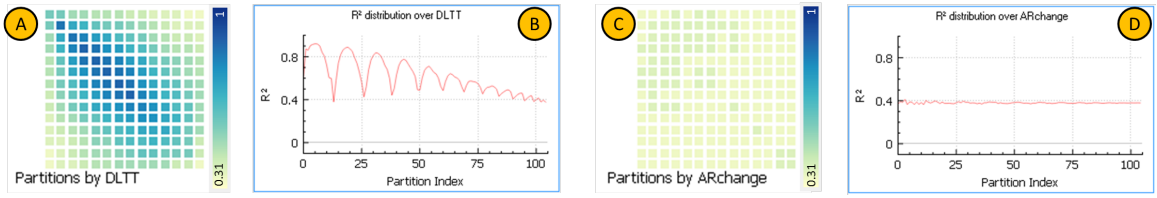

Figure 6: *Plot represents degree of diversities. It shows that the local models isolated by partitioning on DLTT (a,b) have more diversity over the local models isolated by partitioning on ARChange (c,d). ARChange: Account Receivable Change*

where $Div(\mathcal{S})$ denotes the model divergence in $\mathcal{S}$ where $\mathcal{S}$ is a group of partitions. To measure the model divergence, we use a normalized version of Euclidean distance:

$$d_{ij} = \sqrt{\frac{1}{w_a}(a_i - a_j)^2 + \frac{1}{w_b}(b_i - b_j)^2 + \dots}$$

where d_{ij} is the distance between two models m_i and m_j and $a, b, \dots$ are the coefficients. The normalization factor we use is the amplitude of each coefficient: $w_a = max_i(|a_i|)$, $w_b = max_i(|b_i|)$, and so on. To visualize the divergence and the coverage problem, we leverage the idea of *below traversal* in the hierarchical aggregation [EF10]. We also employ a divisive clustering algorithm [KR09] that divides a large cluster of items into smaller clusters in a top-down process. At each iteration it separates clusters of items at a computed cutting location. Icicle plots [KL83] are used to represent the hierarchical group structures. The icicle plots use relative positions of the node instead of edges to infer parents and children thus it is believed to have higher information density than classic tree node graph [MR10]. The model divergence of each cluster is visualized at each node of the icicle plot using a variation of box-plot (Figure 7 right) where bars represent the coefficient statistics of the models. Using the techniques above, the representivity of a model M_R in the partition space $\mathcal{S}$ (a cluster of partitions) can be implied from the divergence of the models in $\mathcal{S}$, the centrality of model M_R in $\mathcal{S}$ and the coverage of $\mathcal{S}$. The divergence of models can be directly read from the box-plot in each node of the icicle plot. We next discuss the interactions needed to learn the centrality of M_R and the coverage of $\mathcal{S}$.

The user can double click on a node to break down a cluster with high divergence or merge smaller clusters with low divergence. The user may find the divergence of a cluster reduces to small values while still covering a set of data partitions (Figure 7). The user can also mouse over the diversity view (Figure 8 left) and examine the centrality of the highlighted partition in a group (Figure 8 right). In this example, it is an outlier trend in the second node at level 3 of the icicle plot (node with red bars in it) because all the three bars are at the boundary of the box-plot (Figure 8 right). Additionally, the divergence of the group is higher than the other three groups at the same level. Another example can be seen in Figure 9l where the divergence of the grouped model is lower than that in the previous example and the coefficients

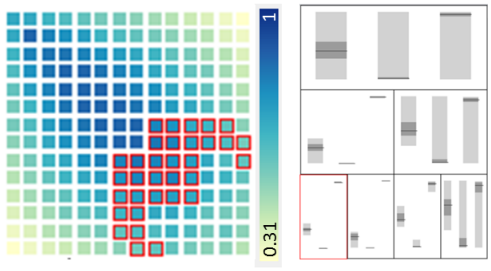

Figure 7: *Visualizing the coverage (cells with red outline on the left) of a selected cluster of data partitions (selected node marked with red rectangle on the right).*

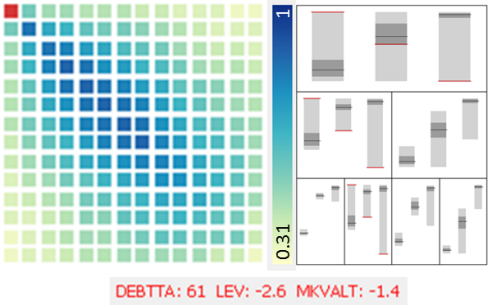

Figure 8: *Visualize the coefficient vector (red horizontal bars in the icicle plot) of the linear trend in the highlighted data partition (left). The red text shows the value of the coefficients and the name of variables.*

of the highlighted model are close to the center of the box-plot. Lastly, the user may want to click on the nodes in the icicle plot (Figure 7) and examine the data coverage of each node. This view space supports:

- *Identify outlier trends:* Coefficient values of a trend that are boundary values comparing to other trends may indicate that it is an outlier trend.
- *Identify a representative trend:* A representative trend can be identified by checking the divergence of the group it belongs to, centrality of the trends in the group and data coverage of the group.

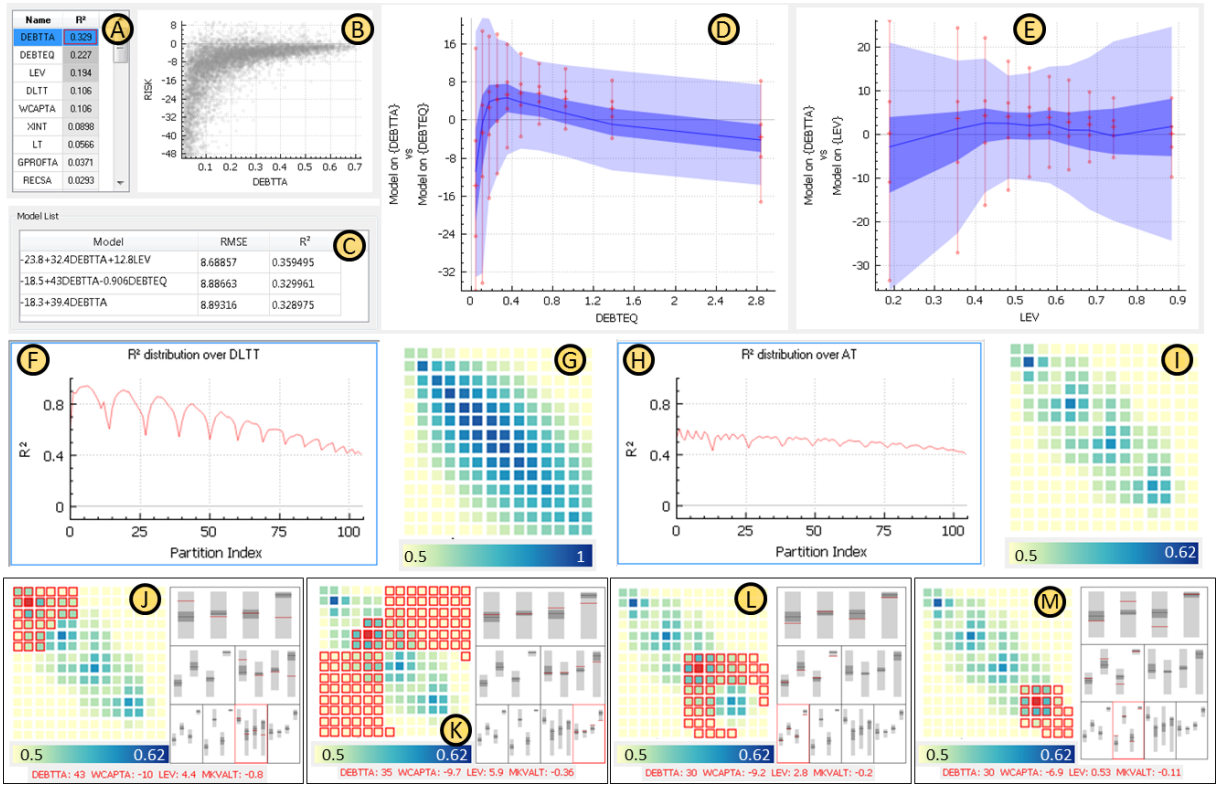

Figure 9: *A case study for modeling risk. a) A ranking list of independent variables. b) Scatterplot of a selected independent variable and the dependent variable. c) A list of built models. d,e) Complementarity analysis. f,g,h,i) Local model diversity analysis. j,k,l,m) Model representivity analysis. Detailed analysis is in Section 4.1.*

4. System Evaluation

In this section, we demonstrate a case study using a financial database. We also report the result of a user study we conducted involving professors and students from the departments of Math, Computer Sciences, and School of Business.

4.1. Case Study: Linear Models of Bankruptcy Risks

The data we use in this work are from Compustat [Poo11], a database of financial, statistical and market information of companies from around the world. Since the database is huge, we focus on only on one sector of the US companies that are active in the year 2010, namely the service sector classified by the SIC standard [sic13]. After cleaning, we acquired 45 variables for 9,483 observed companies.

To build linear models for risk prediction, the analyst first examines the relevance ranking of the independent variables in the relevance view (Figure 9a). The relationship between the highlighted independent variable and the dependent variable is plotted in a scatterplot (Figure 9b). From the relevance ranking list, she identifies that the variables DEBTTA, DEBTEQ, and LEV are most predictive to the dependent variable. However, she would like to figure out which combination is better. Choosing all 3 of them is an option, but it may increase the model complexity unnecessarily. She next examines the model complementarity view (Figures 9d and 9e) to determine which variable *complements* the variable DEBTTA (the first candidate) better. The two models in Figure 9d share a common pattern (up/down and vertical spread, and less complementary). The model represented as red lines in Figure 9e performs better at the right half of the data partitions (smaller error spreading, and more complementary). She confirms that the combination $\{DEBTTA, LEV\}$ is better ($RMSE = 8.68, R^2 = 0.359$) than $\{DEBTTA, DEBTEQ\}$ ($RMSE = 8.89, R^2 = 0.330$) in the model list (Figure 9c) after trying both combinations. Although both of them are better than model with only one variable $\{DEBTTA\}$ ($RMSE = 8.89, R^2 = 0.329$), LEV is the variable that adds more fit. In an automatic model building process, the user does not have direct control over the variable selection, the user knowledge thus cannot be directly applied to help the selection.

Next, the user may examine the local models that are derived from the current best model. The derived local models are based on the same set of variables we identified via

the *complementarity analysis*. Each local model is built on a partition ($R : [l,h]$). By examining the *model diversity* views, the analyst immediately notices two interesting patterns: 1) Figure 9f shows that in some partitions (in Figure 9g cells with darker blue), the local trends are very strong, as R^2 is over 0.9 in some of them. The strong linear trends can be expanded along the orthogonal direction (Figure 9g) to a larger range of partitions at a lower threshold (lighter colors). 2) Another pattern that could be spotted is that the local models show 4 local maxima in Figure 9i, where 4 strong linear trends are isolated in the partitions represented by the darker blue cells. The pattern shows that the domain knowledge of the analyst is partially correct in the sense that the local trends are indeed stronger when isolating them by the variable *total assets*. It suggests that constructing models with a mixture of both small and large companies is less effective because the model with only smaller companies (the dark cell at $R : [1/14, 2/14]$ in Figure 9i) outperforms the model built on all companies (top-right cell at $R : [0/14, 14/14]$ in Figure 9i). The reason she is only partially correct is that the 4 local maxima in Figure 9i suggest modeling the companies at 4 different scales instead of 2.

The next step is to check the *model representivity*. The analyst breaks the local models down hierarchically, and discovers that at level 3 each of the 4 clusters contains one local maximum (Figure 9j, 9k, 9l, 9m). It confirms that using the group of 4 is the right choice, because the directions of the trends in the 4 clusters are different. Specifically, *DEBTTA* and *MKVALT* are more significant in the small company group and the significance decreases with the scale of the companies. *WCAPTA* and *LEV* are less significant in the large medium and large groups, while *WCAPTA* is most significant in the small medium group. Another notable pattern is that the local trend in the small medium group can be represented by the global trend, because the two trends are clustered in the same group that has rather small variances.

The three model spaces in LoVis are additional features that complement the automatic model building process. We compare LoVis to the *LinearRegression* algorithm in Weka from the perspective of model complexity (number of variables) and model fit (R^2). Using the same dataset as input, Weka selects 27 out of the original 45 variables and forms a linear model with R^2 at 0.522. The overall fit is better than the models we formed in LoVis which usually involve fewer variables. However, LoVis has the advantage of modeling the local properties of the dataset. 1) It discovers local data spaces that can form linear models with R^2 at above 0.8 (Figure 9f,g) which is higher than the fit of the automatically formed global model; 2) It also characterizes multiple local models with local maximal fit (Figure 9h,i). With only 4 variables, each model has R^2 of about 0.6 which is higher than the fit of the automatically formed model on 27 variables. (Note: Root Relative Squared Error in Weka is converted to R^2 using: $R^2 = \sqrt{1 - RRSE^2}$)

4.2. Results from a User Study

To validate the usability of the model complementarity, we performed a user study with 20 subjects. The participants answered 3 questions after a short training. In each question, they were asked to choose one option out of two. One option (e.g. Figure 9e) is better than the other (e.g. Figure 9d) measured by Fit Difference (FD). We expected to see the user selected option have have better fit when the variables in the option are combined (set union).

$$FD = |\text{Model Fit}_{\text{variable set 1}} - \text{Model Fit}_{\text{variable set 2}}|$$

In the results, there is a relationship between the *selection accuracy* and the FD between the two options, which is show in the table below:

FD (R^2)	Accurary (%)	Avg time(s)
0.12	90	13.4
0.08	80	24.6
0.03	60	25.3

From the result, more users (90%) made optimal selections when the FD between the two choices is more significant (0.12). When the FD goes down to 0.03 (R^2), the user selection tends to be less accurate (60%) and is more time consuming (25.3s); however, at that point, the performance gain of adding the wrong selection is only 0.03 (measured by R^2) less than the right selection.

5. Conclusion and Future Work

In this work, we presented a system LoVis that integrates three visual spaces, focusing on local pattern discoveries that facilitate the linear model refinement process. We measure the degree of complementarity between a to-be refined model and the candidate variables so that a suitable variable can be selected to compensate for the poor performance of the to-be refined model locally. Local models are built to model the diversity in the dataset in a novel partition space. Divergence of the local models is measured and visualized to investigate the representivity of a group of models.

There are several limitations in our system, and we are planning to address these in the near future. For instance, alternative model discovery is not supported and usually there are some parts of the data that cannot be modeled by adding more variables or using multiple local models. Alternative models in a different subspace may exist and can benefit the process of forming composite models. Another limitation is that we do not support partitioning on multiple variables and we plan to extend our work by utilizing techniques such as Dimensional Stacking [LWW90] and Parallel Sets [KBH06] to address this.

6. Acknowledgement

This work is supported under NSF grant IIS-1117139.

References

[BHO*75] BICKEL P. J., HAMMEL E. A., OŠCONNELL J. W., ET AL.: Sex bias in graduate admissions: Data from berkeley. *Science 187*, 4175 (1975), 398–404. 1, 2

[BWHY05] BROWN G., WYATT J., HARRIS R., YAO X.: Diversity creation methods: a survey and categorisation. *Information Fusion 6*, 1 (2005), 5–20. 2

[CW82] COOK R. D., WEISBERG S.: *Residuals and influence in regression*, vol. 5. Chapman and Hall New York, 1982. 6

[EF10] ELMQVIST N., FEKETE J.-D.: Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics 16*, 3 (2010), 439–454. 7

[GE03] GUYON I., ELISSEEFF A.: An introduction to variable and feature selection. *The Journal of Machine Learning Research 3* (2003), 1157–1182. 1

[GFVS12] GÜNNEMANN S., FÄRBER I., VIROCHSIRI K., SEIDL T.: Subspace correlation clustering: finding locally correlated dimensions in subspace projections of the data. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2012), ACM, pp. 352–360. 2

[GWR09] GUO Z., WARD M. O., RUNDENSTEINER E. A.: Model space visualization for multivariate linear trend discovery. *IEEE Symposium on Visual Analytics Science and Technology* (2009), 75–82. 2, 5

[GWRR11] GUO Z., WARD M. O., RUNDENSTEINER E. A., RUIZ C.: Pointwise local pattern exploration for sensitivity analysis. *IEEE Conference on Visual Analytics Science and Technology* (2011), 129–138. 1, 2

[HFH*09] HALL M., FRANK E., HOLMES G., PFAHRINGER B., REUTEMANN P., WITTEN I. H.: The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter 11*, 1 (2009), 10–18. 1

[HHN00] HAVRE S., HETZLER B., NOWELL L.: Themeriver: Visualizing theme changes over time. In *IEEE Symposium on Information Visualization* (2000), IEEE, pp. 115–123. 5

[Hub11] HUBER P. J.: *Robust statistics*. Springer, Berlin Heidelberg, 2011. 5

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Transactions on Visualization and Computer Graphics 15*, 6 (2009), 993–1000. 1, 2

[KBH06] KOSARA R., BENDIX F., HAUSER H.: Parallel sets: Interactive exploration and visual analysis of categorical data. *IEEE Transactions on Visualization and Computer Graphics 12*, 4 (2006), 558–568. 9

[KL83] KRUSKAL J. B., LANDWEHR J. M.: Icicle plots: Better displays for hierarchical clustering. *The American Statistician 37*, 2 (1983), 162–168. 7

[KR09] KAUFMAN L., ROUSSEEUW P. J.: *Finding groups in data: an introduction to cluster analysis*, vol. 344. John Wiley & Sons, 2009. 7

[LWW90] LEBLANC J., WARD M. O., WITTELS N.: Exploring n-dimensional databases. In *Proceedings of the 1st conference on Visualization'90* (1990), IEEE Computer Society Press, pp. 230–237. 9

[MBD*11] MAY T., BANNACH A., DAVEY J., RUPPERT T., KOHLHAMMER J.: Guiding feature subset selection with an interactive visualization. In *IEEE Symposium on Visual Analytics Science and Technology* (2011), IEEE, pp. 111–120. 2, 4

[MP13] MUHLBACHER T., PIRINGER H.: A partition-based framework for building and validating regression models. *IEEE Transactions on Visualization and Computer Graphics 19*, 12 (2013), 1962–1971. 1, 2, 4, 5

[MR10] MCGUFFIN M. J., ROBERT J.-M.: Quantifying the space-efficiency of 2d graphical representations of trees. *Information Visualization 9*, 2 (2010), 115–140. 7

[PBH08] PIRINGER H., BERGER W., HAUSER H.: Quantifying and comparing features in high-dimensional datasets. In *In Proceedings of the IEEE Symposium on Information Visualisation* (2008), pp. 240–245. 3

[Poo11] POOR'S S. .: Compustat database. www.compustat.com, July, 2011. Accessed: 2013-11-27. 8

[PWR04] PENG W., WARD M., RUNDENSTEINER E.: Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Symposium on Information Visualization* (2004), pp. 89–96. 2

[R C12] R CORE TEAM: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. ISBN 3-900051-07-0. URL: http://www.R-project.org/. 1

[RRF*11] RESHEF D. N., RESHEF Y. A., FINUCANE H. K., GROSSMAN S. R., MCVEAN G., TURNBAUGH P. J., LANDER E. S., MITZENMACHER M., SABETI P. C.: Detecting novel associations in large data sets. *Science 334*, 6062 (2011), 1518–1524. 2

[sic13] Standard industrial classification (sic) system. http://www.census.gov/epcd/www/sic.html, 2013. Accessed: 2013-11-27. 8

[SS04] SEO J., SHNEIDERMAN B.: A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *In Proceedings of the IEEE Symposium on Information Visualization* (2004), pp. 65–72. 2

[TLKT09] TALBOT J., LEE B., KAPOOR A., TAN D. S.: Ensemblematrix: interactive visualization to support machine learning with multiple classifiers. In *Proceedings of the 27th international conference on Human factors in computing systems* (2009), ACM, pp. 1283–1292. 3

[TMF*12] TATU A., MAAS F., FARBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology* (2012), pp. 63–72. 3

[Tuk77] TUKEY J. W.: *Exploratory data analysis*. Addison-Wesley, Reading, Massachusetts, 1977. 4

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. In *Proceedings of the IEEE Symposium on Information Visualization* (2005), pp. 157–164. 2

[WGG10] WU Y., GAUNT C., GRAY S.: A comparison of alternative bankruptcy prediction models. *Journal of Contemporary Accounting & Economics 6*, 1 (2010), 34–45. 4

[YPH*04] YANG J., PATRO A., HUANG S., MEHTA N., WARD M., RUNDENSTEINER E.: Value and relation display for interactive exploration of high dimensional datasets. In *Proceedings of the IEEE Symposium on Information Visualization* (2004), pp. 73–80. 1