Research Replication

Brian G. Peterson

May 20, 2015

This essay provides a process and supporting framework to help in replicating the research of others. Analysis of new trading system hypotheses often starts when an analyst or their management reads an interesting paper which provides the seed of a new idea. To fully understand and make use of the ideas in the paper, the analyst needs to first replicate the paper, since most papers do not publish data and code.

After replicating the paper, the results need to be analyzed to determine whether the model or results are credible, and how the analysis of the paper can be improved. Only after replication is complete, and the analyst has formulated an opinion on the quality and applicability of the research to be done, can s/he move on to creating a backtest for a new strategy incorporating or based on the models or techniques in the paper.

Throughout this document, we will refer to the "original research" or "source paper" interchangeably, though the research to be replicated may be part of a book chapter, or a blog post, or a web page. With practice, after a number of replications of this type are completed, some of the steps in this process may be curtailed or skipped altogether. We believe that the complete process has value as the reproducibility and clarity of thinking of the research will make the deliverables more reusable and useful over time.

Try to go through the entire process at a publication-quality level, as you may choose to publish part or all of your replication work at a later point. You should write in your own words, in order to better develop a deeper understanding of the source material. If you must quote, make sure to properly cite the quotation, as even unintentional or accidental plagiarism can have serious consequences. Any portion of the replication may be edited as you progress through the project; do not consider it to be a completed document until the entire project is finished.[1]

Summarize the Paper

One of the first main steps is to be able to summarize the paper in a structured document which should be 2-3 pages long, devoting no more than one paragraph to each main point. The goal of the summary stage is to make certain that the paper and its research context are understood.

The *précis* model of formal summary is useful in this regard.
Introduction, 2-5 main points, conclusion.
Similar forms are the classic "five paragraph essay".

Why so formal? The concise summary provides a road map for the following stages.
The goal of such a short, formal, summary is to organize your thinking in a structured way which will make replication easier. The summary will help to uncover logical errors or omissions (either in the reader or the paper) much earlier in the process. It will allow you to identify early where you need more information, and where you will be likely to have difficulties with your research.

Candidates for main points are key assertions or findings, major contributions, summary of the main technique(s) used, etc. Each of the main points should describe how the paper supports its points, including what methods, tests, or arguments are made.

[1] Thanks to David Matteson for his additions to this process.

The concise précis model is 4-6 declarative sentences. The analyst should strive to follow the concise model in the introductory paragraph of the paper summary. The introductory paragraph should have the following sentences:

1. a single introductory sentence describing
    - the title, authors, and journal the paper appeared in
    - an appropriate verb such as *assert, argue, deny, demonstrate, disprove, examine, prove, refute*
    - a **that** object clause that describes the thesis of the work
2. one sentence for each major point in the paper
    - choosing *no more* than five points
    - each sentence should describe the technique or claim
    - and the claimed result of applying the concept
3. one sentence that ties it all together: what the reader should be expected to take away from the paper

The rest of the précis document will devote one paragraph to each major point from the introduction, using the structure:

1. a declarative sentence which restates and paraphrases the technique or claim to be covered in this paragraph
2. two to four sentences describing
    - the main methods, tests, or arguments
    - formulas (with appropriate citation) when they add necessary precision
    - results from this portion of the paper
3. one sentence which summarizes any take away conclusions about the idea being described

The summary document will then conclude with a single paragraph which ties together all the main points, and describes the relevance of the work to the analyst's research interests.

Describe the Hypothesis

At this next stage, the analyst should work to expand the understanding of the hypothesis or hypotheses which are to be tested. The hypotheses presented by the source paper need extraction, enumeration, and expansion. Expected tests for the hypotheses also need to be considered and specified in this stage.

The hypothesis descriptions should include:

- what is being analyzed (the subject),
- the dependent variable(s) (the output/result/prediction)
- the independent variables (inputs into the model)
- the anticipated possible outcomes, including direction or comparison
- addresses *how you will validate or refute each hypothesis*

The précis form of structured paragraphs (containing the points above) may be useful in stating the hypotheses, or a less regimented hypothesis/test pairing may be more appropriate. Some judgement will be required at this stage both in extracting hypotheses, and in describing the hypotheses and their tests.

Literature Review

After summarizing the paper, and extracting the hypotheses, it is time to move on to a *literature review.*(Levy and Ellis 2006) A literature review is a formal summary of the literature covering the idea that you are researching. It serves to make certain that you are looking at relevant work, and to keep formal notes on the relevant papers.

We recommend again following the précis model:

1. the bibliographic entry for the paper
2. a single sentence describing the thesis of the work
3. two to four sentences covering the main points or findings of the paper
4. a summary sentence describing this paper's relevance and relation to the current research project

It is not necessary or desirable to gain deep understanding of every paper you will review during the course of your replication research. The main goal is rather to understand the framework that your research goal fits into, and develop resources for deeper understanding when that is required. The literature review is to help organize the tools and knowledge needed to complete the replication. Like a chef measures out ingredients before starting a recipe, or a carpenter gathers tools, the analyst is gathering tools and knowledge to make sure they have enough to work with for the replication project.

organization of the literature review

The analyst will need to decide on an organizational scheme for the literature review. In the social sciences, it is common to present the literature review as a narrative which describes a topic through its available research material almost as a story arc. The quantitative analyst trying to replicate strategy or model research should probably avoid this purely narrative form because it typically lacks enough detail to anchor replication. There are two models which you will return to; which model is appropriate will depend both on how the replication is to proceed, and how the analyst's internal organizational structure is envisioned.

The first model is the *annotated bibliography.* It is organized alphabetically, by author, containing the bibliographic entry plus the paragraph of summary as described above. Methodologically, it has an advantage that all the material for the annotated bibliography may be stored in the reference manager software, and a template may be used to construct the annotated bibliography automatically from all the reference notations in the replication research report. The annotated bibliography has the disadvantage that it has no other organization, so keeping track of multiple references for importance or topic can be very difficult.

The second model organizes the literature review by key topic. The top level for this organization of the literature review should be the two to four main topics of the paper being replicated. It may further be organized into sub-topics or techniques, as appropriate for the research project. The advantage of this model is that it allows the analyst to rapidly locate key references on a topic or idea or technique, along with their summaries. The potential disadvantage is that it is very specific to this research project, and may be less useful in the future for additional research. This risk is usually offset by the topic-based organization, which allows papers to be quickly located by topic.

key references of the paper being replicated

The literature review should start with the key references of the paper you are trying to replicate. It should have become clear while reading and summarizing the paper which papers provided starting material or key techniques for this paper. There will also often be key original references in a particular topic, cited by many other papers. These papers should be located and summarized in a single paragraph each, as described above.

finding similar work

Other papers which should be included in the literature review are similar work. Resources like Google Scholar will automatically recommend similar papers, and even order them by number of papers that reference the paper in the results. Key words and phrases from the source material may be used to find more recent papers. The analyst will want to make sure that they at least review a few papers which are at the top of a literature search on the key topics.

references with implementation hints

Finally, the literature review is not complete without including implementation references that cover the key analytical techniques to be used in the replication. These should be as hands-on as possible. For example, a paper to be replicated which includes a complex multivariate linear model may include a reference to *Multivariate Linear Models in R*(Fox and Weisberg 2011). The paragraph describing this reference can focus on the specific chapter or technique of interest, aiding the analyst in collecting their thoughts, and aiding any readers to find the relevant material.

refuting your hypothesis

The literature review may refute your initial hypothesis. In searching for more information about the topics or models you hope to replicate, you may find sources that refute or challenge those methods or theories. This should be carefully documented, as it is potentially very valuable. If the evidence in the source materials is very detailed andhas a good experimental design, you may choose to document the evidence and scrap or revise your hypotheses. If the new source is suggestive, inconclusive, or refutes a related hypothesis that is not identical to the object of your replication study, then you may choose to identify the argument, evidence, and tests, and add it to what you will test in the replication phase of your project.

Data

The paper to be replicated will have described the data which was used in the paper. Summarize this information in your replication report. Pay particular attention to data sources, time-frames, and specific instruments. If a data vendor is mentioned, be sure to note this in your report. This constitutes the *original source data* which you are trying to replicate.

In rare cases, the original source data will be published exactly online. It may be from some earlier paper, or it may be in the "supplementary materials" included online at the journal publisher or author's website. If you can find this data, it will save you a lot of time. In most cases, the original source data will not be available.

If the source paper did not specifically list the data as private or confidential, and time allows, you may wish to contact the primary author of the source paper via email and (politely) request the data. If the author provides the data, make sure to thank the author and properly cite this in your report.

When you do not have the original source data, you must locate data that is as close as possible to the source data. We recommend working in the following order:

1. see what data vendors you have access to at your organization
2. check and see if you have access to any data vendors that were listed in the original paper
3. if you don't have access to the precise data vendors, or the paper does not state which vendor was used, then make a list of possible sources by subsetting your list of available sources

4. determine what symbols will need to be requested from each vendor
5. if your vendors do not have all the data you need, determine if you can get enough representative data to proceed, or if you need to try another approach

Once you've located the data, you need to download it. Keep notes on the process, the vendor, and all of the symbols. This will aid you later if you come back to this research, or want to extend the data.

If you can download the data directly in *R*, the code to so should be included in your replication report, though you may comment it or mark it to not be run given the time that downloading takes. Store all of the data for your replication with the project archive.

If possible, get data that copies the time-frame used in the original paper, in addition to the specific instruments. In addition, make a data request all the way to the current date. This "newer" data should be reserved as a validation set, it can be used for many tests of out of sample deterioration and overfitting.

Once you have the data, you need to get it into *R* if you were not able to download it directly into your analytical environment. Code for loading the data, labeling it, converting it to **xts**, etc. should be included in your replication report. Add citations for all packages that you use.

Some attention should be given in your report to the checks that you do for data quality. For example:

- is the data complete? are there gaps in it?
- do you have all the instruments from the source paper?
- do you have the same time-frames?
- can you validate that your data looks like the source data by quickly replicating any charts or graphs from the paper?
- do you need to do any data cleaning for outliers?
- was any data cleaning mechanism described in the source paper?

Document all the steps that you take with the data in your replication report, and keep the code for manipulating the data in the replication report as well.

Managing data is a critical step to any quantitative research project. The more information on finding, acquiring, loading, manipulating, and cleaning the data you can include as you do the work, the better.

Building the Model

Replicating a paper will involve several of the same steps as those you would follow to build a backtest. (see e.g. Peterson 2015) It also involves steps unique to paper replication.

replicating key analytical techniques

The summary of the paper identified the key analytical techniques used in paper. You will replicate or build these first after you have data to work with. If you are lucky, the modeling technique will already exist in *R*, and needs only to be applied to your data.

It is more likely that you will need to write code to replicate one or more key techniques from the paper. This will likely be both the hardest part of the replication, and the part most likely to contain errors.

As you work, keep your code in your replication report, with your discussion and analysis of the results. Rmarkdown contains many features for generation and inclusion of charts and tables, so your code should directly generate its outputs.

validate results as you work

At each stage, check to see if you get the same results as the paper. Strive to match, or be close, to every number that the paper publishes, either in the main text or in supplemental materials and appendices. If you can't match, or aren't close, figure out why. Document divergences and the process you are going through as you write code and check your results.

Occasionally, the paper will not contain enough information to understand precisely how the authors implemented the finer details of the technique. As you work, clearly document key assumptions or guesses that you make to try to replicate the technique or analysis. Often, your first (or second . . .) attempt will not get the desired result. Leave the attempt in the replication report, and make another attempt. Eventually, you can usually get close to what you see in the source paper. The "failed" attempts can also be valuable, showing alternative interpretations, and robustness of the methodology. This type of breadth is important if you later move on to using the techniques in a new strategy backtest.

Rarely, you will not be able to replicate one of the techniques in the paper. It may not matter, if you get through enough of the rest of the paper to form conclusions about the work as a whole.

In even more rare cases, you will implement the technique, and get results that are contradictory to the ones published in the paper. *Now what do you do?*

We recommend a series of analytical steps in this case:

- carefully check your work for bugs, reversed signs, etc. This is the most likely cause.
- can you replicate results from a different paper that is the source of the technique?
- can you independently validate the math?
- can data differences or data cleaning/manipulation difference explain the divergence of results?

If, after carefully checking your work, and validating against other data, if possible, you are certain that you have correctly implemented the technique, but the results still don't match, document this too. Add as much detail (and code!) about the checks you performed, things that you tested, and conclusions that you have drawn as you can to your replication report. Sometimes failure or contradicting the paper is the most important result that you could reach, saving you from implementing a bad or non-robust technique in a strategy.

Additional validation beyond replication of the paper's techniques is always valuable. If you identified hypotheses in your paper summary, now is the time to write code to test those hypotheses. If the model you have built has model fit or calibration statistics, you should include the output of these tests or statistics.

choosing a strategy model

If the paper presents a trading or investment strategy, you must choose the model which you will use to replicate it. In our experience, there are three main models which you will encounter:

1. **signal based strategy**: This model implements one or more indicators, signals, and rules to create a trading strategy. This type of model is best replicated in *quantstrat*(Brian G. Peterson, Ulrich, et al. 2015) in **R**.
2. **portfolio strategy**: This model implements some method of choosing, constructing a portfolio in, and rebalancing a portfolio of instruments. This type of model will likely use packages such as *Portfolioanalytics* (Brian G. Peterson, Carl, et al. 2015) or *parma*(Ghalanos and Pfaff 2014).
3. **pricing strategy**: This type of paper may or may not implement a trading strategy at all, or may use a strategy only as an example of the application of the pricing methodology. Needed **R** packages will depend on the exact model to be replicated.

It is also possible to use different methods in the replication than the source used. Sometimes, this will be in addition to a rote replication; in other cases it may be used as a replacement. For example, many "trading strategy"" papers will describe a technique whereby they generate signals based on some indicator, but then "go long" or "go short" by constructing a portfolio containing the desired instruments, ignoring transaction costs, and often ignoring timing (and introducing look-ahead bias). Real trading strategies don't work this way: transaction costs exist, execution is not instant, etc. The analyst may choose to replicate the paper using a complete indicator, signal, and trading rules model which will naturally make the replication more realistic (and thus potentially more useful) rather than blindly replicating the method used in the source material.

Extending the Analysis

Once the analyst has replicated the key techniques and findings from the paper, there is often great value in extending the analysis. In this phase, the analyst is working to tie it all together, validate that the conclusions of the paper are valid, and lay the groundwork for future work.

summary statistics

Every piece of backtesting software (including *quantstrat* and *PortfolioAnalytics*) has summary statistics that will be shown time and time again. It makes sense to present these summary statistics in your replication. Some of them will almost certainly have been presented by the authors of the original paper, so this will further confirm the results of your replication. Others will not have been reported, and will provide an interesting point of comparison, which may warrant some commentary in the replication report.

more data

One of the most obvious places to extend and validate the analysis in the source paper is by adding more data. Three categories suggest themselves:

1. **more recent data, same instruments** : The extension of the replication analysis to more recent data should **always** be part of a replication report. It provides a very clear and simple test of out of sample deterioration, overfitting, or selection bias.
2. **similar instruments** : once the code is done to replicate the paper's techniques and methods, extension to similar instruments is very straightforward. This type of analysis should help the analyst understand selection biases, and begin to draw some conclusions about the general applicability of the analysis.
3. **different asset classes** : extending the analysis to other asset classes is the furthest from pure replication. This type of analysis will often be saved for later work on the ideas contained in the source paper.

going beyond simplifying assumptions

Another necessary place to look for extensions to the replication is in that assumptions used while doing the analysis. Almost all published papers use simplifying assumptions. Many of these assumptions exist to fit in well with other similar literature. Others are made to make the analysis easier on the authors of the paper. In many cases, these simplifying assumptions will make a technique unsuitable for use with real data or real portfolios.

Some examples of common simplifying assumptions and ways to extend or rectify them appear below:

1. **Gaussian assumption** : Probably the worst offender is the use of a Gaussian distribution to model volatility, or errors, or noise, or to sample from. Many authors acknowledge this in their text, and then use it anyway. What would be a better choice for your modeling, and why?

2. **sample moments** : Papers use sample moments because it is easy, under the cover of saying that sample moments introduce less model risk. Many other and better methods exist for estimating moments. Note that with volatility, this may be an extension of the **Gaussian assumption**, above. In other cases, it introduces additional ancillary model choices (and subsequent model testing).

3. **too many/too few parameters**:
   Choice of parameters is a key area where you can tune an analysis. Many papers use a huge number of parameters; others choose a minimal or parsimonious model with few parameters. Either of these choices which made it easier to publish the paper in the first place are unlikely to make a usable model on a real portfolio. Kuhn and Johnson (2013) provides many guidelines on parameter or feature choice, and when and how to increase or decrease the number of parameters under consideration.

similar techniques

It is quite likely that the Literature Review uncovered similar techniques to those used in the source paper. When such similar techniques are readily available in **R**, the analyst should strive to apply them to the data used for the replication. Typically, it is not worth spending a lot of time on this.
Exceptions usually include when a paper claims to have improved a technique, but does an incomplete job of reporting results for the original (theoretically deficient) technique. Another case where you should spend a more time on similar techniques is where one of the Literature Review papers extends, clarifies, or refutes claims made in the paper under replication.

probability of overfitting

Peterson (2015) and Kuhn and Johnson (2013) discuss multiple techniques for detecting biases and overfitting. Most replication reports should contain results of appropriate tests. Specific categories to pay attention to in most cases include selection biases, look-ahead bias, and out of sample deterioration.

Computing Appendix

This essay focuses on replicating published research using *R* (R Core Team 2014). We recommend building a replication report template in *Rmarkdown/knitr* (Xie 2014) (or *Sweave* (Leisch 2002) if you are already proficient with it) which you will expand as you work through the replication. The main advantage of this approach is that you have a living, compilable document which will track the progress of your research. You can edit and fill it in during the course of the replication, compiling the document as you proceed. We additionally recommend the use of a reference manager such as *jabref*. Similar reproducible research processes exist in other languages or environments (e.g. Beaker, mendeley, STATA do-files, iPython notebooks in *python*, etc.) but we have assumed that the reader is using *R* and associated tools.

One good way to structure all this work in **R** is to create a package for the replication project. A further advantage of the package structure is that it can be stored in a version control system such as *bitbucket* .

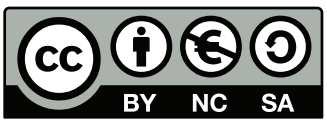
File/Directory	description
DESCRIPTION	defines a package name, and any `Depends` directives
vignettes/	directory to hold Rmarkdown `.Rmd` and `.bib` files
data/	directory will hold the data after it has been downloaded
R/	will hold any `.R` files containing functions used by the analysis
demo/	holds script files called infrequently for things like downloading data
inst/doc/	holds a copy of the source paper, and possibly related papers
man/	documentation auto generated by *roxygen*
NAMESPACE	auto generated by *roxygen*

Reproducibility of research has gained attention in recent years, with prominent papers in journals such as *Science*(Peng 2011), *Nature*(Barnes 2010, Ince, Hatton, and Graham-Cumming (2012)) , and *PLOSOne*(Ioannidis 2005, Moonesinghe, Khoury, and Janssens (2007)). Finance and Economics have entries in Vlaeminck (2013) or the *Economist*("Unreliable Research: Trouble at the Lab" 2013), among others.

Following good citation and computing practices even for routine research is always good policy, and will be particularly useful if work is submitted for publication, reviewed by peers, management, or even "only" your future self.

Acknowledgements

I would like to thank my team for thoughtful comments and questions, and David Matteson at Cornell University for his insightful comments on an early draft of this paper. All remaining errors or omissions should be attributed to the author. All views expressed in this paper are to be viewed as those of Brian Peterson, and do not necessarily reflect the opinions or policies of DV Trading.



References

Barnes, Nick. 2010. "Publish Your Computer Code: It Is Good Enough." *Nature* 467 (7317). Nature Publishing Group: 753–53. http://www.nature.com/news/2010/101013/full/467753a.html.

Fox, John, and Sanford Weisberg. 2011. "Multivariate Linear Models in R." An Appendix to An R Companion to Applied Regression, Sage, Thousand Oaks, CA,

Ghalanos, Alexios, and Bernhard Pfaff. 2014. *Parma: Portfolio Allocation and Risk Management Applications.*

Ince, Darrel C, Leslie Hatton, and John Graham-Cumming. 2012. "The Case for Open Computer Programs." *Nature* 482 (7386). Nature Publishing Group: 485–88. http://www.nature.com/nature/journal/v482/n7386/pdf/nature10836.pdf.

Ioannidis, John PA. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8). Public Library of Science: e124. http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124#s6.

Kuhn, Max, and Kjell Johnson. 2013. *Applied Predictive Modeling.* Springer. http://appliedpredictivemodeling.com/.

Leisch, Friedrich. 2002. "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis." In *Compstat 2002 — Proceedings in Computational Statistics*, edited by Wolfgang Härdle and Bernd Rönz, 575–80. Physica Verlag, Heidelberg. http://www.stat.uni-muenchen.de/~leisch/Sweave.

Levy, Yair, and Timothy J Ellis. 2006. "A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research." *Informing Science: International Journal of an Emerging Transdiscipline* 9 (1). Informing Science Institute: 181–212. http://inform.nu/Articles/Vol9/V9p181-212Levy99.pdf.

Moonesinghe, Ramal, Muin J Khoury, and A Cecile JW Janssens. 2007. "Most Published Research Findings Are False, but a Little Replication Goes a Long Way." *PLoS Medicine* 4 (2). Public Library of Science: e28. http://journals.plos.org/plosmedicine/article?id=info:doi/10.1371/journal.pmed.0040028.

Peng, Roger D. 2011. "Reproducible Research in Computational Science." *Science (New York, Ny)* 334 (6060). NIH Public Access: 1226. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3383002/.

Peterson, Brian G. 2015. *Developing & Backtesting Systematic Trading Strategies.* DV Trading. http://goo.gl/na4u5d.

Peterson, Brian G., Peter Carl, Ross Bennett, and Kris Boudt. 2015. *PortfolioAnalytics: Portfolio Analysis, Including Numerical Methods for Optimization of Portfolios: R Package Version 0.9.3581.* http://r-forge.r-project.org/projects/returnanalytics/.

Peterson, Brian G., Joshua Ulrich, Jan Humme, and Peter Carl. 2015. *Quantstrat: Quantitative Strategy Model Framework: R Package Version 0.9.1667.* http://r-forge.r-project.org/projects/blotter/.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

"Unreliable Research: Trouble at the Lab." 2013. *Economist*, Oct 19. http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

Vlaeminck, Sven. 2013. "Research Data Management in Economic Journals." *American Economic Review, Open Economics.* http://openeconomics.net/resources/data-policies-of-economic-journals/.

Xie, Yihui. 2014. "R Markdown — Dynamic Documents for R." http://rmarkdown.rstudio.com/.