Maximum Mean Discrepancy in High-Dimensional Biology

Frederick Boehm1,✉

11 May, 2021

Text of abstract

1 University of Massachusetts Medical School

✉ Correspondence: [Frederick Boehm <[frederick.boehm@gmail.com](mailto:frederick.boehm@gmail.com)>](mailto:frederick.boehm@gmail.com)

Keywords: keyword 1; keyword 2; keyword 3

Highlights: These are the highlights.

# 1 Specific Aims

The last twenty years have seen an explosion in the volume and dimensionality of molecular biology data. Following the sequencing of the human genome, technologies such as DNA sequencing, RNA sequencing, and single-cell RNA sequencing have proliferated. The new experimental designs and high dimension of modern biological data require parallel development of new statistical methods and computational tools for their analysis. Maximum mean discrepancy (MMD) is a new statistical tool with novel mathematical properties (Borgwardt et al., 2006; Gretton et al., 2012, 2009, 2006). We propose to explore its mathematical and statistical properties, to apply MMD methods in the analysis of high-dimensional biological data, and to build user-friendly R software tools that implement MMD methods.

**Aim 1: Apply MMD to high-dimensional biological data**

Two-sample problems arise in many areas of biomedical research. These include differential gene expression, comparability of data from different laboratories, and classification of cancers. Borgwardt et al. (2006) examined performance characteristics of four multivariate two-sample tests in a collection of four biological settings. They found that MMD outperformed the other three tests. settings of

**Aim 2: Create and maintain an open source R package that implements MMD methods**

Current R software for MMD methods doesn’t provide significance thresholds. Gretton’s website hosts Matlab software that implements the methods in Gretton et al. (2012), Gretton et al. (2009) and Gretton et al. (2006). We’ll implement these methods in R and C++ before assembling the code into a well documented, user-friendly R package (Eddelbuettel and François, 2011; R Core Team, 2021).

# 2 Background and Significance

MMD is a statistical method for determining whether two high-dimensional samples arise from distinct distributions. Throughout this document, we use the term “sample” to refer to measurements on a collection of subjects from a single group. In introductory statistics courses, one often studies Student’s t-test (Student, 1908) when discerning whether two samples come from the same distribution. Implicit to the t-test is the assumption that the two samples arise from normal distributions with a single, common variance parameter.

In high-dimensional settings, the analog of Student’s t-test is Hotelling’s test (Hotelling, 1931). Like its univariate counterpart, Hotelling’s test assumes a normal (multivariate joint normal) distribution of the random errors and a shared covariance matrix for the two groups. Departures from these assumptions may give rise to type I and type II errors in testing.

To combat limitations of existing high-dimensional, two-sample tests, Gretton et al. (2006) proposed a kernel-based test for the two-sample problem. Drawing on reproducing kernel Hilbert space theory, they formulate their test as a measure of distance between two embeddings of probability measures in a RKHS. Gretton et al. (2006) then write the test as a maximization over a subspace of their RKHS, ((2.1)).

# 3 Approach

# 4 Methods

# 5 Results

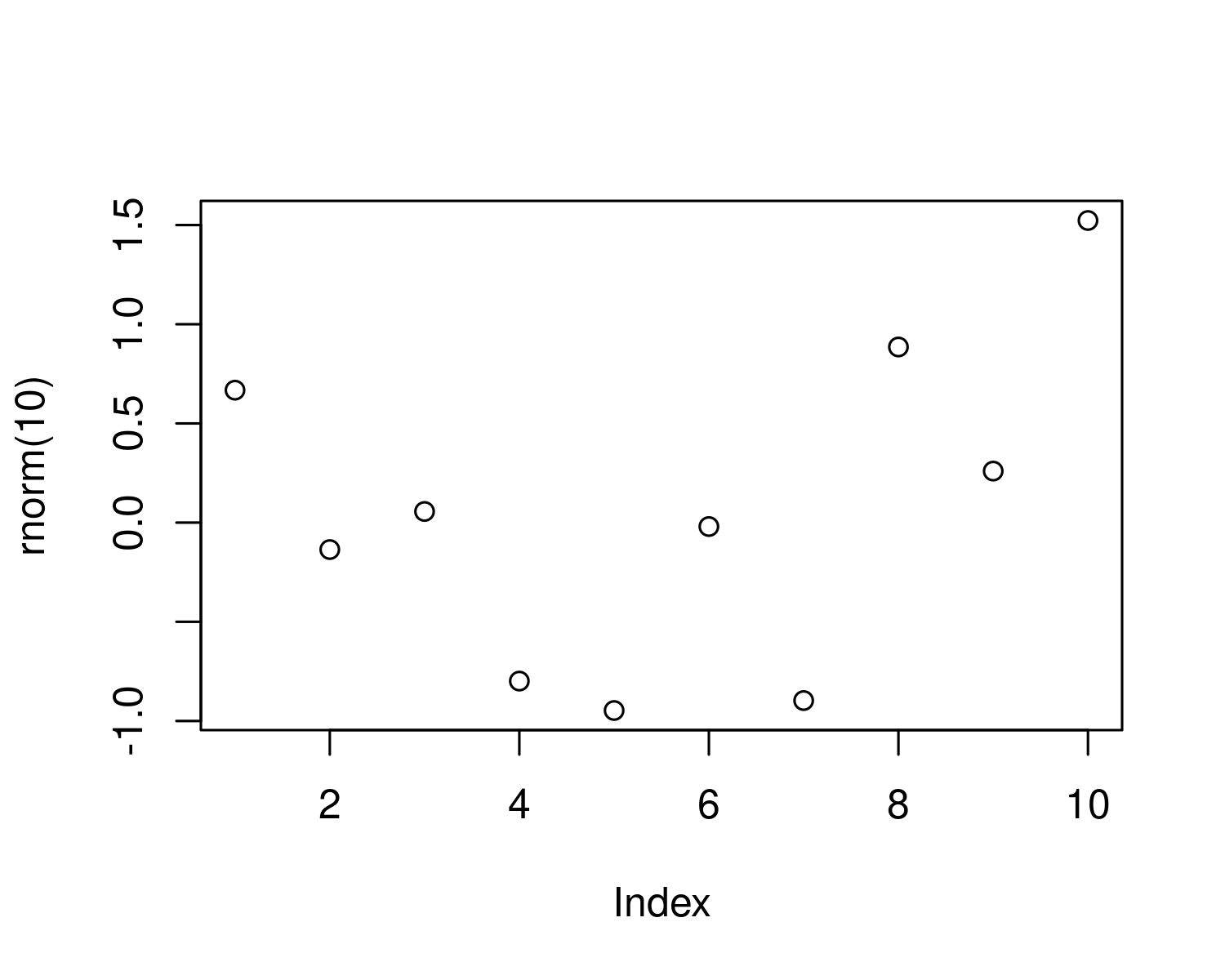


Figure 5.1: A plot of random numbers

Figure 5.1 shows how we can have a caption and cross-reference for a plot

Here is an example of inline code 3.14 in the middle of a sentence.

# 6 Discussion

# 7 Conclusion

# 8 Acknowledgements

# 9 References

Borgwardt, K.M., Gretton, A., Rasch, M.J., Kriegel, H.-P., Schölkopf, B., Smola, A.J., 2006. Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22, e49–e57.

Eddelbuettel, D., François, R., 2011. Rcpp: Seamless R and C++ integration. Journal of Statistical Software 40, 1–18. <https://doi.org/10.18637/jss.v040.i08>

Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., 2012. A kernel two-sample test. The Journal of Machine Learning Research 13, 723–773.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., Smola, A., 2006. A kernel method for the two-sample-problem. Advances in neural information processing systems 19, 513–520.

Gretton, A., Fukumizu, K., Harchaoui, Z., Sriperumbudur, B.K., 2009. A fast, consistent kernel two-sample test., in: NIPS. pp. 673–681.

Hotelling, H., 1931. The generalization of student’s ratio. The Annals of Mathematical Statistics 2, 360–378.

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Student, 1908. The probable error of a mean. Biometrika 1–25.

### 9.0.1 Colophon

This report was generated on 2021-05-11 12:09:28 using the following computational environment and dependencies:

#> ─ Session info ───────────────────────────────────────────────────────────────  
#> setting value   
#> version R version 4.0.4 (2021-02-15)  
#> os Ubuntu 21.04   
#> system x86\_64, linux-gnu   
#> ui X11   
#> language (EN)   
#> collate en\_US.UTF-8   
#> ctype en\_US.UTF-8   
#> tz America/New\_York   
#> date 2021-05-11   
#>   
#> ─ Packages ───────────────────────────────────────────────────────────────────  
#> package \* version date lib source   
#> assertthat 0.2.1 2019-03-21 [1] CRAN (R 4.0.1)   
#> backports 1.2.1 2020-12-09 [1] CRAN (R 4.0.3)   
#> bookdown 0.21 2020-10-13 [1] CRAN (R 4.0.3)   
#> cachem 1.0.4 2021-02-13 [1] CRAN (R 4.0.3)   
#> callr 3.5.1 2020-10-13 [1] CRAN (R 4.0.3)   
#> checkmate 2.0.0 2020-02-06 [1] CRAN (R 4.0.1)   
#> cli 2.5.0 2021-04-26 [1] CRAN (R 4.0.4)   
#> colorspace 2.0-1 2021-05-04 [1] CRAN (R 4.0.4)   
#> commonmark 1.7 2018-12-01 [1] CRAN (R 4.0.1)   
#> crayon 1.4.1 2021-02-08 [1] CRAN (R 4.0.3)   
#> DBI 1.1.1 2021-01-15 [1] CRAN (R 4.0.3)   
#> desc 1.2.0 2018-05-01 [1] CRAN (R 4.0.1)   
#> devtools 2.3.2 2020-09-18 [1] CRAN (R 4.0.3)   
#> digest 0.6.27 2020-10-24 [1] CRAN (R 4.0.3)   
#> dplyr 1.0.6 2021-05-05 [1] CRAN (R 4.0.4)   
#> ellipsis 0.3.2 2021-04-29 [1] CRAN (R 4.0.4)   
#> evaluate 0.14 2019-05-28 [1] CRAN (R 4.0.1)   
#> fansi 0.4.2 2021-01-15 [1] CRAN (R 4.0.3)   
#> fastmap 1.1.0 2021-01-25 [1] CRAN (R 4.0.3)   
#> fs 1.5.0 2020-07-31 [1] CRAN (R 4.0.3)   
#> generics 0.1.0 2020-10-31 [1] CRAN (R 4.0.3)   
#> ggplot2 3.3.3 2020-12-30 [1] CRAN (R 4.0.3)   
#> glue 1.4.2 2020-08-27 [1] CRAN (R 4.0.3)   
#> gt 0.1.0 2021-05-11 [1] Github (rstudio/gt@eff3be7)   
#> gtable 0.3.0 2019-03-25 [1] CRAN (R 4.0.1)   
#> highr 0.8 2019-03-20 [1] CRAN (R 4.0.1)   
#> htmltools 0.5.1.1 2021-01-22 [1] CRAN (R 4.0.3)   
#> knitr 1.31 2021-01-27 [1] CRAN (R 4.0.3)   
#> lifecycle 1.0.0 2021-02-15 [1] CRAN (R 4.0.3)   
#> magrittr \* 2.0.1 2020-11-17 [1] CRAN (R 4.0.3)   
#> memoise 2.0.0 2021-01-26 [1] CRAN (R 4.0.3)   
#> munsell 0.5.0 2018-06-12 [1] CRAN (R 4.0.1)   
#> pillar 1.6.0 2021-04-13 [1] CRAN (R 4.0.4)   
#> pkgbuild 1.2.0 2020-12-15 [1] CRAN (R 4.0.3)   
#> pkgconfig 2.0.3 2019-09-22 [1] CRAN (R 4.0.1)   
#> pkgload 1.2.0 2021-02-23 [1] CRAN (R 4.0.3)   
#> prettyunits 1.1.1 2020-01-24 [1] CRAN (R 4.0.1)   
#> processx 3.4.5 2020-11-30 [1] CRAN (R 4.0.3)   
#> ps 1.5.0 2020-12-05 [1] CRAN (R 4.0.3)   
#> purrr 0.3.4 2020-04-17 [1] CRAN (R 4.0.1)   
#> R6 2.5.0 2020-10-28 [1] CRAN (R 4.0.3)   
#> remotes 2.2.0 2020-07-21 [1] CRAN (R 4.0.3)   
#> rlang 0.4.11.9000 2021-05-11 [1] Github (r-lib/rlang@7cd1f5c)   
#> rmarkdown 2.7 2021-02-19 [1] CRAN (R 4.0.3)   
#> rprojroot 2.0.2 2020-11-15 [1] CRAN (R 4.0.3)   
#> sass 0.3.1.9003 2021-05-11 [1] Github (rstudio/sass@6166162)   
#> scales 1.1.1 2020-05-11 [1] CRAN (R 4.0.1)   
#> sessioninfo 1.1.1 2018-11-05 [1] CRAN (R 4.0.1)   
#> stringi 1.6.1 2021-05-10 [1] CRAN (R 4.0.4)   
#> stringr 1.4.0 2019-02-10 [1] CRAN (R 4.0.1)   
#> testthat 3.0.2 2021-02-14 [1] CRAN (R 4.0.3)   
#> tibble 3.1.1 2021-04-18 [1] CRAN (R 4.0.4)   
#> tidyr 1.1.3 2021-03-03 [1] CRAN (R 4.0.3)   
#> tidyselect 1.1.1 2021-04-30 [1] CRAN (R 4.0.4)   
#> usethis 2.0.1.9000 2021-02-15 [1] Github (r-lib/usethis@aaf79d8)  
#> utf8 1.2.1 2021-03-12 [1] CRAN (R 4.0.3)   
#> vctrs 0.3.8.9000 2021-05-11 [1] Github (r-lib/vctrs@52157b7)   
#> withr 2.4.2 2021-04-18 [1] CRAN (R 4.0.4)   
#> xfun 0.22 2021-03-11 [1] CRAN (R 4.0.3)   
#> yaml 2.2.1 2020-02-01 [1] CRAN (R 4.0.1)   
#>   
#> [1] /home/fred/R/x86\_64-pc-linux-gnu-library/4.0  
#> [2] /usr/local/lib/R/site-library  
#> [3] /usr/lib/R/site-library  
#> [4] /usr/lib/R/library

The current Git commit details are:

#> Local: main /home/fred/work/professional-development/mmdproject  
#> Remote: main @ origin (https://github.com/fboehm/mmdproject.git)  
#> Head: [bda2491] 2021-05-10: added text to spec aims section