Eurographics Conference on Visualization (EuroVis) 2019
M. Gleicher, H. Leitte, and I. Viola
(Guest Editors)

Volume 38 (2019), Number 3

ClustMe: A Visual Quality Measure for Ranking Monochrome Scatterplots based on Cluster Patterns

Mostafa M. Abbas[1], Michaël Aupetit[1], Michael Sedlmair[2], and Halima Bensmail[1]

[1] QCRI, HBKU, Doha, Qatar
[2] VISUS, University of Stuttgart, Germany

Abstract

We propose ClustMe, a new visual quality measure to rank monochrome scatterplots based on cluster patterns. ClustMe is based on data collected from a human-subjects study, in which 34 *participants judged synthetically generated cluster patterns in* 1000 *scatterplots. We generated these patterns by carefully varying the free parameters of a simple Gaussian Mixture Model with two components. and asked the participants to count the number of clusters they could see (1 or more than 1). Based on the results, we form ClustMe by selecting the model that best predicts these human judgments among* 7 *different state-of-the-art merging techniques (*DEMP*). To quantitatively evaluate ClustMe, we conducted a second study, in which* 31 *human subjects ranked 435 pairs of scatterplots of real and synthetic data in terms of cluster patterns complexity. We use this data to compare ClustMe's performance to 4 other state-of-the-art clustering measures, including the well-known Clumpiness scagnostics. We found that of all measures, ClustMe is in strongest agreement with the human rankings.*

CCS Concepts

• ***Human-centered computing*** → ***Visual analytics; Empirical studies in visualization;*** • ***Computing methodologies*** → ***Cluster analysis; Mixture modeling;***

1 Introduction

In visual analytics, users often need to analyze many different views of multidimensional data with the goal to identify interesting "patterns" and unforeseen relations. The most typical way to do so is by using scatterplots, in which the axes are either defined by pairs of the original data dimensions, or as new dimensions synthesized by some dimension reduction technique [SMT13] [NA18].

During the visual analytics process, hundreds of such scatterplots might be generated, and the user needs to visually skim through them one-by-one looking for visual patterns. When an interesting pattern is spot, often a more thorough analysis follows, with advanced interaction and connections with other views or knowledge to explain the identified pattern. The initial skimming task is an *exploration* task as per Brehmer and Munzner's typology [BM13]. It is a perceptual task focusing on patterns without interpretation, but it is tedious when the number of views to be skimmed gets large. To better support this process, the visualization community has proposed various Visual Quality Measures (VQM) [BTK11, BBK*18] that help guiding the users towards perceptually interesting patterns. To do so, VQMs quantify certain visual patterns, which then allow ranking visualizations based on their potential interest for the user. Time and cognitive effort can be saved for later analytical tasks. According to Bertini *et al.* [BS06], VQMs should thus ideally be based on perceptual models rather than heuristics and computational approaches. A perceptual VQM is then a measure that imitates how humans would score views based on perceived visual patterns and that can be used to accurately predict human perceptual judgments in the skimming task.

There has been a considerable amount of work on computational and perceptual VQMs [BBK*18], including some which characterize different patterns in color-coded [SA15, AS16] and monochrome scatterplots [WAG05, MTL18]. So far, however, there has been very little work on modeling the human perception of grouping patterns in monochrome scatterplots and using these models to design VQMs. This is surprising as it is a very typical task in visual analytics [BSIM14] as these groups in 2D scatterplots can reveal important relationships between multidimensional data or dimensions.

In this work, we set out to fill this gap by developing for the first time a VQM based on perceptual data, to rank monochrome scatterplots depending on their grouping patterns. To do so, we collected perceptual data from a human subject experiment and selected the best clustering model of this data to form a perception-based VQM of grouping patterns. We thus call our new VQM: ClustMe (short for Clustering Measure).

Closest to our work is the Clumpiness Scagnostic [TT85, WAG05]. However, Clumpiness is not based on perceptual data;

hence, it is unclear how well it matches with human judgments. Also, in theory any clustering technique could be used to form a computational VQM and to rank monochrome scatterplots. However, most clustering techniques [XW05] come with certain assumptions and parameters that need to be set by the user. In the end, such a process would be tedious and require in-depth knowledge on clustering techniques from the user.

To overcome these issues, we propose the new ClustMe measure, which is based on three contributions:

(1) In the Experiment 1, we collected 34000 human judgments from 34 subjects tasked to count clusters in 1000 monochrome scatterplots. Following a generative approach [SNE*16], the scatterplots were generated from systematically varying parameters of a mixture of two Gaussian distributions. This allowed us to precisely control the visual appearance and collect low-level perceptual judgments [SA15], fulfilling the *precision* criterion described by the McGrath's model of research methodology [Mcg95].

(2) Through modeling these data, we developed the novel ClustMe VQM as a Gaussian Mixture Model (GMM) [BC96, BCRR97, FR02] trained using the Bayesian Information Criterion (BIC) [Sch78] and refined with the *Demp* merging technique [Hen10]. We arrived at ClustMe, by systematically exploring how well different merging techniques model the data collected in Experiment 1. In doing so, we bridge between the perception of simple, measurable mixture of two Gaussian clusters from Experiment 1, and complex cluster structures that can be found in the real world and modeled by a GMM plus a merging technique.

(3) In the Experiment 2, we collected 13485 human judgments from 31 subjects ranking 435 pairs of monochrome scatterplots of real and synthetic data. We compare these human-made rankings with ClustMe, with the Clumpiness Scagnostic [TT85, WAG05], and with 3 standard clustering techniques: X-means [PM00], CLIQUE [AGGR98] and DBSCAN [EKSX96]. The main goal of this study was to evaluate how ClustMe and other approaches perform under realistic conditions, fulfilling the *realism* and *generalizability* criteria of the McGrath's model [Mcg95]. The main results show that ClustMe best agrees with the human raters.

2 Related Work

We discuss related work on visual quality measures based on perceptual data, as well as on visual and algorithmic clustering techniques.

2.1 Visual Quality Measures based on perceptual data

Many *empirical studies* on perceived patterns in scatterplots have been conducted. Closest to our work are studies regarding monochrome cluster patterns. Here, a statistical model of cluster perception in scatterplots based on proximity, density change, and concentration of the points has been proposed [Sad97], based on a controlled study involving 36 subjects and 24 stimuli. Another controlled study explored the perceived principal axis of a set of points when manipulating the separation between a small subset of points and a main ellipsoidal cluster [CSM08]. The human perception of homogeneous dot patterns has been studied varying densities and gaps between two square-shaped clusters [O'C74]. The study involved 8 subjects and 24 patterns. Human perception of the number of dots in scatterplots is also evaluated to analyze how these patterns are perceived as texture shapes or single dots depending on the density [ACB16]. All these studies investigate a small sample of possible cluster shapes, and the data was not collected to serve the design of visual quality measures of cluster patterns.

The survey of Bertini *et al.* [BTK11] gives an overview over *Visual Quality Measures (VQMs)*. VQMs build on research in image quality metrics [LK11] but are more specifically designed to quantify patterns in data visualization. They can support human exploration by automatically filtering, ranking, and mapping visual encodings based on patterns that might be of interest to a human observer [TMF*12, DW14]. Most VQMs are based on heuristics, but a recent trend focuses on designing measures based on perceptual studies. A study by Pandey *et al.* [PKF*16] explored how human subjects perceive and name different patterns in monochrome scatterplots, giving evidence that VQMs should align with visual perception to reproduce human judgments accurately, as it was posited by Bertini *et al.* [BS06, BTK11]. An expansive taxonomy and qualitative analysis more specific to grouping patterns in color-coded and monochrome scatterplots gives insight on the many factors at play in cluster perception [STMT12]. Beyond the number of clusters, other factors are proposed like the shape of the groups, their relative size or their density to name a few. This taxonomy can guide the design of a visual quality measure based on perceptual data for cluster patterns. The first line of work to create such perception-based measures was on correlation in scatterplots [RB10, HYFC14, KH16] and parallel coordinates [LMvW10]. Matute *et al.* [MTL18] proposed a skeletonization process which captures both shape and orientation of the scatterplot to measure similarity between scatterplots, and evaluate it with a human subject experiment.

A more *data-driven approach* was proposed by Demiralp *et al.* [DBH14]. From human subjects, they learned the perceptual similarity between marks with different shape, size and colors and applied this similarity measure to support the choice of mark design in scatterplots. Albuquerque *et al.* [AEM11] describe a perception-based metric derived from pairwise comparison judgment of different reference views by human subjects, and an eigenface decomposition of the images to build a perceptual space based on the reference views. A new view can be projected in the same perceptual space and its similarity to the reference views can be measured. It has been tested on a correlation task in monochrome scatterplots and a class-separation task in color-coded scatterplots. Sedlmair and Aupetit [SA15] proposed a machine learning framework to use data collected from perceptual human subject studies to select computational VQMs that best match the human visual perception. Based on this framework, they designed 2002 new measures optimizing the match between patterns and human perception [AS16]. In both cases, these VQMs were designed to quantify class separation in color-coded scatterplots. In this work, we also follow a data-driven process, but focus on quantifying the grouping patterns in monochrome scatterplots [BSIM14].

2.2 Quantifying Clusters in Scatterplots

We found only two VQMs that directly quantify (monochrome) grouping patterns. Most closely to our work is Clumpiness [TT85]

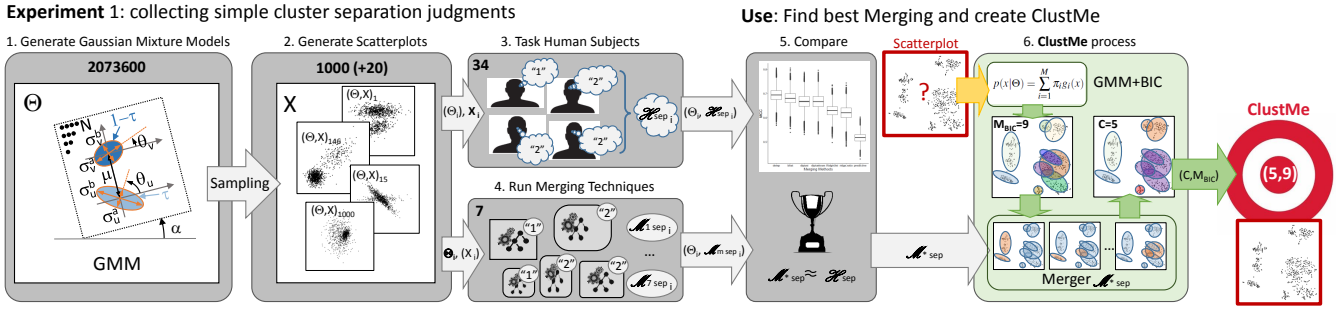

Figure 1: ***Experiment 1****: (1) The space of possible grouping patterns* Θ *is defined by a bivariate Gaussian Mixture Model (GMM) with* 2 *components. (2)* 1000 *scatterplots* X_i *are generated by the GMM* Θ_i*. (3)* 34 *human subjects* $\mathcal{H}$ *are asked to judge if they see* 1 *or more than* 1 *cluster for each* X_i *giving a probability of separation* $\mathcal{H}^i_{sep}$*. (4)* 7 *merging techniques* $\mathcal{M}_m$ *are run to evaluate each scatterplot* (Θ_i, X_i) *giving a separation decision* $\mathcal{M}^i_{sep}$*. (5) Both, human and machine decisions are compared to find the merging technique* $\mathcal{M}^*_{sep}$ *that best matches the human judgment with the Mathew Correlation Coefficient and the Vanbelle Kappa index. The result of this process is our VQM* ***ClustMe****. (6) ClustMe can then be used to automatically quantify grouping patterns in new scatterplots (red frame): it runs a GMM for which we find the optimal number of components* M_{BIC} *according to the Bayesian Information Criterion (BIC). Then, we use the best merger* $\mathcal{M}^*_{sep}$ *approximating the human perception to decide about merging pairs of the* M_{BIC} *components to get the final number of clusters C. The pair* (C, M_{BIC}) *defines the ClustMe Visual Quality Measure of the scatterplot. ClustMe VQM allows ranking scatterplots based on the complexity of their grouping patterns.*

first proposed by Tukey in 1985. It is a computational VQM part of the Scagnostics framework proposed by Wilkinson *et al.* [WAG05]. Clumpiness relies on a statistic of edge lengths of a minimum spanning tree of the data points. It is *parameter-free* and thus does not require hand-tuning of parameters. But it was not intended originally to resemble human perception. Hence, it has limitations in the kind of visual grouping patterns it can detect, and is often not consistent with the human perception of clusters (see Supplementary Material Figure 3).

CLIQUE is a *heuristic*, density-based clustering technique which partitions the data space into a grid and searches for rectangular areas of high density [AGGR98]. It has been used for interactive dimensionality reduction [JJ09], however it is not based on perceptual data and requires setting the number of cuts per dimensions to define the grid, and the density threshold *parameters*.

In Section 6, we will compare ClustMe with both, Clumpiness and CLIQUE, as well as with other standard clustering techniques: X-means [PM00], and DBSCAN [EKSX96].

3 Design Considerations

We first briefly outline the methodological process we follow in this work, and then discuss our choice to take Gaussian Mixture Models as the general modeling framework.

3.1 Design Process: Data-Driven

Our work follows the data-driven approach for designing VQMs based on perceptual data as outlined by Sedlmair and Aupetit [SA15]. This process includes several critical steps. (1) As a first step, we need to gather a large set of scatterplots containing a sufficient variation of the pattern of interest. (2) We then need to task human subjects to judge if (or how) they perceive this pattern in these scatterplots. (3) Afterwards, we look for an automatic technique that can detect this pattern, and compare it with the collected human judgments using statistical approaches. (4) The best of these techniques is declared a "model" of the human perception of this pattern and can be used as a VQM for it.

In this work, we use this approach to build a VQM for visual clustering tasks. Specifically, we will assume a Gaussian mixture model and learn from human data which merging technique to use. This choice allows us to combine the generic nature of GMM to model complex cluster shapes, and the simplicity of merging decision to accurately model the collected human judgments. In the following, we describe the reasoning behind these components in more detail. The overall process is summarized in Figure 1.

3.2 Modeling Approach: Gaussian Mixture Models (GMMs)

The main goal of our work is to design an algorithm to quantify cluster/grouping patterns in monochrome scatterplots on par with human perception. Clustering techniques [XW05] assign points to clusters without supervision. So, a way to design a VQM for grouping patterns could be to use these techniques to group data first by mimicking human perception, and then to quantify the characteristics of these well-defined clusters to get the VQM. However, as explained by Von Luxburg *et al.* [vLWG12], there is not a single way to define a cluster and each clustering technique can detect a formally different family of grouping patterns. Moreover, we do not want to cluster data *per se* in a way that a human would mentally assign points to groups, but we are more flexible in aiming for a score that would allow ranking scatterplots based on *how much* they are clustered according to human visual perception. Therefore, the model we need requires three key characteristics:

- **Being generative**: In order to collect data from human subject studies, we need a model that is able to *generate* interesting stimuli, *i.e.* grouping patterns of various shapes.
- **Being universal**: We need a model that can be trained to *accurately represent* a large range of realistic grouping patterns *from*

the most simple to the most complex, likely to appear in unseen scatterplots.
- **Being parameter-free**: We need a model which could relieve the user from arbitrary hand-tuning of parameters.

Gaussian Mixture Models (GMMs) [BC96, BCRR97, FR02] and their extension with merging techniques [Hen10] have all these characteristics at once. In general, GMMs can model—with desired accuracy—any simple or complex data density distribution by a mixture of multivariate Gaussian distributions. Data points are assigned to Gaussian cluster components with some probability based on their distance to the cluster center. A hard assignment (partition) can be obtained by assigning points to the cluster from which they get maximum probability. Most importantly, however, GMMs have the three key characteristics that are important for our design goals:

GMMs are generative models: GMMs can be used to *generate* sample points in a scatterplots with the desired statistical properties. Thus, they give natural handles for a controlled human subjects experiment by letting us vary the volume, orientation, density and overlap of the generated Gaussian clusters. In doing so, we can control the complexity of the resulting grouping patterns. No other clustering technique allows this control.

GMMs are universal models: GMMs can model any smooth data distribution [PS91]. We can infer their parameters based on sample data points using statistical inference (learning from data) [FR02]. Each component of a GMM identifies a single Gaussian cluster. However, in practice it is possible for a grouping pattern with a complex non-Gaussian distribution to be modeled by several overlapping components of the GMM. Yet, the GMM—being a density function—does not encode explicitly the fact that *several* of its Gaussian components can represent points from what would be perceived as the *same single* non-Gaussian cluster (points forming a banana shape, for instance in Figure 3 right). Merging techniques [Hen10] have been proposed to solve this issue based on the natural assumption that a wide density gap between two components call for leaving them separated, and a strong overlap for merging.

Indeed, merging techniques can automatically decide if pairs of GMM components shall be *merged* or remain *separated*. Merging would mean that the components represent data from the same cluster (all within the banana shape, for instance), while separated would typically indicate a density gap between them. These merging techniques extend the GMM framework, making it a much more generic model (*GMM+Merging*) for clustering. Specifically, it allows to identify and quantify non-trivial grouping patterns in many practical cases. On top of that, deciding to merge clusters with an algorithm is directly linked to a simple cluster-counting task that we can ask to a human, as we analyze in the next section. Hennig studies 7 of these merging techniques [Hen10] that we will evaluate in our experiments.

GMMs are parameter-free models: GMMs allow determining the geometrical characteristics of a cluster (volume, density and orientation), and the optimal number of clusters based on statistical criteria. This process runs fully automatic, making GMMs a *parameter-free* model. A typical approach in GMMs is to use the Bayesian Information Criterion (BIC) [Sch78] to select the model which best compromises between high likelihood and low complexity (a regularization technique to avoid over-fitting, as recommended for any data-driven model).

In summary, GMM+Merging allows us to easily gather large amounts of human judgments on simple patterns, while at the same time the resulting VQM will be able to independently depict complex cluster patterns. In the sequel, we will describe each step of our process in more detail: (1) We fist present the collection process of human judgments. (2) Based on this data, we find the best merging technique to model human cluster perceptions, forming the foundations of ClustMe. And (3) we evaluate ClustMe in a second human subjects experiment.

4 Experiment 1: Gathering Data for Modeling

In this section, we describe Experiment 1, in which we collected data from a cluster-counting task. The resulting data will later be used to find the best merging technique to model human judgments forming our VQM.

4.1 Task

We asked participants to perform a simple and fast perceptual task that requires to observe a single scatterplot at a time and decide if it contains 1 or more-than-1 clusters.

Picking this task was based on a careful consideration of different alternatives. Different tasks regarding pattern perception have been proposed either to judge how two patterns are similar, or to judge some characteristics of a single pattern. The similarity between patterns can be measured by relative ranking using some Likert scale [TBB*10] or by manual grouping of similar scatterplot based on their thumbnail images [DBH14, PKF*16]. Cluster patterns can also be characterized by identifying their elements using a lasso selection, or the overall pattern evaluated by counting the number of clusters [EMdSP*15].

In contrast, we decided to focus on even simpler patterns, characterizing the perception of two clusters only. The key element of our proposition is to use a counting approach with a GMM of 2 Gaussian components as it is amenable to very fast decisions by humans. This allows us to collect amounts of data, which are beneficial for robust learning approaches. Moreover, *the collected data can be directly modeled by the merging stage of a GMM+Merging clustering technique*: indeed, for a merging technique to decide *to merge or not to merge* two Gaussian components of a GMM is equivalent to a human subject deciding if she can see *1 or more-than-1* groups respectively, in a scatter plot generated by such a GMM with 2 components. We call this *2D-Gaussian-Clusters-merging-perception* task, the *2DGCMP* task for short. Hence, our proposal is that a GMM+Merging model for clustering, where the merging part best mimics the way human judge on the *2DGCMP* task, would be a good candidate for a VQM of more complex grouping patterns.

4.2 Generating Scatterplot Stimuli

We generated 1000 scatterplots varying the parameters of a Gaussian mixture of 2 components. Using a stratified sampling approach, we ensured to cover the large space of possible shapes generated by the GMM, including well-separated clusters and more overlapping ones (see Figure 2). In the following, we provide more details on each of these steps.

Table 1: *Selected values of the GMM parameters.*

Param.	Description	Values
N	Number of points	{100, 1000}
τ	Prior probability of component u.	{0.2, 0.3, 0.4, 0.5}
μ	Distance between components u and v	{0, 1, 2, 3, 5, 8, 13, 21}
$\sigma_{u,v}^{a,b}$	Scaling factors	{0.5, 1, 1.5, 2, 2.5, 3}
θ_u, θ_v	Rotation angles	{0, $\pi/8$, $\pi/4$, $3\pi/8$, $\pi/2$ }

Generating data points from a Gaussian Mixture Model: Each scatterplot displays a set of N points $x \in \mathbb{R}^2$ drawn from the distribution generated by a mixture model with M bivariate Gaussian components g_i with distribution $g_i(x) = g(x|\mu_i, \Sigma_i) = \det(2\pi\Sigma)^{-\frac{1}{2}} \exp(-\frac{1}{2}(x-\mu_i)^\top \Sigma_i^{-1}(x-\mu_i))$. μ_i is the 2-dimensional mean vector of the component g_i determining the center position of the Gaussian cluster, and Σ_i is its 2×2 covariance matrix deciding about the elliptic shape of the cluster (orientation, width and length). Σ_i is further decomposed into $\Sigma_i = R_i S_i R_i^\top$ where S_i is a diagonal scaling matrix with independent scales σ_i^a and σ_i^b along a and b orthogonal axes respectively. This gives an elliptic shape to the cluster with width and length driven by a and b, whenever $\sigma_i^a \neq \sigma_i^b$, and R_i is a rotation matrix of angle θ_i which orients the elliptic shape with respect to the a-axis. Figure 1 (stage 1) visually illustrates these parameters.

$$S_i = \begin{pmatrix} \sigma_i^a & 0 \\ 0 & \sigma_i^b \end{pmatrix} \qquad R_i = \begin{pmatrix} \cos\theta_i & -\sin\theta_i \\ \sin\theta_i & \cos\theta_i \end{pmatrix}$$

In our case $M = 2$ and the density function of the GMM is:

$$p(x|\Theta) = \sum_{i=1}^{M} \pi_i g_i(x) = \pi_u g_u(x) + \pi_v g_v(x) \quad \text{with} \quad \sum_{i=1}^{M} \pi_i = 1 \quad (1)$$

where $\tau = \pi_u = 1 - \pi_v$ is the prior probability that component u generates a data point and $\Theta = (\pi_u, \pi_v, \mu_u, \mu_v, \Sigma_u, \Sigma_v)$ is the parameter vector. In order to generate a point of the scatterplot, first the generating component g_i is chosen at random with probability τ to select g_u, then a point is drawn at random from its bivariate Gaussian distribution with parameters μ_i and Σ_i. We keep track of the generating component u or v of each point $x \in X$ in a scatterplot (X, Θ) forming the sets U and V respectively ($U \cup V = X$ and $U \cap V = \emptyset$).

Parameters Setting: Our goal is to cover most of the parameter space, so that a large variety of shapes that could be generated by a Gaussian mixture is covered. To do so, we used the sets of values given in table 1. We set $\mu_u^a = \mu_u^b = \mu_v^a = 0$ only varying $\mu = \mu_v^b$ the Euclidean distance between the two components u and v along the b-axis, which plays a crucial role to make the two components overlap. All combinations of parameter values result in $2,073,600$ unique 9-dimensional parameter vectors, each generating a theoretically distinct grouping pattern (Stage (1) in Figure 1), except for patterns generated with $\sigma_i^a = \sigma_i^b$ which are invariant to any θ_i.

Selecting Diverse Scatterplot Stimuli: To reach a manageable number of stimuli for the study, we sampled 1020 scatterplots from the $2,073,600$, 20 for training and 1000 for evaluation (Stage (2) in Figure 1). We empirically estimated the total time required to complete the experiment to be about 45 minutes. In order to get various cluster patterns that cover the full space of possible shapes (Figure 2), we could not simply uniformly sample the GMM parameter space, as these parameters are not straightforwardly related to the resulting shape's perceived complexity. Instead, we used a stratified sampling approach similar to the one used by Pandey *et al.* [PKF*16]. Section 2 in the supplementary material provides more details about this procedure.

4.3 Experimental Design

Series and task: We split the 1020 scatterplots into 11 series. The first series contained 20 scatterplots (one per K-means group) for the participants to get used to the diversity of grouping patterns and the task, before the actual recording started. For the remaining 10 series, each series contains 100 scatterplots randomly sampled from the 1000 scatterplots without replacement. We showed the 11 series to each participant independently and randomized the order of scatterplots to account for learning effects and other biases [VZS18]. The participants had an optional break of 2 minutes between each series. The task was to push key "1" or "2" of the keyboard depending on the number of clusters perceived in the scatterplot (Stage (3) in Figure 1). When participants perceive 2 or more than 2 clusters they were instructed to hit key "2". For each (participant identifier, scatterplot) pair we recorded the hit key and the time to hit it.

Human Subjects: We recruited 34 adult participants (27 males / 7 females), aged between 19 and 56 (average 31.9), 27 of them with a university degree (BSc, MSc, PhD). All participants reported normal or corrected-to-normal vision, and the experiment was conducted in a lab with daylight on a standard desktop computer and keyboard. Participants also reported about having experience with data analysis applications (1 never, 4 rarely, 27 often). The average completion time was 24 min (stdev: 8 min) excluding breaks.

Process: Before the experiment, we explained the perceptual nature of the task and encouraged the participants not to take more than 2 seconds per task. We also explained that we would not give a clear definition of "cluster" as the very objective of the task was to understand what forms a single cluster for human subjects.

Practical observations: When a key was hit, it was not possible to come back to change the choice. Several participants reported having felt frustrated of pushing the key too fast sometimes, mostly during the first series. This might generate some noise in the collected judgments although we do not expect it to be significant when integrated over 34 participants with randomized plot order.

4.4 Results and Findings

For each scatterplot (X, Θ), we count the number n_1 and n_2 out of 34 participants that perceived 1 or more clusters respectively. We summarize these human judgments as the *human separation* score which is the probability $\mathcal{H}_{sep} = \frac{n_2}{n_1+n_2}$ that a scatterplot contains 2 distinct clusters. The human separation score is high when 2 or more clusters are more frequently perceived than 1 cluster. Figure 2 displays a sample of the scatterplots generated in this experiment together with their human separation score.

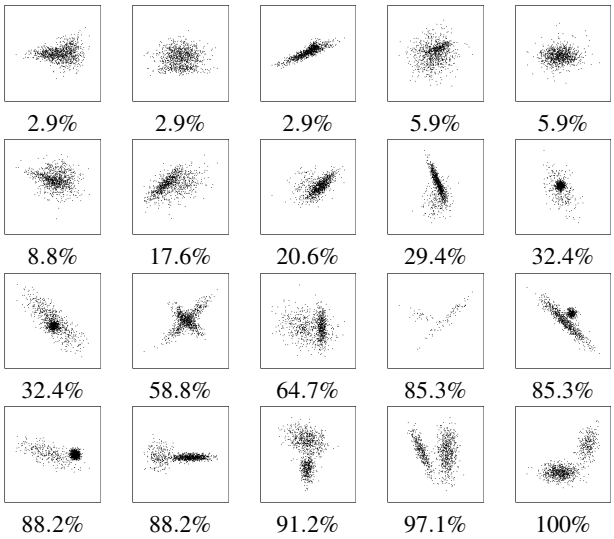

Figure 2: *Subset of the* 1000 *scatterplots judged by the* 34 *human subjects with the percentage of human raters ($\mathcal{H}_{sep}$) judging they display more than a single cluster.*

5 ClustMe

According to our process outlined in Sec. 3.1, we now use the judgments from Experiment 1 to design ClustMe. To do so, we compare this data to automatic decisions from different merging techniques with the goal to find the best one at mimicking human judgments.

5.1 Selecting Merging Techniques

We consider the 7 merging techniques analyzed by Hennig [Hen10] (Stage (4) in Figure 1). These merging techniques are all based on first fitting a GMM to the distribution of the points in the scatterplot. The standard approach for estimating the parameters of a GMM is to fix a number of components M (See Equation 1), and then maximize the likelihood using Expectation Maximization [BCRR97]. Several numbers M of components are tested, and the optimal number is selected as the one maximizing the Bayesian Information Criterion [Sch78] (BIC). Then each pair of components is tested for merging, and the merging decisions are aggregated to get the final clusters. The 7 merging techniques work as follows:

- **Bhat** merges components whose Bhattacharyya [Bha43] distance is lower than a threshold;
- **Demp** merges two components if the misclassification probability between them is higher than a threshold;
- **RidgeUni** merges components if their mixture is unimodal according to Ray and Lindsay's criterion [RL05];
- **RidgeRatio** expand RidgeUni merging nearly-unimodal pairs of components up to some threshold;
- **Dipuni** uses RidgeRatio to decide which clusters to merge but stops merging according to the p-value of the dip statistical test;
- **DipTantrum** is similar to Dipuni but the p-value is computed as in Tantrum *et al.* [TMS03];
- **Predictive** is based on the pairwise prediction strength that a clustering model, computed on one half of the data, has about the other half.

Table 2: *Vanbelle Kappa κ_V index*

DipTan	Demp	Dipu	Bhat	RidgeU	RidgeR	Pred
.788	**.788**	.786	.785	.750	.696	.529

We use the `Rmixmod` R-package [LIL*15] to implement the GMM and BIC, and then the `fpc` R-package [Hen15] proposed by Hennig to merge the components of the GMM and output the final clustering. In our experiment we always used the default parameters given by the `fpc` functions (See Section 1 of the supplementary material).

5.2 Comparison with Human Judgment

We now compare the merging techniques with the human judgments of the 1000 scatterplots (Stage (5) in Figure 1).

We use Vanbelle's Kappa κ_V index [VA09], which quantifies the degree of agreement between an isolated rater (one of the merging techniques) and a group of raters (the 34 subjects) on a nominal scale (the single merging decision and the 34 individual categorical judgments). It ranges from -1 (systematic opposed decisions) to 1 (full agreement). The group of raters is regarded as a whole, a reference or gold-standard group with its own heterogeneity. κ_V comes down to standard Cohen's Kappa κ [Coh60] inter-rater agreement when there is a single rater in the group. Table 2 gives the κ_V indecies for each of the merging techniques, settling *Demp* on par with *DipTantrum* among the best ones with $\kappa_V = 0.788$. As per the scale proposed by Landis and Koch [LK77], values between 0.6 and 0.8 are interpreted to be in *substantial* agreement with the human raters.

As we can see in Table 2, the first four merging techniques are almost equal as per κ_V index. As we want to pick one of them to build a VQM, we need to find another measure that could further distinguish between them. We thus use classification theory [BDA13] to quantify how much the merging technique (predicted class) agrees with the *averaged human decision* (actual class). For this evaluation we first transform the human separation score into a human decision $\mathcal{H}_{mrg}$ by thresholding the probability $\mathcal{H}_{sep}$ at 0.5. For each of the 1000 scatterplots, we end up with a human decision $\mathcal{H}_{mrg}(X,\Theta) \in \{1,2\}$, and the GMM combined with a merging technique gives a merging decision $\mathcal{M}_{mrg}(X,\Theta) \in \{1,2\}$ for the same plot (X,Θ). We then use *Matthews Correlation Coefficient* (MCC), which is recommended by Bekkar *et al.* [BDA13] to account for class imbalance as 81.5% of the 1000 scatterplots are classified 1 by humans (a single cluster is perceived). Figure 3 (top) shows the MCC for the 7 merging techniques, visualized as a boxplot encoding the distribution of 10000 bootstrap samples estimating the variance of this score. Bootstrap sampling [ET93] is used to estimate the sensitivity of the accuracy score to the current sample of 1000 scatterplots and evaluates how different it could be on new samples drawn from the same distribution of scatterplots. *Demp* followed by *Bhat*, stands out again among the best techniques (69.5% median accuracy).

As *Demp* is the only merging technique to rank both times among the first, we select it as the best merging techniques to model the human decision in the *2DGCMP* task, to be used in ClustMe.

5.3 Definition of ClustMe

ClustMe is a multi-scale VQM made of two numbers C and M_{BIC} coming out of a 2-step procedure (Stage (6) in Figure 1):

1. A bivariate GMM is trained with the Expectation Maximization (EM) technique on the 2-dimensional points X of the scatterplot to find the optimal set of parameters Θ^* for each $M \in \mathbb{M} = \{1,\dots,M_{max}\}$. The Bayesian Information Criterion (BIC) is used to select the optimal number $M_{BIC} \in \mathbb{M}$ of components. Setting $M_{max} = N$ ensures to find the optimal M_{BIC}, but in practice, M is incremented until BIC reaches a maximum value which typically occurs at some low value $M << N$.
2. Then, the *Demp* merging technique, which performed best on modeling human judgments in Experiment 1, is applied to all pairs of components of the BIC-optimal GMM. In doing so, it decisions about merging some of them can be made, ending up in the points being assigned to a set of $C \leq M_{BIC}$ clusters.

Johansson and Johansson [JJ09] proposed to use the number of clusters found by the CLIQUE clustering technique as a VQM for cluster patterns. Following the same idea, we propose that ClustMe provides the number C of clusters as a primary complexity measure of grouping patterns, and M_{BIC} as a secondary measure for tie-breaking scatterplots with equal C. For instance, an elliptic pattern and a banana shape pattern can both end quantified by ClustMe as a single $C = 1$ cluster, but the banana shape will likely require several elliptic Gaussian components ($M_{BIC} > 1$) to cover it (see Figure 3 Right), while the elliptic pattern only needs a single component ($M_{BIC} = 1$).

Therefore, we define the ClustMe VQM of a scatterplot X as:

$$ClustMe(X) = (C, M_{BIC})_X$$

Figure 3 (Right) illustrates the different stages of the ClustMe process on a scatterplot with a simple grouping pattern, the ClustMe VQM is $(C, M_{BIC}) = (3,9)$. Also, if applied in Experiment 1 with only 2 components in the GMM and an ideal exact match between *Demp* and the human judgments, ClustMe would be $(1,2)$ when the human perceives a single cluster (merging decision) and $(2,2)$ otherwise.

5.4 How to use ClustMe as a VQM

The ranking based on $ClustMe(X)$ is done first by decreasing C, and then for each equal C by decreasing M_{BIC}. In other words, the higher the value of C and M_{BIC} (for same C), the more complex the grouping pattern would be perceived by a human in the scatterplot X. ClustMe VQM ordered by increasing values look like the following series and quantify less complex (left) to more complex (right) grouping patterns:

$$\text{Simple patterns} \leftarrow (1,1) \prec (1,2) \prec \dots \prec (1;M_{BIC}) \prec (2,2) \prec \dots$$
$$\dots \prec (2,M_{BIC}) \prec (3,3) \prec \dots \prec (M_{BIC},M_{BIC}) \rightarrow \text{Complex patterns}$$

6 Experiment 2: Evaluation of ClustMe

In this section, we discuss the second study. We compared ClustMe to the state-of-the-art Clumpiness [DW14], as well as 3 other clustering techniques, for which we used the number of clusters found as the VQM for grouping patterns complexity (Figure 4):

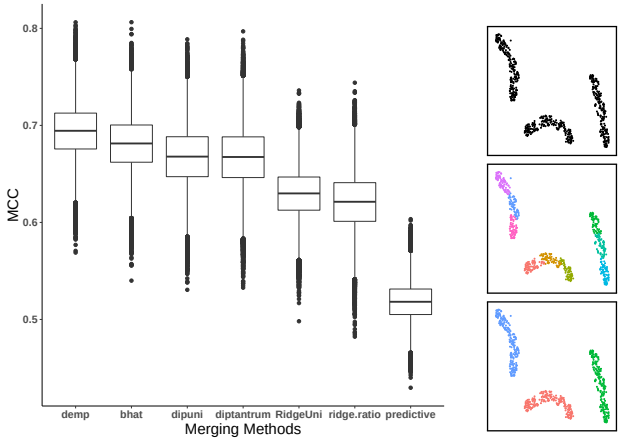

Figure 3: ***Left:*** *median and variance bootstrap estimate of the Matthews Correlation Coefficient (MCC) accuracy of each merging technique decision $\mathcal{M}_{mrg}$ to predict the human merging decision $\mathcal{H}_{mrg}$ from Experiment 1. Demp is the most accurate model of the human merging decision.* ***Right:*** *from top to bottom, example of an original scatterplot, points assigned to color-coded components of the BIC-optimal GMM, and ClustMe output after merging with Demp.*

- **X-means** [PM00] (`RWeka` R-package [HBZ09]) is a parameter-free clustering technique similar to K-means [MJ66] where the number K of clusters is selected automatically, but which can only provide convex-shaped clusters;
- **CLIQUE** [AGGR98] (`subspace` R-package [HH15]) is a grid-based clustering technique which has been used as a VQM for grouping patterns [JJ09], but it requires arbitrary hand-tuning of grid size and density threshold parameters;
- **DBSCAN** [EKSX96] (`dbscan` R-package [HP18]) is a density-based clustering technique which can detect clusters with non-convex shapes, but requires arbitrary hand-tuning of ε and *minPts* parameters.

6.1 Task

Following previous evaluation approaches for VQMs [TBB*10], we set out to compare ClustMe and the other VQMs in terms of how they rank a set of scatterplots in comparison to human rankings. We thus first need a human ranking of scatterplots in terms of the perceived complexity of grouping patterns. A straightforward way to get such a human ranking is to task subjects with pairwise comparisons.

6.2 Selected Scatterplot Stimuli

To increase realism and generalizability as compared to Experiment 1, we this time considered a set of 257 unseen real and synthetic scatterplots from previous work [SMT13]. From this set, we sampled 435 pairs of scatterplots to run a pairwise comparison experiment. We then use these pairwise human rankings as a baseline to assess the quality of the above VQMs.

As comparing all possible pairs of 257 scatterplots would amount to an infeasible number of 32,896 trials per participant, we selected a random subsample of 35 scatterplots from the initial set, 5 for training and 30 for actual evaluation (Figure 5). This leads

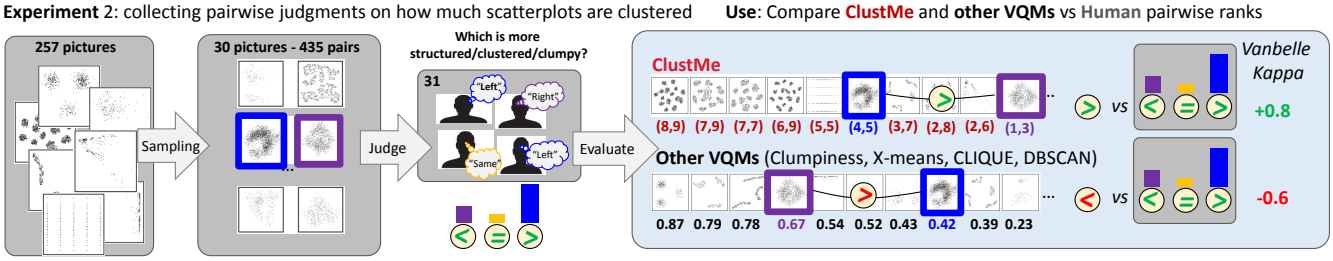

Figure 4: ***Experiment 2****: 31 human subjects decide which one of two scatterplots is more structured/clustered/clumpy for 435 pairs built from 30 out of 257 real and synthetic yet unseen scatterplots. The agreement between ClustMe, Clumpiness, X-means, CLIQUE, and DBSCAN rankings of these 435 pairs to the human rankings is evaluated with Vanbelle's Kappa index.*

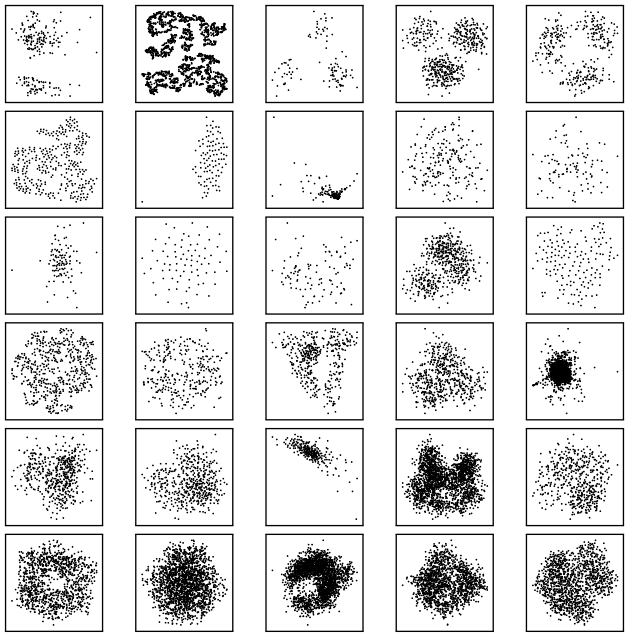

Figure 5: *The 30 scatterplots randomly picked from the 257 benchmark data used for gathering human judegments in Experiment 2.*

to a feasible set of 10 training and 435 evaluation pairs that we can show to each participant, and a total of 13,454 human judgments collected.

6.3 Experimental design

Series and task We first generated the 10 and 435 pairs of pictures for the training and evaluation sets. We then showed these two series to each participant independently, and randomized the pairs in each series as well as the left-right position of the pictures in each pair. After the training series, participants were given a short pause. For each pair, the task was to press the "left" or "right" arrow key based on whether participants perceived the left or right scatterplot as more "structured, clustered or clumpy". Participants could press the "space" bar when they thought both scatterplots were on par.

Human Subjects We recruited 31 participants (19 males and 9 females), 30 with a university degree, aged between 23 and 43 (average 31.1) and conducted the study in the same environment as study 1. 29 participants reported normal or corrected-to-normal vision. 1 reported to be color-blind, and 1 to have a lazy-eye, neither of which is a problem for our study of monochrome 2D scatterplots though. Participants reported about having experience with data analysis applications (4 rarely, 26 often). The average completion time per trial was 1.89 sec (stdev: 1.65 sec) excluding breaks.

Process We followed the same process as in Experiment 1, that is, encouraging participants not to take more than 2s per task, and justifying the reason for not giving an exhaustive definition for "structured, clustered or clumpy". We also decided again not to include a "back" button. Despite some potentially false hits, this choice fosters fast, perceptual judgments, which we deemed much more important for our study design.

6.4 Results and Findings

As in Experiment 1, we use Vanbelle's kappa κ_V to evaluate the degree of agreement between the judgments of the 31 human raters and the results of the 5 different VQM rankings on each of the 435 pairs. We assumed 3 categories of values for the pairwise rankings given by the raters: "left $<$ right" (right arrow key pressed), "left $>$ right" (left arrow key), "left $=$ right" (space bar). Over the 13,454 human judgments, 25% (3,413) were "left $=$ right", the distribution of human raters selecting this option is displayed in Figure 6 (Top). Vanbelle's kappa is the measure best adapted to our setting of categorical judgments, as it takes into account both the agreement between the group of human raters and the isolated VQM rater, together with the group inter-rater agreements.

ClustMe VQM provides 27 "left $=$ right" ranks over the 435 pairs (6.2%), X-means 406 (93.3%), the best CLIQUE 198 (45.5%), the best DBSCAN 110 (25.3%), but Clumpiness provided no equal rank. Indeed, all clustering-based VQM and ClustMe can only take a finite set of possible values based on the number of clusters found. For instance, with a maximum $M_{BIC} = 9$ in this experiment, and given $C \leq M_{BIC}$, the ClustMe measure (C, M_{BIC}) could only take 45 distinct values. In contrast, Clumpiness is based on a ratio of lengths of minimum spanning tree edges [WAG05] ranging from 0 to 1, making equal values even for very similar scatterplots unlikely. Hence, Clumpiness might be in disagreement with human raters only because it gives slightly different scores to scatterplots judged as "equal" by the human raters. For the sake of fairness, we defined an additional Coarse-Clumpiness score.

Coarse-Clumpiness: We allow Coarse-Clumpiness to give 'equal'

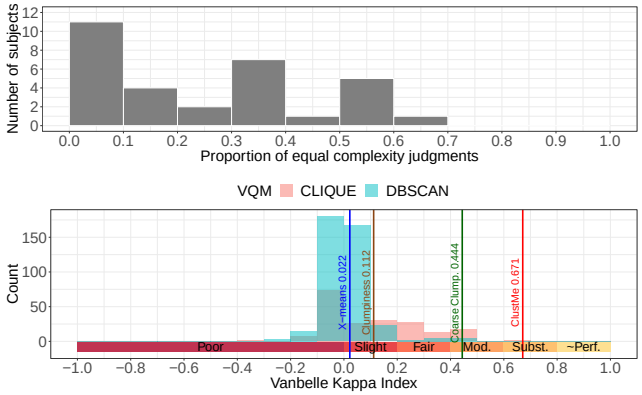

Figure 6: ***Top:*** *Number of subjects selecting equal complexity for a given proportion of the* 435 *trials.* ***Bottom:*** *Vanbelle Kappa index results from Experiment 2 with agreement interpretation as per Landis and Koch [LK77].*

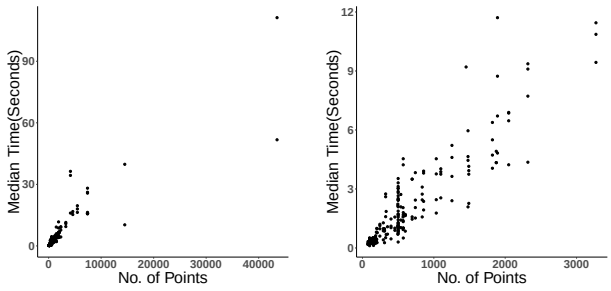

Figure 7: *ClustMe computing time in seconds per number of points N in the scatterplot for all* 257 *benchmark scatterplots (Left) and zoom in on the ones with less than* 4000 *points (Right). The computing time is roughly linear with the number of points*

ranks for pairs with sufficiently similar Clumpiness scores. How much similarity is enough is difficult to decide, so we quantify Clumpiness scale into q levels, q ranging from 1 to 45, the number of ClustMe levels in this experiment. For each number q of quantification levels, we consider as 'equal' (category 'left = right') the pairs of the 435 scatterplots whose Clumpiness values difference was less than $1/q$ (1 being the maximum value for Clumpiness), and compute the Vanbelle's Kappa index κ_V for that q. We report the best *Coarse-Clumpiness* score κ_V^* found among these 45 values, which occurred for $q^* = 24$ levels, to be compared with the standard ClustMe and the other clustering-based VQMs.

All resulting Vanbelle's Kappa scores are displayed in Figure 6 (Bottom) together with the interpretation scale proposed by Landis and Koch for the Kappa index [LK77]. ClustMe has the highest $\kappa_V = 0.671$ (*substantial* agreement) while Coarse-Clumpiness is only in *moderate* agreement with $\kappa_V = 0.444$. Still the quantization process helped to improve significantly over the standard Clumpiness which is in *slight* agreement with $\kappa_V = 0.112$. X-means is among the worst VQM only in slight agreement with $\kappa_V = 0.022$ which can be explained by its inability to model non-convex cluster shapes. For DBSCAN and CLIQUE, Figure 6 shows the distribution of all scores that we obtained through a systematic grid search for optimal clustering parameters. The median score for DBSCAN (0.001) and CLIQUE (0.062) is only in *slight* agreement with human raters. The best DBSCAN reached only a *moderate* agreement ($\kappa_V = 0.491$) with $\varepsilon = 0.02$ and $minPts = 7$. While the best CLIQUE reached a *substantial* agreement ($\kappa_V = 0.651$) with 5 grid intervals and 0.18 density threshold, on a par with ClustMe. However, this systematic exploration is not realistic: the non-expert user cannot spend time in this setting unless transforming the perceptual skimming task into a cognitive task. Moreover, the best setting we found for each clustering technique is *biased* towards the set of scatterplots we used in this experiment (overfitting), *i.e.* it is unlikely to be the best setting for any new scatterplot. By contrast, the choice of the merging technique for ClustMe has been done on a totally different set of scatterplots in the Experiment 1, demonstrating the generalization capacity of ClustMe to rank new scatterplots.

Qualitative analysis: Figure 8 shows the top 7 and bottom 7 scatterplots from the ranking of the whole set of 257 scatterplots ordered by decreasing VQM for all five approaches. CLIQUE and ClustMe rankings appear consistent with most complex patterns on top and least complex ones at the bottom. Other VQMs are not so consistent: strongly structured patterns appear at bottom rank of Clumpiness, DBSCAN and X-means, which would be missed by the user if she used these VQMs to rapidly focus on top-ranked views as they are expected to be the most interesting ones. This qualitative analysis confirms the quantitative agreement scores found in the experiment above.

6.5 ClustMe Scalability

Figure 7 shows the computing time for ClustMe with respect to the number of points for the benchmark scatterplots. All experiments ran on a single core desktop computer (MacOS, 3.4 GHz Intel Core i5 processor, 8 GB 1600 MHz DDR3 RAM). ClustMe takes approximately 15 sec for a scatterplot with 5000 points and about 3 ms per point (333 pts/sec). This means that Clumpiness, X-means, CLIQUE, and DBSCAN, which only need a few miliseconds per scatterplot, are by far faster to compute than ClustMe. A detailed analysis of ClustMe shows that the training of the initial GMM is the bottleneck, while the merging process is extremely fast relying only on the parameters of the GMM but not on the number of points. This is an obvious computational drawback of our method. However, at the same time we must consider its benefits as a VQM that can better mimic human perception of grouping pattern complexity. While there is no obvious way to increase the accuracy of computational VQMs like Clumpiness, X-means, CLIQUE or DBSCAN, there are several possibilities to speed-up ClustMe. Computations of the ClustMe VQM are independent for each scatterplot and could be distributed on multiple processors. Speed-up is also possible using in-memory Myria DBMS [MHT*15] or spatial indexing structures [Moo99, vdM14]. The number of data points could also be reduced by sub-sampling or by using a small set of core representative data built prior to the GMM training [FFK11].

7 Discussion

ClustMe and Experiment 1: The 2DGCMP task of Experiment 1 is actually the simplest setting to assess ClustMe (i.e. $GMM + BIC + Demp$). It corresponds to the optimal and ideal case where

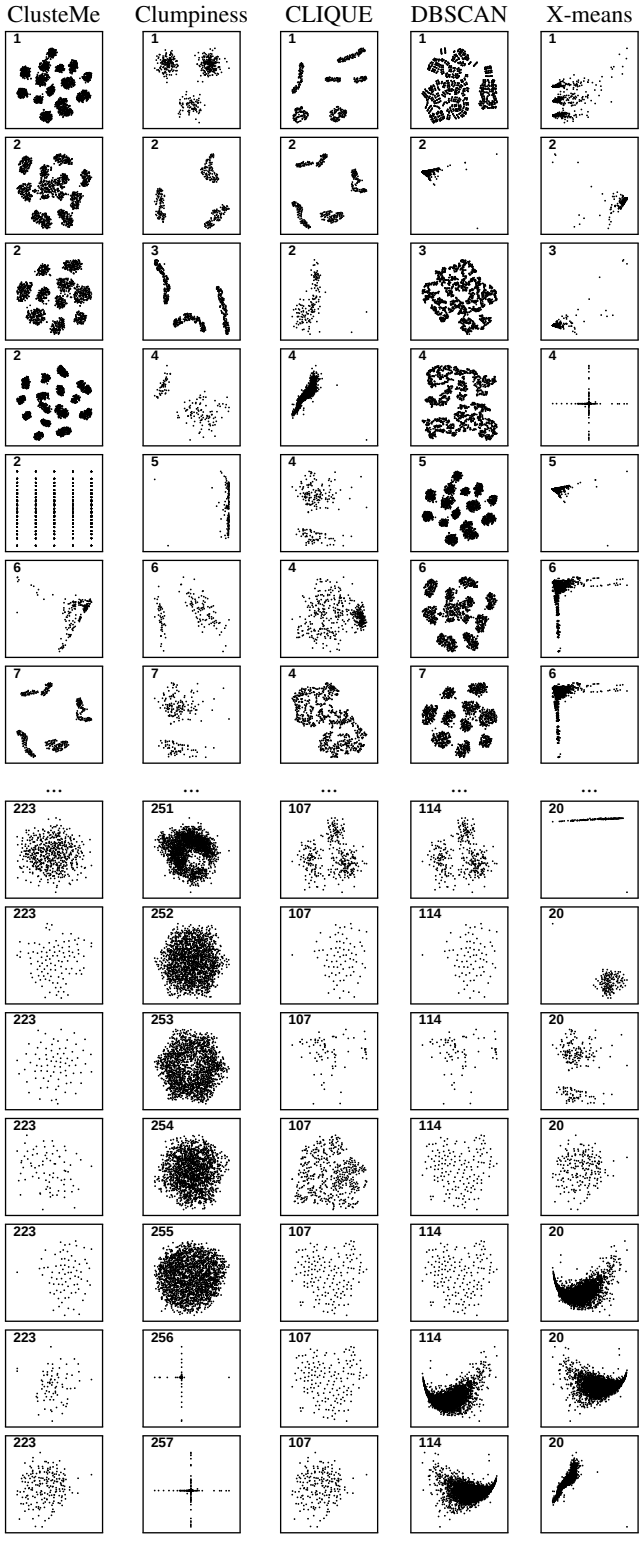

Figure 8: *Top 7 and bottom 7 of the* 257 *scatterplots ranked with ClustMe, Clumpiness, the best settings of CLIQUE and DBSCAN, and X-means. The actual rank is displayed in the corner of the pictures, equally-ranked plots are displayed in random order.*

the GMM+BIC model would have found the 2 Gaussian components with the exact same parameters as the ones generated in the given scatterplot, and only would have to let the *Demp* technique decide about merging these 2 components or not. The ClustMe score is then either (1,2) if Demp merges, or (2,2) otherwise.

Perceptual vs. cognitive tasks: The skimming task we intend to automate with a VQM, is essentially perceptual; it practically takes less than a second to decide if a view contains an interesting pattern or not. This decision allowed us to scale up the number of collected human judgments without loss of accuracy and ecological validity. We do not expect to lose accuracy in human judgments compared to a setting that gives more time for decisions, because the scatterplots we showed have no meaningful axes nor do they display meaningful data for the human subjects. However, we have not empirically validated this assumption.

8 Conclusions and Future Work

The goal of our work was to design a VQM for monochrome grouping patterns in scatterplots. To do so, we collected 34000 human clustering judgments that we used to find the best model of the human perception of 1 or more than 1 Gaussian clusters. We found that the *Demp* merging technique for components of Gaussian Mixture Models is the best model of this counting task, and used it to build ClustMe, the first data-driven visual quality measure for cluster patterns. In a second human subject experiment compared ClustMe to other state-of-the-art VQMs based on a ranking task of scatterplots by their grouping pattern complexity. The results showed that ClustMe is in *substantial agreement* with 31 human raters, while only specific but unrealistic settings of CLIQUE achieved a similar score, and Clumpiness is at best in *moderate* agreement with them.

The time complexity of the GMM is the main current bottleneck to use ClustMe at large scale. Different options exist but finding the most efficient way to compute ClustMe is an important avenue for future work. As VQM for class-separation have been used to guide supervised projection techniques [WFC*18], ClustMe could possibly be used as a criterion for Projection Pursuit [FT74].

Following the Scagnostics approach [WAG05], we see ClustMe as only one among many other VQMs, which complement each other in revealing different patterns. We thus recommend to use ClustMe together with other VQMs to support the exploration of multidimensional data. Which set of VQMs is optimal for exploring multidimensional data is still an open question.

Finally, ClustMe is based on a GMM and modeling the merging decision of humans on simple mixtures of 2 Gaussians. We extended its use to characterize patterns made of more than 2 clusters with good results, but the model itself is still coarse regarding the full human perceptual and cognitive cluster detection pipeline [STMT12]. Our study and evaluation is a first step towards opening the black box of human perception of grouping patterns, but much work remains to be done to get a more complete understanding.

References

[ACB16] ANOBILE G., CICCHINI G. M., BURR D. C.: Number as a primary perceptual attribute: A review. *Perception 45*, 1-2 (2016), 5–31. doi:10.1177/0301006615602599. 2

[AEM11] ALBUQUERQUE G., EISEMANN M., MAGNOR M.: Perception-based visual quality measures. In *Proc. IEEE Symp. on Visual Analytics Science & Technology* (2011), pp. 13–20. doi:10.1109/VAST.2011.6102437. 2

[AGGR98] AGRAWAL R., GEHRKE J., GUNOPULOS D., RAGHAVAN P.: Automatic subspace clustering of high dimensional data for data mining applications. ACM Press, pp. 94–105. 2, 3, 7

[AS16] AUPETIT M., SEDLMAIR M.: Sepme: 2002 new visual separation measures. *Proc. IEEE Pacific Visualization Symp. (PacificVis)* (2016), 1–8. doi:10.1109/PACIFICVIS.2016.7465244. 1, 2

[BBK*18] BEHRISCH M., BLUMENSCHEIN M., KIM N. W., SHAO L., EL-ASSADY M., FUCHS J., SEEBACHER D., DIEHL A., BRANDES U., PFISTER H., SCHRECK T., WEISKOPF D., KEIM D. A.: Quality metrics for information visualization. *Computer Graphics Forum 37*, 3 (2018), 625–662. URL: https://doi.org/10.1111/cgf.13446, doi:10.1111/cgf.13446. 1

[BC96] BENSMAIL H., CELEUX G.: Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association 91*, 436 (1996), 1743–1748. URL: http://www.jstor.org/stable/2291604. 2, 4

[BCRR97] BENSMAIL H., CELEUX G., RAFTERY A., ROBERT C.: Inference in model-based cluster analysis. *Statistics and Computing 7* (1997), 1–10. 2, 4, 6

[BDA13] BEKKAR M., DJEMAA H. K., ALITOUCHE T. A.: Evaluation measures for models assessment over imbalanced datasets. *Journal of Information Engineering and Applications 3*, 10 (2013). 6

[Bha43] BHATTACHARYYA A.: On a measure of divergence between two statistical populations defined by their probability distribution. *Bull. of Calcutta Mathematical Society* (1943). 6

[BM13] BREHMER M., MUNZNER T.: A multi-level typology of abstract visualization tasks. *IEEE Trans. on Visualization & Computer Graphics 19*, 12 (2013), 2376–2385. doi:10.1109/TVCG.2013.124. 1

[BS06] BERTINI E., SANTUCCI G.: Visual quality metrics. In *Proc. Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV)* (2006), pp. 1–5. doi:10.1145/1168149.1168159. 1, 2

[BSIM14] BREHMER M., SEDLMAIR M., INGRAM S., MUNZNER T.: Visualizing dimensionally-reduced data: interviews with analysts and a characterization of task sequences. In *Proc. Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV)* (2014), pp. 1–8. 1, 2

[BTK11] BERTINI E., TATU A., KEIM D.: Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Trans. on Visualization & Computer Graphics 17*, 12 (2011), 2203–2212. doi:10.1109/TVCG.2011.229. 1, 2

[Coh60] COHEN J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20* (1960), 37–46. 6

[CSM08] COHEN E. H., SINGH M., MALONEY L. T.: Perceptual segmentation and the perceived orientation of dot clusters: The role of robust statistics. *Journal of Vision 8*, 7 (2008), 6. doi:10.1167/8.7.6. 2

[DBH14] DEMIRALP C., BERNSTEIN M. S., HEER J.: Learning perceptual kernels for visualization design. *IEEE Trans. on Visualization & Computer Graphics 20*, 12 (2014), 1933–1942. doi:10.1109/TVCG.2014.2346978. 2, 4

[DW14] DANG T. N., WILKINSON L.: Scagexplorer: Exploring scatterplots by their scagnostics. In *Proc. IEEE Pacific Visualization Symp. (PacificVis)* (2014), pp. 73–80. doi:10.1109/PacificVis.2014.42. 2, 7

[EKSX96] ESTER M., KRIEGEL H.-P., SANDER J., XU X.: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD)* (1996), AAAI Press, pp. 226–231. 2, 3, 7

[EMdSP*15] ETEMADPOUR R., MOTTA R., DE SOUZA PAIVA J. G., MINGHIM R., DE OLIVEIRA M. C. F., LINSEN L.: Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE Trans. on Visualization & Computer Graphics 21*, 1 (2015), 81–94. doi:10.1109/TVCG.2014.2330617. 4

[ET93] EFRON B., TIBSHIRANI R. J.: *An Introduction to the Bootstrap*. Chapman et Hall, 1993. 6

[FFK11] FELDMAN D., FAULKNER M., KRAUSE A.: Scalable training of mixture models via coresets. In *Advances in Neural Information Processing Systems (NIPS)* (2011), Shawe-Taylor J., Zemel R. S., Bartlett P. L., Pereira F. C. N., Weinberger K. Q., (Eds.), pp. 2142–2150. 9

[FR02] FRALEY C., RAFTERY A. E.: Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association 97* (2002), 611–631. 2, 4

[FT74] FRIEDMAN J. H., TUKEY J. W.: A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. on Computers C-23*, 9 (1974), 881–890. doi:10.1109/T-C.1974.224051. 10

[HBZ09] HORNIK K., BUCHTA C., ZEILEIS A.: Open-source machine learning: R meets weka. *Computational Statistics 24*, 2 (2009), 225–232. URL: http://doi.org/10.1007/s00180-008-0119-7. 7

[Hen10] HENNIG C.: Methods for merging gaussian mixture components. *Advances in Data Analysis and Classification 4*, 1 (2010), 3–34. doi:10.1007/s11634-010-0058-3. 2, 4, 6

[Hen15] HENNIG C.: *fpc: Flexible Procedures for Clustering*, 2015. R package version 2.1-10. URL: https://CRAN.R-project.org/package=fpc. 6

[HH15] HASSANI M., HANSEN M.: *subspace: Interface to OpenSubspace*, 2015. R package version 1.0.4. URL: https://CRAN.R-project.org/package=subspace. 7

[HP18] HAHSLER M., PIEKENBROCK M.: *dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms*, 2018. R package version 1.1-3. URL: https://CRAN.R-project.org/package=dbscan. 7

[HYFC14] HARRISON L., YANG F., FRANCONERI S., CHANG R.: Ranking visualizations of correlation using weber's law. *IEEE Trans. on Visualization & Computer Graphics 20*, 12 (2014), 1943–1952. doi:10.1109/TVCG.2014.2346979. 2

[JJ09] JOHANSSON S., JOHANSSON J.: Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Trans. on Visualization & Computer Graphics 15*, 6 (2009), 993–1000. doi:10.1109/TVCG.2009.153. 3, 7

[KH16] KAY M., HEER J.: Beyond weber's law: A second look at ranking visualizations of correlation. *IEEE Trans. on Visualization & Computer Graphics 22*, 1 (2016), 469–478. doi:10.1109/TVCG.2015.2467671. 2

[LIL*15] LEBRET R., IOVLEFF S., LANGROGNET F., BIERNACKI C., CELEUX G., GOVAERT G.: Rmixmod: The R package of the model-based unsupervised, supervised, and semi-supervised classification Mixmod library. *Journal of Statistical Software 67*, 6 (2015), 1–29. doi:10.18637/jss.v067.i06. 6

[LK77] LANDIS J. R., KOCH G. G.: The measurement of observer agreement for categorical data. *Biometrics* (1977), 159–174. 6, 9

[LK11] LIN W., KUO C.-C. J.: Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation 22*, 4 (2011), 297 – 312. doi:10.1016/j.jvcir.2011.01.005. 2

[LMvW10] LI J., MARTENS J.-B., VAN WIJK J. J.: Judging correlation from scatterplots and parallel coordinate plots. *Information Visualization 9*, 1 (2010), 13–30. doi:10.1057/ivs.2008.13. 2

[Mcg95] MCGRATH E.: Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000 (2nd. eds.)* (1995), Morgan Kaufman, pp. 152–169. 2

[MHT*15] MAAS R., HYRKAS J., TELFORD O. G., BALAZINSKA M., CONNOLLY A. J., HOWE B.: Gaussian mixture models use-case: In-memory analysis with myria. In *Proc. of the 3rd VLDB Workshop on In-Memory Data Mangement and Analytics* (2015), ACM, pp. 3:1–3:8. doi:10.1145/2803140.2803143. 9

[MJ66] MACQUEEN, J.: Some methods for classification and analysis of multivariate observations. In *Berkeley Symposium on Mathematical Statistics and Probability* (1966). 7

[Moo99] MOORE A. W.: Very fast em-based mixture model clustering using multiresolution kd-trees. In *Advances in Neural Information Processing Systems (NIPS)* (1999), MIT Press, pp. 543–549. 9

[MTL18] MATUTE J., TELEA A. C., LINSEN L.: Skeleton-based scagnostics. *IEEE Trans. on Visualization & Computer Graphics 24*, 1 (2018), 542–552. doi:10.1109/TVCG.2017.2744339. 1, 2

[NA18] NONATO L. G., AUPETIT M.: Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Trans. on Visualization & Computer Graphics* (2018). doi:10.1109/TVCG.2018.2846735. 1

[O'C74] O'CALLAGHAN J. F.: Human perception of homogeneous dot patterns. *Perception 3*, 1 (1974), 33–45. doi:10.1068/p030033. 2

[PKF*16] PANDEY A. V., KRAUSE J., FELIX C., BOY J., BERTINI E.: Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proc. ACM Conf. on Human Factors in Computing Systems (CHI)* (2016), pp. 3659–3669. doi:10.1145/2858036.2858155. 2, 4, 5

[PM00] PELLEG D., MOORE A.: X-means: Extending k-means with efficient estimation of the number of clusters. In *Proc. Int. Conf. on Machine Learning (ICML)* (2000), Morgan Kaufmann, pp. 727–734. 2, 3, 7

[PS91] PARK J., SANDBERG I. W.: Universal approximation using radial-basis-function networks. *Neural Computation 3*, 2 (1991), 246–257. doi:10.1162/neco.1991.3.2.246. 4

[RB10] RENSINK R. A., BALDRIDGE G.: The perception of correlation in scatterplots. *Computer Graphics Forum 29*, 3 (2010), 1203–1210. doi:10.1111/j.1467-8659.2009.01694.x. 2

[RL05] RAY S., LINDSAY B. G.: The topography of multivariate normal mixtures. *Annals of Statistics* (2005), 2042–2065. 6

[SA15] SEDLMAIR M., AUPETIT M.: Data-driven evaluation of visual quality measures. *Computer Graphics Forum 34*, 3 (2015), 201–210. doi:10.1111/cgf.12632. 1, 2, 3

[Sad97] SADAHIRO Y.: Cluster perception in the distribution of point objects. *Cartographica: The International Journal for Geographic Information and Geovisualization 34*, 1 (1997), 49–62. doi:10.3138/Y308-2422-8615-1233. 2

[Sch78] SCHWARZ G.: Estimating the dimension of a model. *The annals of statistics 6* (1978), 461–464. 2, 4, 6

[SMT13] SEDLMAIR M., MUNZNER T., TORY M.: Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Trans. on Visualization & Computer Graphics 19*, 12 (2013), 2634–2643. doi:10.1109/TVCG.2013.153. 1, 7

[SNE*16] SCHULZ C., NOCAJ A., EL-ASSADY M., FREY S., HLAWATSCH M., HUND M., KARCH G. K., NETZEL R., SCHÄTZLE C., BUTT M., KEIM D. A., ERTL T., BRANDES U., WEISKOPF D.: Generative data models for validation and evaluation of visualization techniques. In *Proc. Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV)* (2016), pp. 112–124. doi:10.1145/2993901.2993907. 2

[STMT12] SEDLMAIR M., TATU A., MUNZNER T., TORY M.: A taxonomy of visual cluster separation factors. *Computer Graphics Forum 31*, 3 (2012), 1335–1344. doi:10.1111/j.1467-8659.2012.03125.x. 2, 10

[TBB*10] TATU A., BAK P., BERTINI E., KEIM D. A., SCHNEIDEWIND J.: Visual quality metrics and human perception: an initial study on 2d projections of large multidimensional data. In *Proc. Int. Conf. on Advanced Visual Interfaces (AVI)* (2010), Santucci G., (Ed.), ACM Press, pp. 49–56. doi:10.1145/1842993.1843002. 4, 7

[TMF*12] TATU A., MAASS F., FÄRBER I., BERTINI E., SCHRECK T., SEIDL T., KEIM D. A.: Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proc. IEEE Symp. on Visual Analytics Science & Technology* (2012), pp. 63–72. doi:10.1109/VAST.2012.6400488. 2

[TMS03] TANTRUM J., MURUA A., STUETZLE W.: Assessment and pruning of hierarchical model based clustering. In *Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD)* (2003), pp. 197–205. 6

[TT85] TUKEY J. W., TUKEY P. A.: Computer Graphics and Exploratory Data Analysis: An Introduction. In *Proc. the Sixth Annual Conference and Exposition: Computer Graphics, Vol. III, Technical Sessions* (1985), Nat. Computer Graphics Association, pp. 773–785. 1, 2

[VA09] VANBELLE S., ALBERT A.: Agreement between an isolated rater and a group of raters. *Statistica Neerlandica 63*, 1 (2009), 82–100. doi:10.1111/j.1467-9574.2008.00412.x. 6

[vdM14] VAN DER MAATEN L.: Accelerating t-sne using tree-based algorithms. *Journal of Machine Learning Research (JMLR) 15*, 1 (2014), 3221–3245. 9

[vLWG12] VON LUXBURG U., WILLIAMSON R. C., GUYON I.: Clustering: Science or art? In *Proc. of ICML Workshop on Unsupervised and Transfer Learning* (Bellevue, Washington, USA, 2012), Guyon I., Dror G., Lemaire V., Taylor G., Silver D., (Eds.), vol. 27 of *Proceedings of Machine Learning Research*, PMLR, pp. 65–79. URL: http://proceedings.mlr.press/v27/luxburg12a.html. 3

[VZS18] VALDEZ A. C., ZIEFLE M., SEDLMAIR M.: Priming and anchoring effects in visualization. *IEEE Trans. on Visualization & Computer Graphics 24*, 1 (2018), 584–594. doi:10.1109/TVCG.2017.2744138. 5

[WAG05] WILKINSON L., ANAND A., GROSSMAN R. L.: Graph-theoretic scagnostics. In *Proc. IEEE Information Visualization Symp. (INFOVIS)* (2005), Stasko J. T., Ward M. O., (Eds.), IEEE Computer Society, p. 21. doi:10.1109/INFOVIS.2005.14. 1, 2, 3, 8, 10

[WFC*18] WANG Y., FENG K., CHU X., ZHANG J., FU C., SEDLMAIR M., YU X., CHEN B.: A perception-driven approach to supervised dimensionality reduction for visualization. *IEEE Trans. on Visualization & Computer Graphics 24*, 5 (2018), 1828–1840. URL: https://doi.org/10.1109/TVCG.2017.2701829. 10

[XW05] XU R., WUNSCH D.: Survey of clustering algorithms. *IEEE Trans. on Neural Networks 16*, 3 (2005), 645–678. doi:10.1109/TNN.2005.845141. 2, 3