Journal of Cell Science

Prog-Plot – a visual method to determine functional relationships for false discovery rate regression methods

Nicolás Bello and Liliana López-Kleine
DOI: 10.1242/jcs.260312

Editor: John Heath

Review timeline

Original submission:	6 June 2022
Editorial decision:	25 July 2022
First revision received:	17 October 2022
Editorial decision:	2 November 2022
Second revision received:	25 November 2022
Accepted:	1 December 2022

Original submission

First decision letter

MS ID#: JOCES/2022/260312

MS TITLE: PP-Plot: a visual method to determine functional relationships for FDR regression methods

AUTHORS: Nicolás Bello Reyes and Liliana Lopez-Kleine
ARTICLE TYPE: Tools and Resources

We have now reached a decision on the above manuscript.

To see the reviewers' reports and a copy of this decision letter, please go to: https://submit-jcs.biologists.org and click on the 'Manuscripts with Decisions' queue in the Author Area. (Corresponding author only has access to reviews.)

As you will see, the reviewers raise a number of criticisms that prevent me from accepting the paper at this stage. They suggest, however, that a revised version might prove acceptable, if you can address their concerns. If you think that you can deal satisfactorily with the criticisms on revision, I would be pleased to see a revised manuscript. We would then return it to the reviewers.

Please ensure that you clearly highlight all changes made in the revised manuscript. Please avoid using 'Tracked changes' in Word files as these are lost in PDF conversion.

I should be grateful if you would also provide a point-by-point response detailing how you have dealt with the points raised by the reviewers in the 'Response to Reviewers' box. Please attend to all of the reviewers' comments. If you do not agree with any of their criticisms or suggestions please explain clearly why this is so.

Reviewer 1

Advance summary and potential significance to field

This paper studies the problem of multiple testing with covariates. In traditional multiple testing, the analyst's goal is to reject null hypotheses H_1, ..., H_n based on p-values P_1, ..., P_n (where a

small p-value P_i provides evidence against H_i) subject to controlling false discoveries (e.g., the false discovery rate). In multiple testing with covariates, one also has access to covariates X_1,...,X_n (one for each hypothesis) that may be related to the "prior" probability of a hypothesis being null.

The present paper proposes a visual companion, called the PP-plot ("progressive proportions plot"), that plots an estimate of the proportion of null hypotheses conditionally on a univariate observed covariate X_i as a function of X_i.

The suggested construction has two steps:

A) First, as argued by Boca, and Leek (2018), the proportion of nulls conditionally on covariates is approximately equal to:

$E[1(P_i>\lambda)/(1-\lambda) \mid X_i=x]$

for some $0<\lambda<1$.

B) The above is operationalized through nonparametric regression: Quantile slicing is used to turn X into an ordinal covariate (i.e., to slice X into a fixed number of consecutive strata). Then for each (discretized/categorical) value of x $E[1(P_i>\lambda)/(1-\lambda) \mid X=x]$ is estimated by computing the average of $1(P_i>\lambda)/(1-\lambda)$ over all i such that X_i=x. This is then plotted as a function of x and for different values of λ.

I believe the research area within which the present paper operates is important. Understanding how covariates influence the probability of a hypothesis being null can provide further insights into the nature of the hypothesis tests conducted. Furthermore, the power of multiple testing procedures can substantially increase through covariates.

Comments for the author

I have some concerns regarding this manuscript mostly with regards to novelty, but also with respect to correctness/accuracy of claims (the violation of which could lead to misuses of the underlying statistical methodology).

Novelty:

A) The authors write in the abstract that "there are no available tools to verify it [the relationship between the proportion of p-values and the covariate]". This statement is not true. For example, Figure 2 in Boca, and Leek (2018)---which this paper is based upon---presents a plot that is very similar to the proposed "PP plot".

B) Estimating $E[1(P_i>\lambda)/(1-\lambda) \mid X=x]$ is a traditional problem in nonparametric statistics, see for example Wasserman (2006) for an introductory textbook treatment. This paper reinvents the "regressogram" estimator; a nonparametric regression estimator that is based on local averaging within strata (akin to a histogram), which is the first method typically taught in classes in nonparametric statistics.

There are many methods and software packages that can be used for this nonparametric regression task. For example, the R package "ggplot2" provides some smoothing options through the "geom_smooth" (https://ggplot2.tidyverse.org/reference/geom_smooth.html) functionality that could be used to derive very similar plots. One can also use splines, as done by Boca, and Leek (2018). Or if the authors prefer the regressogram (sometimes also called binscatter), there are existing packages, such as https://cran.r-project.org/web/packages/binsreg/. A benefit of using existing packages is that they can provide "battle-tested" implementations and automated choices for tuning parameters (such as the number of strata for discretization).

Technical issues:

A) There are important assumptions that need to be satisfied before conducting a multiple testing study with covariates: the covariates need to be independent of p-values under the null hypothesis (see, for example, Bourgon, and Huber (2010) for formal statements and explanations). This crucial assumption is never mentioned in this manuscript! But as explained in Bourgon, and Huber (2010), if this assumption does not hold, then using covariates within a multiple testing procedure can lead to excess false discoveries (and would be closely related to e.g., p-hacking). It is essential to provide diagnostic plots for the validity of this assumption and to stress the importance of the assumption to practitioners who may use multiple testing methods with covariates.

B) The authors write (regarding DESeq2) that "the default is a Benjamini-Hochberg". However, note that DESeq2 by default applies a multiple testing procedure with covariates, namely the Independent Filtering procedure of Bourgon and Huber (2010) [this is option "indepentFiltering=TRUE" in the "results" function of DESeq2].

C) "As suggested by Korthauer et al. (2019) we used the mean gene expression as a covariate, ": This covariate had already been suggested in the original DESeq2 publication (Love, Huber, and Anders (2014)).

Reviewer 2

Advance summary and potential significance to field

Bello et. al has created a user-friendly visual R function that identifies the functional relationship between a covariate and the null proportion. This method uses the covariate to address multiple corrections which is superior to the standard B&H (FDR) approach.

Comments for the author

A few suggestions may be necessary before the manuscript can be accepted for publication:

* The context of lambda should be made available in the introduction before usage.
* Authors mention that the smooth estimate of the curve is a determinant of the final estimate. It would be helpful in the example provided to specify this interpretation.
* The authors mention that this method specifically handles a covariate where the B&H method can not. It would be helpful to have a comparison with how the p-p plot visualization is superior or compares to the standard FDR (B&H) correction approach. This could be via a table summarizing the genes selected or metrics using the said GTEx example.
* Authors should include how assumption check is performed via the example provided.

First revision

Author response to reviewers' comments

Reviewer 1:

- **"There are no available tools to verify it [the relationship between the proportion of p-values and the covariate]" claim and Boca and Leek's plot**
  To the best of our knowledge users of the BL method are required to specify either a logistic or a linear regression model and there is no way to check which one could be better for a specific case study. The plots in the BL paper (2018) are all presenting either the true function $\pi_0\ (x)$ or the final estimated curve $\widehat{\pi_0(x)}$ assuming a model for the data. The purpose of the PP-Plot is to be able to plot this covariate / null proportion relationship before a model is assumed, or to assess the goodness of fit of the fitted model.

- **Reinventing the "regressogram" estimator (Wasserman)**
  We are not trying to reinvent the regressogram, we are arguing that such techniche (with a minor adaptation) has not been used for this particular problem, and it solves an issue that arises with the new methodologies that do fdr regression assuming a model for this relationship.

- **There are other implementations for this regression task**
  In order to create a similar plot with geom_smooth you would still need to create a dataset of all the responses for different thresholds lambda (noting that the response is not binary because you have to divide by $1 - \lambda$) and choose a non-parametric function for the "method" argument, as loess would give similar results to the regressogram but it is computationally expensive for +20 000 observations and multiple thresholds as we have. binsreg gives better results but you still need to create the datasets for different thresholds and adjust the response variable yourself before using the function. We are not saying that these other implementations can't be used but rather that for this specific application it is more convenient to have everything in the same function which also gives the estimates for plotting flexibility. ppplot is a simple function but it prevents the user from having to create their own datasets for different thresholds for the visualization.

- **Does not stress enough the independence assumption**
  A paragraph has been added addressing this concern and Figure 1 shows the diagnostic plot.

- **Default for DESeq2 is not BH but BH and IHF**
  The manuscript has been corrected to reflect these default settings. We thank the reviewer for this correction.

- **Mean gene expression was first proposed by Love et al. 2014**
  It was already suggested by Love et al. but we decided to cite Korthauer et al. because their suggestion is based on their experience with these specific methods of correction that did not exist back then and we thought it was the most relevant. But yes, it is not a surprise that the same covariate is also informative here.

Reviewer 2:

- **Give context to the parameter lambda**
  A paragraph has been added in the introduction to address this concern. We talk about the role of lambda in the estimation of the null proportion and how the estimation behaves for different values of lambda.

- **Smooth estimate to final estimate, give example for interpretation**
  A paragraph has been added at the end of section 2.2 explaining how these smoothed estimates of the null proportion are used to obtain the q value in a very descriptive way.

- **Give a comparison with BH by table summarizing genes or metrics**
  We added a table comparing the number of DE genes obtained from the two corrections. A more general comparison of the performance of the two methods was done by Korthauer et al.(2019) and would be beyond the scope of this manuscript.

- **Include assumptions check for the example**
  This concern was addressed and included in the manuscript.

We extend our gratitude to both reviewers for their feedback as we found it very critical and appropriate. We hope that these changes address their concerns and greatly improve the quality of the manuscript, as we believe it did.

Second decision letter

MS ID#: JOCES/2022/260312

MS TITLE: PP-Plot: a visual method to determine functional relationships for FDR regression methods

AUTHORS: Nicolás Bello Reyes and Liliana Lopez-Kleine
ARTICLE TYPE: Tools and Resources

We have now reached a decision on the above manuscript.

To see the reviewers' reports and a copy of this decision letter, please go to: https://submit-jcs.biologists.org and click on the 'Manuscripts with Decisions' queue in the Author Area. (Corresponding author only has access to reviews.)

As you will see, the reviewers gave favourable reports but raised some final critical points that will require editorial amendments to your manuscript. I hope that you will be able to carry these out because I would like to be able to accept your paper, depending on further comments from reviewers.

Please ensure that you clearly highlight all changes made in the revised manuscript. Please avoid using 'Tracked changes' in Word files as these are lost in PDF conversion.

I should be grateful if you would also provide a point-by-point response detailing how you have dealt with the points raised by the reviewers in the 'Response to Reviewers' box. Please attend to all of the reviewers' comments. If you do not agree with any of their criticisms or suggestions please explain clearly why this is so.

Reviewer 1

Advance summary and potential significance to field

See report of initial submission for a summary of the advance made in this paper and its potential significance to the field. The paper has improved after the first round of revision.

Comments for the author

1) It appears that something went wrong in Figure 1. The histograms do not look correct (consider the x-axis and also the gaps at the origin). Perhaps this was uploaded incorrectly? Toward the end of the pdf (at page 4) there is a full page plot of the histograms that looks more reasonable.

2) PP-Plot usually refers to a different type of statistical plot (see, e.g., Wikipedia: https://en.wikipedia.org/wiki/P%E2%80%93P_plot), as such I think calling this paper and the method PP-Plot may be misleading. Please consider renaming.

3) "strong exponential behavior": Exponential does not seem to be the right term here.

Reviewer 2

Advance summary and potential significance to field

The authors have successfully incorporated the revisions and these updates have indeed improved the manuscript.

Comments for the author

Minor comments:
- Table 1 should be referenced in the text appropriately.

Second revision

Author response to reviewers' comments

Reviewer 1:
- **Figure 1 histograms:** Yes, it was uploaded incorrectly. We will make sure that it looks the way it is supposed to.
- **PP-Plot name change:** We agree, and have changed the title and corresponding parts of the manuscript to reflect this change.
- **Exponential behavior:** We only intended this paragraph to be a hypothetical scenario, not to assume that such behavior was real. But we have rephrased that part, in hopes that it is clearer now. .

Reviewer 2:
- **Table 1 not referenced:** It was an oversight, it is now mentioned in the text.

Third decision letter

MS ID#: JOCES/2022/260312

MS TITLE: Prog-Plot: a visual method to determine functional relationships for FDR regression methods

AUTHORS: Nicolás Bello Reyes and Liliana Lopez-Kleine
ARTICLE TYPE: Tools and Resources

I am happy to tell you that your manuscript has been accepted for publication in Journal of Cell Science, pending standard ethics checks.