Functional boxplots for epidemic and other data: extensions of Juul *et al.* (2021)

Ali Gharouni[1] **and Benjamin M. Bolker**[1,2,3]

[1]**Department of Mathematics & Statistics, McMaster University, Hamilton, Canada**
[2]**Department of Biology, McMaster University, Hamilton, Canada**
[3]**Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, Canada**

Corresponding author:
Ali Gharouni[1]

Email address: agharoun@uottawa.ca *

ABSTRACT

The *central set* of an ensemble of epidemic curves (or more generally of any functional data set) helps characterize the variation of the ensemble, which in turn gives insights into disease dynamics and management. Classical fixed-time descriptions of uncertainty such as pointwise confidence envelopes may be unreliable. Juul *et al* 2021 described robust curve-based descriptive methods based on *functional band depth* (FBD) to visualize the uncertainty associated with ensembles of epidemic curves from stochastic epidemic models. Extending and describing this approach further, we studied several different approaches to construct and visualize uncertainty regions. First, we compared curves using Juul *et al*'s FBD approach with a more widely used, computationally efficient implementation of FBD. Second, we developed a new centrality measure that uses a set of features of interest (such as the initial growth rate, peak time and peak height of the epidemic), combined into a Mahalanobis distance from the ensemble's centroid. Both methods provide useful results.

INTRODUCTION

Summarizing the uncertainty for ensembles of curves generated by stochastic epidemic models can provide useful insight into the dynamics and management of epidemics. For these and other sets of *functional data* — data where each observation represents a continuous curve, computing the typical set of observations forms a basis for graphical and quantitative summaries. Computations of central sets rely in turn on robust and computationally efficient estimation methods.

Juul et al. (2021) pointed out shortcomings of the standard methods that researchers use to draw confidence intervals for ensembles of curves, with specific examples drawn from the output of stochastic epidemic models (also, see Appendix 2 in Kiss et al. (2017)'s book). In particular, they showed that fixed-time approaches (e.g., computing pointwise quantiles) can fail to capture the uncertainty in key features of an epidemic such as the timing and magnitude of epidemic peaks. As an alternative to fixed-time approaches, the authors illustrated methods to compute the *central set* of an ensemble of curves, a high-dimensional analogue of interquartile range or confidence interval. A large body of literature addresses this topic under the rubrics of *functional depth* and *functional boxplots* for high dimensional data (Fraiman and Muniz, 2001; López-Pintado and Romo, 2007, 2009; Sun and Genton, 2011; Sun et al., 2012). While Juul *et al.* do cite this literature (Sun and Genton, 2011), exploring it in more depth led us to several useful practical and theoretical points that could be useful for researchers interested in visualizing confidence regions for ensembles of curves.

In univariate data, the central set (region) can be easily represented by summaries based on univariate quantiles. Classical boxplots provide a visual summary of the data that represents the uncertainty (the range of the central set) by the interquartile range. In multivariate data, computing the central set relies on the concept of *statistical depth* (Mahalanobis, 1936; Tukey, 1975; Oja, 1983; Liu, 1990; Singh, 1991;

*Current address: Department of Mathematics and Statistics, University of Ottawa

Vardi and Zhang, 2000; Zuo, 2003). When generalized to functional data, the statistical depth is usually referred to as *functional depth* (Fraiman and Muniz, 2001). Roughly speaking, a functional depth is a bounded non-negative function which measures the average closeness of a function (in practice, one observation in a data set or curve in an ensemble) to all other functions, over a function-valued distribution (Zuo and Serfling (2000) gives formal definitions). One can then order the elements of an ensemble according to decreasing depth values, ranking the observations from the center (the deepest or most central point) outward, and define a central set that includes all the points up to a given depth or rank. The functional boxplot displays an ensemble in a way that highlights this central set (Sun and Genton, 2011; Sun et al., 2012).

López-Pintado and Romo (2007) developed functional band depth (FBD), a sample-based method for determining a curve's centrality. FBD measures the fraction of times that a given curve is completely included within the envelope of a set of other curves randomly sampled from the ensemble. The sample size is determined by a *tuning parameter J*, and one can take any number of samples up to the set of all possible combinations of size J. Juul *et al.* used a version of FBD to create functional boxplots for an ensemble of epidemic curves simulated from a stochastic epidemic model. They chose $J = 50$ (they use the notation N_{curves}), and used $N_{\text{samples}} = 100$ such samples to compute the FBD for each curve. They provided open-source Python code that implements this method, as well as some weighted variants of FBD. For the simple (unweighted) case, however, there are already mature open source implementations available in R (Ramsay et al., 2020; Ieva et al., 2019), Matlab (`https://www.psych.mcgill.ca/misc/fda/downloads/FDAfuns/`), and Python (Seabold and Perktold, 2010). In general these packages use the same functional band depth measure as Juul *et al.*, but substituting $J = 2$, which is robust (López-Pintado and Romo, 2009) and allows the use of a computationally efficient algorithm for large data sets (Sun et al., 2012). It is unclear why Juul *et al.* chose larger values of J (10 and 50), although the dimensions of their examples are small enough that the computational burden is not important.

Juul *et al.* also suggest ranking according to a single, one-dimensional feature of interest such as the maximum values of newly hospitalized cases in a single day (their Fig. 2e). This approach can be extended to incorporate multiple features of interest. FBD can be based on this reduced set of features; here we use the *Mahalanobis distance* (Mahalanobis, 1936), which measures distance from a centroid accounting both for variation in the scales or typical magnitudes of different features and for correlation among features. Our example uses a feature set including the peak value of incidence (new infections), the time at which the peak occurs, and the initial growth rate, duration, and final size of the epidemic. While these are typical epidemiological features of interest, researchers can and should choose the features that are most closely connected to their particular research questions (Probert et al., 2016).

We compare the central set of the epidemic ensemble (provided in Juul *et al.*'s work) by defining various statistical depths and ranking the curves using (i) FBD with different choices of the tuning parameter J to compare classical ($J = 2$) with Juul *et al.*'s 2021 approach ($J = 50$), and (ii) a functional depth measure based on Mahalanobis distances among features of interest. Although we used an epidemic ensemble, our methods are broadly applicable to any functional data set.

METHODS

We present the comparison of 90% central regions computed with different functional band depth methods. We apply our methods on Juul *et al.* dataset (`https://github.com/jonassjuul/curvestat/tree/master/curvestat/tests/test_data`) and compare our results with theirs.

First, we implemented Juul *et al.*'s FBD method in R (R Core Team, 2021). In particular, the algorithm: (i) randomly samples a subset of curves from the ensemble ($J = 50$), (ii) computes the envelopes (pointwise minima and maxima of the sample), (iii) scores all curves in the ensemble based on whether they lie entirely within the envelope (score=1) or not (score=0), and (vi) repeats (i)-(iii) to derive a rank, or depth, for each curve based on the average score. The central set consists of the curves above a specified depth.

Second, we used the function `fda()` in the R package roahd (Ieva et al., 2019) to compute central sets, with the choice of modified band depth (MBD) to break ties which is based on the fast algorithm proposed by Sun et al. (2012).

Third, we computed central sets by defining a feature vector for each curve in the ensemble and computing the average pairwise Mahalanobis distances to every other point in the set, using the `mahalanobis()` function in R and using the covariance matrix from the entire set of features as the scaling factor. The rank of each curve is the rank of its average distance to the rest of the set.

The ensemble of curves, which we took from Juul et al. (2021)'s supplementary data, represents the number of newly hospitalized people over time, which is a version of the epidemic incidence (typically measured by the number of new infections or number of newly detected cases). In keeping with the epidemic-modeling focus, we chose the curve features as (i) the peak hospitalization rate; (ii) the time at which the peak occurs; (iii) the initial growth rate; (iv) the epidemic duration; and (v) the total size of the epidemic (as measured by the total number of people hospitalized). The initial growth rate was estimated by fitting an exponential curve to the section of the curve from the first day of nonzero incidence to the day when the incidence is 10%. The epidemic duration was determined by the time between 10% and 90% of cumulative hospitalizations. The epidemic size was defined as the total number hospitalized (the sum of the number of new daily hospitalizations).

RESULTS

The 90% central regions computed with both pairwise Mahalanobis distance approach and FBD with $J = 2$ are comparable with Juul *et al.*'s result, i.e., FBD with $J = 50$. Both pairwise Mahalanobis distance approach and FBD with $J = 2$ identify an earlier peak as being part of the central set, while Juul *et al.*'s result may be shifted by 1 index point relative to other two.

DISCUSSION

We compared definitions of central sets (and the resulting functional boxplots) based on band depths (based on pointwise 'betweenness' of curves) and on distances among sets of features derived from the curves, or between features of a curve and the centroid of sets of features for a given sample. For band-depth approaches, the only choices in defining a central set are (1) the size of the ensemble to use for "betweenness" — FBDs using pairs, i.e. $J = 2$, can be computed efficiently and seemed to give similar results to other choices for Juul et al's example — and (2) questions of tie-breaking. For distance-based approaches, the primary choices are (1) what distance to use, (2) whether to measure distances among all sample points in a curve or between feature sets and (3) whether to base depth on *average* pairwise distance among curves or distance to a centroid.

For some distance metrics (such as Euclidean distances), it is theoretically easy to define and computationally easy to compute a centroid; in contrast, measuring all pairwise distances among curves (so that depth can be defined as the average pairwise distance to other curves) is computationally expensive unless some efficient algorithm can be implemented. For our Mahalanobis calculations, we used the standard Euclidean centroid.

Mahalanobis distance computations also require a computation of the covariance matrix in order to scale elements of the distance appropriately; as described above, we used a covariance matrix based on the feature vectors for all observations in the data set. Such an approach could be misleading if the feature distribution is strongly bimodal or multimodal, in which case scaling factors derived from the overall data set may not be appropriate for scaling the components of distance between two trajectories whose features put them in the same mode of the distribution.

We briefly explored distance-based methods using all samples from the curve rather than feature vectors; we found that using a standard Euclidean distance (ℓ_2 norm) tended to underestimate key features of epidemic ensemble such as the magnitude of epidemic peaks. Other well-known distances on spaces of functions, such as Fréchet distances and dynamic time warping, intentionally exclude differences based on phase (i.e., two curves that are identical up to a phase shift are considered to be coincident) — this property seems inappropriate for evaluating ensembles of epidemic curves, but could be useful in other settings.

REFERENCES

Fraiman, R. and Muniz, G. (2001). Trimmed means for functional data. *Test*, 10(2):419–440.

Ieva, F., Paganoni, A. M., Romo, J., and Tarabelloni, N. (2019). roahd Package: Robust Analysis of High Dimensional Data. *The R Journal*, 11(2):291–307.

Juul, J. L., Græsbøll, K., Christiansen, L. E., and Lehmann, S. (2021). Fixed-time descriptive statistics underestimate extremes of epidemic curve ensembles. *Nature Physics*, 17(1):5–8.

Kiss, I. Z., Miller, J. C., Simon, P. L., et al. (2017). Mathematics of epidemics on networks. *Cham: Springer*, 598.

Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics*, pages 405–414.

López-Pintado, S. and Romo, J. (2007). Depth-based inference for functional data. *Computational Statistics & Data Analysis*, 51(10):4957–4968.

López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104(486):718–734.

Mahalanobis, P. C. (1936). On the generalized distance in statistics. National Institute of Science of India.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics & Probability Letters*, 1(6):327–332.

Probert, W. J., Shea, K., Fonnesbeck, C. J., Runge, M. C., Carpenter, T. E., Dürr, S., Garner, M. G., Harvey, N., Stevenson, M. A., Webb, C. T., et al. (2016). Decision-making for foot-and-mouth disease control: objectives matter. *Epidemics*, 15:10–19.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsay, J. O., Graves, S., and Hooker, G. (2020). *fda: Functional Data Analysis*. R package version 5.1.9.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with Python. In *9th Python in Science Conference*.

Singh, K. (1991). A notion of majority depth. *Unpublished document*.

Sun, Y. and Genton, M. G. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20(2):316–334.

Sun, Y., Genton, M. G., and Nychka, D. W. (2012). Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked? *Stat*, 1(1):68–74.

Tukey, J. W. (1975). Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians, Vancouver, 1975*, volume 2, pages 523–531.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proceedings of the National Academy of Sciences*, 97(4):1423–1426.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *The Annals of Statistics*, 31(5):1460–1490.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of statistics*, pages 461–482.

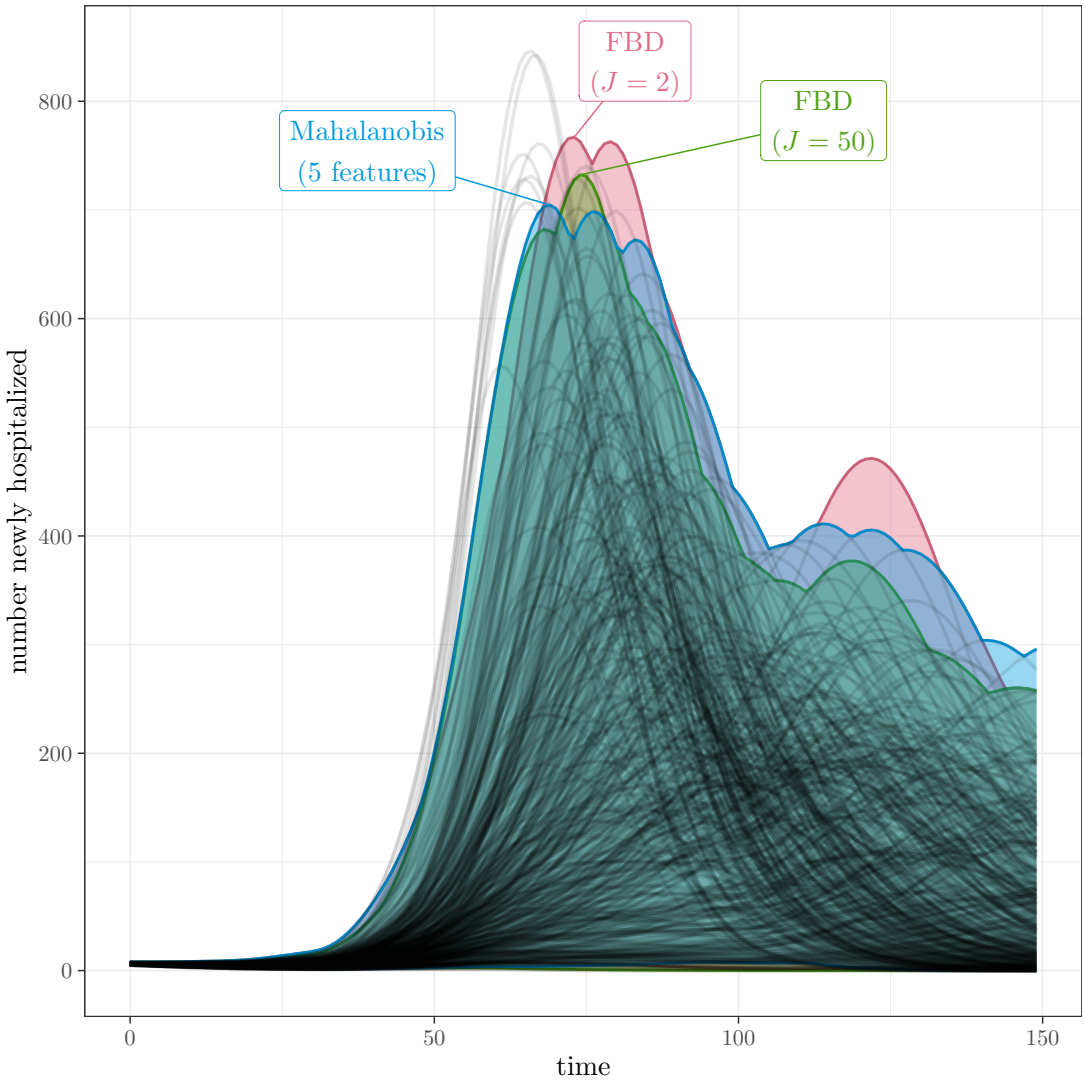

Figure 1. Comparison of 90% central regions computed with different functional boxplot methods, using epidemic curve ensembles from Juul *et al.*. FBD = functional band distance; J = number of curves used for centrality calculation. Curve with $J = 2$ computed via the `roahd` package (Ieva et al., 2019): curve with $J = 50$ used our own implementation of the functional band distance algorithm described by Juul *et al.*, curve with Mahalanobis (5 features) used Mahalanobis distance on features of interest including the peak value of incidence, the time at which the peak occurs, the initial growth rate, epidemic duration, and final size of the epidemic.