Eurographics Conference on Visualization (EuroVis), Posters Track (2017)
A. Puig Puig and T. Isenberg (Editors)

A High-Dimensional Data Quality Metric using Pareto Optimality

Tobias Post[1], Thomas Wischgoll[2], Bernd Hamann[3] and Hans Hagen[1]

[1]Computer Graphics and HCI Group, University of Kaiserslautern, Germany
[2]Advanced Visual Data Analysis Group, Wright State University, U.S.A.
[3]Department of Computer Science, University of California, Davis, U.S.A.

Abstract

The representation of data quality within established high-dimensional data visualization techniques such as scatterplots and parallel coordinates is still an open problem. This work offers a scale-invariant measure based on Pareto optimality that is able to indicate the quality of data points with respect to the Pareto front. In cases where datasets contain noise or parameters that cannot easily be expressed or evaluated mathematically, the presented measure provides a visual encoding of the environment of a Pareto front to enable an enhanced visual inspection.

1. Introduction

The representation of data quality within a high-dimensional dataset was mentioned as one of the top challenges in information visualization [LK07]. This is especially the case when a dataset contains noise or parameters such as aesthetics that cannot easily be evaluated mathematically. Although Pareto optimality is a widely used concept to identify optimal points in a high-dimensional space, hidden dimensions justify to not only consider optimal but also nearly optimal points. So the concept of Pareto optimality has to be extended to obtain a measure on how efficient a data point is.

To address this problem, this work introduce the Pareto factor, describing the relative amount of data points that are more efficient than the evaluated point in the sense of Pareto optimality. By utilizing this new measure, it is possible to evaluate the quality of all given data points and thereby guide users not only to points on the Pareto front, but also to interesting solutions near the Pareto front. The new measure can be embedded into established visualization techniques such as scatterplots or parallel coordinates in a straightforward manner as will be shown in this work.

Therefore, this work contributes:

- A scale-invariant and flexible measure based on Pareto optimality, called Pareto factor
- Visual encoding of Pareto factors in established information visualization techniques

2. Related Work

Pareto optimality is a widely used concept to identify optimal high-dimensional points in an arbitrary space [EM11,FS06]. This section summarizes visualization techniques that are based on the concept of Pareto optimality.

Different applications such as fishery or architecture use Pareto optimality to identify interesting data points [OKPK11,BMPM12]. Although this directly results in a set of optimal data points, these domains are usually confronted with various dimensions not well expressible or previously unknown. Therefore, this paper extends the definition of Pareto optimality.

Ruotsalainen et al. [RMH08] use the gradient of the Pareto front to help the user navigate through this front to find interesting solutions. Although this is a suitable technique to navigate through Pareto optimal points only, hardly to express qualities like aesthetics require to consider solutions that are not quite Pareto optimal as well. Instead of being completely Pareto optimal, these solutions might have other properties such as aesthetics that are improved. This can be accomplished with the measure introduced in this work.

Witowski et al. [WLG09] perform a study on how several known visualization techniques can be applied to visualize the Pareto front. They point out that a combination of several tools is most promising to visualize the Pareto front in a suitable way. These techniques can be extended by the measure introduced in this work, thereby enabling users to evaluate the quality of found solutions with respect to Pareto optimality.

3. Methods and Results

In compensation criteria like the Kaldor-Hicks efficiency [Str01] known from economics, a data point is defined to be more efficient if the sum of all gains is greater than the sum of all losses in comparison to another data point. Here, an optimal point is a data point for that no other data point is more efficient. Fig. 1 **(a)** shows the optimal point B in blue and all other points in dark gray. Point A is no Kaldor-Hicks optimum because there exists a more efficient

DOI: 10.2312/eurp.20171187

point above the diagonal going through point A, and point B is optimal because there exists no such point for B. The problem with this kind of measure is that it is not invariant to anisotropic scaling. Fig. 1 **(b)** shows the same set of data points, this time scaled anisotropically by a factor of three in the horizontal direction. Here, point B is not optimal any more because there exist more efficient points and point C becomes the new optimum.

To avoid this problem, this work is based on Pareto optimality which is scale-invariant [Fle05]. A data point x is more efficient than another point y regarding Pareto optimality if x is greater than y in at least one dimension and not smaller in all other dimensions. Then, a data point is Pareto optimal if there exists no other data point that is more efficient. Fig. 1 **(c)** and **(d)** show the same data points with the same scalings as before, this time evaluated with Pareto optimality. As can be seen, the blue Pareto optimal points remain optimal after anisotropic scaling.

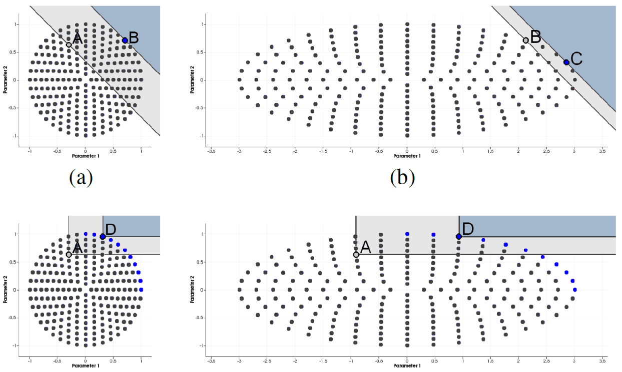

Figure 1: *A set of data points with different scalings.* ***(b)*** *and* ***(d)*** *are anisotropically scaled by a factor of three.* ***(a)*** *and* ***(b)*** *show the optimality towards a compensation criterion and* ***(c)*** *and* ***(d)*** *show Pareto optimality (blue points are optimal). From* ***(a)*** *to* ***(b)*** *the optimum changes from point* B *to point* C*, meaning that this measure is dependent on the (anisotropic) scale. In contrast to that, the Pareto optimality in* ***(c)*** *and* ***(d)*** *is scale-invariant with multiple points being optimal.*

To provide information on how far away from optimality non Pareto optimal points are, this work introduces a novel measure. The complement of the quotient of more Pareto efficient points to all points but one (i.e. $\vec{x}_i$ itself) is used as described in Eq. (1), where the relation $<_p$ means less efficient in the sense of Pareto optimality as described above. Here, $\vec{x}_1, \ldots, \vec{x}_n$ are the given data points and $p(.)$ is the new measure called *Pareto factor*. This measure preserves scale-invariance since $<_p$ is also invariant to anisotropic scaling.

$$p(\vec{x}_i) = 1 - \frac{|\{\vec{x}_j \in \{\vec{x}_1, \ldots, \vec{x}_n\} \mid \vec{x}_i <_p \vec{x}_j\}|}{n-1} \tag{1}$$

Normalizing this measure to range from zero to one for all data points yields what will be called the *normalized Pareto factor* and is useful for evaluation, weighting or visual analysis. Until now, the user cannot choose how far all data points are considered or how far only Pareto optimal points are of interest. To compensate this, the normalized Pareto factor is raised to some user defined power. Fig. 2 **(a)** - **(f)** show the effect of different exponents, where higher exponents focus on Pareto optimal points only, while lower exponents preserve an overview over all data points.

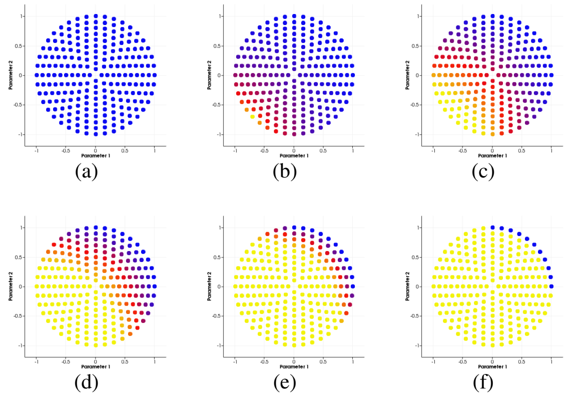

Figure 2: *The introduced Pareto factor is used to evaluate the quality of data points with respect to Pareto optimality. A scatterplot visualization is extended showing the value of this new metric using a color scale ranging from yellow (0.0) over red (0.5) to blue (1.0). The effect of different exponents is shown for the exponents* ***(a)*** *0.0,* ***(b)*** *0.25,* ***(c)*** *1.0,* ***(d)*** *4.0,* ***(e)*** *16.0 and* ***(f)*** ∞.

Fig. 2 shows how the Pareto factor can be applied to established information visualization techniques like scatterplots. A color range from yellow (Pareto factor of 0.0) over red (0.5) to blue (1.0) is used. Users can interactively manipulate the used exponent to focus more or less on the Pareto optimal points only. The applicability of the presented measure is not limited to scatterplots only but can also be applied to scatterplot matrices, parallel coordinate plots, or star plots, for example. Within these techniques, the presented Pareto factor allows to evaluate the quality of data points.

4. Conclusion

This work introduced a novel measure for the quality of data points in a high-dimensional space based on Pareto optimality. It was shown that the introduced measure is scale-invariant and enables the evaluation of the efficiency of data points with respect to Pareto optimality. Based on this measure, it was possible to visually extend established information visualization techniques. This helps to not only consider Pareto optimal data points that might not be the desired solution for problems with noise or hard to evaluate parameters like aesthetics. Instead, also nearly Pareto optimal solutions can be analyzed to find a desired tradeoff.

Acknowledgments

This research was funded by the German Research Foundation (DFG) within the IRTG 2057 "Physical Modeling for Virtual Manufacturing Systems and Processes".

References

[BMPM12] BOOSHEHRIAN M., MÖLLER T., PETERMAN R. M., MUNZNER T.: Vismon: Facilitating Analysis of Trade-Offs, Uncertainty, and Sensitivity In Fisheries Management Decision Making. *Computer Graphics Forum* (2012). doi:10.1111/j.1467-8659.2012.03116.x. 1

[EM11] ESKELINEN P., MIETTINEN K.: Trade-off analysis approach for interactive nonlinear multiobjective optimization. *OR Spectrum* (2011). 1

[Fle05] FLEISCHER M.: Scale invariant pareto optimality: A meta–formalism for characterizing and modeling cooperativity in evolutionary systems. In *Proceedings of the 7th Annual Conference on Genetic and Evolutionary Computation* (2005), GECCO '05, ACM, pp. 233–240. 2

[FS06] FELDMAN A. M., SERRANO R.: *Welfare Economics and Social Choice Theory*, 2. ed. Springere, 2006. 1

[LK07] LARAMEE R. S., KOSARA R.: *Challenges and Unsolved Problems*. Springer Berlin Heidelberg, 2007, pp. 231–254. 1

[OKPK11] OH S., KIM Y., PARK C., KIM I.: Process-driven bim-based optimal design using integration of energyplus, genetic algorithm, and pareto optimality. *Proceedings of the IBPSA building simulation 2011 conference, Sydney, Australia* (2011), 894–901. 1

[RMH08] RUOTSALAINEN H., MADETOJA E., HÄMÄLÄINEN J.: Navigation on a pareto-optimal front utilizing gradient in formation in interactive multiobjective optimization. *International Conference on Engineering Optimization* (2008). 1

[Str01] STRINGHAM E.: Kaldor-hicks efficiency and the problem of central planning. *Quarterly Journal of Austrian Economics 4*, 2 (2001), 41–50. 1

[WLG09] WITOWSKI K., LIEBSCHER M., GOEL T.: Desicion making in multi-objective optimization for industrial applications - data mining and visualization of pareto data. *European LS-DYNA Conference* (2009). 1