Efficient multivariate density estimation: `mvdensity`

Eike Petersen, eike.petersen@uni-luebeck.de

August 24, 2021

If a problem is multivariate, the number of samples is large, and the resulting density estimate must be evaluated efficiently at many points, density estimation is a nontrivial endeavor. To the author's knowledge, no efficient method is readily available in standard software packages. Closest to the fulfillment of these requirements may be the fastKDE method [5], which performs highly efficient kernel density estimation (KDE) in the multivariate setting, but which, however, is inefficient in the evaluation on many query points. (The evaluation complexity is $\mathcal{O}(N^Q N^S)$, where N^S is the number of datapoints and N^Q the number of query points.) For these reasons, a simple custom method inspired by Allison [1] is implemented here, which is described in the following.

The pursued general approach, which is certainly far from new, is to simply smooth a multivariate histogram. The histogram can be computed very efficiently, and the complexity of the smoothing operation then only depends on the number N^B of histogram bins, *not* the number of measured samples. The resulting smoothed density surface can then be evaluated with complexity $\mathcal{O}(N^B N^Q)$, which is sufficiently cheap. There are two methods implemented to obtain a PDF estimation from the histogram: a) RBF-based smoothing, and b) simple interpolation. In both cases, artificial boundary histogram bins with zero counts are first added to force the density surface estimate to decline towards zero outside of the histogram.

RBF smoothing. Radial basis function (RBF)-based smoothing [3] is easily and efficiently extensible to the multivariate setting. First, the center points of the data within each bin are calculated. Next, the *significance* of each bin is calculated using the method proposed by Allison [1], and the most significant bins are selected based on a threshold. The center points of all selected significant bins, all boundary bins, and the bin with the highest histogram count, are then used as the center points of the radial basis functions.[1] Multiquadric RBFs are used (as done in Allison [1]), and the widths of the RBFs are chosen inversely proportional to the significance of the corresponding bin. The weights of the RBFs are determined by solving a ridge-regularized linear least squares problem [3]

[1] It is crucial that not *all* bins are used as center points for RBFs: otherwise, we would perform interpolation, not smoothing.

Interpolation. Interpolation is simply performed between the points specified by the geometrical center of each histogram bin and the corresponding count value. Different methods can be used, such as makima [4] or linear interpolation.

Whereas the RBF method has various appealing properties and seems preferable in general, it has proven stubborn to tune such that it performs well across any dimension and number of datapoints. (Suggestions for improvements are very welcome!) The most stable solution right now is thus to use the (makima) interpolation option.

In both methods, to prevent potentially negative undershoots during the smoothing step and guarantee the (semi-)positivity of the resulting PDF estimate, the histogram counts are transformed using the inverse softplus function [2]

$$y = f^{-1}(x) = \log(\mathrm{e}^{x} - 1).$$

Smoothing or interpolation is performed using these transformed data, and the resulting smoothed PDF estimate is transformed back using the softplus function [2]

$$f(y) = \log(1 + \mathrm{e}^{y}).$$

This guarantees the positivity of the resulting estimate, as was mentioned before. Finally, to obtain a correctly normalized density estimate, the integral of the smoothed surface is estimated using a numerical integration scheme, and the smoothed surface is divided by this constant.

References

[1] J. Allison, "Multiquadric radial basis functions for representing multidimensional high energy physics data," *Computer Physics Communications*, vol. 77, no. 3, pp. 377–395, Nov. 1993. DOI: 10.1016/0010-4655(93)90184-e.

[2] C. Dugas *et al.*, "Incorporating second-order functional knowledge for better option pricing," in *Advances in Neural Information Processing Systems*, T. Leen, T. Dietterich, and V. Tresp, Eds., vol. 13, MIT Press, 2001. [Online]. Available: https://proceedings.neurips.cc/paper/2000/file/44968aece94f667e4095002d140b5896-Paper.pdf.

[3] G. E. Fasshauer, *Meshfree Approximation Methods with MATLAB*. Singapore: World Scientific Publishing Co. Pte. Ltd., Jun. 2007, 520 pp., ISBN: 9812706348.

[4] C. Ionita, *Makima piecewise cubic interpolation*, C. Moler, Ed., Apr. 2019. [Online]. Available: https://blogs.mathworks.com/cleve/2019/04/29/makima-piecewise-cubic-interpolation/.

[5] T. A. O'Brien *et al.*, "A fast and objective multidimensional kernel density estimation method: fastKDE," *Computational Statistics & Data Analysis*, vol. 101, pp. 148–160, Sep. 2016. DOI: 10.1016/j.csda.2016.02.014.