Eurographics Conference on Visualization (EuroVis) 2017
J. Heer, T. Ropinski and J. van Wijk
(Guest Editors)

Volume 36 (2017), Number 3

Interactive Regression Lens for Exploring Scatter Plots

L. Shao[1], A. Mahajan[2], T. Schreck[1], and D. J. Lehmann[3]

[1]Graz University of Technology, Austria
[2]PEC University of Technology, Chandigarh, India
[3]University of Magdeburg, Germany / University Rey Juan Carlos, Madrid, Spain

Abstract

Data analysis often involves finding models that can explain patterns in data, and reduce possibly large data sets to more compact model-based representations. In Statistics, many methods are available to compute model information. Among others, regression models are widely used to explain data. However, regression analysis typically searches for the best model based on the global distribution of data. On the other hand, a data set may be partitioned into subsets, each requiring individual models. While automatic data subsetting methods exist, these often require parameters or domain knowledge to work with. We propose a system for visual-interactive regression analysis for scatter plot data, supporting both global and local regression modeling. We introduce a novel regression lens concept, allowing a user to interactively select a portion of data, on which regression analysis is run in interactive time. The lens gives encompassing visual feedback on the quality of candidate models as it is interactively navigated across the input data. While our regression lens can be used for fully interactive modeling, we also provide user guidance suggesting appropriate models and data subsets, by means of regression quality scores. We show, by means of use cases, that our regression lens is an effective tool for user-driven regression modeling and supports model understanding.

Categories and Subject Descriptors (according to ACM CCS): G.3 [Computer Graphics]: PROBABILITY AND STATISTICS—Correlation and regression analysis

1. Introduction

In the big data era, relevant data is constantly growing in many domains and visual-interactive techniques are becoming more and more important. There already exist techniques that help analysts to explore and explain different types of data, in different application domains. The scatter plot is a well-known basis technique to explore correlations, trends and clusters in bivariate data. Exploration with scatter plots can benefit from interest measures like Scagnostics [WAG05] or regressional features [SBS11], to search and identify informative views in larger sets of scatter plots, e.g., a scatter plot matrix (SPLOM).

An interesting extension for analysis of scatter plots is investigation of *local patterns* in single projection views [SSB*16, JSG16, MG13]. Prior research has shown that multivariate data sets may contain locally valuable information that has to be extracted and properly visualized. In this regard, local patterns can be represented by local regression models that in sum can describe a global scatter plot of a set of local models. The analysis of local regressions, also known as segmented regression [WSZRD02], plays an important role in statistic modeling and is used to find substantial changes in relationships among variables. Therefore one dimension, usually the independent variable, has to be partitioned into intervals for computing the local models. But typically, the partition breakpoints are not known before the analysis and have to be estimated.

Automatic approaches for building regression models are typically limited with respect to incorporating domain knowledge in the process of selecting input variables (also known as feature subset selection). Furthermore, the data must either be labeled or well clustered to compute local regression models automatically. However, there are many clustering algorithms and parameterizations to choose from, which in practice often result in many possible, different cluster segmentations. The challenge here is to choose the best clustering algorithm including the parameter setting for a given data set. Other potential limitations of algorithmic local regression analysis include the identification of local structures, transformations, and interactions between variables.

In this paper, we focus on the visual-interactive extraction and representation of local regression models in scatter plots. We introduce regression lens, a novel concept for visual-interactive regression analysis that allows users to select a portion of data on which they want to conduct regression analysis. The lens can be interactively modified in terms of position or size, to find the best fitting model for areas of interest. Detailed interactive feedback allows to compare different models and data selections effectively. Using our regression lens, users can explore scatter plots in a novel way and are enabled to investigate a plot by their individual constituents of regression models. Based on a best-fit algorithm including data sampling and cross-validation, we determine the best coefficients for the

selected data independent of the model direction i.e., $f_y(x)$ or $f_x(y)$. We provide a set of appropriately defined and visualized statistical measures, for judging the quality of candidate models. Moreover, a selection guidance concept helps to optimize the selection area of the lens by indicating outliers and suggesting translation directions along which regression model quality can be improved.

The remainder of this paper is structured as follows: In Section 2, we discuss related work. Section 3 gives an overview of our regression lens approach and describes challenges in regression-based data analysis. Section 4 introduces our prototype system including an implementation of the regression lens concept and its usability. Next, in Section 5, we apply our approach to different data sets and showcase the exploration benefits. Limitations and possible extensions are discussed in Section 6. We, finally, conclude in Section 7.

2. Related Work

Our work relates to local data pattern analysis, feature extraction and interactive lens techniques for scatter plot visualizations. We present an overview of works next.

2.1. Correlation Analysis and Feature Extraction

In scatter plot analysis advanced data analysis tasks, such as feature computation, pattern extraction or statistical analysis, require important initial steps of assessing correlations, trends and clusters. Regression analysis is widely used to explore statistical relations between selected pairs of variables. In [Ans73], Anscombe explored the importance of graphs, and looked into the usefulness and importance of statistical analysis of scatter plot data using regression analysis. Nowadays, data analysis tools like Tableau[†] are available that provide such analysis possibility but, to date, are limited in focus-plus-context techniques such as interactive lenses.

An interactive framework for building and validating regression models is presented by Mühlbacher et al. in [MP13]. The framework helps analysts to understand relationships between observed variables and a dependent target variable, and explains the most useful feature and its partitions by using a regression model. Another related work is [GWR09]. There, Guo et al. defined model space visualizations including heatmap-based displays, which help identify linear dependencies for multivariate data. Li et al. [LMvW10] conducted a study about the effectiveness of judging correlations in scatter plots. It turned out that scatter plots are more effective in supporting visual correlation analysis than parallel coordinate plots.

For high-level analysis tasks, a combination of techniques from data mining and interactive visualization can be applied to facilitate finding patterns in possibly large data. For instance, Scagnostics [WAG05] characterizes the global distribution of points based on geometrical and topologic properties. Specifically, nine features including density, shape, stringiness and outlier measures are defined. These features can serve to rank and select plots for interactive inspection. In [SBS*14], image-based descriptors are used to search for scatter plot patterns based on user sketch queries. The similarity between a scatter plot pattern and a user sketch is measured by the density of points and the frequency of different edge orientations in the image space. Scherer et al. [SBS11] presented a scatter plot descriptor based on regression features for comparing scatter plots with each other. Specifically, they proposed a feature vector based on the goodness-of-fit of a set of globally applied regression models.

A well-known problem of scatter plots is the degree of overdraw on local regions as the number of points increase. To tackle this problem, hexagonal binning [CLN86] can be applied, encoding point density with a colormap within hexagonal binning regions. Further, Mayorga and Gleicher [MG13] developed an abstraction approach to automatically group dense data points, and used color blending and contour lines to reveal hidden data distributions. In [CCM*14], a hierarchical multi-class sampling technique is used which simplifies the distribution by preserving relative density features.

Besides visual abstraction approaches, an investigation of local patterns can also be helpful to reveal further insights, which may be hidden in the overall view. For instance, scatter plots could be extended by sensitivity coefficients to visualize local variation of one variable with respect to another [CCM10]. They represent the sensitivity information as velocities so that the resulting visualization resembles a flow field. In [JSG16], a method is presented to emphasize a local area of interest based on depth of field and a multidimensional focus selection body. In [SSB*16], a scatter plot interest measure is presented, which is based on an adapted tf$\times$idf approach computed over sets of local clusters. This measure is used to rank scatter plots based on the frequency of local patterns, useful to propose views to a user from a large scatter plot view space.

For high dimensional data sets, scatter plot matrices [CLN86] in combination with the brushing and linking technique [BC87] may be useful for finding related patterns across multiple scatter plot projections. Alternatively, the point distribution of multivariate data can be displayed onto 2D planes by using radial projection-based visualization techniques like Radviz [HGM*97] or Star Coordinates [Kan01]. The effect of judging correlations for these projection-based visualizations are published in [Nv06, Nv09]. In [LKZ*15], guidance pictograms are presented to support standard visual search tasks, such as correlation and distribution analysis, for projections like scatter plots, Radviz and Star Coordinates.

2.2. Interactive Lens Techniques

To interactively explore local scatter plot regions for interesting patterns, virtual lens techniques like magic lens or magnification lens may be used [BSP*93, LHJ01], which provide on demand an alternative visual representation of the underlying data. There already exist a number of different interactive lens techniques for various application and data domains. For instance, there are lenses to show temporally aggregated information of trajectory data (time lens [TSAA12]), to explore multivariate network data (network lens [JDK10]) or to magnify volumetric features in 3D representations (magic volume lens [WZMK05]).

Moreover, there are specific lens techniques to support the analysis of multivariate data in scatter plot visualizations. Ward and Yang [WY04] have presented an overview of interaction operations that can occur in data and information visualization including lens techniques (distortion) for scatter plot matrices. To overcome the overdraw problem Ellis et al. [EBD05] have introduced sampling lens, which estimate a suitable sampling rate for the underlying selection and shows a clutter-reduced representation. SemLens [HLTE11] is another lens technique for scatter plots, which assists local analysis by adding further analytical dimensions to certain regions of

[†] https://www.tableau.com

the scatter plot. A structure-based semantic lens for scatter plots and graph layouts is presented in [HTE11]. This lens technique keeps the selected data records –under the lens surface– unchanged and continuously deforms the data out of the focus in order to maintain the context around the lens. Bertini et al. [BRL09] have introduced an extended excentric labeling lens which dynamically displays labels to the selected data records around the lens. In [LT16], a data-dependent magic lens is presented to minimize the projection related distortions in Radviz and Star Coordinate visualizations. A survey on visual interactive lens and distortion-oriented presentation techniques are given in [TGK*14, LA94]

2.3. Delineation and Our Contribution

Previous works have defined useful methods to visualize correlations and features to rank and select scatter plot patterns. Also, several interactive lenses for scatter plot visualizations introduce distortion or sampling techniques to support the data exploration process by preserving an overview of the data during drill-down operations. Our approach is novel in that we extend the scatter plot lens concept to compute and visualize in interactive time candidate regression models for user-selected subsets of data. Our tool supports modeling of data by sets of local regression models, hence contributing a novel tool for scatter plot analysis.

3. Concept of the Regression Lens

Next, we describe basic concepts and issues of regression-based data analysis. We will then derive our interactive regression lens concept and its visualization building on these concepts.

3.1. Regression in Practice

We start with relevant background knowledge for univariate regression analysis of data: Given a 2D set of data elements $\mathbf{p}_i = (x_i\ y_i)^T; i = 1,\ldots,m$ with $\mathbf{P} = (\mathbf{p}_1\ \ldots\ \mathbf{p}_m)$. A regression model is a function $y = f(x)$ explaining the data elements $(x_i\ y_i)^T$ best. Figure 1 illustrates this regression model concept.

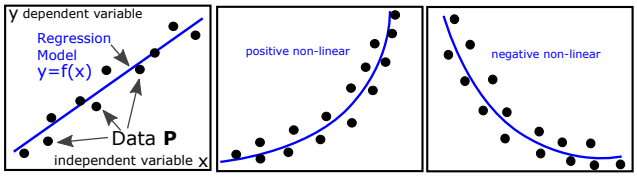


Figure 1: Regression models: (left) linear and (middle-right) non-linear models (blue) to explain the data (black).

The idea of univariate regression data analysis is to find a functional relation in the data. It allows to compare different data sets with each other, it supports to have options to compress or to classify the data, and it mutually binds two (or more) variables with each other, which were unrelated and independent before. Finding these phenomenological relations between variables/dimensions is the most relevant aspect of a regression-based data analysis.

Typically used univariate regression models are exponential models

$$y(x)_e = \beta_{e_1} \cdot e^{\beta_{e_2} x},$$

logarithmic regression models

$$y(x) = \beta_{l_1} + \beta_{l_2} \cdot \ln(x),$$

power fit regression models

$$y(x)_p = \beta_{p_1} \cdot x^{\beta_{p_2}},$$

or polynomial regression models of degree n

$$y(x)_n = \sum_{i=0}^{n} \beta_i \cdot x^i = \boldsymbol{\beta} \cdot \mathbf{x},\ \ \boldsymbol{\beta} = (\beta_0 \ldots \beta_n),\ \mathbf{x} = (x\ x^2 \ldots x^n), \quad (1)$$

where the parameters β_i are called regression coefficients, while x is the independent and y the dependent variable.

The polynomial regression model is appropriate to substitute or mimic a set of further models. In fact, the power fit $y(x)_p$ is a subset of $y(x)_{n=\beta_{p_2}}$, positive exponential models $y(x)_e$ can be approximated with $y(x)_3$, etc. Since a wide area of regression cases is covered, we focus on the family of polynomial regression models in this work. In this regard,

$$y(x)_1 = \sum_{i=0}^{1} \beta_i \cdot x^i = \beta_0 + \beta_1 \cdot x^1 \text{ is called a } \textit{linear model},$$

$$y(x)_2 = \sum_{i=0}^{2} \beta_i \cdot x^i = \beta_0 + \beta_1 \cdot x^1 + \beta_2 \cdot x^2 \text{is called a } \textit{quadratic model}, \text{ and}$$

$$y(x)_3 = \sum_{i=0}^{3} \beta_i \cdot x^i = \beta_0 + \beta_1 \cdot x^1 + \beta_2 \cdot x^2 + \beta_3 \cdot x^3 \text{ is called a } \textit{cubic model}.$$

The estimation of a good regression model comes with a set of issues to be handled, which are denoted as the issue of *underfitting vs. overfitting*, *independent vs. dependent* variable, and *uniformity vs. clumpiness*. Subsequently, we introduce and discuss these issues.

3.2. The Regression Issues

We describe common issues when using regression in practice.

Underfitting vs. Overfitting: A model is well chosen if it fits the data and if it is the simplest model for doing so. To fit the data, the in-sample error $e(f)$ for model f is

$$e(f) = \sum_{i=1}^{m} (y_i - f(x_i))^2, \quad (2)$$

also known as sum of squared errors (SSE). An appropriate model f minimizes $e(f)$ to prevent data underfitting. See Figure 2 (left-middle), where e is stressed by red lines.

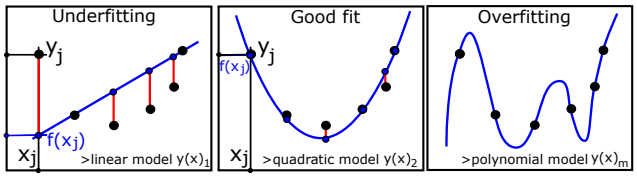


Figure 2: Underfitting vs. Overfitting

In contrast, choosing the simplest of such in-sample error minimizing models prevents overfitting: "simple" is such a model with a small complexity, e.g., a small number of regression coefficients β_i. In fact, considering a number of m data points, with $y_i(x_i) \neq y_j(x_j)$; there is always a polynomial $y(x)_n$ of degree m fitting the data perfectly (i.e., it interpolates the data), with $e(y(x)_{n=m}) = 0$ (see Figure 2 (right)). Does it mean that $y(x)_{n=m}$ is still the optimal model?

For obvious reasons, this is not the case: this model is not simple but complex with a large number of m parameters $\beta_i, i = 0,\ldots,m$; the model has a large waviness and thus it is also geometrically complex; and for each new data element, every time the model requires one new term and one more regression coefficient –which does not seem to be plausible at all– known as overfitting.

To prevent overfitting, a model $y(x)_n = f_n$ of degree n is assigned with the out-of-sample error $o(f_n)$ given as:

$$o(f_n) = \frac{1}{2} \cdot \sum_{j=0}^{m} (f_n^{\mathbf{P}_1}(x_j) - f_n^{\mathbf{P}_2}(x_j))^2, \tag{3}$$

where $\mathbf{P}_1, \mathbf{P}_2$ are subsets of $\mathbf{P}$. These subsets are mutually disjoint with $\mathbf{P}_1 \cap \mathbf{P}_2 = \emptyset$, have a similar number of elements –ideally each set has a number of $m/2$ data representatives, i.e., $\mathbf{P}_1 \cup \mathbf{P}_2 = \mathbf{P}$– and each subset should have a similar distribution behavior as $\mathbf{P}$ in order to mimic its statistical properties, making a comparison fair. When using more than two disjoint subsets, i.e. $\mathbf{P}_1, ..., \mathbf{P}_k$, the out-of-sample error o is given by averaging. Then, a good model f_n preventing overfitting minimizes $o(f_n)$, known as cross-validation. In total, an optimal model f is given by the optimization process for in-sample error e and out-sample error o as

$$f_n \text{ with } \underset{\boldsymbol{\beta},n}{arg\,min} \; (o(f_n) + e(f_n)). \tag{4}$$

Independent vs. Dependent Variable: The choice of the independent variable for model f is arbitrary. Clearly, both options are reasonable: either choosing $y = f_y(x)$ or choosing $x = f_x(y)$ as dependent or independent variable. A concept to describe the influence of the chosen independent variable is the *correlation*.

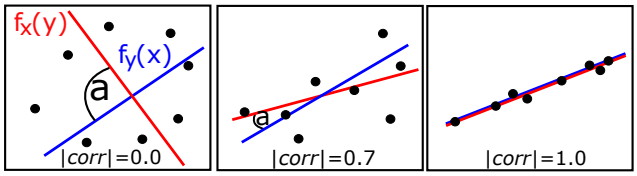


Figure 3: Correlation: influence of the chosen independent and dependent variable to the appearance of a linear regression model.

The correlation $corr(x,y)$ is a measure describing how unimportant the choice of direction is, so $f_y(x)$ or $f_x(y)$, because both choices lead to the same image of the function. If $|corr(x,y)| = 1$, it means that $f_y(x)$ and $f_x(y)$ look identical, while $|corr(x,y)| = 0$ means that both directions are unrelated and look different. For a linear model $y(x)_1$, the correlation is described by the angle α spanned in between $f_y(x)$ and $f_x(y)$, giving the correlation measure $corr_{y_1}(x,y) = cov(x,y)/(\sigma(x)\sigma(y))$, with the covariance cov and the standard deviation σ. So, if $\alpha = 0$ then $corr_{y_1}$ is 1, if $\alpha = \pi/2$ then $corr_{y_1} = 0$. Figure 3 illustrates this. A generalization of this correlation concept for higher order models is known by

$$corr(x,y)_{y_i} = \sqrt{1 - \frac{SSE}{SST}} = \sqrt{1 - \frac{e}{SST}} = \sqrt{1 - \frac{\sum_{j=1}^{m}(y_j - f(x_j))^2}{\sum_{j=1}^{m}(y_j - \bar{y})^2}}$$

with

$$\bar{y} = \frac{1}{m} \sum_{i=1}^{m} f(x_i).$$

The correlation measure is an important information in order to judge the quality of a regression model.

Uniformity vs. Clumpiness: One last issue is that the data along a good model ought to be uniformly distributed. In fact, a model may run through different clusters in the data, raising the question how good a model is that would connects clusters of data. Not quite good we argue, as (i) such a model is locally influenced by a varying information density (dense areas and sparse areas), which sophisticate the model, and as (ii) two or more clusters are generally not well described by one model (see Figure 4).

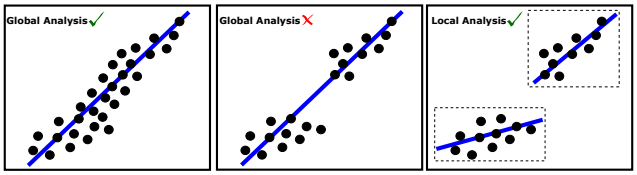


Figure 4: Uniformity vs. clumpiness or global vs. local analysis.

For two clusters, two models (one per cluster) seem to be a better choice in terms of fitting the data, which also motivates our concept of a local analysis with our regression lens. However, to measure the distribution of the involved data along the model's pathway, we define a distribution measure $h(f)$ of the non-uniformity for the data elements of model f as

$$h(f) = \frac{1}{2} \cdot (h_x + h_y) \tag{5}$$

where h_x and h_y are the deviations of the discrete uniform distribution $P(X = x_i) = \frac{1}{n}$ for $i \in \{1,...,n\}$. The deviation of the uniform distribution is defined by the goodness of fit measure, which is the sum of differences between observed and expected outcome frequencies of an interval i as

$$h_x = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{6}$$

where O_i is the number of observations in interval i and E_i is the expected number in interval i on the x-axis and y-axis respectively – known as χ^2-distribution. A model is considered to be more appropriate if it minimizes $h(f)$.

Note that our lens concept handles all these issues by using appropriate visualization concepts, allowing to observe and compare the quality and applicability of regression models for interactively selected data. The following section describes how our lens is constructed, handling issues of overfitting/underfitting, dependent/independent variables, and uniformity/clumpiness.

3.3. Construction of Regression Models for our Approach

By considering our data $(x_i\ y_i), i = 1,\ldots,m$, we define the $(n+1 \times m)$ power data matrix $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \ldots & x_m^n, \end{pmatrix} \tag{7}$$

and write the polynomial model $y(x)_n$ with $\mathbf{y} = (y_1\ y_2 \ldots y_m)^T$ as the linear system

$$\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{e} \tag{8}$$

for which the in-sample error $e(f) = \sum(y_i - f(x_i))^2 = \sum e_i^2$, $e_i \in \mathbf{e}$ is minimized by solving the linear system [Fre05], e.g., by choosing the regression coefficients $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{9}$$

Please note that the inverted matrix, $\mathbf{X}_{inv} = \mathbf{X}^T\mathbf{X}$, is a $(n+1) \times (n+1)$ matrix, is not growing with the number m of considered data but only with the degree n of the polynomial regression model $y(x)_n$. For instance, $\mathbf{X}_{inv}$ is a 2×2 for linear models $y(x)_1$, 3×3 for quadratic models $y(x)_2$, 4×4 for cubic models $y(x)_3$ etc. Thus, inverting $\mathbf{X}_{inv}$ and solving Eq. (9) is a cheap operation as long as the polynomial degree n is not too big. From our experience, a regression model fitting the data well has usually a degree less than $n \leq 5$.

To find an optimal model, our approach solves the optimization process of Eq. (4) within a two-step process.

Step 1: To minimize the in-sample error e, our technique considers a set of k different polynomial regression models $y(x)_1, \ldots, y(x)_k$ and ranks them regarding their minimized in-sample errors e, $e_1 < \cdots < e_k$. Figure 5 (left) illustrates the first step.

Step 2: From the ranked regression models, we consider a number k_T of the best ranked models as candidates and select the candidate as optimal which has the minimized out-sample error $o(f)$ in the list of candidates. To compute $o(f)$ for the different candidates, our approach uniformly samples the data $\mathbf{P}$ to get the subsets $\mathbf{P}_1$ and $\mathbf{P}_2$, as is seen in Figure 5 (right).

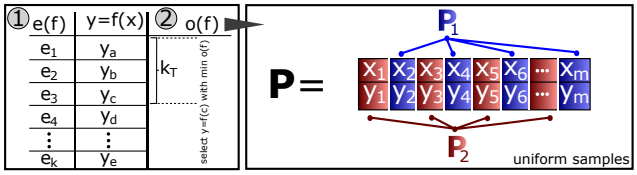


Figure 5: Two step process to find the best fitting regression model.

If $k = k_T$, the ranking process in step 1 is not required. For proof of concept, we do so by choosing $k = k_T = 4$, i.e., we consider only polynomial regression models up to degree 4, from which we choose the one with the smallest out-sample error as optimal.

Finally, for the purpose of later use, from the chosen model $y(x) = f_y(x)$ our approach calculates the correlation $corr(x,y)$, the model $f_x(y)$ by switching dependent & independent variables and uniformity properties, as described above. By having the regression model and the attributes, we are prepared to subsequently explain our visual design for the regression lens.

3.4. Visual Design of our Regression Lens

To select a subset of data elements for our regression lens, a box, circle, free-form, or a manual selection can be basically used, illustrated in Figure 6 (a-c). In that regard, a circle or free-form selection does not naturally yield axially parallel edges that are required to further visualize dimension-wise aligned (statistical) information. Moreover, a free-form selection causes an additional cognitive effort and a lot of further interaction steps, which may be exhausting. On

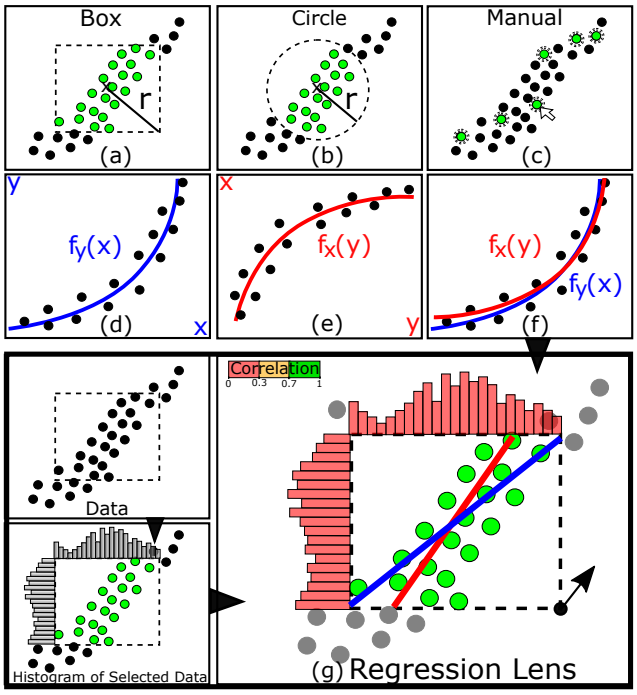


Figure 6: Visual Design of the Regression Lens.

the other side, a manual selection is too expensive and time consuming if a large number of data points need to be selected. We like to keep it simple for proof of concept. Thus, considering these reasons, we rely in this work on a rectangle selection by a simple user mouse drag operation. However, integration of further interaction schemes, if needed is straightforward to do.

With the rectangle selection scheme, the user selects a subset of data elements, to which regression analysis, quality computation and optionally, user guidance regarding improvement possibility is applied. Please note that by the choice of a rectangle, the "locality vs. globality" level of the analysis is implicitly user-defined.

As already explained, both variables x and y can be seen as a correct choice for the independent variable for the univariate polynomial regression, and both models $f_y(x)$ and $f_x(y)$ are correct in a way. This is also justified by the fact that a plot p_{xy} for any dimension x and y of the SPLOM is equivalent to the transposed version of the plot for the dimensions y and x, i.e., $p_{yx}^T = p_{xy}$ where variable x and y are interchanged. Consequently, an order of the variables does not exist by nature. Figure 6 (d,e) illustrates this. Thus, our approach needs to draw both found regression models in the lens selection area, to allow a complete insight for the users. While $f_y(x)$ can be drawn with standard techniques in the x-y-space, an inverse of $f_x(y)$ (which is given in y-x-space) does not necessarily exist. Thus, our approach exploits the symmetry that point (y,x) for $f_x(y)$ is equivalent to point (x,y) in the space of $f_y(x)$, to draw $f_x(y)$ in x-y-space: $(y,x)_{f_x(y)} \rightarrow (x,y)_{f_y(x)}$ (see Figure 6 (f)).

To judge distribution properties, our approach visualizes a normalized histogram for each variable on the boundaries of the selection box (see Figure 6 (g)). Since the boundaries of the selection box are axis-parallel and thus assigned to the variable directions of the plot, the histograms can easily be mentally connected with the variables. This also motivates to use a box selection instead of other options.

Due to the fact that the color for the histogram bins and the boundaries of the box are free usable visualization parameter, our approach maps correlation values to discrete colors. Specifically, $corr(x,y) < 0.3$ is mapped to red, $0.3 \leq corr(x,y) < 0.7$ is mapped to orange, and $0.7 \leq corr(x,y) < 1.0$ is mapped to green, to visually stress the level of correlation. Note that this is a pragmatic choice and other color mapping schemes, including continuous mappings, are possible in principle.

As part of our concept, both directions $f_y(x)$ and $f_x(y)$ of the respective optimal polynomial regression model are drawn, as well as both univariate histograms for x and y. In addition, we show the correlation information $corr(x,y)$, as can be seen in Figure 6 (g).

To distinguish the different directions of the drawn models, we color models of $f_y(x)$ in blue and models of $f_x(y)$ in red (shown in Figure 6 (d - f)). Furthermore, we provide another optional color coding to directly visualize the in-sample error on the models' pathway. By means of this visual feature, users can quickly identify the quality of the regression model according to the selected points. Therefore, we compute the Euclidean distance of the model's pathway to the nearest point and map the distance by using a diverging red-green color coding, as demonstrated in Figure 7(b) – bottom lens. Moreover, users can compare the in-sample errors of different models (e.g., linear vs. quadratic) and spot unsuitable parts on the models' pathway. This support users e.g., to split inappropriate selections (colored in red) into individual subsets for modeling.

4. System Overview & Selection Guidance

In this section, we present the design implementation of our regression lens concept and introduce a guidance component that supports finding appropriate lens selections. Figure 7 shows our design.

4.1. System Design

To analyze scatter plots from high-dimensional data sets, we present all pairwise combinations of dimension variables in a SPLOM, as shown in Figure 7(a). Data points which belong to a particular class label (if available) are visualized by different colors and can be filtered out for further investigation. The user selects one cell (plot) from the SPLOM which is shown in detail in (b). The result of regression analysis for interactively selected data subsets is shown directly on the lens in this view (Figure 7(b)). Individual settings for local lenses, such as model selection, activating distribution histograms or class label filtering, can be performed in the setting view (Figure 7(c)). Moreover, the user can save interesting findings and previous settings of lenses in this view. Detailed information about the current lens selection is shown in Figure 7(d). This information includes selected points, boundaries of the area and statistical measures like $corr(x,y)$, $e(f)$ and $h(f)$ values.

For specific analytical tasks, users can limit the degree of the model and manually switch between the polynomial models as well as the direction of the model $f_y(x)$ and $f_x(y)$. Alternatively, users can let the system choose the best fitted model and the direction according to the selected points. If both directions are simultaneously drawn, the system automatically reduces the saturation of the less fitting direction to highlight the better model, as shown in Figure 7(b). To measure the best model and direction respectively, the $e(f)$ values of each combination can be compared.

Since the computation of the distribution measure $h(f)$ depends on the histogram bin size, we not only allow users to adjust this parameter setting but also provide an automatic selection based on the equal-width binning approach. For automatic selection, we determined the bin size by the square-root choice that takes the square root of the number of samples in the lens. By this way, the comparison of different local lenses is not influenced by the sizes of the lens.

By default, we color-code points selected inside the lens box in green. The cross-validation subset is coded in dark blue, and the regression models in red and blue, respectively. All color-schemes, for the figures in this paper and the application, are taken from Colorbrewer [HB03]. These color settings can be changed individually, e.g., to circumvent color perception disabilities, if present.

4.2. Guidance Concept

One major advantage of our regression lens approach is that it offers users the possibility to freely define an area of interest for the regression analysis. Mouse interactions for moving and resizing the lens selection facilitates an exploratory analysis procedure and creates the desired lens effect with real-time feedback of regression models. Thus, users will be able to exclude specific data points, such as clusters or points which belong to a particular class, which may have negative influence on the statistical computation. However, finding a good selection area for the lens can be a challenging task, given the many possible positions and sizes of a regression lens.

During the development of our regression lens approach, we demonstrated its functionality to a smaller number of members of our research group, and invited them to informally test the system and specifically, comment on the interactivity of the lens operation. During these tests, we observed that often after a lens position was found, the researchers applied small local repositioning of the lens, to see if the chosen model would change noticeably or not. This observation inspired us to include an automatic search step that mimics this user behavior, with the goal of improving the local regression model by small changes and offloading the user from fine-grained selection tasks. Specifically, after a lens is dropped by the user, we apply tentative horizontal and vertical translations of 5% of plot width and height, respectively, and test if the regression quality is improving as measured via the in-sample error $e(f)$ for any of these translations. Note that the selection of 5% is a parameter set heuristically and can be easily adapted to user requirements. We visually indicate the potential for improvement in model precision, thus guiding the user through the space of local regression models. To inform the user about improving directions, we provide a visual hint in terms of an arrow that points to the most significant direction of improvement (if given). This procedure optionally applies after each interactive adjustment of the lens, and thus creates an iterative feedback loop that helps users to find better fitting models.

Furthermore, we include an outlier detection to our guidance approach for indicating the points that may negatively influence the model computation. The used outlier detection is a distance-based approach that considers for each point selected by the given lens, the sum of the distances from its k-nearest neighbors as outlier score [AP02]. Points are detected as outliers if their outlier score is larger than three times the average distance of all points to its k-nearest neighbors. For a fast and smooth computation, we auto-

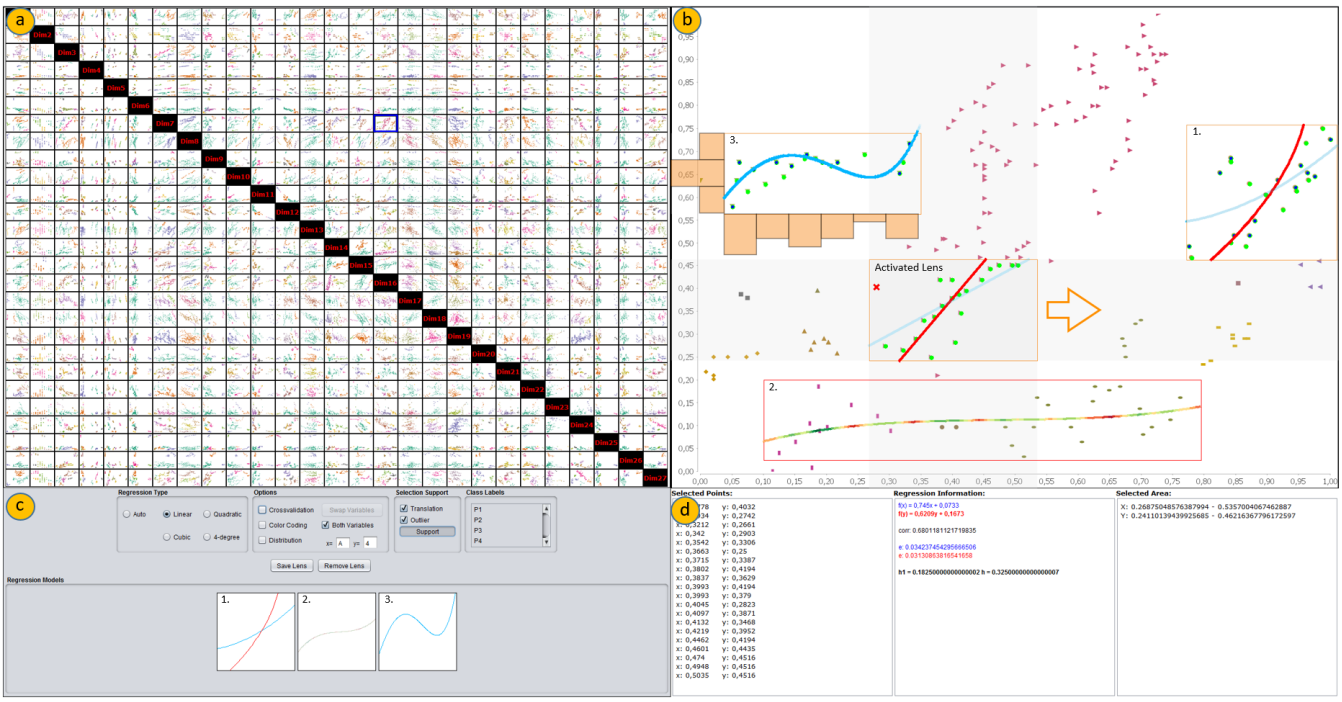


Figure 7: Our prototype is separated into four views: (a) visualizes multivariate data by means of a SPLOM; (b) is the interactive analysis area for the investigation of local regression models for a chosen cell from the SPLOM; (c) is a settings window to control regression computation; (d) shows additional information like selected points, regression coefficients and correlation measure for the current selection to be investigated.

matically set k by using the heuristic $\sqrt{\frac{n}{2}}$ with n being the number of selected points.

A demonstration of this guidance concept is depicted in Figure 7(b) – activated lens in the middle. In this case, the guidance concept suggests to move the lens in right direction to improve the current $e(f)$ value of the linear regression model $f_x(y)$ (colored in red) from $e(f) = 0.031$ to $e(f) = 0.021$. If we take a closer look at the current selection, one can see that this selection involves an outlier (indicated as red cross) that impairs the linear regression in this example. By moving the lens to the right, this outlier gets excluded from the selection, resulting in the above mentioned improvement.

4.3. Implementation Details

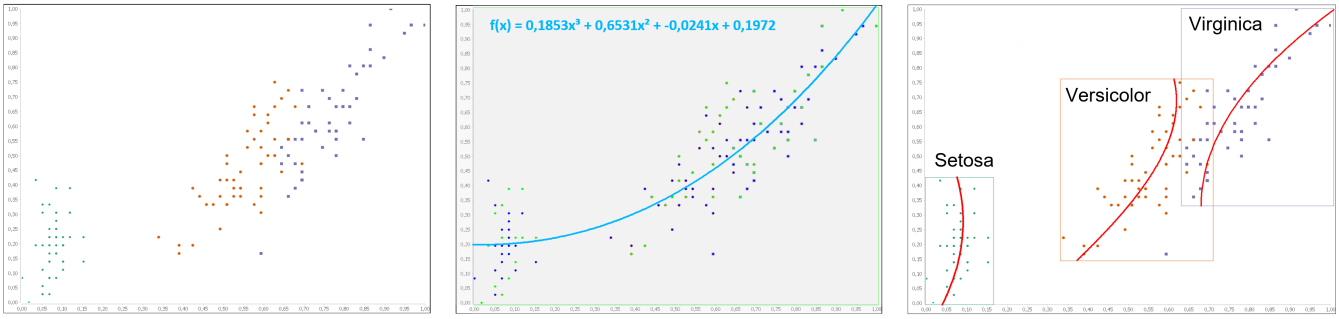
We implemented the user interface of the regression lens in Java, and integrated an R-Environment for the statistical computations. To render scatter plots, the JFreeChart library is used. To provide a smooth exploration, we implemented our system using two threads, which separate foreground and background computations. The foreground thread handles user interactions and display updates. The background thread translates screen coordinates to plot coordinates, calls R to compute the candidate regressions and other needed information. For computing the regression models, we used the standard linear models function `lm()` in R.

5. Case Study

Next, we demonstrate the usability of our approach and evaluate the different regression issues by using well-known data sets from the UCI Machine Learning Repository [Lic13].

Global vs. Local Analysis: One primary goal is to enable an interactive exploration for local regression models in scatter plot data. Figure 8 exemplifies this intention with the help of the Iris data set and shows an interactive outcome compared to a general global regression analysis. The first figure on the left (Figure 8a) shows the distribution of petal length against sepal length, and highlights the different iris species (setosa in green, versicolor in orange and virginica in purple). The plot shows clear separated patterns of the different classes. However, if we apply the regression lens over the whole data (Figure 8b), it returns a cubic model $f_y(x) = 0.185x^3 + 0.653x^2 + 0.024x + 0.197$ as best fitting model (cf. Section 3.3). By interactively positioning three local regression lenses for the different classes (Figure 8c), one can see that none of the three models has a similar trend compared to the global one. In fact, the main directions of the models have changed to the direction $f_x(y)$ and describe different models for the three classes.

Underfitting vs. Overfitting: Next, we consider the underfitting/overfitting issue with respect to our regression lens concept. Figure 9 demonstrates the effect of including and excluding cross-validation for the regression computation. In the background, the actual lens selection is shown, which at first glance seems to capture only little data. Actually, the selection contains 49 data points which are mostly overdrawn due to their similarity. The data point framed in red actually contains 30 overdrawn points and highly influences the regression model computation. This is the reason why the lens tries to match this point in a very precise manner in the previous example (Overfitting). By activating cross-validation, we decrease the weighting of this single area and receive a quadratic model, which seems to be more suitable for this selection.



(a) Iris data set – petal length/sepal length. (b) Global regression analysis. (c) Local regression analysis.

Figure 8: Comparison between global and local regression analysis with the Iris data set. (a) Shows a scatter plot that visualizes petal length as independent variable (x) and petal length as dependent variable (y). The color coding denotes the different class labels of the data (species). (b) By applying a usual global regression analysis, we obtain a cubic function as best fitting model according to the test data subset (colored in dark blue). (c) Shows the result of local regression analysis for the different flower species setosa, versicolor and virginica.

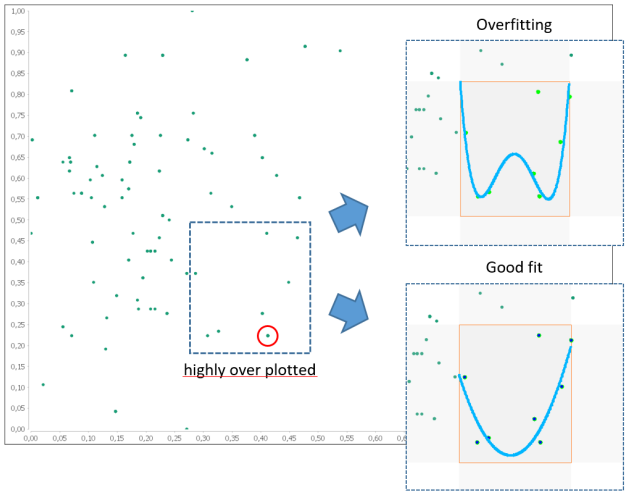


Figure 9: Overplotting issue – Bosting Housing data set: By using the raw selected data points our regression lens returns a polynomial model of degree 4 as best fitted model (Overfitting). However, if we include the out-of-sample error into the regression determination, it changes the model to a quadratic model (Good fit).

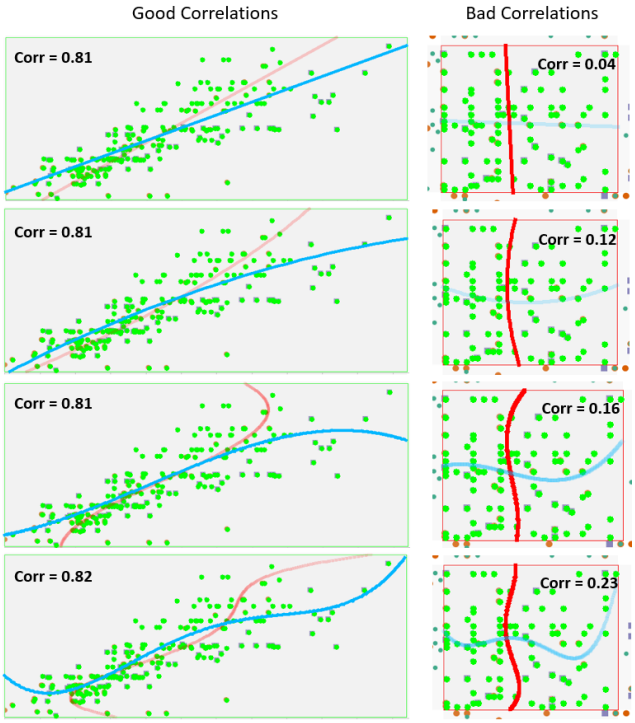


Figure 10: Independent vs. Dependent Variable – Auto MPG data set: Impact of correlation measure demonstrated on extreme examples for each polynomial model. Good correlation here means that the choice of direction is unimportant, whereas a bad correlation indicates that the models show to be very different.

Independent vs. Dependent Variable: Since our approach provides two different directions ($f_y(x)$ and $f_x(y)$) for each polynomial model, the number of possible models increase. To reveal how important the choice of a direction is, the correlation measure $corr(x,y)$ as defined in Section 3.2 is used. Figure 10 depicts the information content of the correlation measure. It shows the extreme examples for each model by using the Auto MPG data set. On the left, we compare a local area of the plot car weight against displacement, which has a strong positive correlation. In this case, it is obvious that it does not play a major role in which direction the user is going to choose, since all models describe the positive trend very well. Moreover, one can see that the $corr(x,y)$ value stays stable for each model and remains at around 0.81. On the other hand, if the correlation is weak, the choice of direction may influence the analysis process, or worse, lead to improper hypotheses.

Uniformity vs. Clumpiness: The last example covers the issue of finding uniformly distributed selections for the lens. To judge a selection, the user can compare the quality by the distribution measure $h(f)$ and the shown normalized histograms on the variable axes of the regression lens. Figure 11 exemplifies this functionality on the Wine data set. In the background, we show a bad selection example of two different wine clusters (class 1 and class 3). This is indicated by the relatively high $h(f)$ value of 0.65 and the unequal distributed histograms on the variable axes (i.e., two peaks on the x and y axis). To improve the selection, the user can gradually downsize the selection box and compare visual and computational

improvements on the histograms and $h(f)$ value. Gaps along the histogram axes are good split indicators. If we split the two clusters (red and blue selection), one can see an improvement in the results. The red selection examples show an iterative improvement of the $h(f)$ value by minimizing the height of the box. In the end, we improve over the initial single selection (total distribution measure $h(f) = 0.65$) to a two-fold sub-selection (with $h(f) = 0.45$ and $h(f) = 0.44$ distribution values).

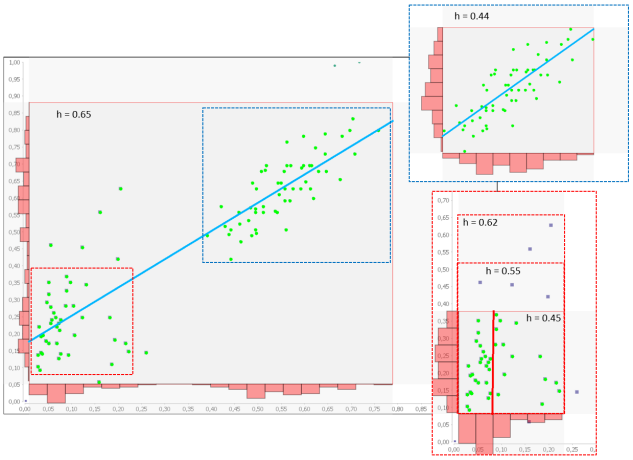


Figure 11: Uniformity issue – Wine data set: Lens selections can be verified by comparing the axes histograms (visually) or by the $h(f)$ measure (analytically). In the background a bad selection with a relatively high h value is shown, which can be improved by using two separated lenses for the two clusters (cutouts on the right).

6. Discussion and Extension Possibilities

Our regression lens concept allows to define and compare models of user-selectable locality. Local modeling involves the segmentation of data which is typically a hard problem to do automatically, since often parameter settings are required that do not fit for all data or user interests. Our lens approach allows to easily factor in user background knowledge. Compared to fully-automatic analysis approaches, our lens technique provides the degree of freedom for exploring local patterns in classified or unclassified data. It can complement automatic approaches of searching for local regression models, which typically require information about clusters or class labels, or need to specify an automatic data segmentation step.

We use a set of plausible statistical scores to select regression models. A possible extension is to include additional scores or domain-dependent quality measures. For example, one could use Scagnostics features [WAG05] for quantification of patterns. Features like clumpiness or monotonicity can be integrated for validating user selection, and thus help to identify good local selections.

Another important aspect to be considered is the scalability of our approach with respect to the number of selected data items. Therefore, we empirically evaluated the performance regarding different models and data samples. The evaluation was performed on a notebook with an Intel i7-6500U CPU and 16 GB RAM, the results are shown in Table 1. The tested sample size includes 100, 1.000 and 10.000 data points. We observe that the response time increases by taking models of higher degrees and, of course, by increasing the data size. Since the automatic lens selection computes all models to find the best fitting model, its response time is accordingly the longest. Our implementation can process lens selections up to 1.000 data items for a pre-selected model at response times in the range of 100-200 milliseconds, which can be considered fully interactive. For larger data size and automatic model selection, we observe response times between 0.7 and 22 seconds. A speedup may be achieved by data sampling or increasing implementation efficiency.

	100	1000	10000
Linear	43ms	104ms	1941ms
Quadratic	30ms	152ms	4189ms
Cubic	31ms	220ms	7287ms
Degree 4	44ms	294ms	10589ms
Automatic	83ms	738ms	22193ms

Table 1: Computation time for the regression models.

A recurring problem in data analysis is the issue of dealing with outliers, a well-known problem in practice. Approaches exist to detect and handle outliers in data analysis. We provide a guidance approach including a simple outlier detection to indicate selected points that may negatively influence the model computation. More advanced or domain-specific outlier detection methods can be thought of. Our guidance uses a translation model with a predefined shift size to suggest local repositioning of the lens. More advanced guidance may involve more exhaustive search over the translation, or other transformations of the lens selection like rotation.

It will also be interesting to devise methods for higher-dimensional regression. To this end, definition of variable and data selection is expected to become more difficult. A first idea is to define a process by which the user incrementally adds more variables. Then, we note our approach shows a smaller number of models in-place in the lens. We may also think about using comparative visualization to compare more models against each other. Finally, we note that modeling with the regression lens is an interactive process. It may be useful to record the interaction operations or intermediate models considered by the user. This sequence of operations or models could be shown using provenance techniques, to make plausible how a particular choice of models was obtained by an analyst.

7. Conclusion

We introduced regression lens, a visual-interactive approach for the exploration of global and local regression models in scatter plot data. This approach allows users to interactively select a portion of data on which the regression computation is run. It provides statistical measures and visual feedback features to judge the quality of a given selection as well as the output model. We introduced a guidance concept that supports the interactive process of finding well distributed selections based on the in-sample error of slight translations. Furthermore, we pointed out and evaluated important analysis issues for our regression lens concept, and demonstrated the applicability and benefits of our approach with use cases on example data sets. We also discussed several extension possibilities.

Acknowledgment

We thank Dieter Schmalstieg of Graz University of Technology for valuable comments made to an earlier version of the regression lens approach.

References

[Ans73] ANSCOMBE F. J.: Graphs in statistical analysis. *The American Statistician 27*, 1 (1973), 17–21. 2

[AP02] ANGIULLI F., PIZZUTI C.: Fast outlier detection in high dimensional spaces. In *Proc. of the 6th European Conference on Principles of Data Mining and Knowledge Discovery* (2002), pp. 15–26. 6

[BC87] BECKER R., CLEVELAND W.: Brushing scatterplots. *Technometrics 29, 2* (1987), 127–142. 2

[BRL09] BERTINI E., RIGAMONTI M., LALANNE D.: Extended excentric labeling. In *Proceedings of the 11th Eurographics / IEEE - VGTC Conference on Visualization* (2009), pp. 927–934. 3

[BSP*93] BIER E. A., STONE M. C., PIER K., BUXTON W., DEROSE T. D.: Toolglass and magic lenses: The see-through interface. In *Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques* (1993), ACM, pp. 73–80. 2

[CCM10] CHAN Y.-H., CORREA C., MA K.-A.: Flow-based scatterplots for sensitivity analysis. *Proceedings of IEEE VAST Symposium* (2010). 2

[CCM*14] CHEN H., CHEN W., MEI H., LIU Z., ZHOU K., CHEN W., GU W., MA K. L.: Visual abstraction and exploration of multi-class scatterplots. *IEEE Transactions on Visualization and Computer Graphics 20*, 12 (Dec 2014), 1683–1692. 2

[CLN86] CARR D. B., LITTLEFIELD R. J., NICHLOSON W. L.: Scatterplot matrix techniques for large n. *Proceedings of the Seventeenth Symposium on the Interface of Computer Sciences and Statistics on Computer Science and Statistics* (1986), 297–306. 2

[EBD05] ELLIS G., BERTINI E., DIX A.: The sampling lens: Making sense of saturated visualisations. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems* (2005), ACM, pp. 1351–1354. 2

[Fre05] FREEDMAN D.: *Statistical Models: Theory and Practice*. Cambridge University Press, 2005. 5

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

[HB03] HARROWER M., BREWER C. A.: Colorbrewer.org: An online tool for selecting colour schemes for maps. *The Cartographic Journal 40*, 1 (2003), 27–37. 6

[HGM*97] HOFFMAN P., GRINSTEIN G., MARX K., GROSSE I., STANLEY E.: Dna visual and analytic data mining. In *Proceedings of the 8th Conference on Visualization '97* (1997), IEEE Computer Society Press, pp. 437–ff. 2

[HLTE11] HEIM P., LOHMANN S., TSENDRAGCHAA D., ERTL T.: Semlens: Visual analysis of semantic data with scatter plots and semantic lenses. In *Proceedings of the 7th International Conference on Semantic Systems* (2011), ACM, pp. 175–178. 2

[HTE11] HURTER C., TELEA A., ERSOY O.: Moleview: An attribute and structure-based semantic lens for large element-based plots. *IEEE Transactions on Visualization and Computer Graphics 17*, 12 (Dec 2011), 2600–2609. 3

[JDK10] JUSUFI I., DINGJIE Y., KERREN A.: The network lens: Interactive exploration of multivariate networks using visual filtering. In *14th International Conference Information Visualisation* (2010), pp. 35–42. 2

[JSG16] J. STAIB S. G., GUMHOLD S.: Enhancing scatterplots with multidimensional focal blur. *Computer Graphics Forum (In Proc. EuroVis)* (2016). 1, 2

[Kan01] KANDOGAN E.: Visualizing multi-dimensional clusters, trends, and outliers using star coordinates. *In Proceedings of the seventh ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (2001). 2

[LA94] LEUNG Y. K., APPERLEY M. D.: A review and taxonomy of distortion-oriented presentation techniques. *ACM Trans. Comput.-Hum. Interact. 1*, 2 (June 1994), 126–160. 3

[LHJ01] LAMAR E., HAMANN B., JOY K. I.: A magnification lens for interactive volume visualization. In *Proceedings Ninth Pacific Conference on Computer Graphics and Applications* (2001), pp. 223–232. 2

[Lic13] LICHMAN M.: UCI machine learning repository, 2013. URL: http://archive.ics.uci.edu/ml. 7

[LKZ*15] LEHMANN D., KEMMLER F., ZHYHALAVA T., KIRSCHKE M., THEISEL H.: Visualnostics: Visual guidance pictograms for analyzing projections of high-dimensional data. *ComputerGraphicsForum(Proc.EuroVis) 34*, 3 (2015). 2

[LMvW10] LI J., MARTENS J.-B., VAN WIJK J. J.: Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization 9*, 1 (Mar. 2010), 13–30. 2

[LT16] LEHMANN D. J., THEISEL H.: General projective maps for multidimensional data projection. *Computer Graphics Forum (Proc. Eurographics) 35*, 2 (2016). 3

[MG13] MAYORGA A., GLEICHER M.: Splatterplots: Overcoming overdraw in scatter plots. *IEEE Transactions on Visualization and Computer Graphics 19*, 9 (Sept. 2013), 1526–1538. 1, 2

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

[Nv06] NOVÁKOVÁ L., ŠTEPÁNKOVÁ O.: Multidimensional clusters in radviz. In *Proceedings of the 6th WSEAS International Conference on Simulation, Modelling and Optimization* (2006), pp. 470–475. 2

[Nv09] NOVÁKOVÁ L., ŠTĚPÁNKOVÁ O.: Visualization of trends using radviz. In *ISMIS '09: Proceedings of the 18th International Symposium on Foundations of Intelligent Systems* (2009), pp. 56–65. 2

[SBS11] SCHERER M., BERNARD J., SCHRECK T.: Retrieval and exploratory search in multivariate research data repositories using regressional features. In *Proc. of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (2011), pp. 363–372. 1, 2

[SBS*14] SHAO L., BEHRISCH M., SCHRECK T., VON LANDESBERGER T., SCHERER M., BREMM S., KEIM D. A.: Guided sketching for visual search and exploration in large scatter plot spaces. *Proc. EuroVA International Workshop on Visual Analytics* (2014). 2

[SSB*16] SHAO L., SCHLEICHER T., BEHRISCH M., SCHRECK T., SIPIRAN I., KEIM D. A.: Guiding the exploration of scatter plot data using motif-based interest measures. *Journal of Visual Languages & Computing 36* (2016), 1 – 12. 1, 2

[TGK*14] TOMINSKI C., GLADISCH S., KISTER U., DACHSELT R., SCHUMANN H.: A Survey on Interactive Lenses in Visualization. In *EuroVis - STARs* (2014), Borgo R., Maciejewski R., Viola I., (Eds.), The Eurographics Association. 3

[TSAA12] TOMINSKI C., SCHUMANN H., ANDRIENKO G., ANDRIENKO N.: Stacking-based visualization of trajectory attribute data. *IEEE Transactions on Visualization and Computer Graphics 18*, 12 (Dec 2012), 2565–2574. 2

[WAG05] WILKINSON L., ANAND A., GROSSMAN R.: Graph-theoretic scagnostics. In *Proceedings of the IEEE Symposium on Information Visualization* (Oct 2005), pp. 157–164. 1, 2, 9

[WSZRD02] WAGNER A. K., SOUMERAI S. B., ZHANG F., ROSS-DEGNAN D.: Segmented regression analysis of interrupted time series studies in medication use research. *Journal of Clinical Pharmacy and Therapeutics 27*, 4 (2002), 299–309. 1

[WY04] WARD M., YANG J.: Interaction spaces in data and information visualization. In *Proceedings of the Sixth Joint Eurographics - IEEE TCVG Conference on Visualization* (2004), Eurographics Association, pp. 137–146. 2

[WZMK05] WANG L., ZHAO Y., MUELLER K., KAUFMAN A.: The magic volume lens: an interactive focus+context technique for volume rendering. In *VIS 05. IEEE Visualization* (Oct 2005), pp. 367–374. 2