Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

A Capitalized Title: Something about a Package GGally

Barret Schloerke
Purdue University

Second Author
Plus Affiliation

Abstract

The abstract of the article.

Keywords: keywords, comma-separated, not capitalized, Java.

1. Introduction

The R package **ggplot2** is a plotting system based on the grammar of graphics. **GGally** extends **ggplot2** by adding several functions to reduce the complexity of combining geoms with transformed data. Some of these functions include a pairwise plot matrix, a scatterplot plot matrix, a parallel coordinates plot, a survival plot, and several functions to plot networks.

The focus of this paper explores the function ggduo, which makes a two grouped plot matrix.

generalized pairs plot expanded the traditional scatter plot matrix to work for both continuous and categorical columns.

ggduo's older brother, ggpairs, has already been been used in Cook, Lee, and Majumder (2016) `ggs_pairs` from Fernández-i-Marín (2016) and a reverse dependency in over twenty other packages on CRAN and BioConductor.

2. Methodology

A scatterplot matrix shows the same collection of data in each panel of a plot matrix, however each panel has different continuous axes. Similarly, the generalized pairs plot displays the same collection of data using different axes but allows for a mix of plotting methods and both continuous and discrete plot axes. ggduo removes the restriction plotting all pairs of points and displays the combination of two groups of the same dataset.

With ggduo, a scatterplot can be next to a boxplot, which can be next to a mosaic plot.

2.1. What is ggduo

ggduo is function in the GGally package that produces a ggmatrix plot matrix with one column of data plotted against another column of data for every panel of a plot matrix. For a given row, the same y axis variable is used, and for a given column, the same x axis variable is used. Unlike ggpairs, the restriction of every column being paired against every column is removed. ggduo only requires a set of 'Y' columns and a set of 'X' columns to be paired against each other.

Column types

ggduo inspects and displays the data columns according to their variable type: continuous or discrete. There are three plot type groups that an be made from these two options: continuous vs. continuous, continuous vs. discrete, and discrete vs. discrete. ggduo's default plotting behavior for continuous vs. continuous is to produce a scatterplot with a loess smooth curve displayed on top of the points. The default plotting behavior for discrete vs. discrete, is to summarize the data and display it as a mosaic plot.

The third group, continuous vs. discrete, will be refered to as a 'combination plot'. ggduo makes a distinction between the two possible combination plots: continuous vs. discrete (vertical combination plot) and discrete vs. continuous (horizontal combination plot). By default, ggduo displays grouped histograms for a horizontal combination plot and grouped box plots for a vertical combination plot. This distiction between a horizontal and vertical combination plot is made as there are no plot matrix sections (upper, lower, diagonal) in a ggduo plot.

2.2. User defined functions

The default plotting functions are provided by the GGally package, however the user may use their own plotting functions for each panel. Example!!

Each of these panels are full ggplot2 and are stored in a ggmatrix object which will be talked about in the next section.

2.3. Plot matrix

Winston Chang explains facet'ing as "[ploting] subsets of data into separate panels" Chang (2013). With ggplot2's faceting, each piece of data is displayed only once per plot matrix with the same plotting method and same axes. On the other hand, GGally's ggduo displays the same piece of data in every panel but with different axes and different plotting methods. As stated earlier, each column of the plot matrix shares the same x axis variable and each row of the plot matrix shares the same y axis variable

ggplot2 prevents discrete scales from being mixed with continuous scales. This idea makes sense when looking at a single panel plot or a multiple panel plot with different subsets of the same data. The ability to display different axes natively in the same plot matrix is not possible with ggplot2. However, only one scale type is used per ggmatrix panel, so the original 'per panel' logic is still kept. Before a ggmatrix object, to have two, related mixed axes plots in the same display could only be done with **gridExtra** or **grid**. Neither of these packages would produce output with the native ggplot2 presence.

`ggmatrix` extends the `facet_grid` idea but allows for a different scale per row and column. This allows for a cohesive plot matrix structure where all axes are shared among common parts.

Mixing different plot types allows the user the digest their data with multiple ways with the same display.

To be able to mix axes arbitrarily, the GGally package uses a ggmatrix object. It contains a collection of plot objects and other plotting features such as number of columns and rows, title, labels, etc. Each plot is displayed in a cell of a plot matrix with the same contruction as a ggplot2's `facet_grid()` plot matrix. There are strips on the top and right side of the plot matrix, and plot axes are only displayed on left and bottom side of the plot matrix. Like with ggplot2, space can be made available for the legend, title, x label, and y label.

ggplot2 displays data with the same axis size even if the type of plot changes. Leveraging this fact, each plot is displayed with independent axis knowledge.

that can be displayed to

limitation of ggplot2. all continuous scale or all same category. no mixing able to create the same plot for different subsets of data with ggplot2 unable to combine two related plots in the same display without use of **gridExtra** or **grid**

plots may be retrieved individually. not possible with ggplot2

core of plot matrix is a ggmatrix object generalized plot matrix. plots are arbitrary composite plot Emerson, Green, Schloerke, Crowley, Cook, Hofmann, and Wickham (2013) allows for plots with completely different scales to be shared in the same matrix

output is modeled directly after **ggplot2**'s `facet_grid` columns rows strips (titles) general plot behavior allows for legends and uses the ggplot2 formatting similar to ggplot2, meta information is stored in the plot matrix until print time and may be altered after it's inception.

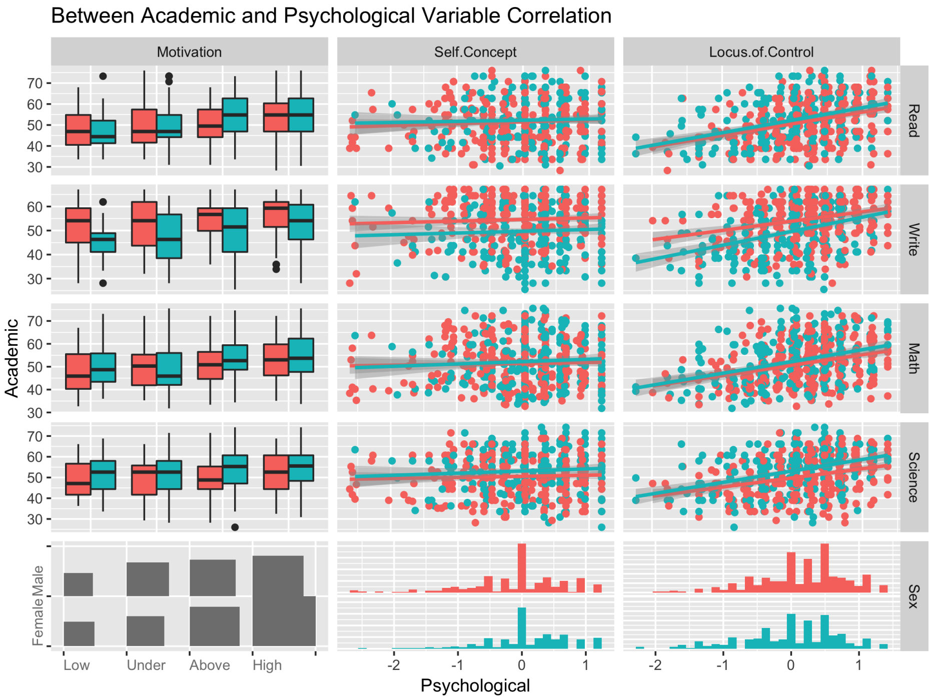
similar to ggpairs, ggduo formats the plot objects to be displayed with ggmatrix inherits all the functionality of ggmatrix

3. Examples

3.1. Canonical correlation analysis

Canonical correlation analysis is a method to analyize the correlation between two matrices Hotelling (1936). Canonical correlation analysis can be directly displayed with `ggduo`. Before `ggduo`, canonical correlation analysis did not have a cohesive plotting mechanism to visually display the associations of two sets of mixed type variables. Current examples use 'ggpairs' to display all pairs of columns when only a subset of combinations are needed. 'ggpairs' can be used to check the within correlation of both the explanatory variables and response variables, whereas `ggduo` can be used to check the correlation between the explanatory and response variables. Using ggmatrix's power of independent plots, each cell can display custom plots to show any information.

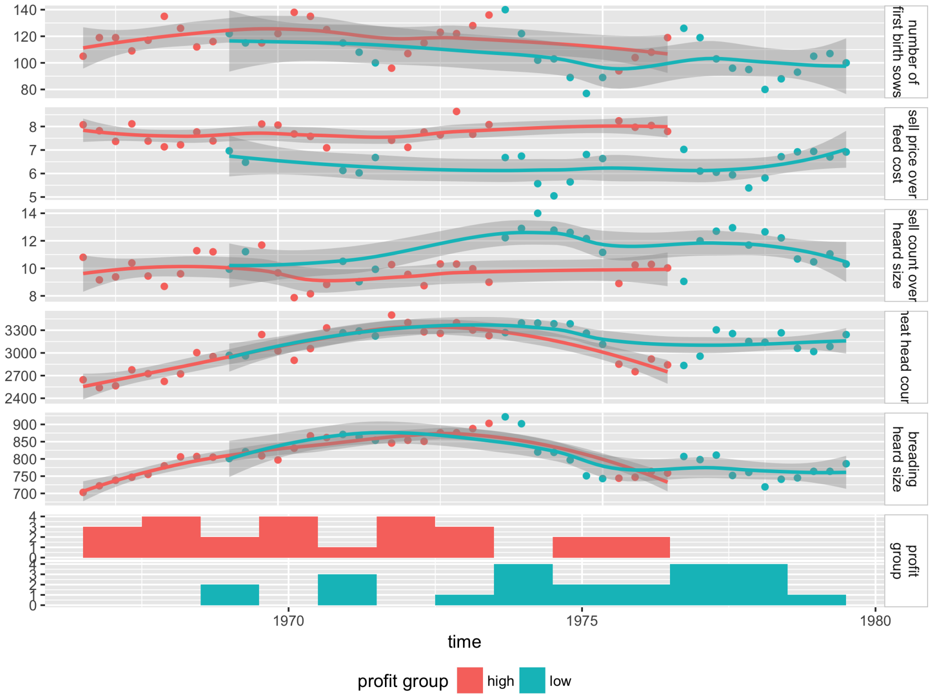
Explain academic data here



3.2. Multiple Time Series Analysis

The **stats** package has a `ts.plot` function that currently allows for multiple time series to be printed in a single plot. Display the time axis on the X axis with multiple columns on the Y axis. `ts.plot` displays the data in the same plot panel with a shared Y axis. Displaying the data on the same axis does not make sense in all cases. Splitting the multiple time series plot along the Y axis, we can display multiple panels with different Y axes with a shared X axis using `ggduo`. This can be done with `ggts` which wraps to `ggduo` with the X column label turned off by default and a plot X label of 'time'.

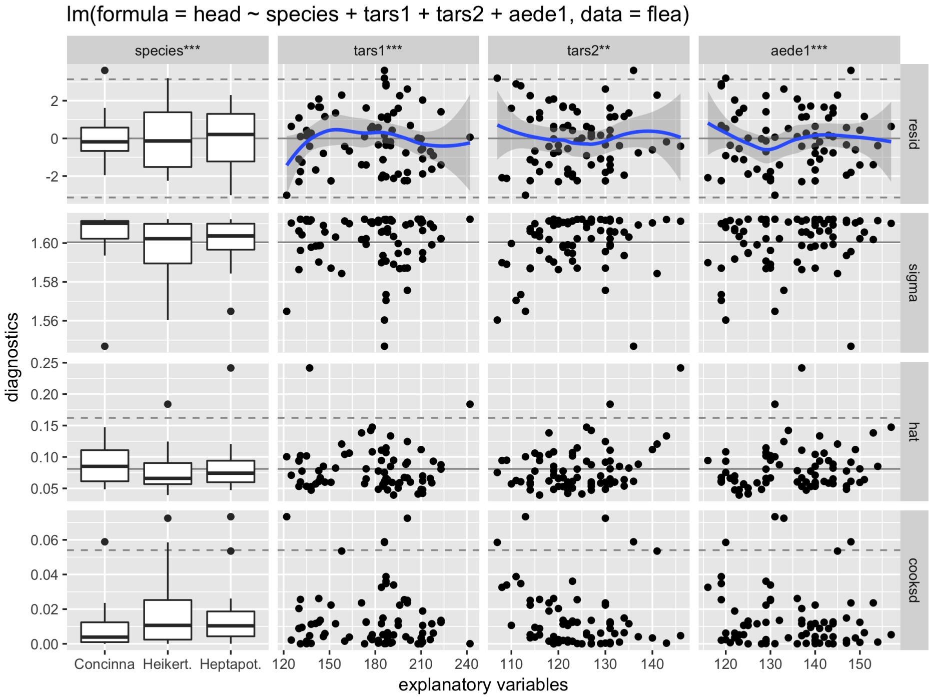
Explain pigs data here

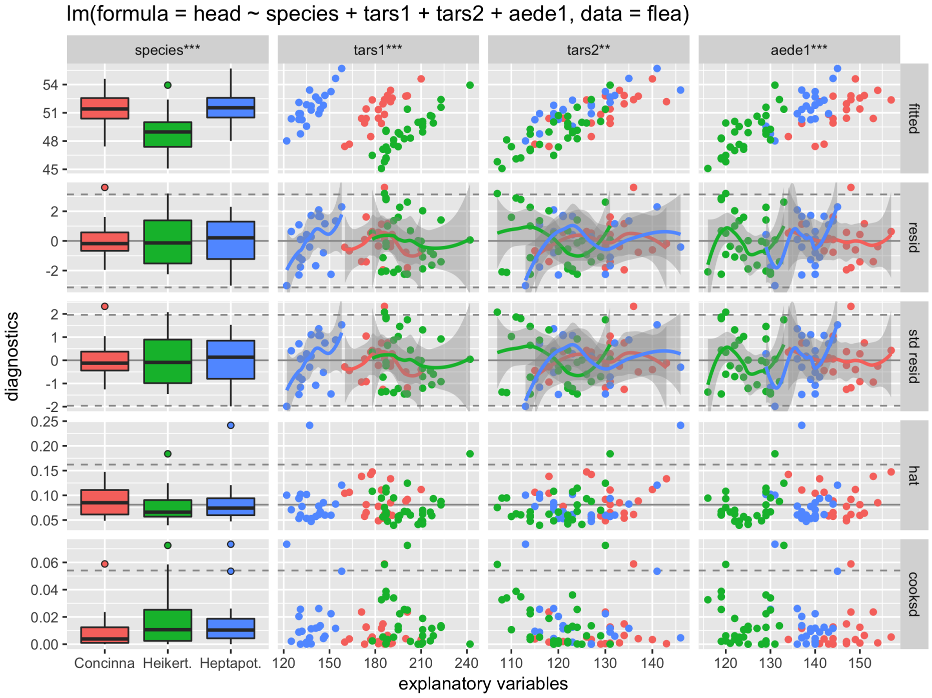


3.3. Multiple regression diagnostics

Multiple regression analysis is currently being done using ggpairs and only needs to display a subset of the pairs of columns. With the basis of ggduo displaying each row of the data in every panel with different functions, ggduo quickly extends to model diagnostics. There are many diagnostics that can be calculated for each row of explanatory data. By default, `ggnostic` (a function that wraps around `ggduo`) looks at the residuals, leave one out sigma value, leverage points, and Cook's Distance. Each diagnostic information is plotted against all explanatory variables used in the model.

Using stats::step to determine a best fitting model, the default diagnostics are displayed against 'species', 'tars1', 'tars2', and 'aede'. The residuals have a 95% confidence interval in dashed lines and a solid line at 0. The leave one out sigma value displays a solid line for the current model's sigma value. The leverage points (hat) are centered around the solid line at p/n and have a dashed line at $2 * p/n$. Finally, the Cooks's distance has a grey dashed line at $4/n$. Each solid line corresponds to the expected value and each dashed line corresponds to a 'signifigance' cuttoff value. The asterisks in the X axis strips corresponding to the significance of an anova F test.





4. Future Direction

Currently all individual plots are rendered at run time. **ggplot2** is known for having slower print speeds. currently printing n*m plots takes time.

Future ggplot2 versions will allow for custom faceting. this could potentially allow ggmatrix to print a single ggplot2 object with custom facet scales for each plot. This would dramatically reduce the amount of time a ggmatrix takes to render. Printing time would still remain the same.

Link up with the javascript packages for interactive plot matricies. Would be great to add a one more line of code to turn it into an interactive plot.

5. Discussion

Pros single line of code to produce a composite plot matrix works well with wide data Cons takes longer time to print with larger data prints m*n ggplots to produce a single m*n plot matrix

ggmatrix was a feature request to handle different length and different column plot matrix users wanted to place custom plots in a custom arrangement ggduo came about was a feature request to have a ggmatrix version of the TeachingDemos::pairs2 function.

Can achieve custom plots. Can not achieve a forced cohesive scale to be used on all plots in a row or column is not a native ggplot2 object, but trying hard to be like one!

6. References

References

Chang W (2013). *R Graphics Cookbook*. O'Reilly Media, Inc. ISBN 1449316956, 9781449316952.

Cook D, Lee EK, Majumder M (2016). "Data Visualization and Statistical Graphics in Big Data Analysis." *Annual Review of Statistics and Its Application*, **3**, 133–159.

Emerson JW, Green WA, Schloerke B, Crowley J, Cook D, Hofmann H, Wickham H (2013). "The generalized pairs plot." *Journal of Computational and Graphical Statistics*, **22**(1), 79–91. doi:10.1080/10618600.2012.694762.

Fernández-i-Marín X (2016). "ggmcmc: Analysis of MCMC Samples and Bayesian Inference." *Journal of Statistical Software*, **70**(1), 1–20. ISSN 1548-7660. doi:10.18637/jss.v070.i09. URL https://www.jstatsoft.org/index.php/jss/article/view/v070i09.

Hotelling H (1936). "Relations Between Two Sets of Variates." *Biometrika*, **28**(3/4), 321–377. ISSN 00063444. URL http://dx.doi.org/10.2307/2333955.

Affiliation:

Barret Schloerke
Department of Statistics
Purdue University
250 N. University St.
West Lafayette, IN 47906 USA
E-mail: schloerke@gmail.com
URL: http://ggobi.github.io/ggally

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd