Robust Kernel Density Clustering of Highly Incomplete Data: Supplementary Material

Richard Leibrandt* Stephan Günnemann†

Abstract

Real-world datasets are frequently incomplete, e.g., due to data loss or sampling errors. Often, it is expensive or impossible to recover missing values. Warp-clustering is a novel method that clusters incomplete datasets without imputation or marginalization. This document supplies supplementary information.

1 Introduction

Warp-clustering is a clustering method that utilizes the concept of virtual objects. Multiple types were introduced like virtual objects obtained by recursive projection [1], from a grid, by aggregation [2]. This document details aspects of the latter two types.

2 Virtual Objects Obtained from a Grid: Number of Grid Vertices

As established [2], the number of virtual objects obtained by projection can get rather large. Thus we suggested to span a rectilinear grid over $\mathbb{M}_u$ for each undetermined object $\boldsymbol{o}_u$ and obtain determined virtual objects $\boldsymbol{o}_v$ by placing one on each vertex of the grid. In the following we discuss the number of grid vertices (and with that the number of initials) in relation to the number of objects $|\mathcal{O}|$, the number of features F and the missing values M.

2.1 Analytical approximation.

While the number of initials I increases linearly with the number of objects $|\mathcal{O}|$, we cannot expect the same for the number of features F or missing values M, since $|\mathcal{O}_u^{\leftarrow}|$ produced by an undetermined $\boldsymbol{o}_u$ increases exponentially with $|\mathcal{M}_u|$, which is the number of missing values of object $\boldsymbol{o}_u$. However, we cannot assume exponential dependence of I on F and/or M, because in a MCAR setting the probability to have a value deleted is larger for an object with a smaller $|\mathcal{M}_u|$ than for an object with a larger $|\mathcal{M}_u|$. To estimate the dependence of I on $|\mathcal{O}|$, F, and M we assumed a worst and a best case MCAR scenario. The worst case scenario is that, while deleting values, the first object loses values, until $(F-1)$ are missing. Then the next object loses values until it has $(F-1)$ missing values and so on. The best case scenario is that, while deleting values, the first feature loses values, until $(|\mathcal{O}|-1)$ objects miss values in this feature. Then the next feature loses values and so on. For both scenarios we approximated an upper bound to

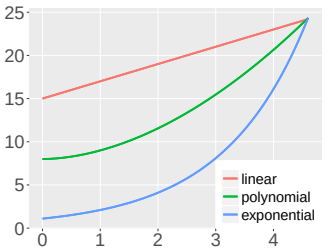

Fig. 1. worst-case, expected, best-case

$$I_{\mathrm{w}}(|\mathcal{O}|, F, M) < \frac{|\mathcal{O}|\cdot F + (M+F-1)\cdot G^{F-1}}{F-1}\,, \tag{2.1}$$

$$I_{\mathrm{b}}(|\mathcal{O}|, F, M) < 2\cdot|\mathcal{O}|\cdot G^{\frac{M}{|\mathcal{O}|}+1}\,, \tag{2.2}$$

where G is the number of grid points per dimension. The worst-case approximation is linear, while the best-case approximation is exponential in respect to M because of larger constants as visualized in Fig. 1. The expected number of I is

$$\mathrm{E}[I] = \sum_{\omega_l}\left(\mathrm{P}(\omega_l)\cdot\sum_{i=1}^{|\mathcal{O}|} G^{|\mathcal{M}_i|}\right)\,, \tag{2.3}$$

$$\mathrm{P}(\omega_l) = \frac{\binom{F}{|\mathcal{M}_1|}\cdot\binom{F}{|\mathcal{M}_2|}\cdot\ldots\cdot\binom{F}{|\mathcal{M}_{|\mathcal{O}|}|}}{\binom{|\mathcal{O}|\cdot F}{M}} \tag{2.4}$$

where $\mathrm{P}(\omega_l)$ is the probability measure for the outcome

$$\omega_l = \Big\{\{\mathsf{U}_1 = |\mathcal{M}_1|, \ldots, \mathsf{U}_{|\mathcal{O}|} = |\mathcal{M}_{|\mathcal{O}|}|\} \mid |\mathcal{M}_1| + \ldots + |\mathcal{M}_{|\mathcal{O}|}| = M\Big\} \tag{2.5}$$

where $|\mathcal{M}_i| \in \{0, \ldots, F-1\}$ is a specific number of missing values for $\boldsymbol{o}_i$ and $\mathsf{U}_i = \mathrm{U}(\hat{\omega}_i)$ is the random variable that maps the outcome $\hat{\omega}_i$ (the value constellation of $\boldsymbol{o}_i$) to U_i (the number of missing values of $\boldsymbol{o}_i$). With the Stirling approximation and a few minor approximations we receive Eq. 1, derived in Eq. 2.

2.2 Experimental approximation.

Both numerator and denominator of Eq. 1 are difficult to interpret.

*Technical University of Munich, r.leibrandt@tum.de
†Technical University of Munich, guennemann@in.tum.de

Equation 1 Analytical approximation of the number of grid vertices

$$\mathrm{E}[I] \approx \sum_{\omega_l} \left(\frac{|\mathcal{O}|^{(|\mathcal{O}|\cdot F-|\mathcal{O}|)} \cdot F^{(|\mathcal{O}|\cdot F-F)} \cdot \mathrm{e}^{F+|\mathcal{O}|-|\mathcal{O}|\cdot F} \cdot M^{M+0.5}}{\sqrt{2\cdot\pi}^{|\mathcal{O}|} \cdot |\mathcal{M}_1|^{|\mathcal{M}_1|+0.5} \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|^{|\mathcal{M}_{|\mathcal{O}|}|+0.5}} \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right)$$

Thus we approximated $\mathrm{E}[I]$ experimentally: First we generated different datasets by varying $|\mathcal{O}|$, F, M, and the seed of the MCAR value deletion and determined I for them. Then we used symbolic regression [3] to uncover a function $|\mathcal{O}| \times F \times M \to I$. Symbolic regression is an established method based on genetic programming for searching the space of mathematical expressions. To ensure the validity of the results, we ran multiple repetitions of symbolic regression. In some of these repetitions we included $(|\mathcal{O}|, F, M, I)$-sets with large I values. Since in such repetitions, a lot of emphasis is put on reducing the modeling error for these high-I-sets, neglecting low-I-sets, we also ran repetitions for low-I-sets, to model these correctly as well. Nearly all of the resulting formulas were polynomial functions, often of quadratic or cubic nature, with F and M having a larger influence than $|\mathcal{O}|$. Examples are shown in Eq. 3.

3 Virtual objects obtained by aggregation: Phase 1 for Area Detailed

Two types of virtual objects obtained by aggregation were presented [2]. The more sophisticated version uses principles of dynamic programming in order to find the maximal reachable area from a maximum in the space of missing features for an undetermined object. Fig. 2 shows in yellow such an area for the orange maximum. First we detail aspects concerning a single maximum, then aspects concerning the entire phase 1.

3.1 Reduction to directed acyclic graph. The reduction of phase 1 to the problem to find all the nodes reachable from a start node in a directed acyclic graph (DAG) can be performed by considering the grid vertices $\boldsymbol{o}_g$ as the nodes of the DAG. Two DAG nodes get connected if the grid vertices they represent are neighbors. If two objects are connecting in the DAG, the direction of the edge between them points from the object with the larger WDF to the object with the lower WDF. Finally, DAG nodes to which no edges point are maxima from where we start to traverse the DAG. We are only allowed to traverse the DAG in the direction of the edges.

3.2 Optimal structure and overlapping sub-problems. For a single maxima, the problem has optimal substructure and overlapping sub-problems.

Optimal structure means, that an optimal solution to a problem can be obtained by using optimal solutions of its subproblems. Here, "optimal" means "maximal number of nodes that can be reached from current nodes". While traversing the DAG, the nodes the algorithm can reach from a specific node $\boldsymbol{o}_g$ is this node $\boldsymbol{o}_g$ itself plus the union of the nodes the algorithm can reach from $\boldsymbol{o}_g$.

Overlapping subproblems means, that a problem can be broken down into subproblems which are reused. Two neighbor nodes of a specific node $\boldsymbol{o}_g$ can share another neighbor that is not $\boldsymbol{o}_g$. Fig. 3 shows how the cyan paths overlap for the paths taken from the orange maximum.

3.3 Memoization for multiple maxima. Fig. 3 shows how paths of a single maximum can overlap; however, we can also reuse calculated paths across maxima. Fig. 4 shows how the cyan sub-path is shared by the green paths of the orange 9 and the orange 10 (under the assumption that the depth-first algorithm moved from the green 6 to the cyan 5 first and not to the cyan 2). If we memoize the mass accumulated by the "cyan 5"-node during the processing of the "orange 9"-maximum, we can reuse this information during the processing of the orange 8.

However, additional aspects need to be considered, as shown in Fig. 5. Lets assume the orange 10 maximum is processed first and the displayed paths are taken; first the top path and later the bottom path until the yellow 6. The "yellow 5"-node memoizes that it reached five nodes (including itself), while the "yellow 6" node returns to it neighbor (the yellow 7) that it reached only one node (itself). The "yellow 6"-node does not continue to traverse the graph, because the "yellow 5"-node has already been traversed. When the orange 8 is processed and the "yellow 6"-node is queried regarding the number of nodes that can be reached from it, it returns "one" from the previous maximum's processing. But as we can see, the answer should have been that six nodes (including itself) were reached. Thus, during the processing of the orange 10, we need to establish a link between the "yellow 5"-node and the "yellow 6"-node and need to take this connection into account for the processing of later maxima.

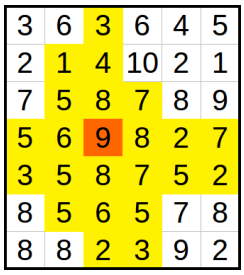

Fig. 2. Area of a maximum

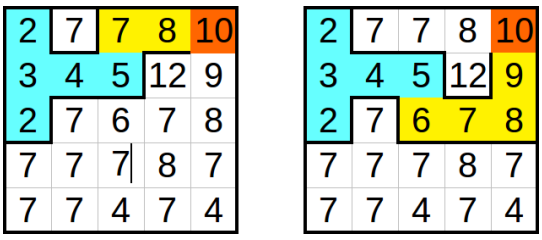

Fig. 3. Overlapping subproblems

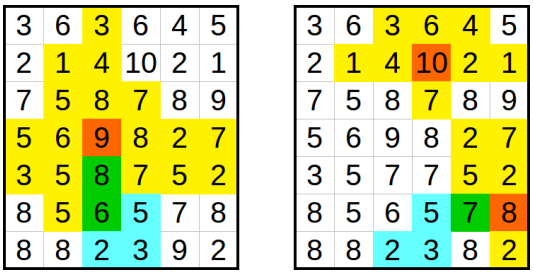

Fig. 4. Memoization for multiple maxima

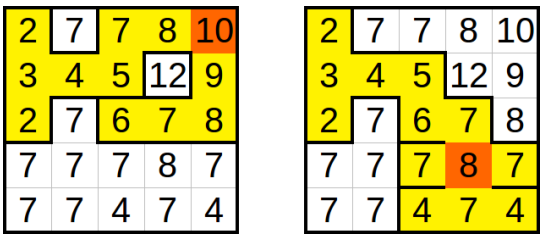

Fig. 5. Issues of memoization for multiple maxima

3.4 Overlapping nodes for multiple maxima. Fig. 5 shows that paths for the two different maxima overlap, resulting in the masses of the overlapping nodes being assigned to both maxima. This needs to be considered when calculating the final mass. For that we consider to how many maxima a node is associated with, divide the node's mass by the number of these maxima and associate the resulting fraction of the node's mass to each maxima. This way we ensure that each grid vertex contributes equally to the weights w_m of the maxima.

References

[1] R. Leibrandt and S. Günnemann, *Making Kernel Density Estimation Robust Towards Missing Values in Highly Incomplete Multivariate Data Without Imputation*, in Proc. of SIAM SDM, 2018.

[2] R. Leibrandt and S. Günnemann, *Robust Kernel Density Clustering of Highly Incomplete Data*, submitted to Proc. of SIAM SDM, 2019.

[3] M. Schmidt and H. Lipson, *Distilling Free-Form Natural Laws from Experimental Data*, Science, 324.5923 (2009), pp. 81–85.

Equation 2 Derivation of the analytical approximation of the number of grid vertices

The expected number of initials is

$$\mathrm{E}[I] = \sum_{\omega_l} \left(\mathrm{P}\left(\mathsf{U}_1 = |\mathcal{M}_1|, \ldots, \mathsf{U}_{|\mathcal{O}|} = |\mathcal{M}_{|\mathcal{O}|}|\right) \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right)$$

$$\mathrm{P}\left(\omega_l\right) = \mathrm{P}\left(\mathsf{U}_1 = |\mathcal{M}_1|, \ldots, \mathsf{U}_{|\mathcal{O}|} = |\mathcal{M}_{|\mathcal{O}|}|\right) = \frac{\binom{F}{|\mathcal{M}_1|} \cdot \binom{F}{|\mathcal{M}_2|} \cdot \ldots \cdot \binom{F}{|\mathcal{M}_{|\mathcal{O}|}|}}{\binom{|\mathcal{O}|\cdot F}{M}}$$

$$\omega_l = \left\{ \left\{ \mathsf{U}_1 = |\mathcal{M}_1|, \ldots, \mathsf{U}_{|\mathcal{O}|} = |\mathcal{M}_{|\mathcal{O}|}| \right\} \mid |\mathcal{M}_r| \in \{0, \ldots, F-1\} \ \wedge \ |\mathcal{M}_1| + \ldots + |\mathcal{M}_{|\mathcal{O}|}| = M \right\}$$

We calculate

$$\begin{aligned}
\mathrm{E}[I] =& \sum_{\omega_l} \left(\frac{\binom{F}{|\mathcal{M}_1|} \cdot \binom{F}{|\mathcal{M}_2|} \cdot \ldots \cdot \binom{F}{|\mathcal{M}_{|\mathcal{O}|}|}}{\binom{|\mathcal{O}|\cdot F}{M}} \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right) \\
=& \sum_{\omega_l} \left(\frac{\frac{F!}{|\mathcal{M}_1|!\cdot(F-|\mathcal{M}_1|)!} \cdot \ldots \cdot \frac{F!}{|\mathcal{M}_{|\mathcal{O}|}|!\cdot(F-|\mathcal{M}_{|\mathcal{O}|}|)!}}{\frac{|\mathcal{O}|!\cdot F!}{M!\cdot(|\mathcal{O}|\cdot F-M)!}} \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right) \\
=& \sum_{\omega_l} \left(\frac{F!^{(|\mathcal{O}|-1)} \cdot M! \cdot (|\mathcal{O}| \cdot F - M)!}{|\mathcal{M}_1|! \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|! \cdot (F-|\mathcal{M}_1|)! \cdot \ldots \cdot (F-|\mathcal{M}_{|\mathcal{O}|}|)! \cdot |\mathcal{O}|!} \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right) \\
\approx& \sum_{\omega_l} \left(\frac{M! \cdot (|\mathcal{O}| \cdot F)!}{|\mathcal{M}_1|! \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|! \cdot F! \cdot |\mathcal{O}|!} \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right) \\
\approx& \sum_{\omega_l} \left(\frac{M! \cdot (|\mathcal{O}| \cdot F)!}{|\mathcal{M}_1|! \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|! \cdot F! \cdot |\mathcal{O}|!} \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right)
\end{aligned}$$

and with the Stirling approximation

$$n! \approx \left(\sqrt{2 \cdot \pi \cdot n} \cdot \frac{n^n}{\mathrm{e}^n} \right)$$

we get

$$\begin{aligned}
& \frac{M! \cdot (|\mathcal{O}| \cdot F)!}{|\mathcal{M}_1|! \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|! \cdot F! \cdot |\mathcal{O}|!} \\
\approx& \frac{\sqrt{M} \cdot M^M \cdot (|\mathcal{O}| \cdot F)^{(|\mathcal{O}|\cdot F)} \cdot \mathrm{e}^M \cdot \mathrm{e}^{F+|\mathcal{O}|}}{\mathrm{e}^{M+|\mathcal{O}|\cdot F} \cdot \sqrt{2\cdot\pi}^{|\mathcal{O}|} \cdot \sqrt{|\mathcal{M}_1| \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|} \cdot |\mathcal{M}_1|^{|\mathcal{M}_1|} \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|^{|\mathcal{M}_{|\mathcal{O}|}|} \cdot F^F \cdot |\mathcal{O}|^{|\mathcal{O}|}} \\
=& \frac{M^{M+0.5} \cdot |\mathcal{O}|^{(|\mathcal{O}|\cdot F-|\mathcal{O}|)} \cdot F^{(|\mathcal{O}|\cdot F-F)} \cdot \mathrm{e}^{F+|\mathcal{O}|-|\mathcal{O}|\cdot F}}{\sqrt{2\cdot\pi}^{|\mathcal{O}|} \cdot |\mathcal{M}_1|^{|\mathcal{M}_1|+0.5} \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|^{|\mathcal{M}_{|\mathcal{O}|}|+0.5}} \\
=& \frac{|\mathcal{O}|^{(|\mathcal{O}|\cdot F-|\mathcal{O}|)} \cdot F^{(|\mathcal{O}|\cdot F-F)} \cdot \mathrm{e}^{F+|\mathcal{O}|-|\mathcal{O}|\cdot F}}{\sqrt{2\cdot\pi}^{|\mathcal{O}|}} \cdot \frac{M^{M+0.5}}{|\mathcal{M}_1|^{|\mathcal{M}_1|+0.5} \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|^{|\mathcal{M}_{|\mathcal{O}|}|+0.5}}
\end{aligned}$$

and finally

$$\mathrm{E}[I] = \sum_{\omega_l} \left(\mathrm{P}\left(\mathsf{U}_1 = |\mathcal{M}_1|, \ldots, \mathsf{U}_{|\mathcal{O}|} = |\mathcal{M}_{|\mathcal{O}|}|\right) \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right) \tag{3.6}$$

$$\approx \sum_{\omega_l} \left(\frac{|\mathcal{O}|^{(|\mathcal{O}|\cdot F-|\mathcal{O}|)} \cdot F^{(|\mathcal{O}|\cdot F-F)} \cdot \mathrm{e}^{F+|\mathcal{O}|-|\mathcal{O}|\cdot F}}{\sqrt{2\cdot\pi}^{|\mathcal{O}|}} \cdot \frac{M^{M+0.5}}{|\mathcal{M}_1|^{|\mathcal{M}_1|+0.5} \cdot \ldots \cdot |\mathcal{M}_{|\mathcal{O}|}|^{|\mathcal{M}_{|\mathcal{O}|}|+0.5}} \cdot \sum_{i=1}^{|\mathcal{O}|} g^{|\mathcal{M}_i|} \right) \tag{3.7}$$

Equation 3 Examples for equations uncovered by symbol regression

For the percentage of missing values $M\%$:

$$I = |\mathcal{O}| + 20.3 \cdot |\mathcal{O}| \cdot M\% + 13.8 \cdot |\mathcal{O}| \cdot F^3 \cdot M\%^2 - 118 \cdot |\mathcal{O}| \cdot M\%^2$$

$$I = |\mathcal{O}| + 9.93 \cdot F \cdot |\mathcal{O}| \cdot M\% + 186 \cdot F^3 \cdot M\%^2 + 9.93 \cdot |\mathcal{O}| \cdot F^3 \cdot M\%^2 - 2070 \cdot M\%^2 - 83.4 \cdot |\mathcal{O}| \cdot M\%^2$$

$$I = 38400 \cdot F \cdot |\mathcal{O}| \cdot M\% + 0.0629 \cdot F \cdot |\mathcal{O}| \cdot \left(2.72 \cdot F \cdot M\% - 0.00773 \cdot F \cdot |\mathcal{O}| \cdot M\%^4\right)^{6.89} - 2.43 \cdot 10^7$$

$$I = 0.0515 \cdot F \cdot |\mathcal{O}| \cdot \left(2.75 \cdot F \cdot M\% - M\%^3 \cdot \sqrt{0.0657 \cdot F \cdot |\mathcal{O}|}\right)^7$$

For the number of missing values $M\%$:

$$I = |\mathcal{O}| + \frac{110 \cdot F \cdot M \cdot \sqrt{M + 0.00171 \cdot |\mathcal{O}|^2}}{|\mathcal{O}| + F} - M$$

$$I = |\mathcal{O}| + \frac{105 \cdot F \cdot M \cdot \sqrt{M + 0.00165 \cdot |\mathcal{O}|^2}}{|\mathcal{O}| + F}$$

$$I = |\mathcal{O}| + \frac{91.3 \cdot F \cdot M \cdot \sqrt{M + 0.00231 \cdot |\mathcal{O}|^2}}{|\mathcal{O}|}$$

$$I = |\mathcal{O}| + 60.9 \cdot F \cdot M \cdot \sqrt{\frac{0.0201 \cdot M}{|\mathcal{O}|}}$$

$$I = |\mathcal{O}| + 4.38 \cdot F \cdot M + \frac{2010 \cdot M}{|\mathcal{O}|}$$

$$I = |\mathcal{O}| + 8.85 \cdot M + \frac{(16.5 \cdot F - 31) \cdot M^2}{|\mathcal{O}|}$$

$$I = |\mathcal{O}| + 9.39 \cdot M + \frac{39.6 \cdot M^2}{9.39 + |\mathcal{O}|} - \frac{110 \cdot M^2}{|\mathcal{O}| \cdot M} + \frac{7.24 \cdot F \cdot M^2 - 66.9 \cdot F \cdot M}{|\mathcal{O}| - 19.8}$$

$$I = \frac{1.26 \cdot 10^5 \cdot F \cdot M^{3.39}}{|\mathcal{O}|^3} - 4.96 \cdot 10^6 \cdot F$$