Available online at www.alphanumericjournal.com

alphanumeric journal
The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 4, Issue 2, 2016

Received : April 13, 2016
Accepted : August 11, 2016
Published Online : September 26, 2016

AJ ID: 2016.04.02.ECON.01
DOI : 10.17093/aj.2016.4.2.5000185408

A New Descriptive Statistic for Functional Data: Functional Coefficient Of Variation

İstem Köymen Keser | Econometrics Department, Dokuz Eylul University, Turkey, istem.koymen@deu.edu.tr
İpek Deveci Kocakoç | Econometrics Department, Dokuz Eylul University, Turkey, ipek.deveci@deu.edu.tr
Ali Kemal Şehirlioğlu | Econometrics Department, Dokuz Eylul University, Turkey, kemal.sehirlioglu@deu.edu.tr

ABSTRACT

In this study, we propose a new descriptive statistic, coefficient of variation function, for functional data analysis and present its utilization. We recommend coefficient of variation function, especially when we want to compare the variation of multiple curve groups and when the mean functions are different for each curve group. Besides, obtaining coefficient of variation functions in terms of cubic B-Splines enables the interpretation of the first and second derivative functions of these functions and provides a stronger inference for the original curves. The utilization and effects of the proposed statistic is reported on a well-known data set from the literature. The results show that the proposed statistic reflects the variability of the data properly and this reflection gets clearer than that of the standard deviation function especially as mean functions differ.

Keywords: Coefficient Of Variation Function, Descriptive Statistics, Functional Data Analysis

Fonksiyonel Veri İçin Yeni Bir Tanımlayıcı İstatistik: Fonksiyonel Değişkenlik Katsayısı

ÖZET

Bu çalışmada fonksiyonel veriler için yeni bir tanımlayıcı istatistik olan değişkenlik katsayısı fonksiyonunu önermekteyiz. Özellikle her bir eğri grubunun ortalama fonksiyonları farklı olduğu durumda, çoklu eğri gruplarının değişkenliklerini karşılaştırmada önerilen değişkenlik katsayısı fonksiyonunun kullanılmasını tavsiye ediyoruz. Değişkenlik katsayısı fonksiyonunu elde ederken kübik B-Splaynlar kullanıldığından dolayı, bu fonksiyonların birinci ve ikinci türev fonksiyonlarının da yorumlanabiliyor olması, orijinal eğriler hakkında daha güçlü çıkarımlar yapabilmemizi sağlamaktadır. Önerilen istatistiğin kullanımı ve etkileri literatürdeki iyi bilinen bir veri seti üzerinde raporlanmıştır. Sonuçlar göstermektedir ki önerilen istatistik verinin değişkenliğini düzgün bir şekilde yansıtmaktadır ve bu durum ortalama fonksiyonları farklılaştıkça standart sapma fonksiyonundan daha üstün bir hale gelmektedir

Anahtar Kelimeler: Değişkenlik Katsayısı Fonksiyonu, Tanımlayıcı İstatistik, Fonksiyonel Veri Analizi

1. Introduction

The increase in storage power of computers has led to new types of data which require new tools for analysis. Functional data analysis is one of those new tools which emerged to meet this requirement. In the last 15-20 years, studies on functional data analysis began to spread after the seminal work of Ramsay and Silverman (1997).

A functional datum is not a single observation but rather a set of measurements along a continuum that, taken together, are to be regarded as a single entity, curve or image. Usually the continuum is time, and in this case the data are commonly called "longitudinal." But any continuous domain is possible, and the domain may be multidimensional (Levitin et al., 2007).

Functional data analysis techniques such as functional principle component analysis, functional canonical correlation analysis, and functional regression analysis are utilized for a vast domain spreading from medical data to psychological data. A detailed review of applications of functional data analysis can be found in Ullah and Finch (2013).

Our aim in this paper is to propose a new descriptive statistic, "coefficient of variation" for functional data. Although standard deviation function and mean function themselves are not new, a new concept of coefficient variation (CV) function is necessary to compare the variation between curve groups especially when the mean curves are different between curve groups.

The paper is organized as follows: Section two presents common descriptive statistics of functional data: standard deviation function and mean function. The concept and formulation of proposed coefficient of variation function is introduced. In section three, the well-known Berkeley Growth Curve Data is investigated by both common functional descriptive statistics and functional CV. Then, comments on the benefits and effects of proposed statistic are presented. Finally, section four deals with some conclusions and suggestions.

2. Descriptive Statistics for Functional Data

2.1. Mean and Standard Deviation Functions

Classical descriptive statistics for univariate data can similarly be applied to functional data with minor modifications. The mean function is a simple analogue of classical mean for univariate data. It can be calculated by averaging the functions pointwise across replications.

In functional data, N observation curves are assumed to be observed for t_1, t_2, ..., t_n points. The j^{th} point on i^{th} curve is denoted by $x_i(t_j)$, where i = 1,2, ... , N, j=1,2,...,n.

Briefly, mean curve which corresponds to the mean of each observation point is obtained by Eq1.

$$\bar{x}(t) = \frac{1}{N}\sum_{i=1}^{N} x_i(t_j) \qquad (1)$$

Mean curves can be compared to each other for all time points across curve groups to see the time ranges which an evident decrease or increase of means occurs.

Alternatively, if the functions have a large noise component or other undesirable non-smooth components, it may be preferable to first approximate them by suitable splines and then take the average of the approximations (Ramsay, 1982). In other words, the accuracy of the estimations can be increased by smoothing (Rice&Silverman, 1991).

In functional data analysis, variation is measured by variance and covariance functions. Variation function is defined as

$$\operatorname{Var} x(t) = \frac{1}{N-1}\sum_{i=1}^{N}\left(x_i(t) - \bar{x}(t)\right)^2 \qquad t \in [a,b] \qquad (2)$$

[a,b] is the real interval that the function is defined. Positive square root of this function is named as standard deviation function. As in the mean function, variance function or standard deviation function can be smoothed if required.

These functions are also simple analogues of classical univariate measures and have similar interpretations. Standard deviation function shows the variation between curves across time points. Detailed information on descriptive statistics of functional data can be found in Shang (2015).

2.2. A New Statistic: Coefficient of Variation Function

For univariate data, if the need is to compare the degree of variation from one data series to another, especially if the means and/or scales are drastically different from each other, coefficient of variation is more useful than standard deviation. Since it is the ratio of the standard deviation to the mean, coefficient of variation measures the variation of a series of data independent from its measurement units, so it can be used to compare data with different scales or different means. Ratio scale measurement is more plausible for coefficient of variation since it possesses a meaningful zero value. Therefore, interpretation of mean becomes more rational. Non-negative data can be preferred to provide a non-negative mean.

Functional data, too, needs similar interpretations. If variability of multiple functional data groups needs to be compared, especially when mean functions of these data groups are different, coefficient of variation functions can be used instead of standard deviation functions. However, in literature, such a statistic does not exist so far. In this study, the following coefficient of variation (CV) function is proposed:

$$\text{CV function} = \frac{\text{Standard deviation function}}{mean\ function}$$

In order to find the CV function, firstly, standard deviation coefficients and mean function coefficients are computed. Then, standard deviation coefficients are divided by mean function coefficients and CV function coefficients are obtained. Later, CV functions are formed via basis functions. However, as it is the case for univariate CV, CV function is very sensitive near point zero. Especially when the curves has a mean function very close to zero, CV function gets affected. Hence, usage of proposed CV function is not recommended in that case. Beside other approaches, basis function approach is primarily used when estimating the curves in functional data analysis because of its differentiability. Complete computational procedure for the estimation

of mean, standard deviation, and proposed CV function according to basis function approach are given in Appendix-A. Calculated coefficients for CV functions can be found in Appendix B. Matlab code files of Ramsay (2015) are partly used and modified in order to calculate the proposed statistic. Modified code files can be found on the corresponding author's website (Keser&Deveci Kocakoç, 2015).

3. An Application to Height Data

Berkeley Growth Data (Tuddenham&Snyder, 1954) is analyzed in order to present the benefits of the proposed CV function. Growth data is a well-known data set which is used and investigated in many studies on functional data, such as Ramsay and Silverman (2005), Clarkson et.al (2005), Ramsay et.al. (2009), Sun and Genton (2011), and Zhang (2013). This data contains the heights of 54 girls and 39 boys observed at 31 not equally spaced ages from year 1 to year 18.

Here, we first examine the individual functions, mean functions and derivative functions for girls and boys. Subsequently, difference of mean functions are tested and presented by functional t-test. The effects of this difference on standard deviation function and CV function are investigated.

3.1. Examination of Individual Functions, Mean Functions and Derivative Functions

Individual functions for girls and boys are constructed by cubic B-Splines and given in Figure 1a and 1b. When these figures are examined, it can be seen that individual functions for both girls and boys have a regular increase.

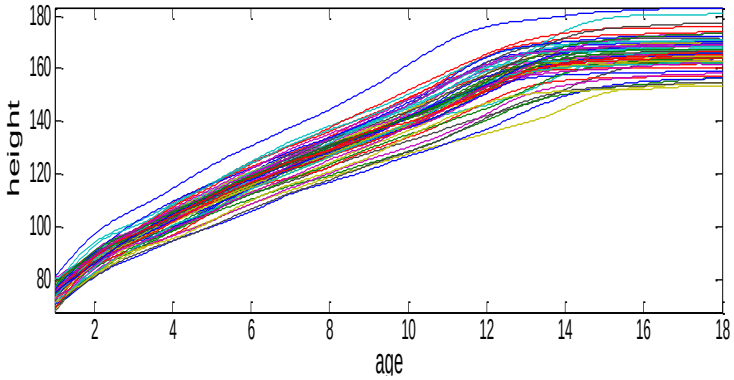

Figure 1a. Individual height functions for 54 girls

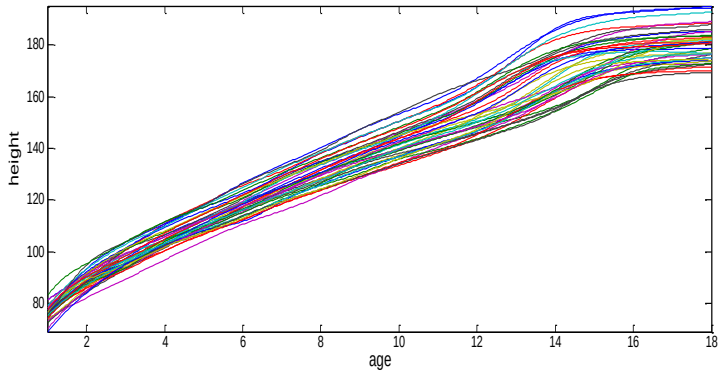

Figure 1b. Individual height functions for 39 boys

Since cubic B-Splines provide us the first and the second derivative functions, we can also examine the growth rate and acceleration. First derivative functions (Figure 2a and 2b) which give the growth rate show the puberty growth spurts for both groups. As expected, first derivative functions have a negative slope just the opposite of individual functions. Growth spurts seems to happen with a jump between ages 10 and 12 for girls and ages 12-15 for boys. Besides, slight individual shifts can be seen for each of the girls and boys.

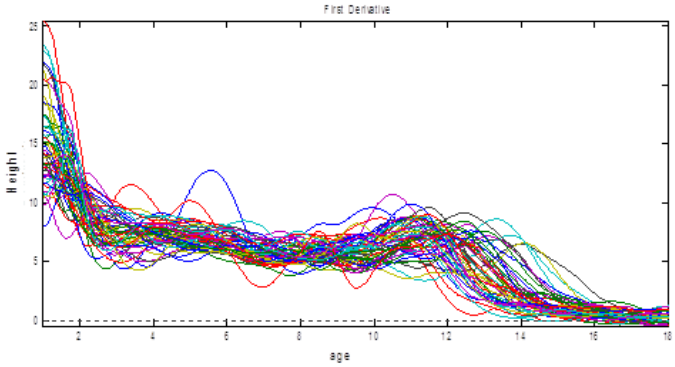

Figure 2a. First derivative functions for girls

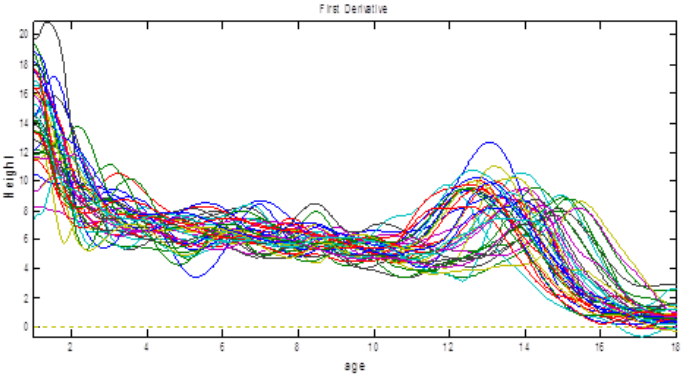

Figure 2b. First derivative functions for boys

Mean functions and their first derivatives given in Figure 3a and 3b may be more useful to see the group differences. When mean functions in Figure 3a are examined, we can see that means of boys and girls follow each other very closely, boys ahead of girls, to the age of 11. After 11, girls get ahead of boys, and the difference in means gets higher after age 13, boys again get ahead of girls.

Although group mean functions gives insights of each gender's growth model, it is the first derivatives of mean functions that give the beginning ages of growth spurts thus enables us to investigate data thoroughly. The first derivatives of mean functions in Figure 3b shows that the growth spurts of boys follows the growth spurts of girls very closely. Especially, growth spurts in ages 10-12 for girls and 12-15 for boys can be very distinctly recognized. The real difference begins around age 12, when boys' growth spurt actually gets started. After age 13, boys mean function clearly starts inclining. All these interpretations are in accordance with reality and with other studies using this data set.

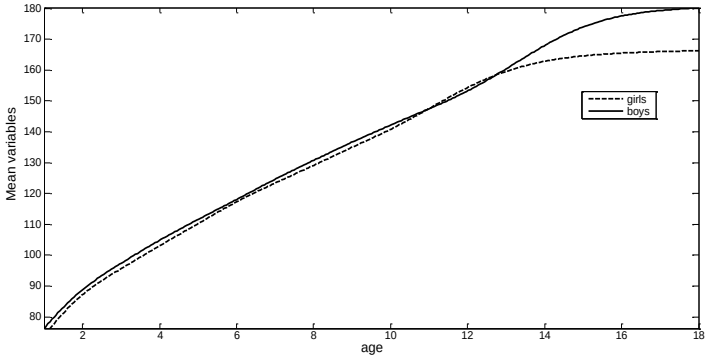

Figure 3a. Mean functions for girls and boys

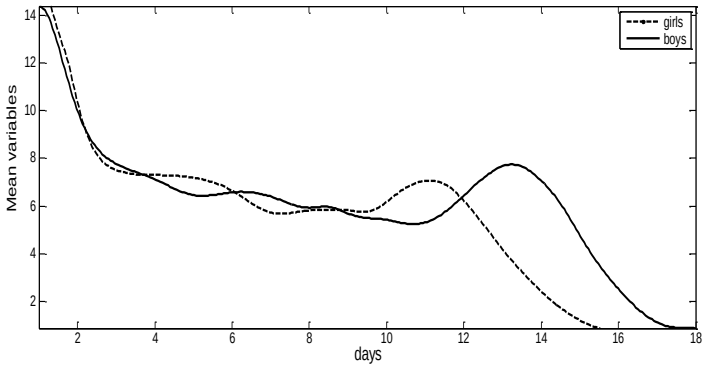

Figure 3b. Derivative mean functions for girls and boys

3.2. Examination of Standard Deviation Functions and CV Functions

In order to see whether the difference in mean functions are statistically significant or not, functional t-test may be utilized. Ramsay and Silverman (2005) suggest using a pointwise test approach based on a permutation method. This method is based on randomly changing curve labels and calculating the test statistic each time. This statistic is referred to as permuted test statistic. This process is repeated ten thousand times in order to obtain a null distribution and it gives a reference in order to evaluate maximum of T(t). At the same time, test statistics are also calculated from original data. Obtained statistics is referred to as observed test statistics. A p-value is obtained by permuted test statistic's ratio that is higher than or equal to observed test statistic. It is assumed that null hypotheses will be rejected for high values of the test statistic (Lee, 2005, Ramsay et all., 2009, Coffey&Hinde, 2011).

This procedure is advantageous as it is distribution independent. On the other hand, it is an exact level α test because of the features of permutation test, so, it gives valid p values (Lee, 2005). For the steps of functional t-tests based on a permutation method, Cox and Lee (2008), Yaree (2011) and Keser (2014) can be referred.

In this study, our observations in both girls and boys are analyzed for the range 1-18 years on the 31 time points. The result of functional t-test is given in Figure 4.

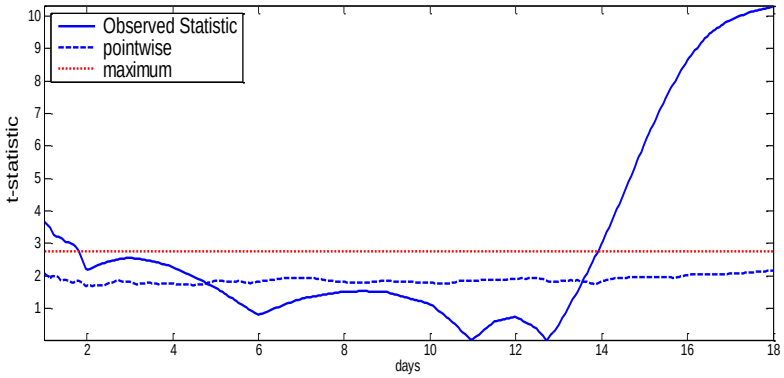

Figure 4. Functional t-test

With pointwise functional t-test, the time when girls and boys started to differ can be revealed. The time points where observed statistic exceeds the pointwise statistic shows difference in means. If observed statistic exceeds the conservative reference line (maximum critical value), then the difference is statistically significant with p=0.05. When pointwise critical value is taken as reference, means of girls and boys tend to differ approximately starting from age 14. But when conservative reference line (maximum critical value) is taken into consideration, main strong differences start at approximately the half way of boys' puberty growth spurt and last until age 18. This situation exists around the period between ages 14-18 and this is completely consistent with the expectations of the study. So, based on this t-test, it can be concluded that gender is effective on the group means of height data.

Figures 5a and 5b show standard deviation and CV functions respectively. Both standard deviation function and CV function clearly show the change of variation for boys and girls during their growth spurts. However, CV function in Figure 5b shows this change more distinctively since it also accounts for the change in the mean. Especially the effect of the group mean after the significant difference can be seen better in CV function. As can be seen from Figure 5b, the variation of girls exceeds the variation of boys after age 16 when we introduce the group mean information into the statistic. In other words, as group mean changes, group variation also changes comparatively. Standard deviation function does not give this information. Hence, CV function provides a better insight on the effect of mean function and the variation changes between groups.

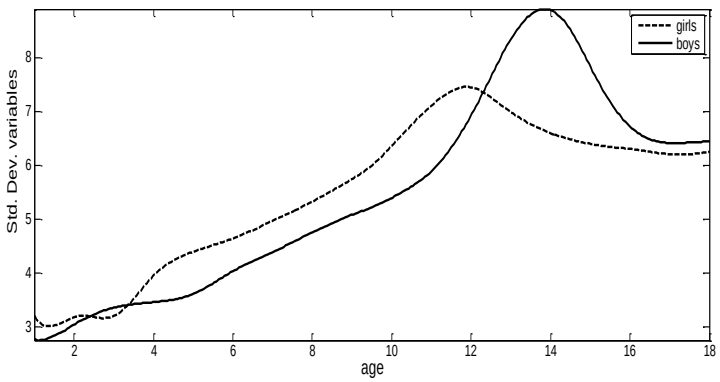

Figure 5a. Standard deviation functions for girls and boys

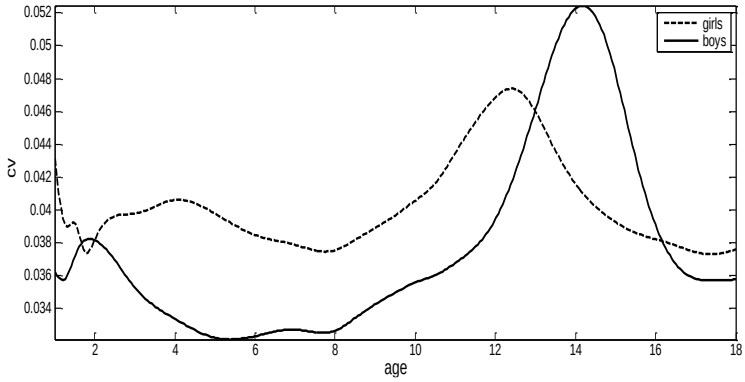

Figure 5b. CV functions for girls and boys

4. Conclusions

In this study, we propose a coefficient of variation function for functional data analysis as a descriptive statistic. Height data of girls and boys is analyzed in this study in order to prove the usefulness of the proposed CV function in detecting variation changes during growth spurts. CV function is found to be a practical descriptive statistic especially when the mean curves of functional data groups are different. We found that as group mean changes, group variation may also change comparatively, but standard deviation function does not extract this information. Hence, when we compare more than one group of functional data, CV function can be utilized to have a better insight on the effect of mean function and the variation changes between groups. The availability of first and second derivatives of CV function also strengthens its utilization.

This study gives a hint of new opportunities in analysis of functional data. Simulation studies or comparative studies may be conducted to show different uses of CV function in special cases. Comparative studies with standard deviation may also be conducted.

References

Clarkson, D.B., Fraley, C., Gu, C., Ramsay, J.(2005). *S+Functional Data Analysis User's Manual for Windows*®, Springer-Verlag, New-York.

Coffey, N., Hinde, J.(2011). Analyzing time-course microarray data using functional data analysis-a review. *Statistical Applications in Genetics and Molecular Biology,* 10(1),1-32.

Cox, D.D., and Lee, J.S. (2008). Pointwise testing with functional data using the Westfall-Young randomization method, *Biometrika*, 95(3), 621-634.

Keser, I.K. (2014). "Comparing two mean humidity curves using functional t-tests: Turkey case", *Electronic Journal of Applied Statistical Analysis,* 7(2), 254-278.

Keser, I.K, Deveci Kocakoç, I. (2015), *FDAPackage, software*. Available at http://people.deu.edu.tr/istem.koymen/fda.html

Lee, J.S. (2005). *Aspects of Functional Data Inference and Its Applications*. Doctor of Philosophy, Houston, Texas.

Levitin, D.J., Nuzzo, R.L., Vines Bradley W., Ramsay J.O., (2007). Introduction to Functional Data Analysis, *Canadian Psychology,* 48(3), 135-155.

Ramsay, J.O. (1982). When the data are functions, *Psychometrika* 47, 379-396.

Ramsay, J.O. (2015). *FDAPackage, software*. Available at: http://www.psych.mcgill.ca/misc/fda/downloads/FDAfuns/Matlab/.

Ramsay, J.O., Hooker, G., Graves, S. (2009). *Functional Data Analysis with R and MATLAB*, Springer-Verlag, New-York.

Ramsay J.O, Silverman B.W. (1997). *Functional Data Analysis*, Springer-Verlag, New-York.

Ramsay, J.O, Silverman, B.W. (2005). *Functional Data Analysis*, Second Edition, Springer-Verlag, New-York.

Rice, J. A., Silverman, B.W. (1991). Estimating the Mean and Covariance Structure When the Data are Curves, *Journal of the Royal Statistical Society*. Series B. 53(1), 233-243.

Shang, H.L. (2015). Resampling Techniques for Estimating the Distribution of Descriptive Statistics of Functional Data, *Communications in Statistics - Simulation and Computatio*n, 44:3, 614-635.

Sun, Y., Genton, M.G. (2011) Functional Boxplots, *Journal of Computational and Graphical Statistics*, 20:2, 316-334, DOI: 10.1198/jcgs.2011.09224

Tuddenham, R., Snyder, M. (1954). Physical growth of California boys and girls from birth to age 18. *California Publications on Child Development*, 1, 183-364.

Ullah, S., Finch, C. F. (2013). Applications of functional data analysis: a systematic review, *BMC Medical Research Methodology*, 13(43), 539–572.

Yaree, K. (2011). *Functional data analysis with application to ms and cervical vertebrae data*. Master of Science in Statistics, Edmonton, Alberta.

Zhang, J-T. (2013). *Analysis of Variance for Functional Data*, CRC Press.

Appendix A

In this appendix, detailed formulations for **CV functions** are given. Here, **B** is the (nxK) basis function matrix which is consisting of $B_i(t_j)$, i=1,2,..., K, j=1,2,...,n values. **C** is the variance-covariance matrix of coefficients which are obtained by roughness penalty method or the least sum of squares method. According to the basis function approach, variance-covariance matrix for n curves is $\mathbf{V} = \mathbf{B}*\mathbf{C}*\mathbf{B}^T$.

1. Calculation of the coefficient vector of standard deviation function (*std*):
   As *s*=sqrt(diag(V)), *s*=B**std*. In order to find *std*, the coefficient vector of standard deviation function, the calculations below are carried out.
   Since **B** matrix may not be a square matrix, it is converted into a square matrix by multiplying both sides by $\mathbf{B}^T$ because of the inverse problem.
   $B^T*s = B^T*B*std$
   B^T*B matrix may not be invertible by singular value decomposition. In this case, Cholesky decomposition or Ridge regression may be used.
   $\mathbf{B}^T*s = \mathbf{D}$
   $\mathbf{B}^T*\mathbf{B}=\mathbf{E}$
   While **R** is an upper triangular matrix, according to Cholesky decomposition:
   $\mathbf{E}= \mathbf{R}^T*\mathbf{R}$.
   $\mathbf{D}= \mathbf{R}^T*\mathbf{R}*std$
   $(\mathbf{R}\text{-}1)^T\mathbf{D} = (\mathbf{R}\text{-}1)^T*\mathbf{R}^T*\mathbf{R}*std$
   $(\mathbf{R}\text{-}1)*(\mathbf{R}\text{-}1)^T*\mathbf{D}= (\mathbf{R}\text{-}1)*(\mathbf{R}\text{-}1)^T*\mathbf{R}^T*\mathbf{R}*std$
   So, coefficient vector of standard deviation function is
   $std= (\mathbf{R}\text{-}1)* (\mathbf{R}\text{-}1)^T*\mathbf{D}$.
2. Calculation of the coefficient vector of mean function $\bar{c}$:
   While $\bar{y}$ is the mean coordinate vector, $\bar{c}$ may be obtained in a similar way as the *std*:
   $\bar{c}= (B^T*B)*B^T*\bar{y}$.
3. Calculation of the coefficient vector of **CV function**:
   We propose that the coefficients of **CV function** to be calculated as:
   $$\mathbf{CV}= \frac{\mathbf{B}*std}{\mathbf{B}*\bar{c}}$$

Appendix B.

	Girls			Boys		
Knots	$\boldsymbol{B}^{*}std$	$\boldsymbol{B} * \bar{c}$	CV	$\boldsymbol{B}^{*}std$	$\boldsymbol{B} * \bar{c}$	CV
1.0000	3.2149	73.8139	0.0436	2.7624	76.2446	0.0362
1.2500	3.0684	77.8339	0.0394	2.8500	79.8017	0.0357
1.5000	3.3774	81.3018	0.0415	3.0910	83.1423	0.0372
1.7500	3.2145	84.3852	0.0381	3.2856	86.1408	0.0381
2.0000	3.2864	87.2957	0.0376	3.3881	88.7906	0.0382
3.0000	3.8114	95.5438	0.0399	3.4330	97.3713	0.0353
4.0000	4.2208	103.0589	0.0410	3.4959	104.8315	0.0333
5.0000	4.3924	110.2152	0.0399	3.5914	111.5847	0.0322
6.0000	4.5040	117.3070	0.0384	3.8110	118.0486	0.0323
7.0000	4.6766	123.3705	0.0379	4.0739	124.6228	0.0327
8.0000	4.8499	129.1130	0.0376	4.2650	130.7314	0.0326
8.5000	5.0424	132.0278	0.0382	4.4653	133.7134	0.0334
9.0000	5.2501	134.9666	0.0389	4.6760	136.6391	0.0342
9.5000	5.4667	137.8605	0.0397	4.8712	139.4234	0.0349
10.0000	5.7211	140.7676	0.0406	5.0555	142.1612	0.0356
10.5000	5.9819	144.0496	0.0415	5.2153	144.8326	0.0360
11.0000	6.4194	147.5096	0.0435	5.4182	147.4634	0.0367
11.5000	6.8465	151.0443	0.0453	5.6682	150.2114	0.0377
12.0000	7.2259	154.3572	0.0468	6.0387	153.2493	0.0394
12.5000	7.4671	157.2223	0.0475	6.6264	156.6553	0.0423
13.0000	7.3471	159.5794	0.0460	7.3943	160.3793	0.0461
13.5000	7.0320	161.4133	0.0436	8.2132	164.2413	0.0500
14.0000	6.7754	162.8515	0.0416	8.7791	167.9415	0.0523
14.5000	6.5885	163.8652	0.0402	8.8698	171.2526	0.0518
15.0000	6.4599	164.6108	0.0392	8.3988	173.9824	0.0483
15.5000	6.3688	165.1059	0.0386	7.5969	176.0432	0.0432
16.0000	6.3220	165.5016	0.0382	6.9362	177.5510	0.0391
16.5000	6.2759	165.7743	0.0379	6.5528	178.5975	0.0367
17.0000	6.2072	165.9923	0.0374	6.4192	179.2888	0.0358
17.5000	6.1937	166.1243	0.0373	6.4179	179.7815	0.0357
18.0000	6.2511	166.2944	0.0376	6.4419	180.2250	0.0357