Analysis of high-dimensional data using local input space histograms

Article *in* Neurocomputing · April 2015

DOI: 10.1016/j.neucom.2014.12.094

2 authors:

Jochen Kerdels
FernUniversität in Hagen
36 PUBLICATIONS **131** CITATIONS
SEE PROFILE

Gabriele Peters
FernUniversität in Hagen
88 PUBLICATIONS **565** CITATIONS
SEE PROFILE

Jochen Kerdels and Gabriele Peters, Analysis of High-Dimensional Data Using Local Input Space Histograms, Neurocomputing (2015), DOI: 10.1016/j.neucom.2014.12.094, in press, accepted manuscript (unedited version), available online: 4-APR-2015

Analysis of High-Dimensional Data Using Local Input Space Histograms

Jochen Kerdels*, Gabriele Peters

University of Hagen - Faculty of Mathematics and Computer Science
Human-Computer Interaction
Universitätsstrasse 1, D-58084 Hagen, Germany

Abstract

The idea of *local input space histograms* was recently introduced as a means to augment prototype-based vector quantization methods in order to gather more information about the structure of the respective input space. Here we investigate the utility of this new idea for analysing and clustering high-dimensional data. Our results demonstrate, that the additional information gained about the input space structure can be used to enable and improve visualization and hierarchical clustering. Furthermore, we show that contrary to common view the Minkowski distance with $p > 1$ can be a meaningful distance measure for high-dimensional data.

Keywords: Local Input Space Histograms, Prototype-based Vector Quantization, Growing Neural Gas, Curse of Dimensionality, Minkowski Distance

1. Introduction

The analysis of data on a large scale is a challenging task. Commonly there is only few apriori knowledge available about structures contained within the data, e.g., information about possible classes the data could be partitioned into. In such a case methods that utilize forms of *unsupervised competitive learning*

*Corresponding author
Email address: Jochen.Kerdels@FernUni-Hagen.de (Jochen Kerdels)

like the self-organizing map (SOM, [1]) or neural gas (NG, [2]) can be used to discover potential structures in the data. Both, the SOM and NG are prototype-based vector quantization methods that use a set of prototypes to cover the particular input space as well as possible, i.e, to minimize the quantization error based on a given dissimilarity measure.

If there is little information about the structure of the data the euclidian distance is often choosen as a "default" dissimilarity measure. In that case the individual prototypes can only represent local regions of the input space as convex polyhedrons and more complex structures must be approximated piecewise by multiple prototypes. In order to gather more information about the input space structure between prototypes the idea of *local input space histograms* [3] was introduced recently. As a proof of concept it has been shown that augmenting a growing neural gas (GNG, [4]) with local input space histograms can improve the discovery of non-convex clusters in two-dimensional datasets.

In this paper we investigate the utility of local input space histograms for analysing and clustering high-dimensional data. Section 2 introduces the methods and materials used in the subsequently described experiments. In particular, the section describes how a prototype-based vector quantization method – here a GNG – can be augmented by local input space histograms. In section 3 the behavior of local input space histograms is analysed for high-dimensional random data as well as high-dimensional color histogram data. Section 4 discusses a number of interesting aspects of our results. Finally, a short conclusion and suggestions for further research are provided in section 5.

2. Materials and methods

Growing Neural Gas Revisited. To investigate the utility of local input space histograms for the analysis of high-dimensional data we extended a GNG as an exemplary prototype-based method. The GNG is a *topology representing network* [5], i.e., it uses a data-driven growth process to approximate the topology of the input space instead of using a fixed network topology like, e.g., a SOM

does. Here we summarize the operation of the growing neural gas algorithm as described by Fritzke [4]. The growing neural gas is a network that consists of a set A of units and a set C of edges. Each unit $a \in A$ can be described by a tuple[1] (w, e) with the *prototype* $w \in \mathbb{R}^n$, n being the dimension of the input space, and the accumulated error variable $e \in \mathbb{R}$. Each edge $c \in C$ can be described by a tuple (a, b, t) with the units $a, b \in A \wedge a \neq b$ that are connected by the edge and the variable $t \in \mathbb{N}$ which stores the current age of the edge. The direct neighborhood D_a of a unit a is defined as $D_a := \{b | \exists\, (a, b, t) \in C,\, b \in A,\, t \in \mathbb{N}\}$. The network is initialized with two units that have random prototypes and accumulated error variables set to zero.

A given input $\xi \in \mathbb{R}^n$ is processed by the network in the following way:

- Find the two units s_1 and s_2 whose prototypes are closest to the input ξ:

  $$s_1 := \operatorname{argmin} \left\{ a^{(1)} - \xi | a \in A \right\}, \quad s_2 := \operatorname{argmin} \left\{ a^{(1)} - \xi | a \in A \setminus \{s_1\} \right\}.$$

- Increment the age of all edges connected to s_1:

  $$c^{(3)} := 0, \quad c \in C \ \wedge\ c^{(1)} = s_1 \ \wedge\ c^{(2)} = b,\, \forall b \in D_{s_1}.$$

- If no edge exists between s_1 and s_2, create one:

  $$C := C \cup \{(s_1, s_2, 0)\}.$$

- Reset the age of the edge between s_1 and s_2 to zero:

  $$c^{(3)} := 0, \quad c \in C \ \wedge\ c^{(1)} = s_1 \ \wedge\ c^{(2)} = s_2.$$

- Add the squared distance between the input ξ and the prototype of unit s_1 to the accumulated error of s_1:

  $$s_1^{(2)} := s_1^{(2)} + \| s_1^{(1)} - \xi \|^2$$

- Adapt the prototype of s_1 and all prototypes of its direct neighbors $b \in D_{s_1}$:

  $$\Delta s_1^{(1)} := \epsilon_b \left(\xi - s_1^{(1)} \right), \quad \Delta b^{(1)} := \epsilon_n \left(\xi - b^{(1)} \right),\ \forall b \in D_{s_1}.$$

[1] We use the notation $a^{(i)}$ to reference the ith element of a tuple beginning with index 1.

- Remove all edges with an age above a given threshold $t_{\max}$ and remove all units that no longer have any edges connected to them.

- If an integer-multiple of λ inputs was presented to the network insert a new unit r. The new unit is inserted between the unit $q \in A$ with the maximum accumulated error and the unit $f \in D_q$ which has the largest accumulated error among the neighbors of q, i.e., the prototype of unit r is set to:

  $r^{(1)} := (q^{(1)} + f^{(1)})/2.$

  Create edges between q and r as well as f and r, and remove the edge betweem the units q and f. Decrease the accumulated errors of q and f by a factor α and set the accumulated error of the new unit r to the decreased accumulated error of unit q.

- Finally, decrease the accumulated error of all units in A by a factor β.

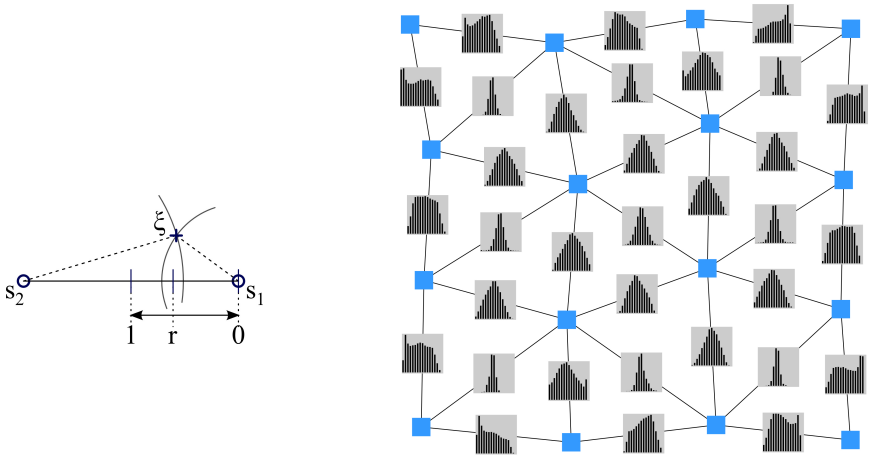
Typically, the inputs ξ are randomly choosen from a set of training data and fed into the network until a given halting criterion (e.g., a maximum network size) is met.

In all experiments the following parameter values were used:

$$\begin{array}{lllllllll} \epsilon_b & = & 0.01, & \epsilon_n & = & 0.0001, & t_{\max} & = & 500, \\ \lambda & = & 2000, & \alpha & = & 0.5, & \beta & = & 0.0005. \end{array}$$

The parameters deviate from the values proposed by Fritzke [4]. They result in a slower development of the GNG which turns out to be more robust with respect to high-dimensional inputs. A slower development compensates for possible inhomogeneities in the training data, which are in general more likely to occur in high-dimensional data as the ratio between the number of available training data points and the size of the input space typically diverges with increasing dimension.

Local Input Space Histograms. As described above, edges in a GNG network are created between the first and second best matching units (BMUs) s_1 and s_2 of



(a) (b)

Figure 1: **(a)** Geometric interpretation of the distance ratio r. **(b)** Example of local input space histograms for a small, two-dimensional GNG receiving uniform, random input.

each input ξ and are maintained as long as they are used regularly. Thus, the neighborhood relations among units represented by the GNG network indicate that the input space between connected units is not empty. However, the mere existence of an edge does not provide any further information about the underlying input space structure. The core idea of local input space histograms is to increase the available information in this regard by adding a small histogram $H = \{h_0, \ldots, h_{k-1}\}$, e.g., with $k = 16$ bins, to each edge $c \in C$, $c = (a, b, t, H)$ of the GNG network and to update this histogram for those inputs ξ that are mapped to the corresponding edge using a distance ratio r:

$$r := \frac{\|s_1^{(1)} - \xi\| - \|s_2^{(1)} - \xi\|}{\|s_1^{(1)} - s_2^{(1)}\|} + 1,$$

with $s_1^{(1)}$ and $s_2^{(1)}$ the prototypes of the first and second BMUs for the given input ξ.

The ratio r lies in the interval $[0, 1]$ and describes how close the prototype of the best matching unit s_1 is to the input ξ in relation to the prototype of the

second best matching unit s_2. A geometric interpretation of the distance ratio is depicted in figure 1a. As a local input space histogram $c^{(4)}$ is part of an edge $c \in C$ it is shared by the two units $c^{(1)}$ and $c^{(2)}$. Thus, the ratio r is used to either update the upper or the lower half of the histogram depending which of the units is the BMU s_1:

$$\Delta h_u = 1, \quad u = \begin{cases} \lfloor k\,(r/2) \rfloor & \text{if } c^{(1)} = s_1, \\ \lfloor k\,(1 - r/2) \rfloor & \text{if } c^{(2)} = s_1, \end{cases} \quad h_u \in c^{(4)} = \{h_0, \ldots, h_{k-1}\}\,.$$

The resulting histogram represents the distribution of the approximate, relative positions of those inputs that are located somewhere around the two connected units. Figure 1b provides an example of local input space histograms occuring in a two-dimensional GNG that received uniform, random input.

The additional information provided by the local input space histograms allows to characterize the input space in more detail. For example, it can be estimated if the input space between two connected units is sparse or dense. One measure to quantify this property is the average bin error[2] $\bar{e}_H$ of a histogram H:

$$\bar{e}_H := \frac{1}{k} \sum_{i=0}^{k-1} e_i, \quad e_i := \begin{cases} \sqrt{h_i}/h_i & \text{if } h_i > 0, \\ 1 & \text{if } h_i = 0, \end{cases} \quad h_i \in H = \{h_0, \ldots, h_{k-1}\}\,.$$

In case of a local input space histogram $c^{(4)}$ the value of $\bar{e}_{c^{(4)}}$ will be near 1 if the corresponding region of input space is sparse and it will be close to 0 in case the input space is dense.

Distance Measures. The analysis of high-dimensional data spaces is accompanied by a number of problems that are commonly referred to as the "curse of dimensionality" [6]. In this context a major problem is that the ability to discriminate data points by their relative distances diminishes with increasing dimensionality [7]. To observe the impact of different distance measures on the GNG and the local input space histograms we use the Minkowski distance d_p

[2]Note: the definition of the average bin error given here differs from [3].

in our analysis with varying values for p:

$$d_p(x,y) := \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}, \quad x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n).$$

By choosing the Minkowski distance a range of popular distance measures can be covered: for $p = 1$ it is equivalent to the Manhattan distance, for $p = 2$ it is equivalent to the Euclidian distance , and for $p \to \infty$ it approaches the Chebyshev distance.

Data. Experiments performed on random data use uniformly distributed values in the interval $[0, 1]$ for each vector component of the inputs ξ. The random values are generated by the *Mersenne Twister* pseudorandom number generator [8] using the implementation provided by the ROOT data analysis framework [9].

Experiments performed on color histogram data of images use the oxford 102 flowers image dataset[3], which contains 8189 images of flowers from 102 categories [10]. Color histograms were generated by transforming the images into HSV color space and using the resulting hue values weighted pixelwise by either saturation or value depending on which was smaller. The use of weighted hue values is motivated by the fact that neither very bright (small saturation) nor very dark (small value) image regions contribute much to the color information in an image. Each histogram has 360 bins corresponding to the 360 possible hue values of the HSV color space. The number of entries in the histogram is normalized to an image size of 1000×1000 pixels. Depending on the particular experiment the number of bins was scaled down using linear interpolation to provide comparable input vectors that have the same dimensionality as the random data vectors used. In general, this linear interpolation is possible because neighboring bins in the color histogram represent similar hues.

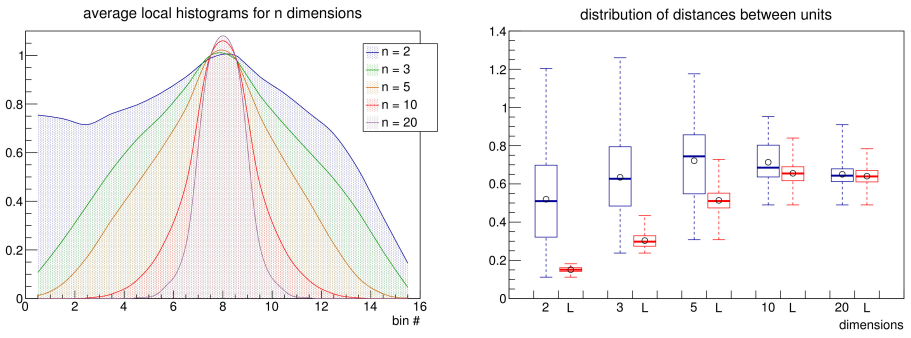
[3] available at http://www.robots.ox.ac.uk/~vgg/data/flowers/102/

3. Results

Random Data Experiments. In order to establish a baseline on how the appearance of the local input space histograms changes with respect to either a dense or a sparse input space and varying parameter p of the Minkowski distance a set of experiments using random input data was conducted. For each run a GNG with a maximum of 50 units was used. To characterize the appearance of the local input space histograms arising at the edges of the GNG the histograms themselves were clustered as well. As no further information about the potential characteristics of the histogram clusters, e.g., an expected number of clusters, was available, it was decided to also use a GNG to cluster the histograms. This secondary GNG had 20 units and used as distance measure the euclidian distance, i.e., the Minkowksi distance with $p = 2$ as originally proposed by Fritzke [4]. For each input to the primary GNG a local input space histogram was choosen randomly among those linked to the respective BMU and fed as input vector into the secondary GNG.

To test the influence of an increasingly sparse input space on the characteristic shapes of the local input space histograms a set of seven experiment runs with one million inputs of n-dimensional, random data for $n = \{2, 3, 4, 5, 10, 20, 40\}$ and fixed parameter $p = 2$ in the primary GNG were performed[4]. With increasing dimension n the primary GNG has to cover an exponentially growing volume of input space with a constant number of units. Similarly, the constant number of inputs is uniformly spread across this exponentially growing volume, too. As a consequence, the input space as represented by the constant number of input samples becomes increasingly sparse and the particular inputs to the GNG approach an equidistant position between their respective first and second BMUs. Likewise, the distances between the GNG units themselves become more and more similar and the locality given by the GNG edges gets essentially lost as the distributions of pairwise distances between all units and the distributions

[4]Data for $n = \{4, 40\}$ omitted in figure 2.

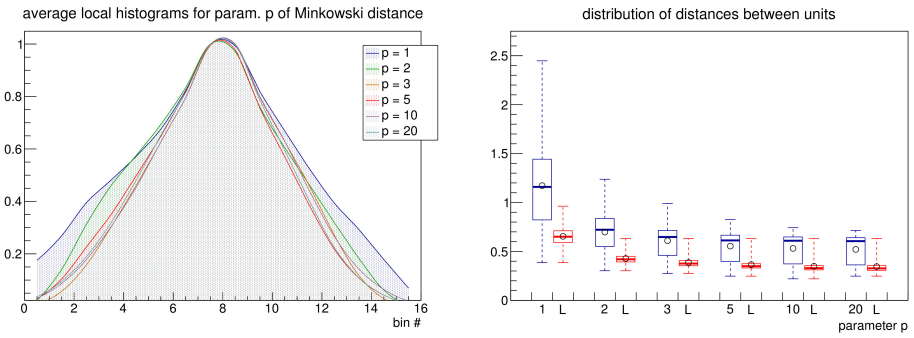


(a) (b)

Figure 2: **(a)** Change of (approximated) average local input space histograms with increasing dimension n in GNGs of 50 units receiving random input (histograms drawn with a smoothed curve). **(b)** Distributions of pairwise distances between the units of the GNGs in (a). Blue boxes describe the pairwise distances between all units, red boxes (L-columns) describe the pairwise distances between all units connected by edges. Bottom and top of dashed lines represent minimum and maximum values, bottom and top of each box represent lower and upper quartiles, thick lines represent medians, and circles represent mean values of the distributions.

of pairwise distances between connected units converge (see figure 2b). Consequently, the average degree of the GNG units increases as well (see figure S6). This dynamic is reflected in the shape of the local input space histograms. Supplementary figure S1 shows the prototypes of all secondary GNG units for each run. The prototypes represent the typical shapes of local input space histograms that emerge in primary GNGs with uniformly distributed units. The variation in appearance which can be observed in the two-dimensional case (see also fig. 1b) is reduced with increasing dimension and blends into a common triangular shape that gets increasingly narrower. Therefore figure 2a compares the change of the average local input space histograms[5] arising in the primary GNGs with increasing dimension n. The central, sharp peak for $n = 20$ represents the typical shape of a local input space histogram of a GNG edge that

[5]The average local input space histograms shown in figure 2a were approximated using the prototypes of the secondary GNG.



(a) (b)

Figure 3: **(a)** Change of (approximated) average local input space histograms with varying parameter p of the Minkowski distance in GNGs of 50 units receiving 4-dimensional, uniform, random input. **(b)** Distributions of pairwise distances between the units of the GNGs in (a). Same format as in fig. 2b.

spans a sparse region of input space.

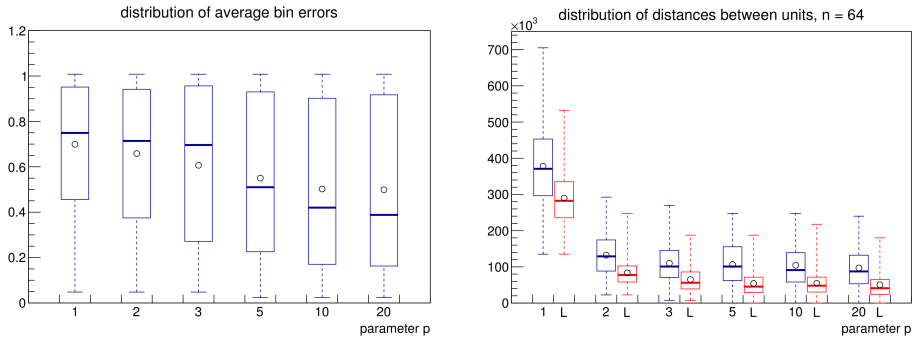
Based on the previous results we tested the influence of the choosen distance measure in two scenarios with a dense and a sparse input space, respectively. Two sets of experiment runs with varying parameter $p = \{0.5, 1, 2, 3, 5, 10, 20\}$ of the Minkowski distance and fixed dimensions $n = \{4, 64\}$ were performed[6]. The two dimension values were choosen to obtain data for a scenario where the input space is still sufficiently dense and the average local input space histograms have already a uniform triangular shape ($n = 4$), and for a scenario where the input space is guaranteed to be sparse ($n = 64$). The increase of parameter p leads in either scenario to a compression of distances between the GNG units similar to the effect caused by an increase of input space dimension (figures 3b and S4). However, in case of a dense input space ($n = 4$) the locality given by the GNG edges stays intact with increasing values of p, i.e., the pairwise distances of connected units differ clearly from the pairwise distances between all units. The median value of all pairwise distances between the units

[6]Data for $p = 0.5$ and $n = 64$ omitted in figure 3.

gets pushed towards the upper quartile while the bulk of all pairwise distances between connected units stays below the lower quartile of distances between all units. This behavior can be explained by the fact that with increasing values of p differences between vectors that are only distributed across few vector components are emphasized whereas differences that are spread across many vector components are deemphasized. Thus, despite the absolute compression of distance values, the *relative contrast* between near and far distances of inputs that share a common BMU actually improves with increasing values of p. This interesting property will be discussed in more detail in section 4.

The appearance of the local input space histograms is virtually unaffected by the described changes in the distance distributions (figures 3a, S2, and S3) as the distance ratio r on which the histograms are based depends only on the *relative* distances between an input vector and its respective first and second BMUs. This assertion, as will be shown in the next section, holds true for high-dimensional, but locally dense data as well.

Color Data Experiments. The experiments using random data demonstrated two properties of local input space histograms: they are robust with respect to the choosen distance measure, and they take on a distinct, spike-like shape if the corresponding GNG edge spans a region of sparse input space. It remains to be shown that local input space histograms can be used to identify locally dense regions in a high-dimensional input space. In general, such a locally dense region *must* be a low-dimensional submanifold in order to be detectable through a limited number of input samples. In this regard color histograms of images represent a suitable test case as they provide a high-dimensional input space that most likely will contain locally dense regions. Natural images – in this case images of flowers – contain usually only a small number of main hues while the resulting color histograms themselves are high-dimensional. The basic experimental setup for the color data experiments was equal to the random data experiments described above. A primary GNG with a maximum of 50 units processed color histogram inputs. A secondary GNG with a maximum



(a) (b)

Figure 4: **(a)** Distributions of the local input space histograms' average bin error $\bar{e}$ in GNGs of 50 units receiving 64-dimensional color histogram input with varying parameter p of the Minkowski distance. **(b)** Distributions of pairwise distances between the units of the GNGs in (a). Same format as in fig. 2b.

of 20 units was again used to assess the appearance of the local input space histograms arising in the primary GNG. The color histograms were generated once from the images in the 102 flower dataset and were then fed into the primary GNG in random order one million times per run.

A set of experiment runs with varying parameter $p = \{0.5, 1, 2, 3, 5, 10, 20\}$ of the Minkowski distance and 64-dimensional color histograms as input were performed. The color histograms were scaled down from the original 360-dimensional color histograms as described in section 2. The dimensionality was choosen to be 64 in order to be comparable to the high-dimensional, sparse input space scenario of the previously described random data experiments.

The shapes of local input space histograms discovered by the secondary GNG clearly indicate that the color histogram input space has locally dense regions (figure S5). Interestingly, the variation of the local input space histogram shapes increases with bigger values of parameter p. Particularly those shapes that indicate an underlying dense input space are appearing more frequently with increasing p. This trend is also reflected in the distributions of the local input space histograms' average bin error $\bar{e}$ shown in figure 4a. Another indication

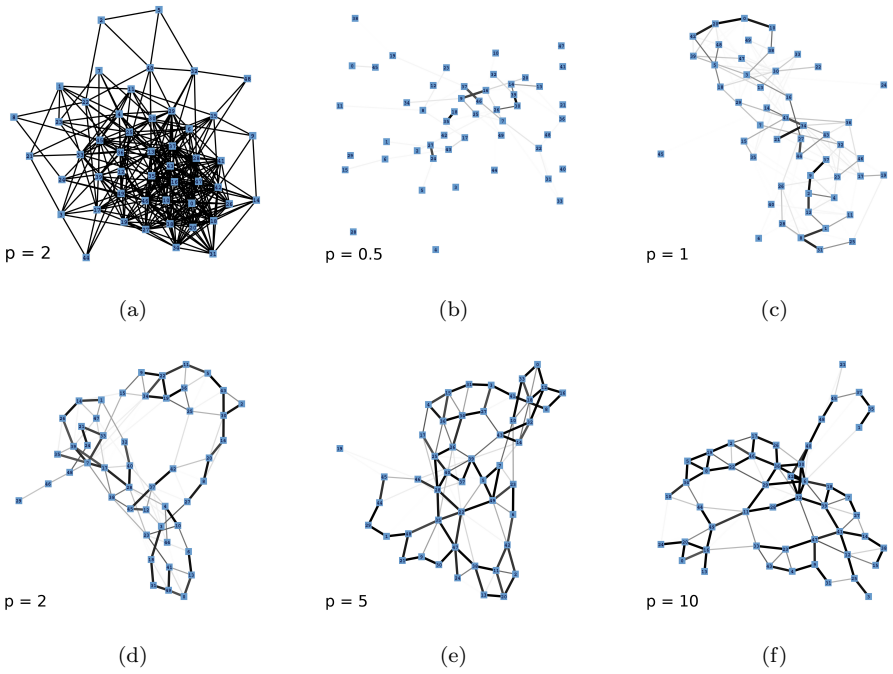


Figure 5: **(a)** Force-based graph drawing of a GNG (50 units, 64-dimensional color histogram input, fixed parameter $p = 2$) using all edges of the GNG indiscriminately. **(b-f)** Force-based graph drawings of GNGs (50 units, 64-dimensional color histogram input, varying parameter p) using the edges of the GNGs weighted according to their local input space histograms.

that the color histogram input space has locally dense regions is given by the distributions of pairwise distances shown in figure 4b which are similar to the distributions of dense, 4-dimensional input space shown in figure 3, i.e., the pairwise distances of connected units differ clearly from the pairwise distances between all units. Despite the high-dimensionality of the color histograms the locality given by the GNG edges appears to remain intact and even improve with increasing values of p. The latter assertion is supported by the fact that the degree of the GNG units, i.e., the number of incident GNG edges, decreases with increasing values of p (figure S7) as it would be expected of GNG units lying in a low-dimensional submanifold.

In order to check if these indicators are actually corresponding to some structure of the input space or are just artefacts, we adapted the force-based graph

drawing approach of Fruchterman and Reingold [11] to visualize the structures in question. The drawing approach of Fruchterman and Reingold uses two forces to control the position of the individual nodes of the graph: a repelling force that all nodes exert on each other, and an attractive force that each node exerts on all nodes connected to it. We modified the latter force to be weighted by an edge strength s defined as:

$$s := 1 - \frac{1}{\left(1 + 0.5\,\mathrm{e}^{-25(\bar{e}-0.25)}\right)^2},$$

where $\bar{e}$ is the average bin error of the local input space histogram of the particular edge. In addition, the edge color and thickness in the drawing is also controlled by the edge strength (higher strength = thicker and darker). The resulting graph drawings for GNGs with 50 units that received 64-dimensional color histogram input are shown in figure 5b-f. For comparison, figure 5a shows the result of the unmodified drawing algorithm. The drawings seem to support the previous observations. With increasing value of p the locality among neighboring units appears to increase as more and more units get connected by strong edges, i.e., edges that cover dense regions of the input space.

To examine if the resulting structures are also semantically meaningful we mapped the images whose color histograms were closest to the prototypes of the GNG onto the graph drawing shown in figure 5f. The mapping shown in figure 6 demonstrates, that the strong edges selected on the basis of their local input space histograms indeed represent meaningful neighborhood relations.

As an alternative way to assess the information provided by the local input space histograms a hierarchical clustering of GNG units was implemented using a bottom-up approach. In case two GNG units were connected by an edge, the distance between the two units was defined as the average bin error $\bar{e}$ of the corresponding local input space histogram, otherwise a constant distance of 1 was assumed. Single-linkage was used as linkage criterion. Figure 7b-f shows the resulting dendrograms for GNGs with 50 units that received 64-dimensional color histogram input. For comparison, figure 7a shows the dendrogram for a hierarchical clustering that uses the euclidian distance between the units as

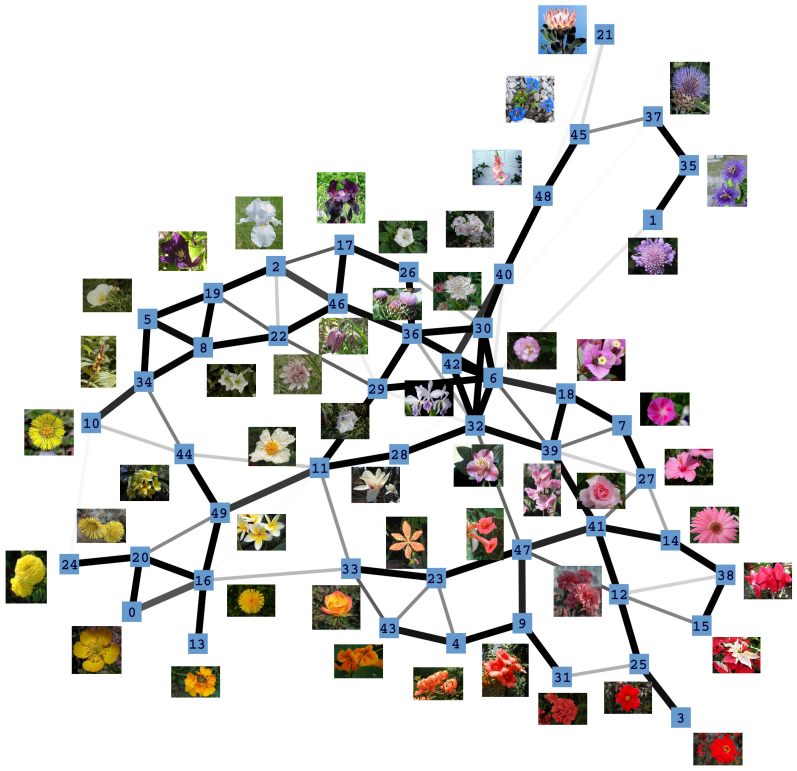


Figure 6: Mapping of closest input images to corresponding GNG units in a force-based graph drawing approach of a GNG (50 units, 64-dimensional color histogram input, fixed parameter $p = 10$).

distance measure. Similar to the force-based graph drawing approach the hierarchical clustering too shows an increase in locality among units with increasing value of p. In contrast, the hierarchical clustering using the euclidian distance does not reveal much structure and is afflicted by chaining.

Analogous to the force-based graph drawing we tested the semantic meaningfulness of this approach by mapping the images whose color histograms were closest to the prototypes of the GNG onto the hierarchical clustering dendrogram of a GNG. The mapping shown in figure 8 illustrates, that in this case too the local input space histograms provide usefull information about the structure of the input space.

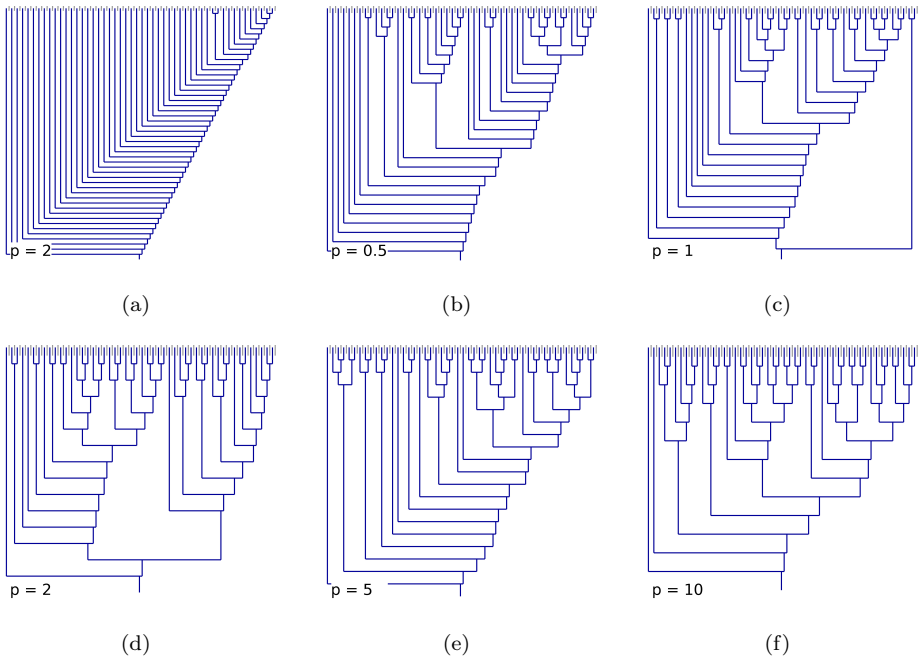


Figure 7: **(a)** Hierarchical clustering a GNG (50 units, 64-dimensional color histogram input, fixed parameter $p = 2$) using Minkowski distance with $p = 2$ as element-wise distance measure and single linkage as linkage criterion. **(b-f)** Hierarchical clustering of GNGs (50 units, 64-dimensional color histogram input, varying parameter p) using the average bin error of the local input space histograms as element-wise distance measure and single linkage as linkage criterion.

4. Discussion

We investigated the utility of local input space histograms as an extension to prototype-based vector quantization methods for the analysis and clustering of high-dimensional data. The obtained results indicate that local input space histograms can provide useful information to support the characterization of input space.

One interesting – and to some degree surprising – aspect of our results is the increased visibility of input space structures with increasing values of parameter p of the Minkowski distance. This result contradicts the common view [7, 12, 6] that for high-dimensional data it is favorable to use the Minkowki distance with

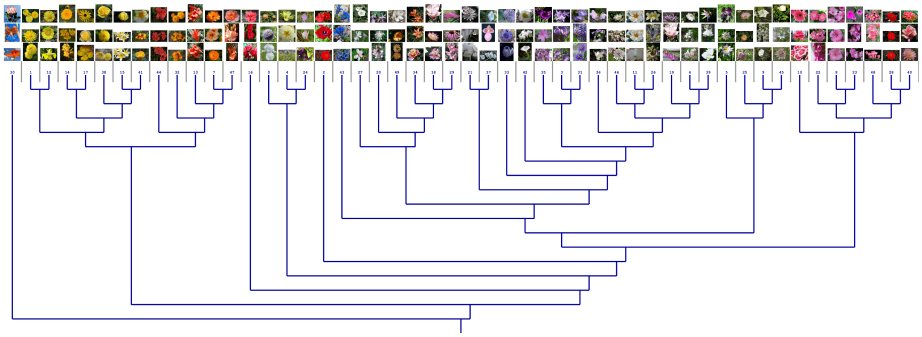
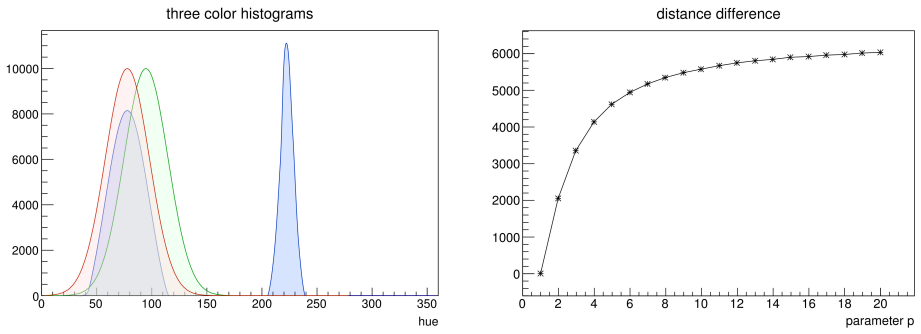


Figure 8: Mapping of closest input images to corresponding GNG units for a hierarchical clustering of a GNG (50 units, 64-dimensional color histogram input, fixed parameter $p = 10$). A larger version of this figure is provided in the supplementary material (figure S8).

$p = 1$ or even fractional values with $p < 1$. In case of Hinneburg et al. [12] this view is based on the analysis of nearest neighbor search in databases.

Our results indicate that in the case of prototype-based methods like the GNG this view on the Minkowski parameter p is not applicable in general and the best value for the parameter depends on the particular type of data that is processed. The behavior of the Minkowski distance with increasing values of its parameter p can be illustrated with a simple example. Figure 9a shows three idealized color histograms. The red colored histogram describes an image with predominantly orange, yellow, and green colors. The green colored histogram is a slightly shifted copy of the red colored histogram and it describes an image with a similar color distribution. In contrast, the blue colored histogram describes an image that has less orange, yellow, and green content and contains additional blue colors. Intuitively, one would expect the red and green colored histograms to be more similar than the red and blue colored histograms. However, the histograms are crafted in such a way, that for $p = 1$ the distance d_{rg} between the red and green colored histograms and the distance d_{rb} between the red and blue colored histograms are equal. With increasing value of p, the distances d_{rg} and d_{rb} diverge. The graph shown in figure 9b illustrates this by plotting the difference of the distances, $d_{rb} - d_{rg}$, for increasing values of p. This



(a) (b)

Figure 9: **(a)** Example of three color histograms. The red colored histogram describes an image with predominantly orange, yellow, and green hues. The green colored histogram is a slightly shifted copy of the red colored histogram. In contrast, the blue colored histogram has an additional sharp peak at the blue hues. **(b)** The graph describes the difference of the distance between the red and green colored histograms and the distance between the red and blue colored histograms from (a) for increasing values of parameter p of the Minkowski distance.

divergence of the distances can be explained by the fact that the difference between the red and the green colored histogram is spread over many bins whereas the difference between the red and the blue colored histogram is concentrated in only a few bins (mainly the blue hues). With increasing p large differences in individual bins get emphasized whereas small differences in individual bins get deemphasized. In the extreme case, for $p \to \infty$, the Minkowski distance approaches the Chebyshev distance where the distance is determined exclusively by the bin with the highest difference. Thus, the common view that pairwise distances are *generally* no longer meaningful in very high-dimensional spaces is not entirely correct. If the difference between classes in a particular type of data corresponds to large changes in a small number of data elements rather than small changes in a large number of data elements the Minkowski distance with high value p can actually improve the contrast between intra- and inter-class pairwise distances.

As a consequence for prototype-based methods like the GNG, the *relative*

contrast between near and far inputs with respect to a common BMU can, depending on the particular type of input data, increase with p and improve the locality between first and second BMUs. In such a case, the average degree of the GNG units decreases (see, e.g., figure S7) and, as a further consequence, the influence of indirect adaptation decreases, too. In case of histogram data, e.g., color histograms, this behavior appears to be favorable and implies the use of higher values for p. If this holds true for other kinds of data, e.g., in case of a "bag of features", remains to be determined.

5. Conclusion

The utility of local input space histograms for analysing and clustering high-dimensional data was investigated. It could be shown that they provide useful, additional information about the structure of the input space that can be used, e.g., for visualization and hierarchical clustering of the data. Furthermore, our results demonstrate that contrary to common view the Minkowski distance with $p > 1$ can be a meaningful distance measure for high-dimensional data.

Based on these promising, early results a number of interesting questions can be identified for future research:

- How do other distance measures affect the behavior of local input space histograms, e.g., crossbin distances for histograms like the earth movers distance [13, 14] or the cosine distance for sparse feature vectors?
- How do the parameters of the GNG, especially the maximum number of units, affect the results? Preliminary data indicate, that with an increasing number of GNG units chains of units connected by edges with low average bin error emerge. This behavior may be used to define a robust criterion to determine the maximum number of units automatically.
- Which kind of information can be gained from other types of data, e.g., when a "bag of features" approach is used?

- How can other clustering methods such as density-based clustering be supported by the information contained in local input space histograms?

- Can local input space histograms support classification? For example, if an input is mapped onto an edge with high average bin error, it could be identified as outlier. Alternatively, the average bin error of an edge could be used as some form of uncertainty measure for the classification of an input.

Furthermore, the concept of local input space histograms should be easily adaptable to other prototype-based vector quantization methods. For example, they could be used in SOMs to identify borders between regions in the resulting, two-dimensional mapping.

References

[1] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1) (1982) 59–69.

[2] T. Martinetz, S. Berkovich, K. Schulten, 'Neural-gas' network for vector quantization and its application to time-series prediction, Neural Networks, IEEE Transactions on 4 (4) (1993) 558–569.

[3] J. Kerdels, G. Peters, Supporting GNG-based clustering with local input space histograms, in: M. Verleysen (Ed.), Proceedings of the 22nd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Louvain-la-Neuve, Belgique, 2014, pp. 559–564.

[4] B. Fritzke, A growing neural gas network learns topologies, in: Advances in Neural Information Processing Systems 7, MIT Press, 1995, pp. 625–632.

[5] T. M. Martinetz, K. Schulten, Topology representing networks, Neural Networks 7 (1994) 507–522.

[6] H.-P. Kriegel, P. Kröger, A. Zimek, Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering, ACM Trans. Knowl. Discov. Data 3 (1) (2009) 1:1–1:58.

[7] C. Aggarwal, A. Hinneburg, D. Keim, On the surprising behavior of distance metrics in high dimensional space, in: J. Van den Bussche, V. Vianu (Eds.), Database Theory — ICDT 2001, Vol. 1973 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2001, pp. 420–434.

[8] M. Matsumoto, T. Nishimura, Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator, ACM Trans. Model. Comput. Simul. 8 (1) (1998) 3–30.

[9] R. Brun, F. Rademakers, ROOT - an object oriented data analysis framework, in: AIHENP'96 Workshop, Lausanne, Vol. 389, 1996, pp. 81–86.

[10] M.-E. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 2008, pp. 722–729.

[11] T. M. J. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement, Softw. Pract. Exper. 21 (11) (1991) 1129–1164.

[12] A. Hinneburg, C. C. Aggarwal, D. A. Keim, What is the nearest neighbor in high dimensional spaces?, in: Proceedings of the 26th International Conference on Very Large Data Bases, VLDB '00, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2000, pp. 506–515.

[13] Y. Rubner, C. Tomasi, L. Guibas, A metric for distributions with applications to image databases, in: Computer Vision, 1998. Sixth International Conference on, 1998, pp. 59–66.

[14] S. Shirdhonkar, D. Jacobs, Approximate earth mover's distance in linear time, in: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, 2008, pp. 1–8.

Supplementary Material for "Analysis of High-Dimensional Data Using Local Input Space Histograms"

Jochen Kerdels and Gabriele Peters

University of Hagen - Faculty of Mathematics and Computer Science
Human-Computer Interaction
Universitätsstrasse 1, D-58084 Hagen, Germany

November 14, 2014

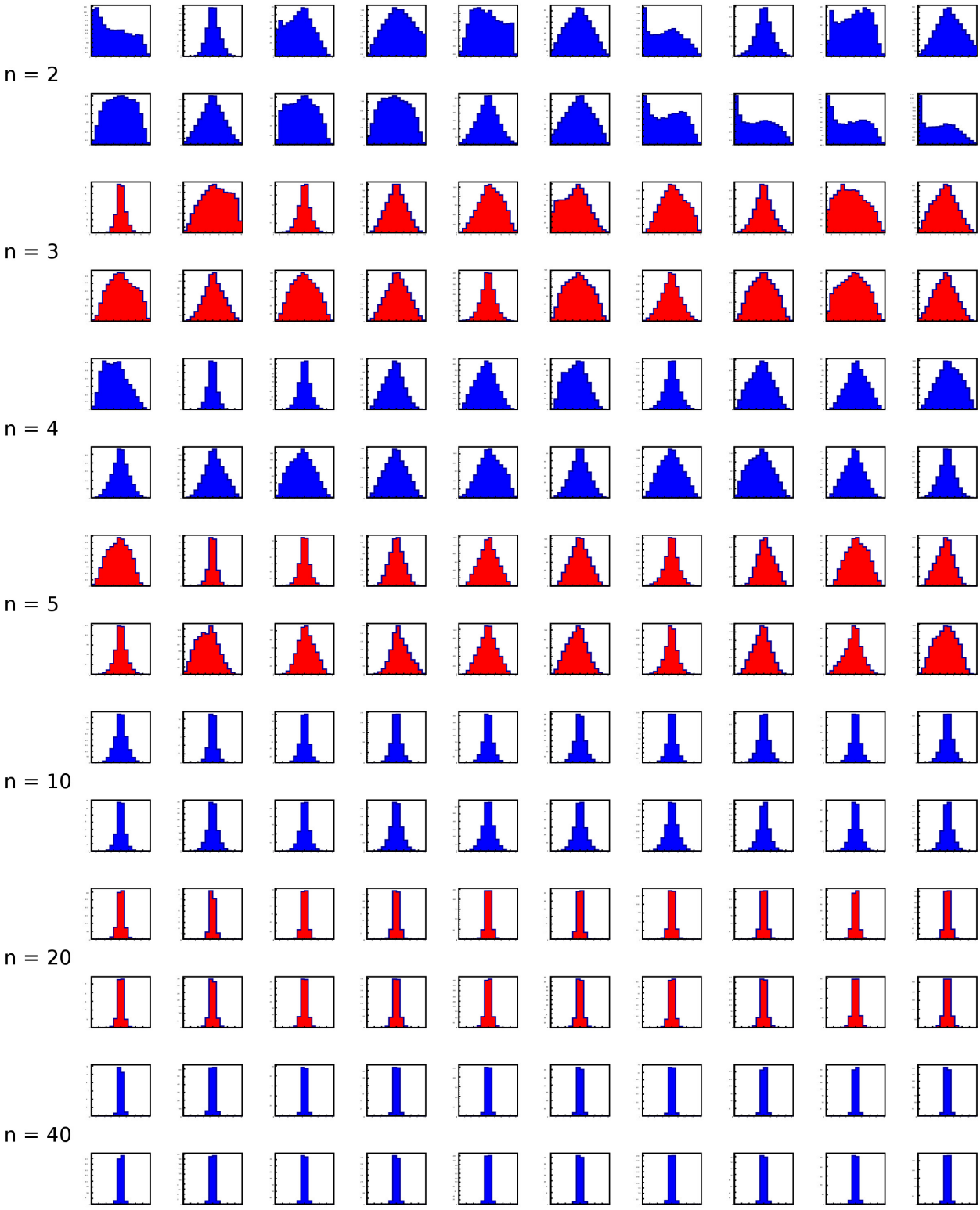


Figure S1: Comparison of local input space histograms of GNGs with 50 units, fixed Minkowski parameter $p = 2$, and increasing dimensions n of input space.

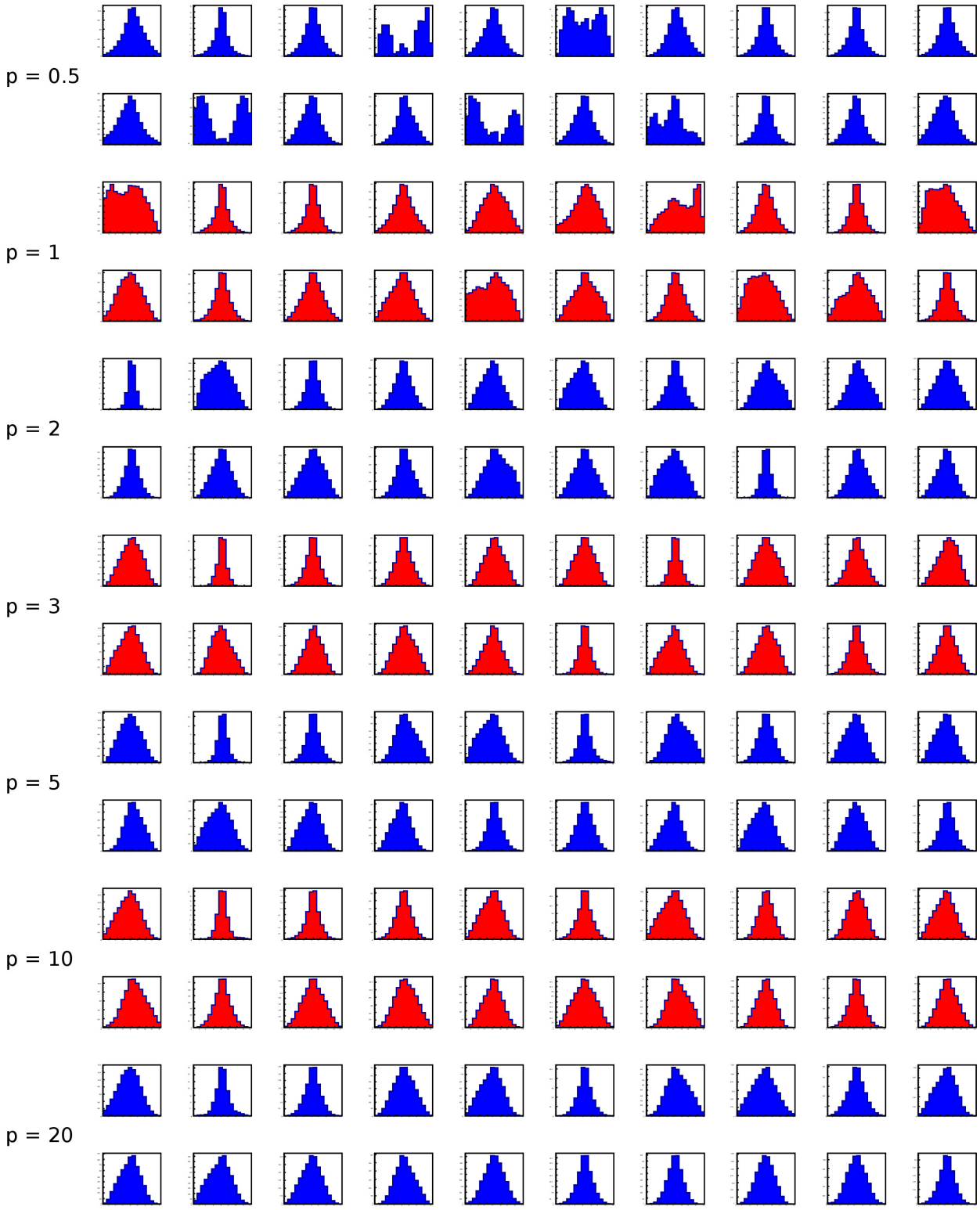


Figure S2: Comparison of local input space histograms of GNGs with 50 units, random 4-dimensional input, and increasing values for the Minkowski parameter p.

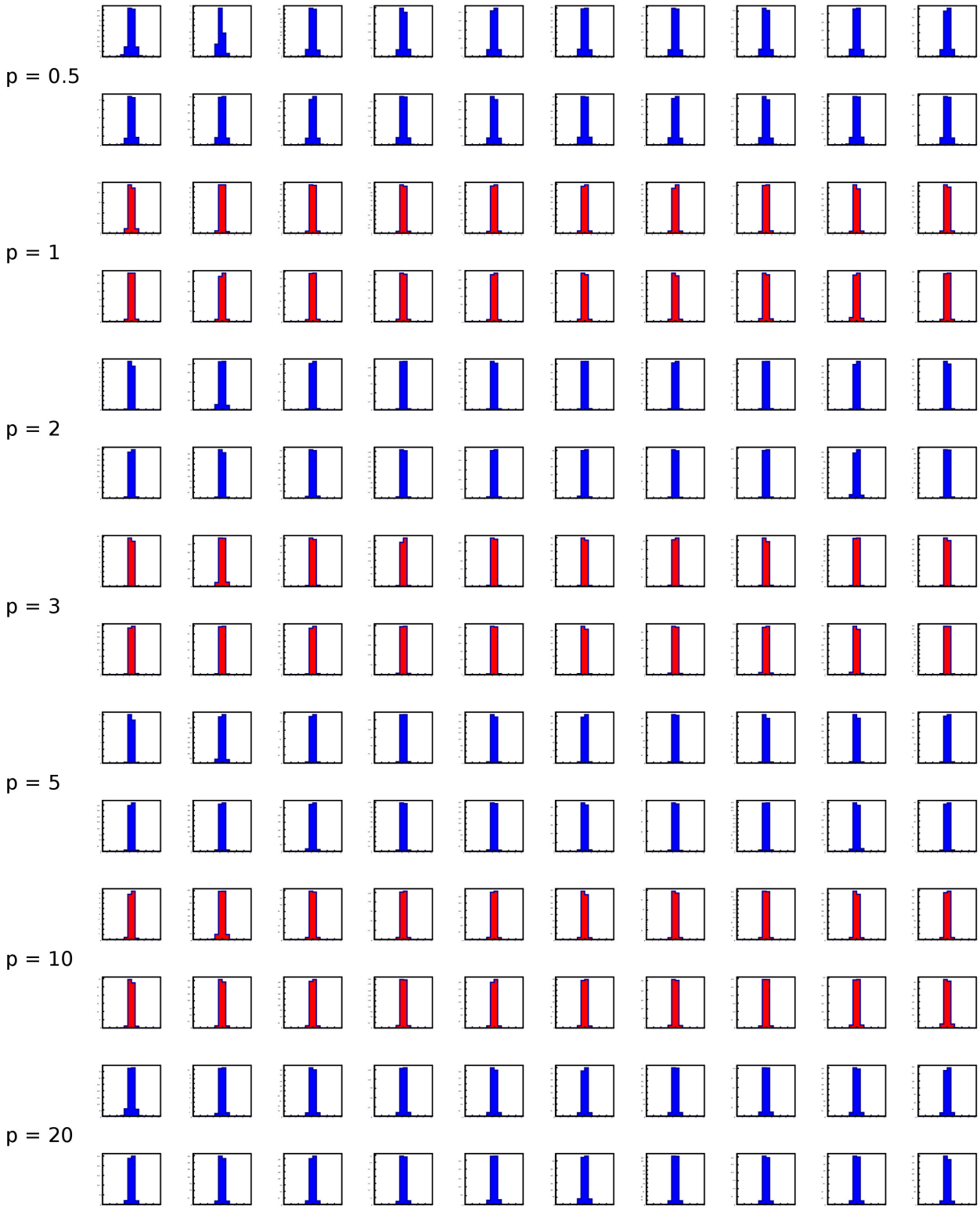


Figure S3: Comparison of local input space histograms of GNGs with 50 units, random 64-dimensional input, and increasing values for the Minkowski parameter p.

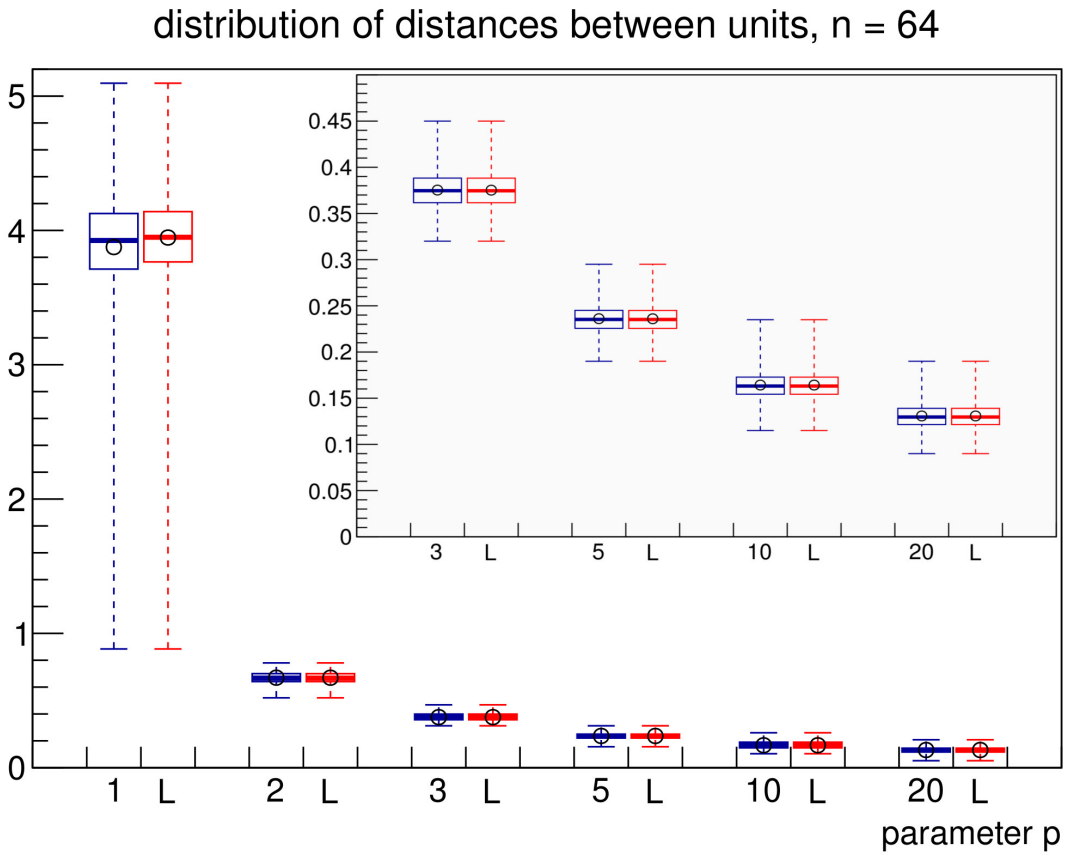


Figure S4: Box plot of the distributions of pairwise distances between the units of GNGs with fixed input space dimension $n = 64$ and varying Minkowski parameters p. Blue boxes describe the pairwise distances between all units, red boxes (L-columns) describe the pairwise distances between all units connected by edges. Circles represent the mean values of the distributions. Inset: Magnification of entries for $p = \{3, 5, 10, 20\}$.

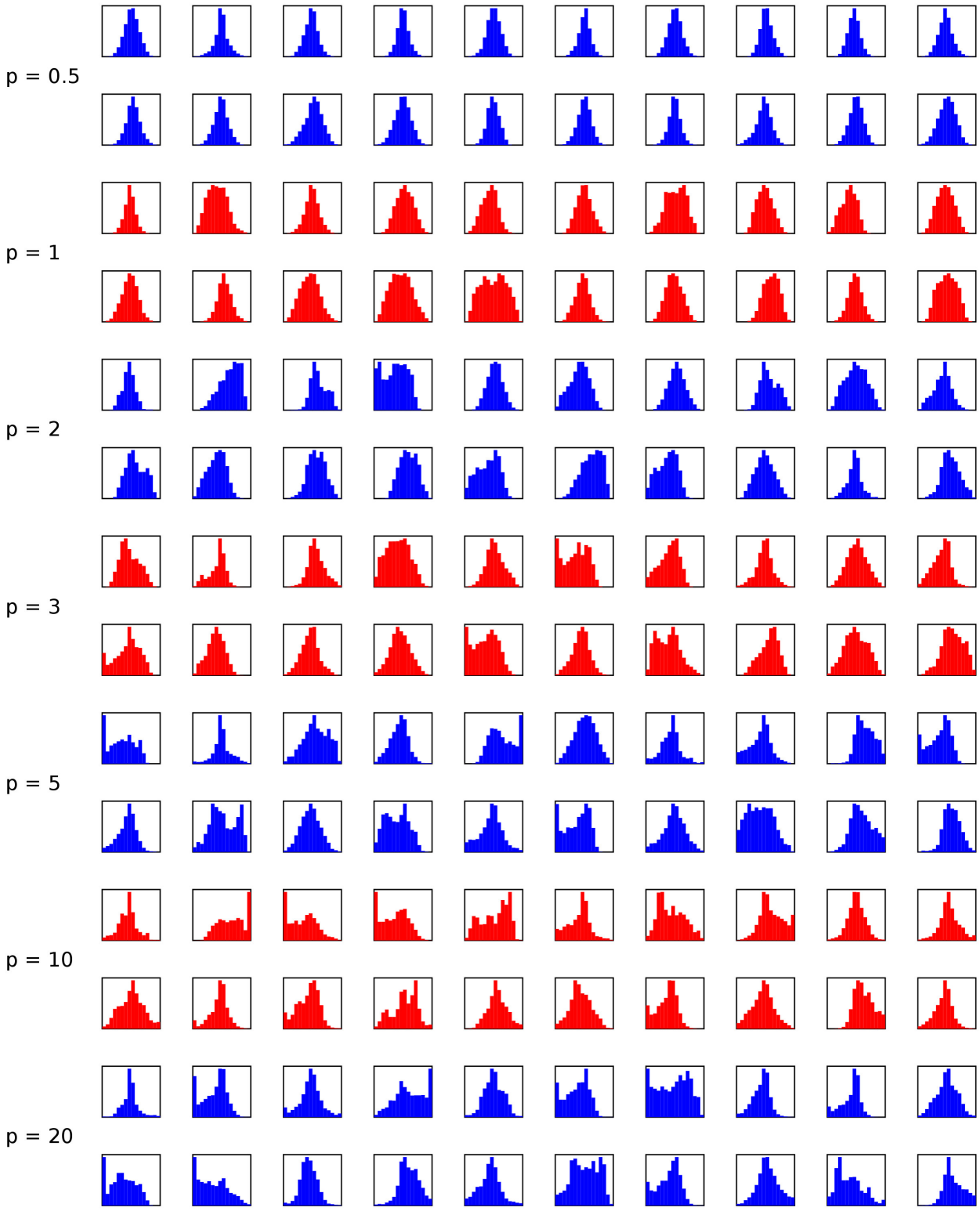


Figure S5: Comparison of local input space histograms of GNGs with 50 units, 64-dimensional color histogram input, and increasing values for the Minkowski parameter p.

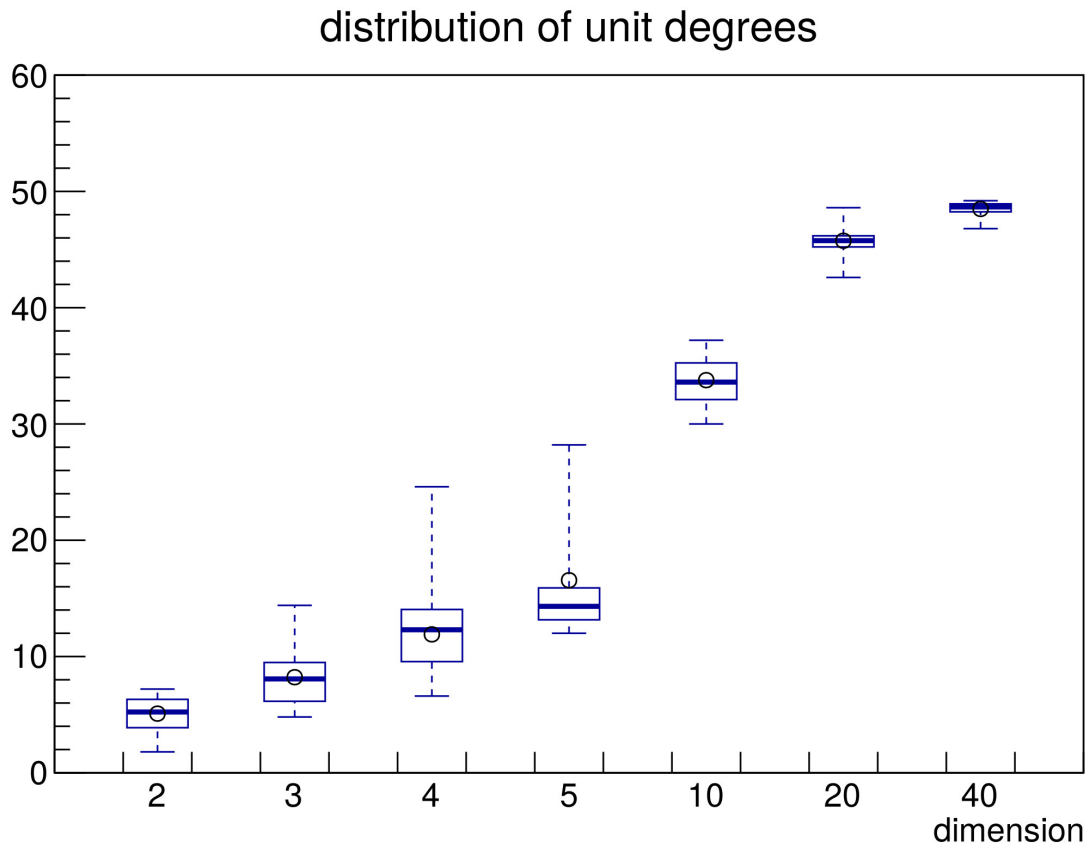


Figure S6: Box plot of the distributions of unit degrees of GNGs with 50 units, random input with increasing dimension, and fixed Minkowski parameter $p = 2$.

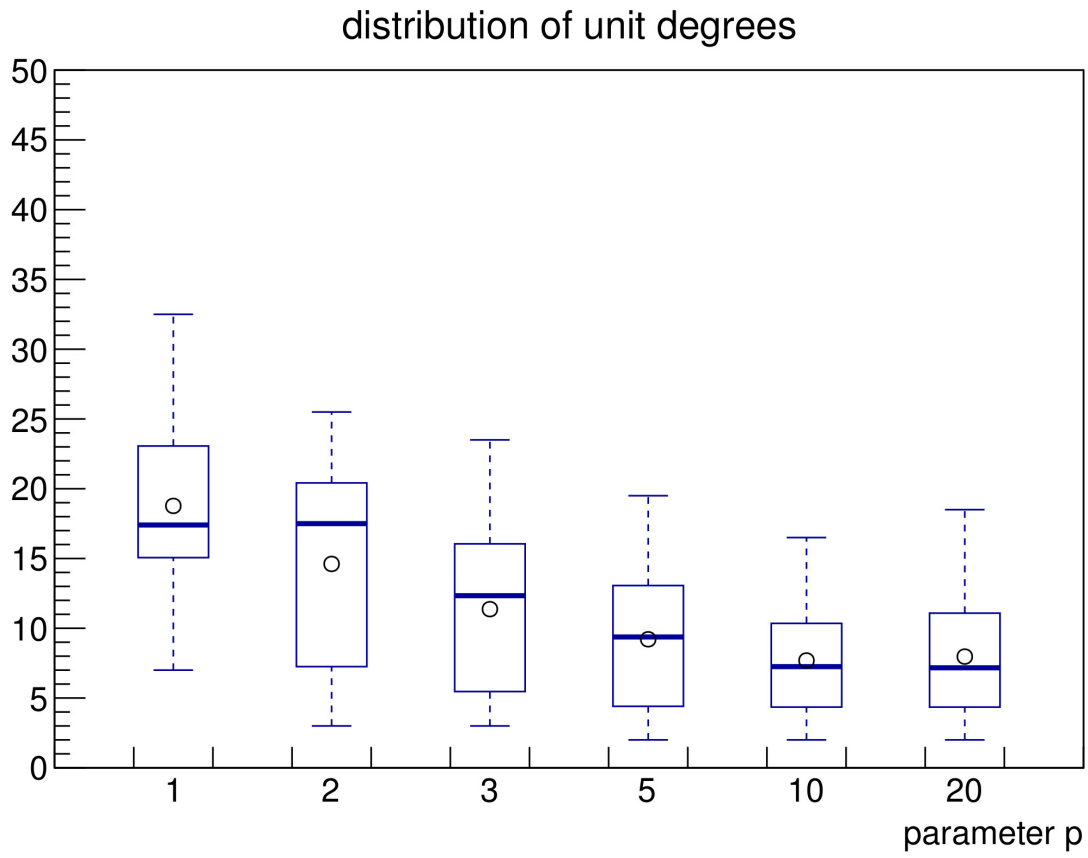


Figure S7: Box plot of the distributions of unit degrees of GNGs with 50 units, 64-dimensional color histogram input, and increasing values for the Minkowski parameter p.

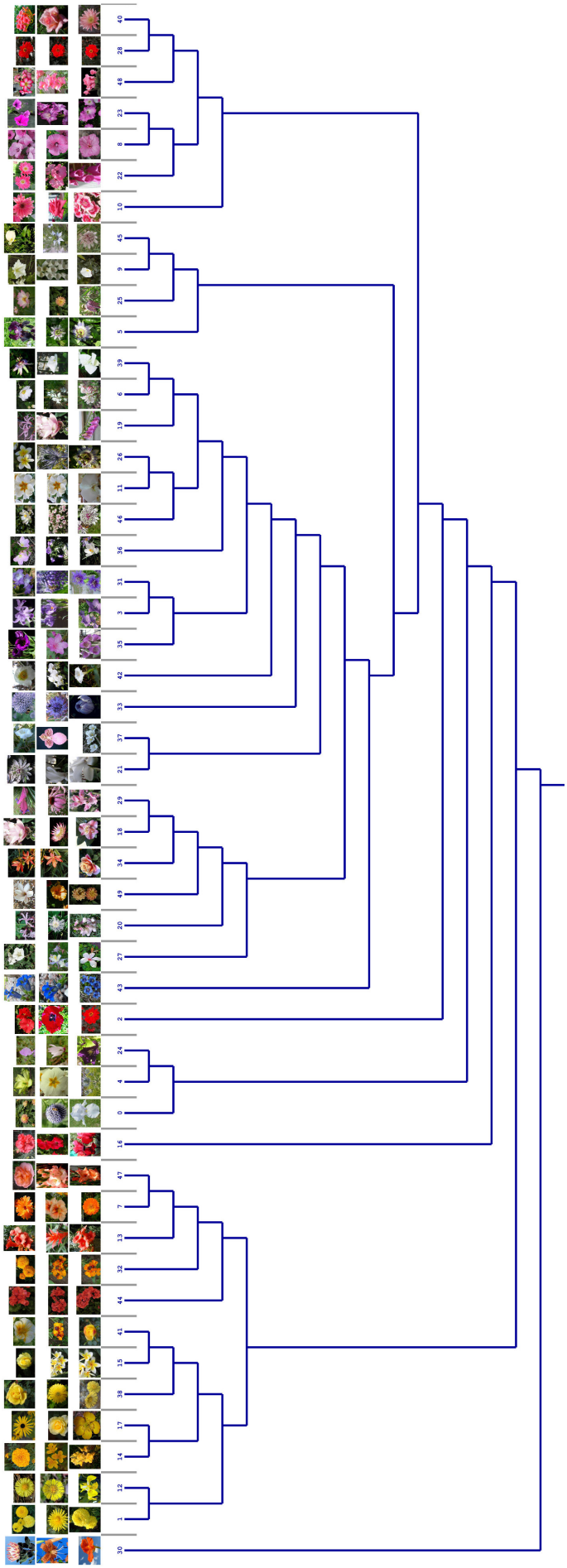


Figure S8: Mapping of closest input images to corresponding GNG units.