Overlapping: a R package for Estimating Overlapping in Empirical Distributions

Massimiliano Pastore[1]

1 Department of Developmental and Social Psychology, University of Padova

Summary

Overlapping can be defined as the area intersected by two or more probability density functions. The idea of overlapping was introduced in a formal way by Gini & Livada (1943) and, more recently, it has been applied in several research problems involving, for instance, data fusion (Moravec, 1988), information processing (Viola & Wells III, 1997), applied statistics (Inman & Bradley Jr, 1989), economics (Milanovic & Yitzhaki, 2001) and psychology, as a basis for Cohen's U index (Cohen, 1988), McGraw and Wong's CL measure (McGraw & Wong, 1992), and Huberty's I degree of non-overlap index (Huberty & Lowman, 2000).

`overlapping` is an R package for estimating the overlapping area of two or more kernel density estimations from empirical data. The main idea of the package is to offer an easy way to quantify the similarity (or the difference) between two or more empirical distributions. In addition, the package allows to plot density distributions, highlighting the overlapped area by using the `ggplot2` R package (Wickham, 2009).

The package is available from GitHub (https://github.com/masspastore/overlapping) and CRAN (https://cran.r-project.org/package=overlapping). A full reference manual can be found at https://cran.r-project.org/web/packages/overlapping/overlapping.pdf.

A recent R package, `overlap` (Ridout & Linkie, 2009), offers an implementation of the overlapping index which can be used to analyse temporal activity patterns of animals and species in echology. Compared to this latter, `overlapping` package offers a more general approach where overlapping can be computed for any type of numerical variable, and it allows for computations with more than two variables.

Examples

Suppose we have collected data in two groups of 100 subjects each, with respect to a generic variable Y, expressed by scores ranging between 0 and 30, and to be interested in assessing whether the two groups can be considered samples from populations with the same average.

We can simulate the groups' scores as follows:

```
set.seed( 1 )
n <- 100
G1 <- sample( 0:30, size = n, replace = TRUE )
G2 <- sample( 0:30, size = n, replace = TRUE, prob = dbinom( 0:30, 31, .55 ) )
```
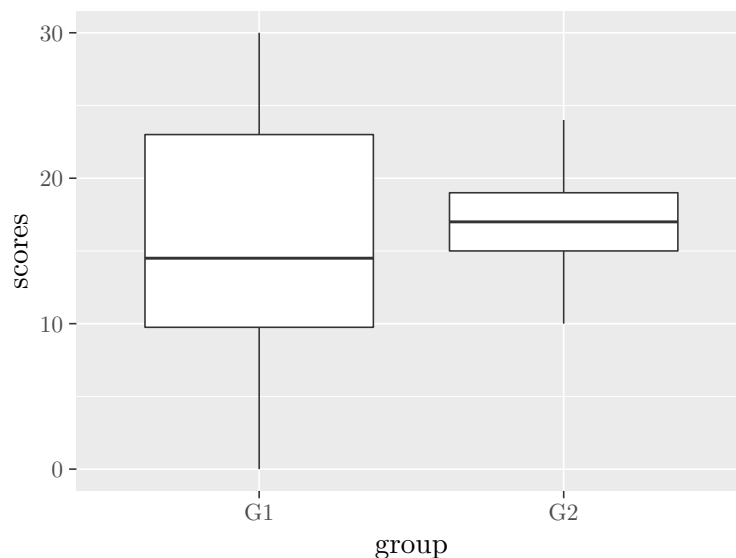
Figure 1: Scores distribution of simulated groups of 100 subjects each.

For Group 1 (`G1`) we randomly sampled `n` $= 100$ values from a uniform distribution; for Group 2 (`G2`) we randomly sampled 100 values from a binomial distribution. In the first group, scores range between 0 and 30 with mean 15.55 and standard deviation 8.32. In the second group, scores range between 10 and 24 with mean 16.72 and standard deviation 2.74.

We can display the scores distribution as follows:

```
library( ggplot2 )
Data <- data.frame( y = c(G1,G2), group = rep(c("G1","G2"),each=n) )
ggplot( Data, aes( x=group, y=y ) ) + geom_boxplot() + ylab("scores")
```
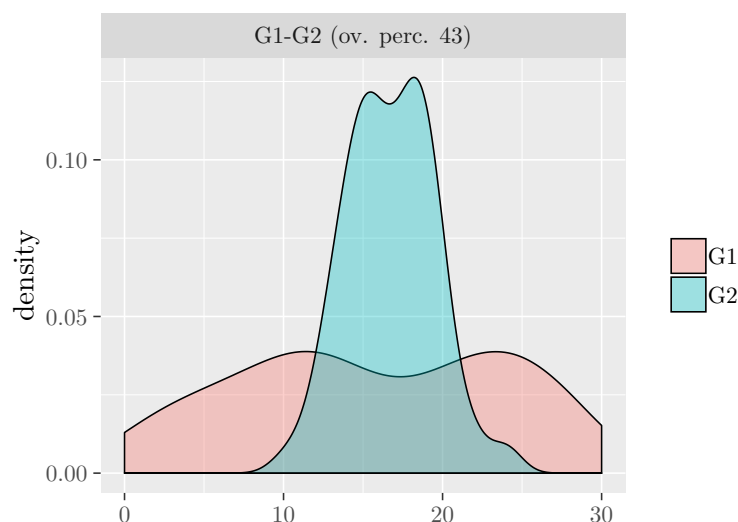
obtaining Figure 1. From this figure it is evident the heterogeneity of the variances in the two groups. In such a case, the statistical comparison between means can be biased and not very informative; for example, with a t-test, corrected for heterogeneity, we obtain the following result: $t(120.24) = -1.34$, $p = 0.18$, from which we cannot draw any conclusion (Wilkinson & Task Force on Statistical Inference, 1999).

So, let us assume a different perspective: Rather than assessing the similarity between the two groups on the basis of averages (and standard deviations) only, we use all the information available in the data. In practice, we estimate the degree of overlap between groups as the overlap between their kernel density estimates. We expect 0% to indicate the absence of overlapping (i.e., maximum distance between groups), and 100% to indicate the perfect overlap between the two distributions (i.e., groups are identically distributed). We can use the **overlapping** package in the following way:

```
library( overlapping )
dataList <- list( G1 = G1, G2 = G2)
overlap( dataList )$OV * 100

##    G1-G2
## 43.21998
```

With the command `library()` we load the **overlapping** package, next we create a `list` containing the two groups' scores, and finally, by using the `overlap()` function, we compute the overlap index. The index value (43.22) is an estimate of the percentage of

Pastore, (2018). Overlapping: a R package for Estimating Overlapping in Empirical Distributions. *Journal of Open Source Software*, 3(32), 1023. https://doi.org/10.21105/joss.01023

Figure 2: Comparison between densities of two groups. The overlap (43%) is represented by the shaded area.

overlapping between estimated densities. We can obtain a graphical representation by adding the option `plot = TRUE` as follows:

```
overlap( dataList, plot = TRUE )
```

obtaining Figure 2. In the figure are represented the estimated densities of the two groups' scores, with different colors. The shaded region is the overlapping area of densities.

Examples of real-world analysis

`overlapping` package has already used in different publications for many purposes, such as: 1) evaluating group invariance in questionnares, by using parameters bootstrap distributions (Lionetti, Mastrotheodoros, & Palladino, 2018, Marci et al. (2018)); 2) for computing a distance index in antropological measures (Altoè, D'Amore, & Scalfari, 2018); 3) for identifying cut-off scores in questionnaires, estimating the intersection points of density distributions (Pluess et al., 2018, Lionetti et al. (2018)).

References

Altoè, G., D'Amore, G., & Scalfari, F. (2018). Skulls and transvariation. In M. D. Bacco & F. Scalfari (Eds.), *Biostat at 25 invited essays in theoretical, biomedical and social statistics*. Edizioni ETS, PISA.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Psychology Press.

Gini, C., & Livada, G. (1943). *Nuovi contributi alla teoria della transvariazione*. Atti della VI Riunione della Società Italiana di Statistica, Roma, gennaio 1943.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement*, *60*(4), 543–563. doi:10.1177/0013164400604004

Inman, H. F., & Bradley Jr, E. L. (1989). The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two

normal densities. *Communications in Statistics-Theory and Methods*, *18*(10), 3851–3874. doi:10.1080/03610928908830127

Lionetti, F., Aron, A., Aron, E. N., Burns, G. L., Jagiellowicz, J., & Pluess, M. (2018). Dandelions, tulips and orchids: Evidence for the existence of low-sensitive, medium-sensitive and high-sensitive individuals. *Translational psychiatry*, *8*(1), 24. doi:10.1038/s41398-017-0090-6

Lionetti, F., Mastrotheodoros, S., & Palladino, B. E. (2018). Experiences in close relationships revised child version (ecr-rc): Psychometric evidence in support of a security factor. *European Journal of Developmental Psychology*, *15*(4), 452–463. doi:10.1080/17405629.2017.1297228

Marci, T., Lionetti, F., Moscardino, U., Pastore, M., Calvo, V., & Altoé, G. (2018). Measuring attachment security via the Security Scale: Latent structure, invariance across mothers and fathers and convergent validity. *European Journal of Developmental Psychology*, *15*(4), 481–492. doi:10.1080/17405629.2017.1317632

McGraw, K. O., & Wong, S. (1992). A common language effect size statistic. *Psychological bulletin*, *111*(2), 361.

Milanovic, B., & Yitzhaki, S. (2001). *Decomposing world income distribution: Does the world have a middle class?* The World Bank. doi:10.1596/1813-9450-2562

Moravec, H. P. (1988). Sensor fusion in certainty grids for mobile robots. *AI magazine*, *9*(2), 61. doi:10.1609/aimag.v9i2.676

Pluess, M., Assary, E., Lionetti, F., Lester, K. J., Krapohl, E., Aron, E. N., & Aron, A. (2018). Environmental sensitivity in children: Development of the highly sensitive child scale and identification of sensitivity groups. *Developmental psychology*, *54*(1), 51. doi:10.1037/dev0000406

Ridout, M., & Linkie, M. (2009). Estimating overlap of daily activity patterns from camera trap data. *Journal of Agricultural, Biological, and Environmental Statistics*, *14*(3), 322–337. doi:10.1198/jabes.2009.08038

Viola, P., & Wells III, W. M. (1997). Alignment by maximization of mutual information. *International journal of computer vision*, *24*(2), 137–154. doi:10.1023/A:1007958904918

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. doi:10.1007/978-0-387-98141-3_6

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals - guidelines and explanations. *American Psychologist*, *54*(8), 594–604.