Planar projection of the principal components of fractal Brownian functions

Pavel V Moskalev, Lyudmila I Fedulova and Irina V Gridneva
Voronezh State Agricultural University, 1 Michurin street, Voronezh, 394087, Russia
E-mail: moskaleff@mail.ru

Abstract. The error in the planar projection of the principal components of a multidimensional data array strongly depends on the number and dimension of the column vectors of this array and the correlation between the vectors. In this work, we investigate the dependence of the error on the planar projection of the principal components of fractal Brownian functions on the number of statistically independent realizations in the multidimensional data array. As a result, we show that, under certain assumptions, the number of statistically independent realizations of fractal Brownian functions coincides with the effective dimension of the projection of the principal components.

1. Introduction
One of the tasks of statistical data analysis is to estimate the smallest number of variables for modeling the process under study with an error not exceeding some allowable value. Such problems often arise in mathematical modeling of physical, chemical, and biological phenomena using the theory of similarity and dimensions. In this case, we are talking about the smallest number of dimensionless combinations (similarity criteria) made up of dimensional parameters that determine the structure of the mathematical model [1, 2]. The formal solution to this problem comes down to determining the rank of the dimension matrix and allows us to write in dimensionless form any equation of a mathematical model that has physical meaning (criterion equation). For the practical application of this approach, the choice of universal algorithms is very important: (a) for constructing similarity criteria; (b) for estimating the error in the dimensionality reduction of mathematical models.

2. Materials and methods
Effective statistical analysis and visualization of multidimensional data arrays usually requires their planar projection. Let us try to estimate the error of dimensionality reduction for an array formed by a series of columns — realizations of fractal Brownian functions. The correlation between the columns of different series is less significant, and the correlation between the columns within each series is more significant.

Such a structure of sample data is typical for chemometric problems in the analysis of multicomponent mixtures. We will associate highly correlated implementations in each series with different concentrations of each analyte, and low correlated implementations from different series with different analytes [3, 4].
All algorithms used in chemometrics for processing multidimensional sensory data can be divided into qualitative and quantitative analysis algorithms. Typical qualitative analysis algorithms classify test sample data by comparing it with previously classified training sample data. Typical quantitative analysis algorithms estimate unknown function values at test points using known values of this function at training points. Note that in both cases the task of these algorithms is to estimate the values of a function. But in quantitative analysis, the set of values of this function is continuous, and in qualitative analysis, the set is discrete [5].

2.1. Dimension reduction

To reduce the dimension of correlated data, we use the hypothesis of the existence of a small number of latent variables $z_j$ (principal components) that can be represented as linear orthonormal combinations of observed variables $x_i$ [6]:

$$ z_j = \sum_{i=1}^{p} l_{ij} x_i, \quad \text{where} \quad \sum_{i=1}^{p} l_{ij}^2 = 1, \quad \text{and} \quad \sum_{i=1}^{p} l_{ki} l_{ji} = 0,$$

for all $k \neq j = 1, 2, \ldots, p$.

Then the projection of the matrix of observed variables $X$ into the space of principal components $Z$ will be determined by the product $Z = XL$, where $L = (l_{ij})$ is the transformation matrix. To find the elements of the transformation matrix $l_{ij}$, we use the condition of the monotonicity of the dispersion of the main components $D(z_1) \geq D(z_2) \geq \ldots \geq D(z_p)$, which leads to a series of conditional optimization problems $D(z_j) = (l_{j})^T l_j \rightarrow \max$, the solution of which is determined by the ranked spectrum $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ of the covariance matrix of the observed variables $\Sigma = XLX^T$, where $l_j$ is the eigenvector of the matrix $\Sigma$ corresponding to the eigenvalue $\lambda_j$.

It is known that the principal component method (PCA) gives the best results with a significant correlation of the observed variable $X = (x_i)$, when the determinant of the covariance matrix $\det(\Sigma) = \lambda_1 \lambda_2 \ldots \lambda_p$ is close to zero. When the eigenvalues $\lambda_j$ are ranked in magnitude, this is possible due to the convergence to zero of both the spectrum of eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \rightarrow 0$ and the variances of the corresponding principal components $D(z_1) \geq D(z_2) \geq \ldots \geq D(z_p) \rightarrow 0$ [7]. Then we have the opportunity to reproduce a significant part of the variance of the observed variables $X$ with effective reduction of the data dimension using the first principal components $(z_1, z_2, \ldots, z_m)$, where $m \ll p$. In fact, using PCA, we can project the observed variables $X \in \mathbb{R}^p$ into a subspace of lower dimension $Z \in \mathbb{R}^m$ (for example, projection onto a plane with $m = 2$) while maintaining the maximum variance of the analyzed data.

To project an array of observable variables $X$ into the space of principal components $Z$, the authors of this article used the “prcomp()” function, which implements PCA algorithms for the free statistical analysis system $R$ [8].

2.2. Fractal Brownian functions

We use the model of a fractal Brownian function, obtained using the Voss sequential random addition algorithm [9], in the one-dimensional case determined by the iterative function:

$$ F_k = F_{k-1} + \Delta F_k, $$

where $\Delta F_k \sim N(0, \sigma_k)$ — centered normally distributed increment at the $k$-th iteration. If we double the number of intervals at each iteration, then the current number of realization points will be determined by the exponential function $N_k = 2^k + 1$. For the convergence of the iterative process, the standard deviation of the normal increments is usually assumed to be an
exponentially decreasing function [10]: \( \sigma_k = 2^{-H} \sigma_{k-1} \), where \( H \in [0, 1] \) is the Hurst exponent, and the initial value \( \sigma_0 \) is normalized by the sum of the series \( \sum_{k=0}^{\infty} \sigma_k = 1 \Rightarrow \sigma_0 = 1 - 2^{-H} \).

The described algorithm allows us to simulate statistically self-affine realizations of fractal Brownian functions that differ from classical Brownian functions by nonzero correlation between normally distributed increments \( \Delta F_1 = F(x) - F(0) \) and \( \Delta F_2 = F(x + \Delta x) - F(x) \). It is easy to verify that

\[
D(\Delta F_1 \Delta F_2) = \frac{\sigma^2}{2} |(x + \Delta x)^{2H} - x^{2H} - (\Delta x)^{2H}| \]

will be equal to zero zero only for \( H = \frac{1}{2} \).

Under condition \( H > \frac{1}{2} \), the probability that the increments \( \Delta F_1 \) and \( \Delta F_2 \) will have the same signs will be greater than \( \frac{1}{2} \) and the function \( F(x) \) will exhibit the persistence property, that is, the preservation of the current trend is more likely than its violation. Then, under condition \( H < \frac{1}{2} \), the probability that the increments \( \Delta F_1 \) and \( \Delta F_2 \) will have the same signs will become less than \( \frac{1}{2} \) and the function \( F(x) \) will exhibit the antipersistency property, that is, a violation of the current trend is more likely than its preservation [11].

Figure 1 shows one-dimensional correlated realizations of fractal Brownian functions \( F_k(x_k) \) after eight iterations with decreasing values of the Hurst exponent: \( H = \left( \frac{5}{9} \right)^2, \left( \frac{5}{9} \right)^4, \ldots, \left( \frac{5}{9} \right)^7 \). Light gray color corresponds to realizations of fractal Brownian functions \( F_k(x_k) \) with the Hurst exponent above half \( H = \left( \frac{5}{9} \right)^2 > \frac{1}{2} \), and black color corresponds to the same realization with a significantly lower value of the Hurst exponent \( H = \left( \frac{5}{9} \right)^7 < \frac{1}{2} \).

![Figure 1. Correlated realizations of fractal Brownian functions \( F_k(x_k) \) with various values of the Hurst exponent \( H = \left( \frac{5}{9} \right)^2, \ldots, \left( \frac{5}{9} \right)^7 \), which are shown in colors from light gray to black.](image-url)

An a priori estimate of the similarity dimension of realizations of fractal Brownian functions follows from the scaling relation

\[
N(\Delta x) \propto \Delta x^{H-2},
\]

which implies to \( d_S = 2 - H \). The Hurst exponent is \( H \in [0, 1] \), from which the similarity dimension converges to \( d_S \to 2- \) for \( H \to 0^+ \), or to \( d_S \to 1+ \) for \( H \to 1^- \). With an increase in the Hurst exponent \( H \to 1^- \), increments of fractal Brownian functions \( \Delta F_k \) decrease (light gray lines in figure 1) and all realizations converge to a straight line \( F = 0 \), which has a similarity dimension \( d_S = 1 \) and is shown by a dashed line in figure 1.

Correlation within each series of realizations of fractal Brownian functions \( F_k(x_k) \) was achieved by fixing a pseudorandom sequence of normally distributed increments \( \Delta F_k \). Table 1 shows that the correlation between realizations \( F_k(x_k) \) from different series increases as the
Hurst exponent $H$ decreases. To quantify this relationship, the following table shows sample correlation coefficients between realizations with the same Hurst exponents.

**Table 1.** Sample correlation coefficients between paired realizations of fractal Brownian functions $F_k(x_k)$ from different series with the same Hurst exponent $H = \frac{5}{9}, (\frac{2}{9})^2, \ldots, (\frac{2}{9})^7$.

| $H$   | $r_{12}$ | $r_{13}$ | $r_{23}$ |
|-------|----------|----------|----------|
| 0.5556 | 0.7658   | 0.6825   | 0.7688   |
| 0.3086 | 0.6375   | 0.5681   | 0.6821   |
| 0.1715 | 0.5114   | 0.4350   | 0.5787   |
| 0.0953 | 0.4227   | 0.3436   | 0.6821   |
| 0.0529 | 0.3698   | 0.2913   | 0.4586   |
| 0.0294 | 0.3400   | 0.2625   | 0.4333   |
| 0.0163 | 0.3234   | 0.2467   | 0.4193   |

Analyzing table 1, we can say that when the Hurst exponent decreases, the correlation between realizations from different series also decreases from high on the Cheddock scale at $H = \frac{5}{9}$ to low at $H = (\frac{2}{9})^7$.

To generate the realizations of fractal Brownian functions $F_k(x_k)$ shown in figure 1, we used the “voss1d()” function from the “Voss” library [12], released in 2012 by the author of this paper for the free statistical analysis system R under the GNU GPL-3 license.

```r
voss1d <- function (g = 7, H = 0.5, r = 0.5, center = TRUE) {
  N <- r^-seq(0, g) + 1
  s <- H * log(1/r) * r^H * seq(0, g)
  voss <- list(x = c(0, 1), y = c(0, 0))
  for (i in seq(2, g + 1)) {
    voss <- approx(voss, N = N[i])
    voss$y <- rnorm(N = N[i], mean = voss$y, sd = s[i])
  }
  if (center) voss$y <- voss$y - mean(voss$y)
  return(voss)
}
```

Arguments “g”, “H” and “r” correspond to the number of iterations, the Hurst exponent and the grid refinement coefficient. Vectors “N” and “s” in this listing contain the number of grid points $N_k$ and the standard deviation $\sigma_k$ of pseudorandom increments $\Delta F_k$ at the k-th iteration, and the returned list “voss” contains abscissas “x” and ordinates “y” for the realization of a fractal Brownian function $F_k(x_k)$.

### 3. Results and discussion

To analyze the applicability of the principal component method for reducing the dimension of fractal Brownian functions, we create three arrays containing one, two, and three series of correlated realizations, shown above in figure 1. Projections of these arrays on the plane of the first pair of principal components ($z_1, z_2$) shown below in figure 2. Individual points on these projections correspond to realizations of fractal Brownian functions. Points corresponding to each series of high correlated realizations are connected by a broken line. In parentheses, the percentage of the sample variance corresponding to this component is indicated. It can be seen that an increase in the number of low correlated realizations of fractal Brownian functions...
leads to a decrease in the cumulative sample dispersion of the first pair of principal components: (1) 100 %, (2) 97 %, (3) 86.7 %.

Despite the significantly different sample structure, we can note that an increase in the Hurst exponent $H \to 1-$ contracts the points to the origin $O(0,0)$. Moreover, the average distances between points in each series are reduced when they approach the origin $O$ and vice versa. However, the latter pattern can be violated the more strongly, the smaller the percentage of the total sample variance is displayed by the planar projection $(z_1, z_2)$. Such is the influence of the third and subsequent principal components $z_3, z_4, \ldots$ that are not displayed on the planar projection $(z_1, z_2)$.

The cumulative sample dispersion of the first five principal components is shown more clearly in figure 3. The left, center and right diagram in figure 3 corresponds to the cumulative sample dispersion for one, two and three series of correlated realizations of fractal Brownian functions shown above in figure 1.
4. Conclusion
The fraction of cumulative sample variance displayed by the planar projection of the principal components is important for solving the problems of analyzing heterogeneous sample data. A heterogeneous structure of sample data can be formed in chemometrics when analyzing conventionally one-, two-, and multicomponent mixtures. The above examples show that a quantitative analysis of such samples using planar projections of their principal components is possible for conditionally one- or two-component mixtures. The cumulative fraction of the sample variance displayed by the planar projection of the principal components for conventionally three- and multicomponent mixtures is sufficient only for qualitative analysis algorithms.

5. Acknowledgments
This work was funded by the Russian Foundation for Basic Research (RFBR) according to the research project No. 18-29-24128.

6. References
[1] Bridgman P W 1922 Dimensional Analysis (Yale University Press)
[2] Boucher D F and Alves G E 1960 Chemical Engineering Progress 55 55–64
[3] Shaposhnik A, Moskalev P, Sizae E, Rybatsev S and Vasiliev A 2019 Sensors 19 1135
[4] Chegereva K L, Shaposhnik A V, Moskalev P V and Zyrygin A A 2019 Izvestiya Vysshikh Uchebnykh Zavedenii. Seriya Khimiya i Khimicheskaya Tekhnologiya 62 76–81
[5] Rodionova O E and Pomerantsev A L 2006 Russian Chemical Reviews 75 271–287
[6] Mardia K V, Kent J T and Bibby J M 1979 Multivariate Analysis (Academic Press)
[7] Venables W N and Ripley B D 2002 Modern Applied Statistics with S (Springer-Verlag)
[8] R-core@R-project.org 2019 Pcomp: Principal Components Analysis (URL: https://www.rdocumentation.org/packages/stats/versions/3.6.1/topics/prcomp)
[9] Voss R F 1985 Fundamental Algorithms in Computer Graphics (Springer-Verlag) 805–835
[10] Moskalev P V 2018 Percolation Modeling of Porous Structures (URSS, Moscow)
[11] Feder J 1988 Fractals (Plenum Press)
[12] Moskalev P V 2012 Voss: Generic Voss Algorithm (Random Sequential Additions) (URL: https://cran.r-project.org/package=Voss)