























with increasing dimensionality. In this case, the data set series was denoted by  $D(100).d(x).\theta(2).U(1.6)$ , where  $x$  denotes the changing variable. The dimensionality is illustrated on the X-axis, whereas the error is illustrated on the Y-axis. The errors (per dimension) on the unsharpened data do not change much with increasing data dimensionality. However, the errors reduce significantly with increasing data dimensionality, because a greater number of dimensions are now available in order to reduce the errors and uncertainty in the correlation-based sharpening process. Since high dimensional data sets are quite common in real applications, this is quite promising for the utility of the sharpening method.

In Figure 18, we have illustrated the error level of the data set with increasing number of data points. In this case, the data set series was denoted by  $D(x).d100.\theta(2).U(1.6)$ , where  $x$  denotes the changing variable. The number of data points are illustrated on the X-axis, whereas the error is illustrated on the Y-axis. As in the previous case, the errors on the unsharpened data do not increase with data set size. However, the errors on the unsharpened data reduce considerably with data size. This is because of the intermediate steps is approximate covariance estimation which is most accurate with increasing data size. While the accuracy increased considerably with increasing data size, it is interesting to see that the sharpening process continues to be effective even for very small data sets containing only 500 points. This illustrates the robustness of the sharpening method.

#### 4 Conclusions and Summary

In this paper, we presented a method for multidimensional sharpening of uncertain data sets. The process of sharpening improves the quality of the underlying representation by using the correlation information which is usually available in the underlying information. Since most real data sets have underlying inter-attribute correlations, this means that the technique is usually quite helpful in improving the accuracy of representation. Our results show that the technique not only improves the mean of the representation, but also reduces the variance of the underlying uncertainty. Furthermore, the technique continues to be effective on relatively small data sets, whose statistical parameters are often difficult to estimate accurately because of the underlying uncertainty. Our results show that the technique is extremely effective on a wide variety of data sets, and improves with increasing data set size and dimensionality.

#### Acknowledgement

This research is continuing through participation in the Network Science Collaborative Technology Alliance sponsored by the U.S. Army Research Laboratory under contract number W911NF-09-2-0053.

#### References

- [1] C. C. Aggarwal, *On Density Based Transforms for Uncertain Data Mining*, ICDE Conference, (2007), pp. 866–875.
- [2] C. C. Aggarwal, *On Unifying Privacy and Uncertain Data Models*, ICDE Conference, (2008), pp. 386–395.
- [3] C. C. Aggarwal, and P. S. Yu, *A Framework for Clustering Uncertain Data Streams*, ICDE Conference, (2008), pp. 150–159.
- [4] C. C. Aggarwal, and P. S. Yu, *A Survey of Uncertain Data Algorithms and Applications*, IEEE Transactions on Knowledge and Data Engineering, 21(5), May 2009, pp. 609–623.
- [5] C. C. Aggarwal, *Managing and Mining Uncertain Data*, (2009), Springer.
- [6] D. Burdick, P. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan, *OLAP Over Uncertain and Imprecise Data*, VLDB Conference, (2005), pp. 970–981.
- [7] R. Cheng, Y. Xia, S. Prabhakar, R. Shah, and J. Vitter, *Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data*, VLDB Conference, (2004), pp. 876–887.
- [8] R. Cheng, D. Kalashnikov, and S. Prabhakar, *Evaluating Probabilistic Queries over Imprecise Data*, SIGMOD Conference, (2003), pp. 551–562.
- [9] N. Dalvi, and D. Suciu, *Efficient Query Evaluation on Probabilistic Databases*, VLDB Conference, (2004), pp. 864–875.
- [10] A. Das Sarma, O. Benjelloun, A. Halevy, and J. Widom, *Working Models for Uncertain Data*, ICDE Conference, (2006), pp. 7.
- [11] Z. Huang, W. Du, and B. Chen, *Deriving Private Information from Randomized Data*, ACM SIGMOD Conference, (2005), pp. 37–48.
- [12] H.-P. Kriegel, and M. Pfeifle, *Density-based clustering of uncertain data*, KDD Conference, (2005) pp. 672–677.
- [13] L. V. S. Lakshmanan, N. Leone, R. Ross, and V. S. Subrahmanian, *ProbView: A Flexible Database System*, ACM Transactions on Database Systems, 22(3), 1997, pp. 419–469.
- [14] S. Singh, C. Mayfield, S. Prabhakar, R. Shah, and S. Hambrusch, *Indexing Uncertain Categorical Data*, ICDE Conference, (2007), pp. 616–625.
- [15] Y. Tao, R. Cheng, X. Xiao, W. Ngai, B. Kao, and S. Prabhakar, *Indexing Multi-dimensional Uncertain Data with Arbitrary Probability Density Functions*, VLDB Conference, (2005), pp. 922–933.