Mathematical Details of Functions in dplR

Mikko Korpela

Processed in R version 3.1.2 Patched (2014-11-03 r66928) on November 10, 2014

1 Introduction

This document presents mathematical details about the Dendrochronology Program Library in R (dplR) [3, 4] which is an add-on package for R [10]. Section 2 deals with the spline smoothing function `ffcsaps` whereas section 3 covers the computation of Gini coefficients in `gini.coef`.

The original implementations of the functions covered here were not written by the author of this document. Therefore the functions were analyzed with a reverse engineering approach. At the time of writing, dplR was at version 1.6.0. Although any changes affecting the mathematical details of the functions are unlikely, the reader is advised to check that the document file originated from a current version of dplR.

2 Spline Smoothing Parameters in ffcsaps

The `ffcsaps` function fits a cubic smoothing spline to a given data vector. In the manual (Rd file) of the function [2], it is stated that the frequency response of the spline is `f` at a wavelength (period) of `nyrs` years[1], where these two are parameters of the function. We aim to clarify how they relate to the single smoothing parameter of the spline and what that parameter stands for. The smoothing parameter is denoted by `p` in the source code of the function.

The manual of the `ffcsaps` function cites [5]. On page 111, they give the following frequency (amplitude) response function for the spline:

$$u(f) = 1 - \frac{1}{1 + \frac{p(\cos(2\pi f)+2)}{6(\cos(2\pi f)-1)^2}} \quad , \tag{1}$$

where f is frequency and p is stated to be the Lagrange multiplier of the spline, the single parameter that determines the frequency response. However, the exact definition of the optimization problem is absent. Neither is it given in [6], the reference used by [5]. I did not find a copy of [9] when trying to follow the chain of references further.

Note that the relationship between frequency and period using mixed notation of `ffcsaps` and (1) is $f = 1/\texttt{nyrs}$. Setting parameters `f` and `nyrs` in `ffcsaps` is equivalent to the following directive: set the smoothing parameter

[1] assuming that the sampling rate is once per year

to a value that fulfills $u(1/\texttt{nyrs}) = \texttt{f}$. By making the variable substitutions and rearranging (1) we get the following equation for p:

$$p = \frac{6\texttt{f}(\cos(2\pi/\texttt{nyrs}) - 1)^2}{(1 - \texttt{f})(\cos(2\pi/\texttt{nyrs}) + 2)} \tag{2}$$

or for the inverse of p:

$$\frac{1}{p} = \frac{(1 - \texttt{f})(\cos(2\pi/\texttt{nyrs}) + 2)}{6\texttt{f}(\cos(2\pi/\texttt{nyrs}) - 1)^2} \quad . \tag{3}$$

The source code of `ffcsaps` contains code lines that correspond to the equation

$$\texttt{p.inv} = \frac{1}{\texttt{p}} = \frac{(1 - \texttt{f})(\cos(2\pi/\texttt{nyrs}) + 2)}{12\texttt{f}(\cos(2\pi/\texttt{nyrs}) - 1)^2} + 1 \quad , \tag{4}$$

where `p` and its inverse `p.inv` are variables used in the code. We find that (3) and (4) are connected by

$$\frac{1}{p} = 2\left(\frac{1}{\texttt{p}} - 1\right) \tag{5}$$

or equivalently

$$\frac{\texttt{p}}{1 - \texttt{p}} = 2p \quad . \tag{6}$$

Figure 1 shows the results of a test where the frequency response of `ffcsaps` (blue circles) is compared to the theoretical result (green line) obtained using (1) and (2). We see that theory meets practice very well, particularly for low frequencies. It must be noted that the theoretical result does not take into account the effect of having a series of finite length. The orange crosses show what happens if one pretends that p and `p` are the same quantity, forgetting (5) and (6).

MATLAB (version 8.3.0.532 (R2014a)) contains a function called `csaps` in the "Curve Fitting Toolbox". The name bears a resemblance to `ffcsaps`. The smoothing parameter of `csaps` is called `p` which makes it a namesake of the internal variable of `ffcsaps` derived from the parameters `nyrs` and `f`. To compare the results of the two functions, I modified `ffcsaps` slightly so that it can take `p` as an argument and omit `nyrs` and `f`. I created a segment of a noisy sine wave and smoothed it with both functions using all values of `p` in the set $\{0, 0.01, 0.02, \ldots, 1\}$, covering the whole range of useful values [8][2]. Using the R function `all.equal` to compare each pair of smoothed series, I found that the results of `csaps` and `ffcsaps` always matched. The result was reproduced when this document was compiled. Figure 2 shows the input series and four smoothed series of the experiment.

According to the documentation of `csaps` [8], it fits the spline by minimizing the following sum, presented here in a simplified form:

$$\texttt{p} \times \text{Error} + (1 - \texttt{p}) \times \text{Roughness} \quad . \tag{7}$$

[2]The Mathworks web site openly provides access to the documentation of the latest Matlab version. Older documents are available after a login to the website or by running `doc csaps` in the command prompt of a particular Matlab version. The author has checked that the documentation of `csaps` agrees between versions 2012b, 2013a, 2013b, 2014a and 2014b.

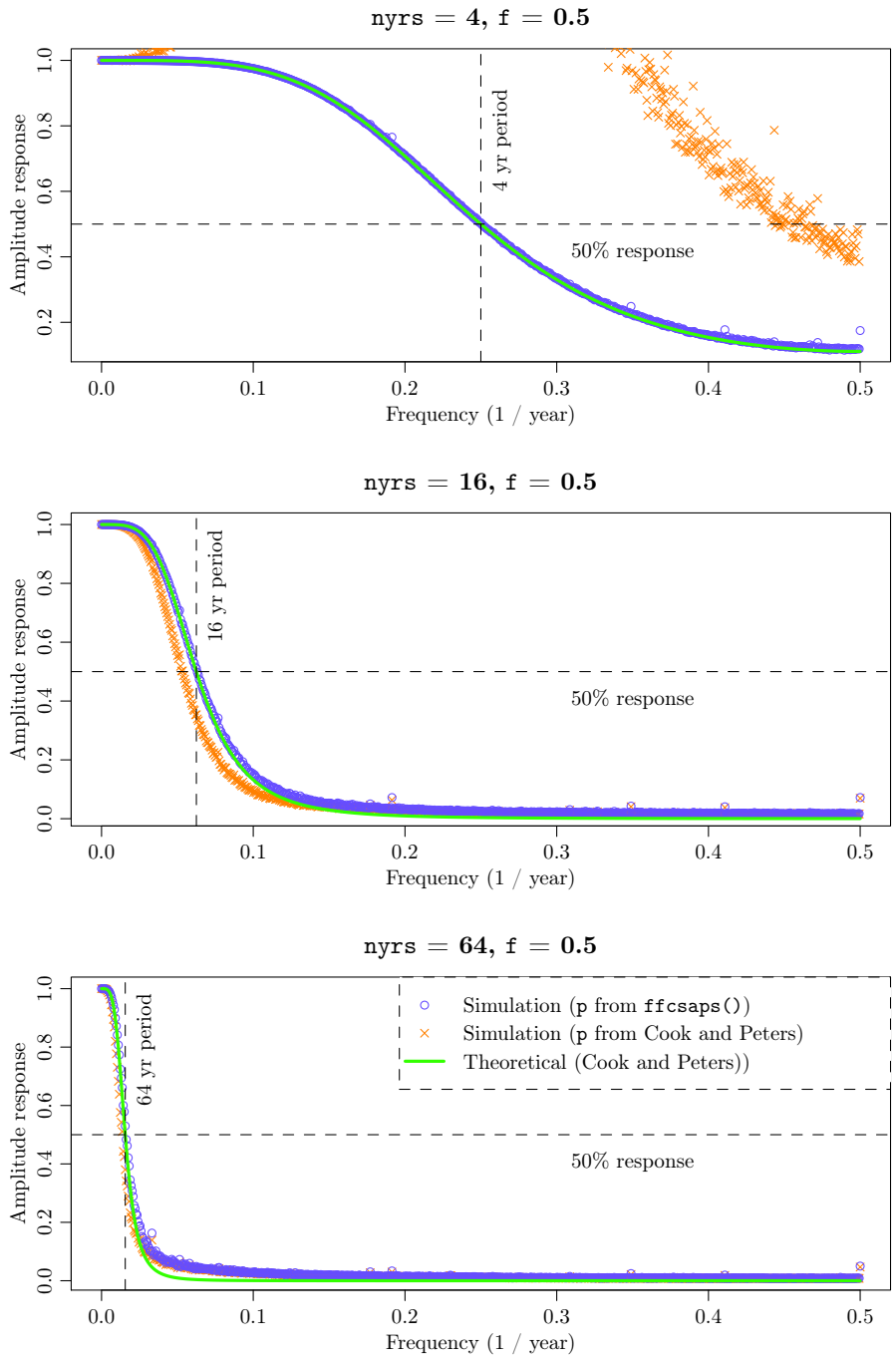

Figure 1: Theoretical frequency response of spline filter vs response with i.i.d. normal series of 1536 samples (mean of 500 repeats) using `ffcsaps`. The legend on the bottom panel applies to all panels. The blue circles were obtained by using (4) for computing (inverse) `p` in `ffcsaps`. The orange crosses show the results when (3) is used instead.

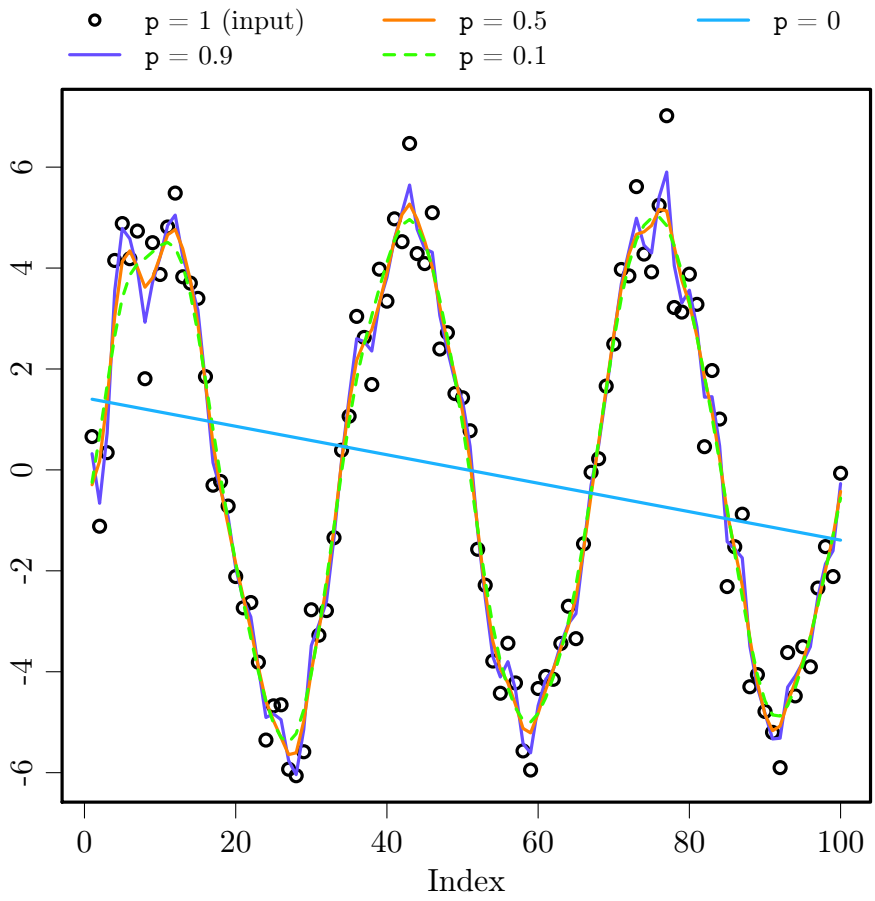

Figure 2: Spline with different values of smoothing parameter p fitted to a noisy sine wave

Having demonstrated that the results of `csaps` and `ffcsaps` match when using any chosen value of `p`, we can safely say that (7) is also the function minimized by `ffcsaps`, with the same definitions of Error and Roughness, details of which are omitted here. A more direct analysis would require one to completely reverse engineer the scarcely documented source code of `ffcsaps`. Following from (6) and (7), the splines described in [5] seem to be the result of minimizing

$$2p \times \text{Error} + \text{Roughness} \quad . \tag{8}$$

3 Formulation of Gini Coefficient in gini.coef

The `gini.coef` function computes the Gini coefficient (Gini index) of a given data vector. The manual (Rd file) of the function has a reference to [1] which uses the following formula for the Gini coefficient (G):

$$G = \frac{1}{2n \sum_{i=1}^{n} x_i} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| \quad . \tag{9}$$

In (9), the Gini coefficient is defined in terms of pairwise differences between all pairs of observations (x_i, $i \in 1, \ldots, n$). More specifically, the Gini coefficient is one half of the relative mean difference, which is defined as the mean of the absolute pairwise distances divided by the mean of the observations.

The C source code of the `gini.coef` function uses the following formula for the Gini index:

$$G = \left(X_n(n-1) - 2 \sum_{i=1}^{n-1} X_i \right) / (X_n n) \quad , \tag{10}$$

where n is the number of observations and X_i is the i:th cumulative sum

$$X_i = \sum_{j=1}^{i} x_j \tag{11}$$

of sorted observations x_j:

$$\forall i : i < j \Rightarrow x_i \leq x_j \quad . \tag{12}$$

(10) can be reformulated as

$$G = 1 - \frac{1}{n} - \frac{2}{X_n n} \sum_{i=1}^{n-1} X_i \tag{13}$$

or as

$$G = \left(\frac{1}{2} - \left(\frac{1}{2n} + \frac{1}{X_n n} \sum_{i=1}^{n-1} X_i \right) \right) / \frac{1}{2} \quad . \tag{14}$$

When we assign

$$A + B = \frac{1}{2} \tag{15}$$

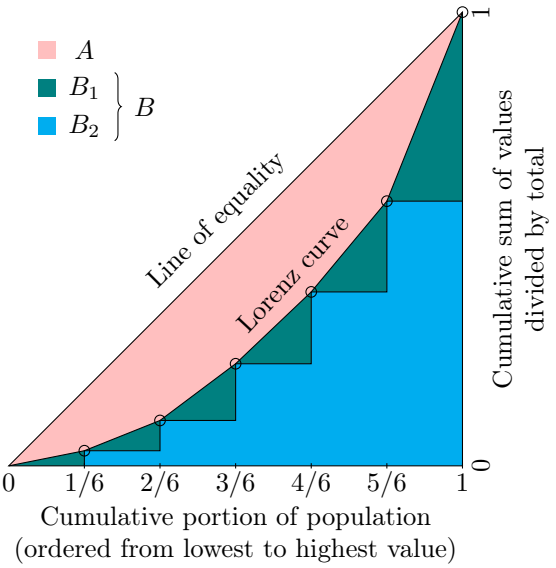

Figure 3: Graphical representation of the Gini coefficient based on areas defined by the Lorenz curve ($n = 6$). See equations (15), (16), (17) and (18).

and

$$B = B_1 + B_2 = \frac{1}{2n} + \frac{1}{X_n n} \sum_{i=1}^{n-1} X_i \quad , \tag{16}$$

(14) becomes

$$G = A/(A + B) \tag{17}$$

or equivalently

$$G = 1 - 2B \quad . \tag{18}$$

Figure 3 is a graphical representation of the Gini coefficient using an example data set of the following six observed values: $\{0.2, 0.4, 0.75, 0.95, 1.2, 2.5\}$. It shows the definition of the Gini coefficient as the ratio of the area above the Lorenz curve [7] to the total area of the triangle [11]. The Lorenz curve is defined by the cumulative distribution function of the empirical probability distribution of the observations. The sides of the triangle corresponding to the axes are normalized to length 1.

Comparing Figure 3 to (16), $B_2 = \sum_{i=1}^{n-1} X_i/(X_n n)$ is obviously the sum of the areas of the cyan bars. Summing the areas of the teal triangles, we get

$$\sum_{i=1}^{n} \left(\frac{1}{2} \frac{1}{n} \frac{x_i}{X_n} \right) = \frac{1}{2nX_n} \sum_{i=1}^{n} x_i = \frac{1}{2n} = B_1 \quad . \tag{19}$$

Note that B_1 only depends on the number of observations, not on their values. From (16) and (18) we find that the value of the Gini coefficient at maximum inequality (winner takes all) is $G_{\max}(n) = 1 - 1/n$. When all observed values are equal, the Lorenz curve matches the line of equality (Figure 3), and the Gini coefficient is $G_{\min} = 0$. We have assumed that all values x_i are non-negative.

The equivalence of different definitions of the Gini coefficient is reviewed in [11]. One of the results shown in the paper is that the geometric definition (17) used by the `gini.coef` function is equivalent to the definition based on the relative mean difference (9). This can be experimentally verified by comparing the results of the following R function to those of `gini.coef`.

```
## Gini index is one half of relative mean difference.
## x should not have NA values.
gini.rmd <- function(x) {
    mean(abs(outer(x, x, "-"))) / mean(x) * 0.5
}
```

References

[1] F. Biondi and F. Qeadan. Inequality in paleorecords. *Ecology*, 89(4):1056–1067, 2008.

[2] A. Bunn, M. Korpela, F. Biondi, F. Campelo, P. Mérian, M. Mudelsee, F. Qeadan, M. Schulz, and C. Zang. *dplR: Dendrochronology Program Library in R*, 2014. R package version 1.6.0.

[3] A. G. Bunn. A dendrochronology program library in R (dplR). *Dendrochronologia*, 26(2):115–124, 2008.

[4] A. G. Bunn. Statistical and visual crossdating in R using the dplR library. *Dendrochronologia*, 28(4):251–258, 2010.

[5] E. R. Cook and L. A. Kairiukstis. *Methods of dendrochronology: applications in the environmental sciences*. Springer, 1990.

[6] E. R. Cook and K. Peters. The smoothing spline: a new approach to standardizing forest interior tree-ring width series for dendroclimatic studies. *Tree-ring bulletin*, 41:45–53, 1981.

[7] M. O. Lorenz. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association*, 9(70):209–219, 1905.

[8] MathWorks. csaps: Cubic smoothing spline. `http://www.mathworks.com/help/curvefit/csaps.html`. Accessed: 2014-11-10. Documentation of the latest version at that time (Matlab R2014b).

[9] K. Peters and E. R. Cook. The cubic smoothing spline as a digital filter. Technical report, Lamont-Doherty Geological Observatory of Columbia University, Tree-Ring Laboratory, 1981.

[10] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[11] K. Xu. How has the literature on Gini's index evolved in the past 80 years. *China Economic Quarterly*, 2:757–778, 2003.