Detection of changes in mean of vectors with application to hydrology and meteorology

Daniela JARUŠKOVÁ

CzechTechnical University of Prague, Dept. of Mathematics

Change point analysis

At time points $t = 1, \ldots, n$ we observe a sequence of independent variables (vectors) $\{X_1, \ldots, X_n\}$.

Decision problem:

Is there a change in a stochastic model at unknown time point(s)?

off-line approach - statistical hypotheses testing

Data - daily values

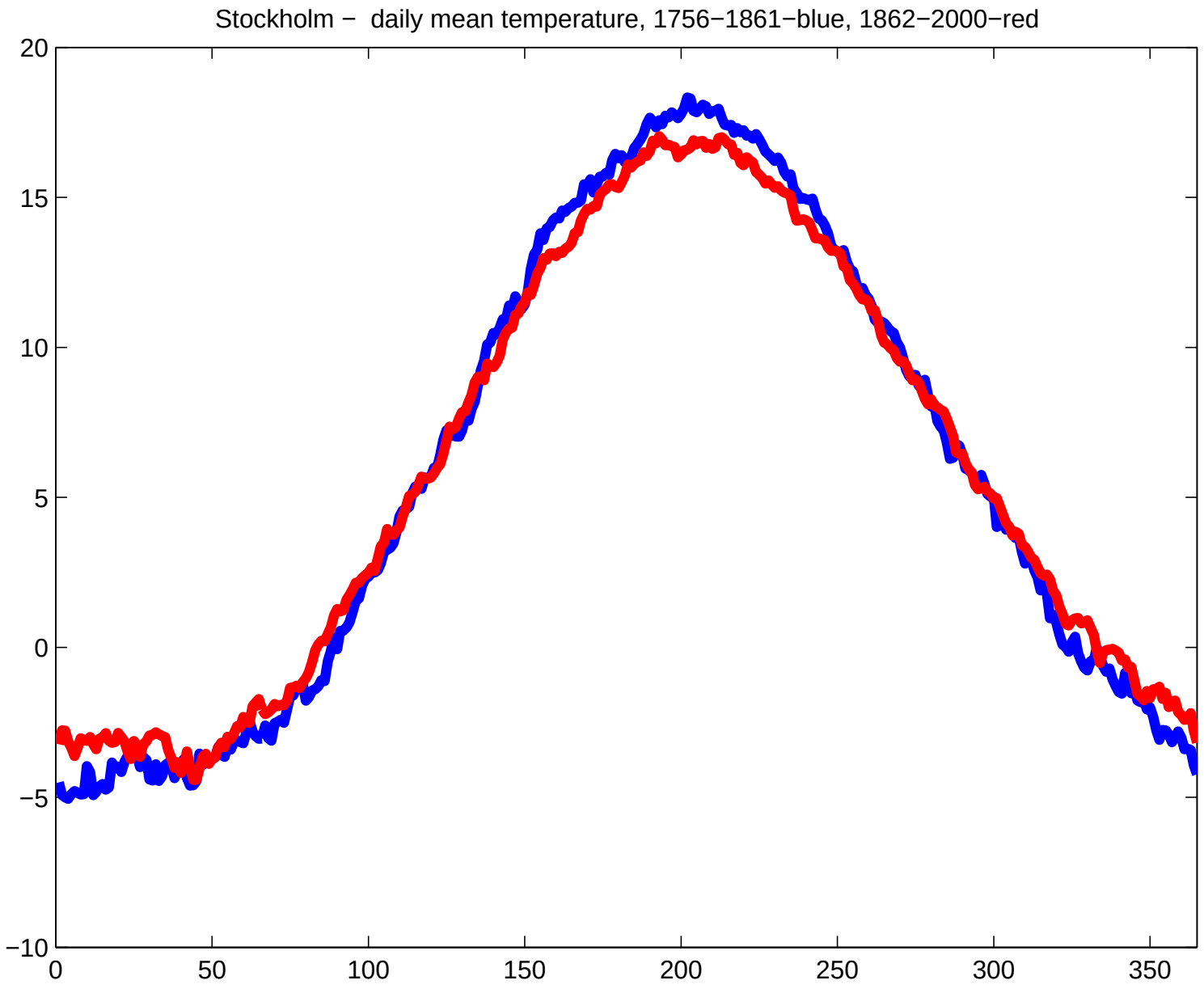
$$\mathbf{x}_1 = \left(X_{1,1}, \ldots, X_{365,1}\right)^T$$

$$\vdots$$

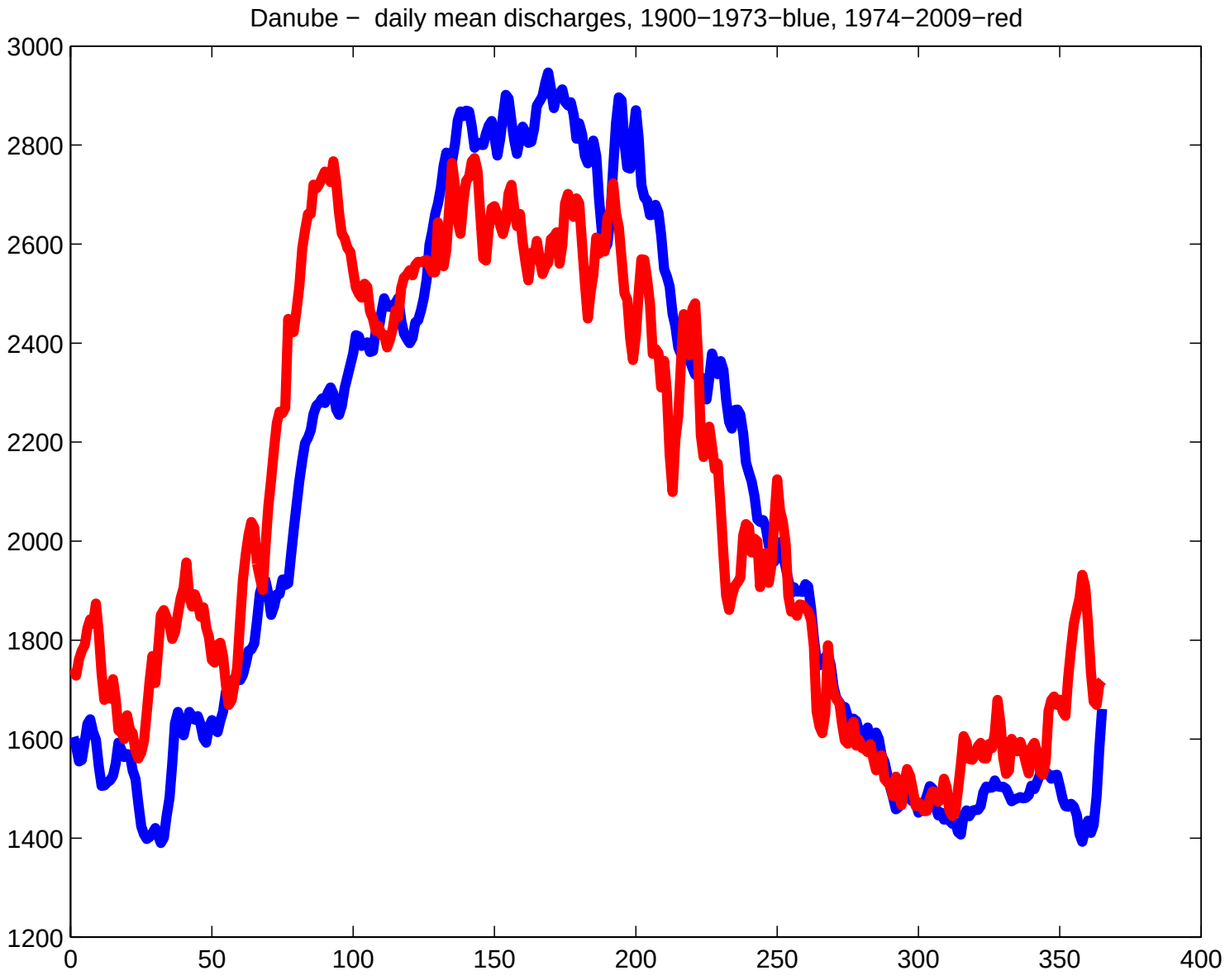
$$\mathbf{x}_n = \left(X_{1,n}, \ldots, X_{365,n}\right)^T$$

Has the annual cycle changed?

Stockholm-change in daily temperatures



Danube-change in daily discharges



Change point detection methods for detecting change(s) in annual cycle based on daily observations

MY LECTURE: Suggest methods for detecting a change (changes) in annual cycle and present experience with them.

Hypotheses testing problem

$$
\begin{aligned}
&H_0 : \mathbf{X}_i = \boldsymbol{\mu} + \mathbf{e}_i, && i = 1, \ldots, n, \\
&A_p : \exists\, k \in \{1, \ldots, n-1\} \quad \text{such that} \\
&\qquad \mathbf{X}_i = \boldsymbol{\mu}_1 + \mathbf{e}_i, && i = 1, \ldots, k \\
&\qquad \mathbf{X}_i = \boldsymbol{\mu}_2 + \mathbf{e}_i, && i = k+1, \ldots, n,
\end{aligned}
$$

$\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ are $p-$ dimensional unknown vectors. The vectors $\{\mathbf{e}_i\}$ are i.i.d. and $E\,\mathbf{e}_i = \mathbf{0}$, $Var\,\mathbf{e}_i = \boldsymbol{\Sigma}_{p\,x\,p}$, $E||\mathbf{e}_i||^{2+\delta} < \infty$.

$$
\begin{aligned}
&H_0 : \mathbf{X}_i = \boldsymbol{\mu} + \mathbf{e}_i, && i = 1, \ldots, n, \\
&A_p : \exists\, k \in \{1, \ldots, n-1\} \quad \text{such that} \\
&\qquad \mathbf{X}_i = \boldsymbol{\mu} + \mathbf{e}_i, && i = 1, \ldots, k \\
&\qquad \mathbf{X}_i = \boldsymbol{\mu} + \boldsymbol{\Delta} + \mathbf{e}_i, && i = k+1, \ldots, n,
\end{aligned}
$$

$\boldsymbol{\mu}$ and $\boldsymbol{\Delta} \neq \mathbf{0}$ are $p-$ dimensional unknown vectors. The vectors $\{\mathbf{e}_i\}$ are i.i.d. and $E\,\mathbf{e}_i = \mathbf{0}$, $Var\,\mathbf{e}_i = \boldsymbol{\Sigma}_{p\times p}$, $E||\mathbf{e}_i||^{2+\delta} < \infty$.

For $k = 1, \ldots, n-1$

$$\frac{1}{\sigma}\frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{1}{k} + \frac{1}{n-k}}} = \frac{1}{\sigma}\sqrt{\frac{k(n-k)}{n}}\left(\overline{X_1} - \overline{X_2}\right) \overset{as}{\sim} N(0,1)$$

Moreover

$$\overline{X_1} - \overline{X_2} = \frac{1}{k}\sum_{i=1}^{k} X_i - \frac{1}{n-k}\sum_{i=k+1}^{n} X_i =$$

$$= \frac{1}{k}\sum_{i=1}^{k} X_i - \frac{1}{n-k}\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{k} X_i\right) =$$

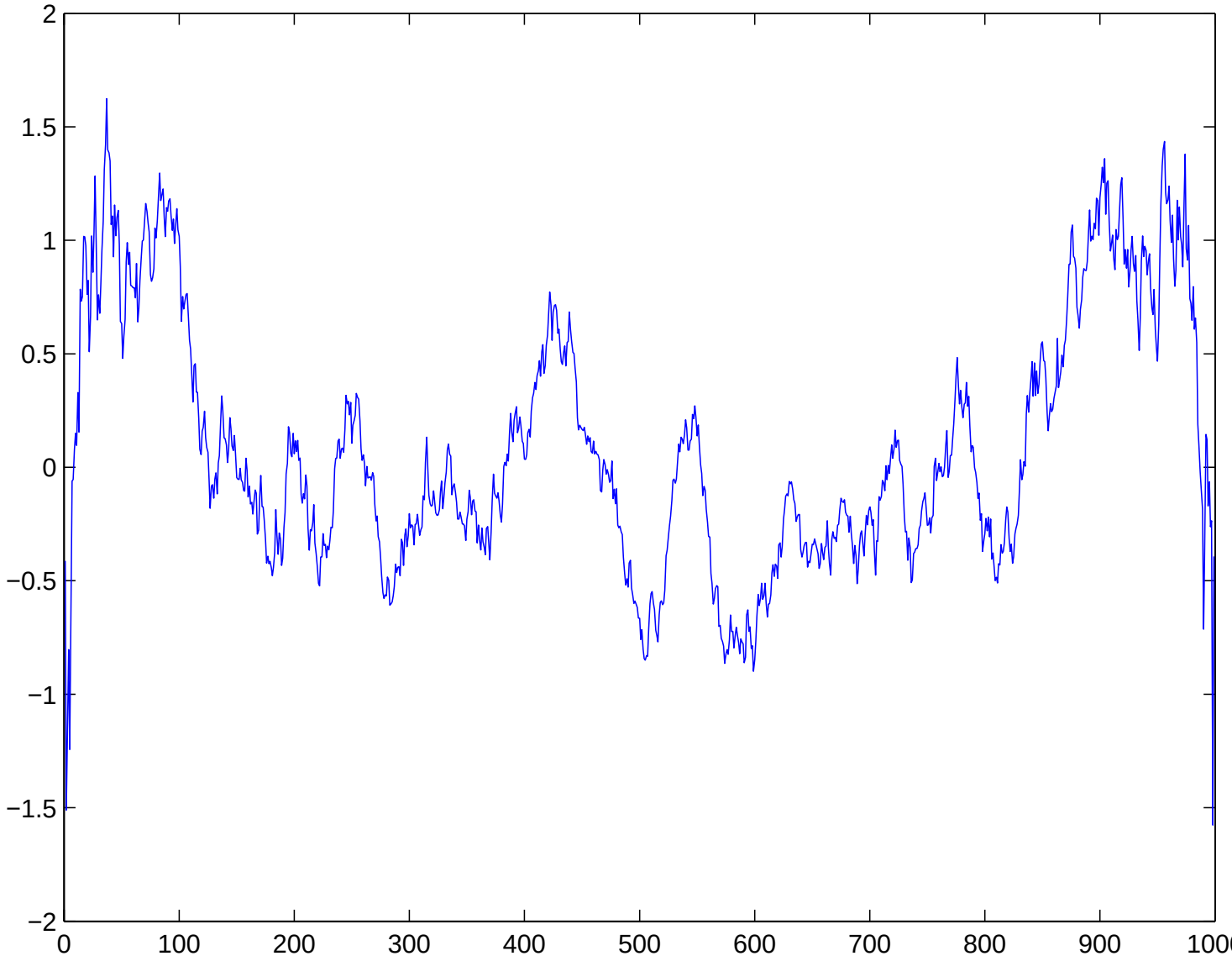
$$= \left(\frac{1}{k} - \frac{1}{n-k}\right)\sum_{i=1}^{k} X_i - \frac{n}{n-k}\overline{X} = \frac{n}{k(n-k)}\sum_{i=1}^{k}(X_i - \overline{X})$$

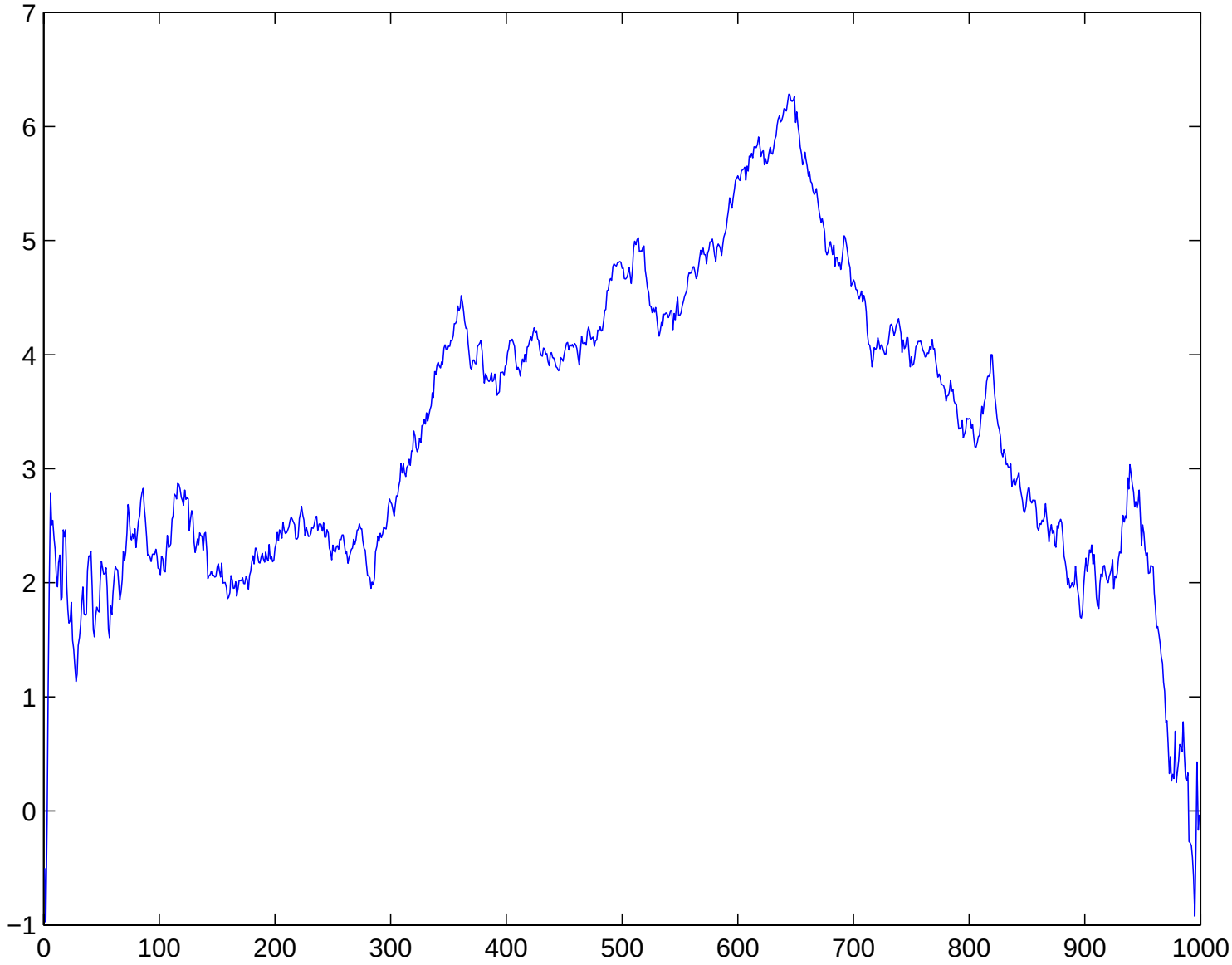
For a fixed (known) k the test statistic for a two-sided alternative $\mu_1 \neq \mu_2$

$$\frac{1}{\sigma}\sqrt{\frac{k(n-k)}{n}}|\overline{X_1} - \overline{X_2}| = \frac{1}{\sigma}\sqrt{\frac{n}{k(n-k)}}\Big|\sum_{i=1}^{k}(X_i - \overline{X})\Big|$$

or

$$\frac{1}{\sigma^2}\frac{k(n-k)}{n}\Big(\overline{X_1} - \overline{X_2}\Big)^2 = \frac{1}{\sigma^2}\frac{n}{k(n-k)}\Big(\sum_{i=1}^{k}(X_i - \overline{X})\Big)^2$$


2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
0
100
200
300
400
500
600
700
800
900
1000


7
6
5
4
3
2
1
0
−1
0
100
200
300
400
500
600
700
800
900
1000

For $k = 1, \ldots, n-1$

$$\frac{k(n-k)}{n} \left(\bar{\mathbf{X}}_1(k) - \bar{\mathbf{X}}_2(k)\right)^T \hat{\mathbf{\Sigma}}_n^{-1} \left(\bar{\mathbf{X}}_1(k) - \bar{\mathbf{X}}_2(k)\right) =$$

$$\frac{n}{k(n-k)} \left({}^n\mathbf{S}^X(k)\right)^T \hat{\mathbf{\Sigma}}_n^{-1} \left({}^n\mathbf{S}^X(k)\right)$$

where ${}^n\mathbf{S}^X(k) = \left({}^nS_1^X(k), \ldots, {}^nS_p^X(k)\right)^T$ with
${}^nS_j^X(k) = \sum_{i=1}^k \left(X_{i,j} - \bar{X}_{\cdot,j}\right)$, $j = 1, \ldots, p$.

Test Statistics

$$TNS = \max_{\epsilon n \le k \le (1-\epsilon)n} \frac{n}{k(n-k)} \left({}^n\mathbf{S}^X(k)\right)^T \hat{\mathbf{\Sigma}}_n^{-1} \left({}^n\mathbf{S}^X(k)\right)$$

$$TN = \max_{1 \le k \le n} \frac{1}{n} \left({}^n\mathbf{S}^X(k)\right)^T \hat{\mathbf{\Sigma}}_n^{-1} \left({}^n\mathbf{S}^X(k)\right)$$

$$MNS = \frac{1}{n} \sum_{k=1}^{n} \frac{n}{k(n-k)} \left({}^n\mathbf{S}^X(k)\right)^T \hat{\mathbf{\Sigma}}_n^{-1} \left({}^n\mathbf{S}^X(k)\right)$$

$$MN = \frac{1}{n} \sum_{k=1}^{n} \frac{1}{n} \left({}^n\mathbf{S}^X(k)\right)^T \hat{\mathbf{\Sigma}}_n^{-1} \left({}^n\mathbf{S}^X(k)\right)$$

For $k = 1, \ldots, n-1$

$\left\{ \frac{n}{k(n-k)} \left({}^n\mathbf{S}^X(k)\right)^T \hat{\mathbf{\Sigma}}_n^{-1} \left({}^n\mathbf{S}^X(k)\right) \right\} \ldots$ weights $\left\{ \frac{k(n-k)}{n^2} \right\}$.

Limit variables

$$TNS \xrightarrow{D} \max_{\beta \le t \le (1-\beta)} \frac{B_1^2(t) + \cdots + B_p^2(t)}{t(1-t)}$$

$$TN \xrightarrow{D} \max_{0 \le t \le 1} B_1^2(t) + \cdots + B_p^2(t)$$

$$MNS \xrightarrow{D} \int_0^1 \frac{B_1^2(t) + \cdots + B_p^2(t)}{t(1-t)}\, dt$$

$$MN \xrightarrow{D} \int_0^1 B_1^2(t) + \cdots + B_p^2(t)\, dt$$

$\{(B_1(t), \ldots, B_p(t))\}$ . . . multivariate Gaussian process with components being independent Brownian bridges

Critical values obtained with the help of

- asymptotic distribution
- simulations
- random permutations

Approximation of a distribution (survival) function of the limit variables

$\max_{\beta \le t \le (1-\beta)} \frac{B_1^2(t)+\cdots+B_p^2(t)}{t(1-t)}$ - Theorem A.3.3. of Csörgő M. and Horváth L. (1997), *Limit Theorems in Change Point Analysis*.

$\int_0^1 \frac{B_1^2(t)+\cdots+B_p^2(t)}{t(1-t)}\,dt$ - Scholz and Stephens (1987) in *JASA*.

$\max_{0 \le t \le 1} B_1^2(t) + \cdots + B_p^2(t)$ and $\int_0^1 B_1^2(t) + \cdots + B_p^2(t)\,dt$ - Kiefer (1959) in *AMS*.

Simulations

Due to the Donsker invariance principle whose consequence is an existence of limit variables we can generate samples of independent standard normally distributed vectors and calculate values of the considered test statistics. The critical values are obtained as quantiles of their empirical distributions.

Permutations

Application of permutation principle in change point detection methods was suggested by Antoch J. and Hušková M. (2001) in JSPI.

Permute randomly the vector of observations. For every permutation compute the value of the considered statistic. Empirical quantiles serve as critical values.

My experience: Critical values obtained by simulations and by permutations are very close.

Reduction of dimensionality

We replace the original observed vectors

$$\mathbf{Y}_i^T = \left(Y_{i,1}, \ldots, Y_{i,365}\right),$$

by vectors

$$\mathbf{V}_i^T = \left(V_{i,1}, \ldots, V_{i,p}\right) = \left(Y_{i,1}, \ldots, Y_{i,365}\right) \mathbf{U}_{365*p}.$$

For $i = 1, \ldots, n$ we introduce new vectors

$$\mathbf{V}_i^T = \left(V_{i,1}, \ldots, V_{i,p}\right) = \left(Y_{i,1} - \overline{Y}_{\cdot,1}, \ldots, Y_{i,365} - \overline{Y}_{\cdot,365}\right) \mathbf{U}_{365*p}$$

and we are looking for a change in the sequence
$\{\mathbf{V}_i^T = (V_{i,1}, \ldots, V_{i,p}), i = 1, \ldots, n\}$.

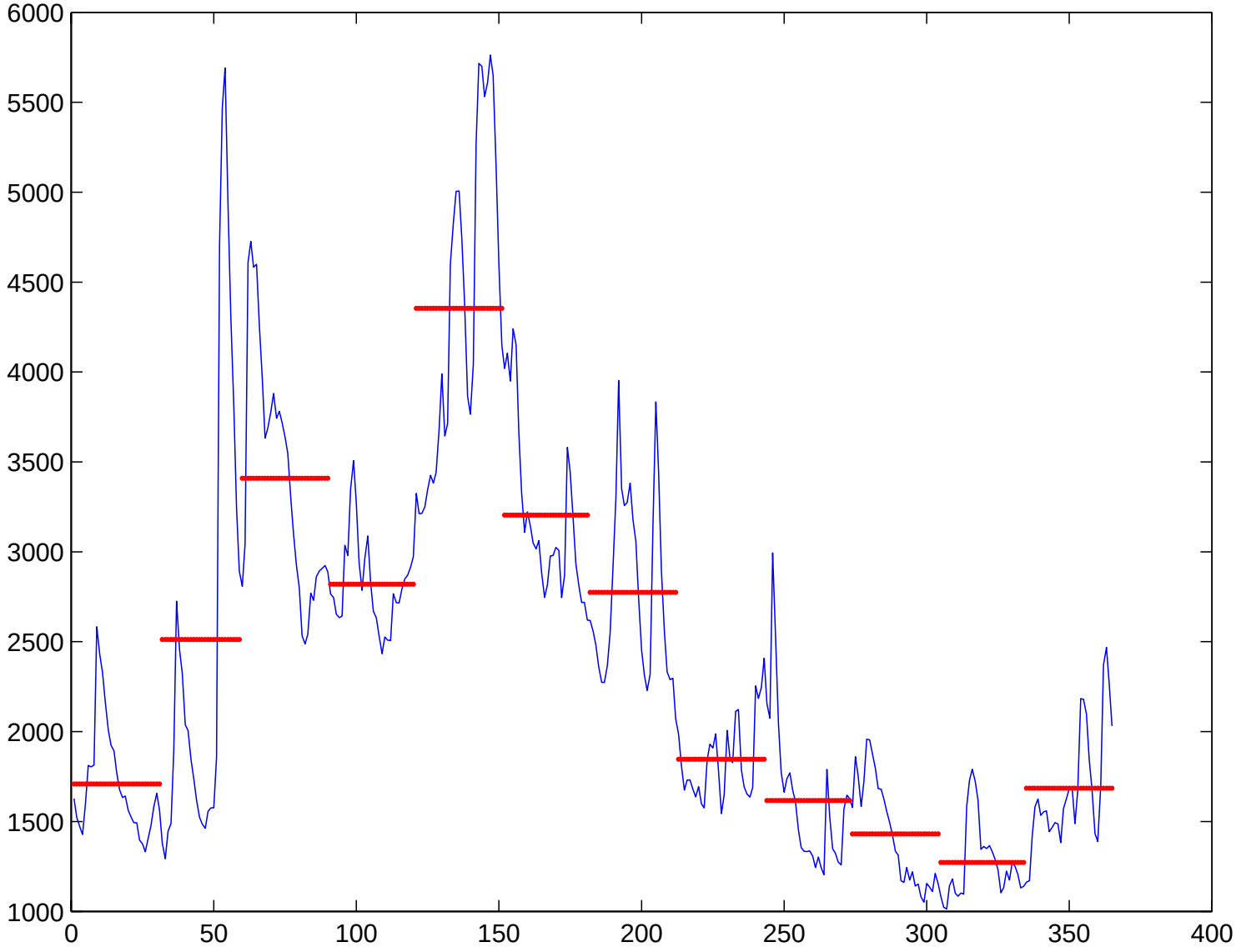
For $i = 1, \ldots, n$ we replace $\mathbf{Y}_i^T = (Y_{i,1}, \ldots, Y_{i,365})$ by the vector

$$\left(\widehat{Y}_{i,1}, \ldots, \widehat{Y}_{i,365}\right) = \left(V_{i,1}, \ldots, V_{i,p}\right) \mathbf{U}^T + \left(\overline{Y}_{\cdot,1}, \ldots, \overline{Y}_{\cdot,365}\right)$$

Replace daily averages by monthly averages

$$\mathbf{U} = \begin{pmatrix} 1/31 & & & \\ \vdots & \mathbf{0} & \dots & \mathbf{0} \\ 1/31 & & & \\ & 1/28 & & \\ \mathbf{0} & \vdots & \dots & \mathbf{0} \\ & 1/28 & & \\ & & & \\ \mathbf{0} & \mathbf{0} & \vdots & \mathbf{0} \\ & & & \\ & & & 1/31 \\ \mathbf{0} & \mathbf{0} & & \vdots \\ & & & 1/31 \end{pmatrix}$$

Danube-year 1998 plus monthly averages



Expand the annual cycle by Fourier series with the frequencies $2\pi/365, \ldots, 2\pi r/365$

$$\widehat{Y}_{i,j} = \mu_i + \sum_{k=1}^{r} a_{i,k} \cos \frac{2\pi kj}{365} + b_{i,k} \, sin \frac{2\pi kj}{365}, \quad i = 1, \ldots, n, \, j = 1, \ldots, 365,$$

$$V_i = (\mu_i, a_{i,1}, b_{i,1}, \ldots, a_{i,r}, b_{i,r})^T.$$

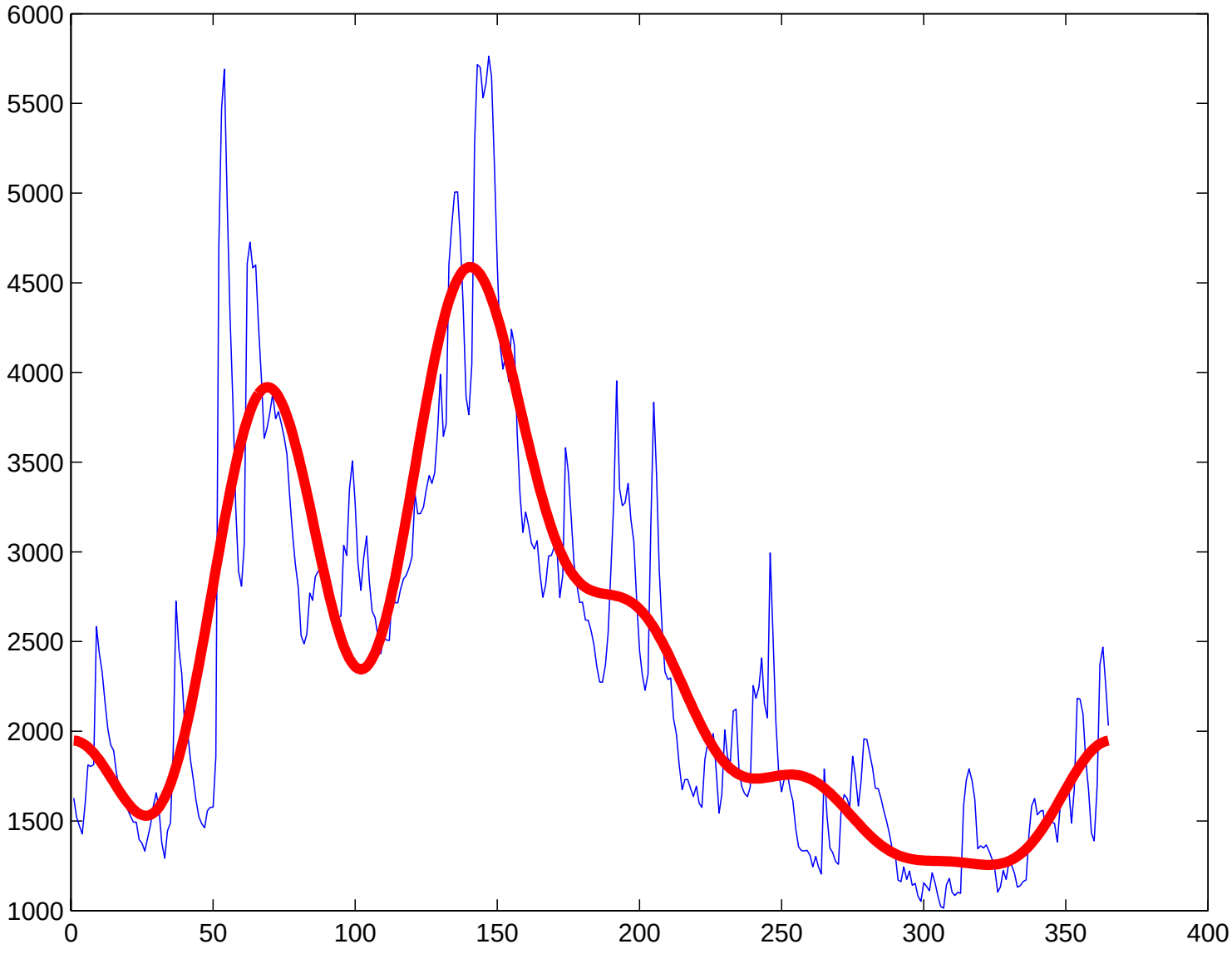
$$\mathbf{U} = \begin{pmatrix} 1 & \cos \frac{2\pi}{365} & \sin \frac{2\pi}{365} & \ldots & \cos \frac{2\pi r}{365} & \sin \frac{2\pi r}{365} \\ 1 & \cos \frac{2\pi 2}{365} & \sin \frac{2\pi 2}{365} & \ldots & \cos \frac{2\pi 2r}{365} & \sin \frac{2\pi 2r}{365} \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 1 & \cos 2\pi & \sin 2\pi & \ldots & \cos 2\pi r & \sin 2\pi r \end{pmatrix}$$

Why does it work?

1. The mean function $\mu(t)$, resp. the mean functions $\mu_1(t)$, $\mu_2(t)$ and $\mathbf{\Delta}(t) = \mu_2(t) - \mu_1(t)$ are periodic functions.

2. The mean function $\mu(t)$, resp. the mean functions $\mu_1(t)$, $\mu_2(t)$ and $\mathbf{\Delta}(t) = \mu_2(t) - \mu_1(t)$ are smooth functions.

Danube-year 1998 plus approx. by Fourier series with 6 frequencies



Principal components approach - expand the annual cycle in p eigenvectors (corresponding) to p largest eigenvalues of the variance-covariance matrix of daily values $\widehat{\mathbf{\Sigma}} = ||cov(\mathbf{Y}_{\cdot,j}, \mathbf{Y}_{\cdot,j'}||_{j,j'=1,365}$

The eigenvectors $\{\phi_k,\ k = 1, \ldots, p\}$ are the solutions of the problem

$$\widehat{\mathbf{\Sigma}}\boldsymbol{\phi} = \lambda\,\boldsymbol{\phi},$$

with $\lambda_1 > ... > \lambda_p$.

The matrix $\mathbf{U}$ has a form:

$$\mathbf{U} = \begin{pmatrix} \phi_1(1) & \ldots & \phi_p(1) \\ \vdots & \ldots & \vdots \\ \phi_1(365) & \ldots & \phi_p(365) \end{pmatrix}$$

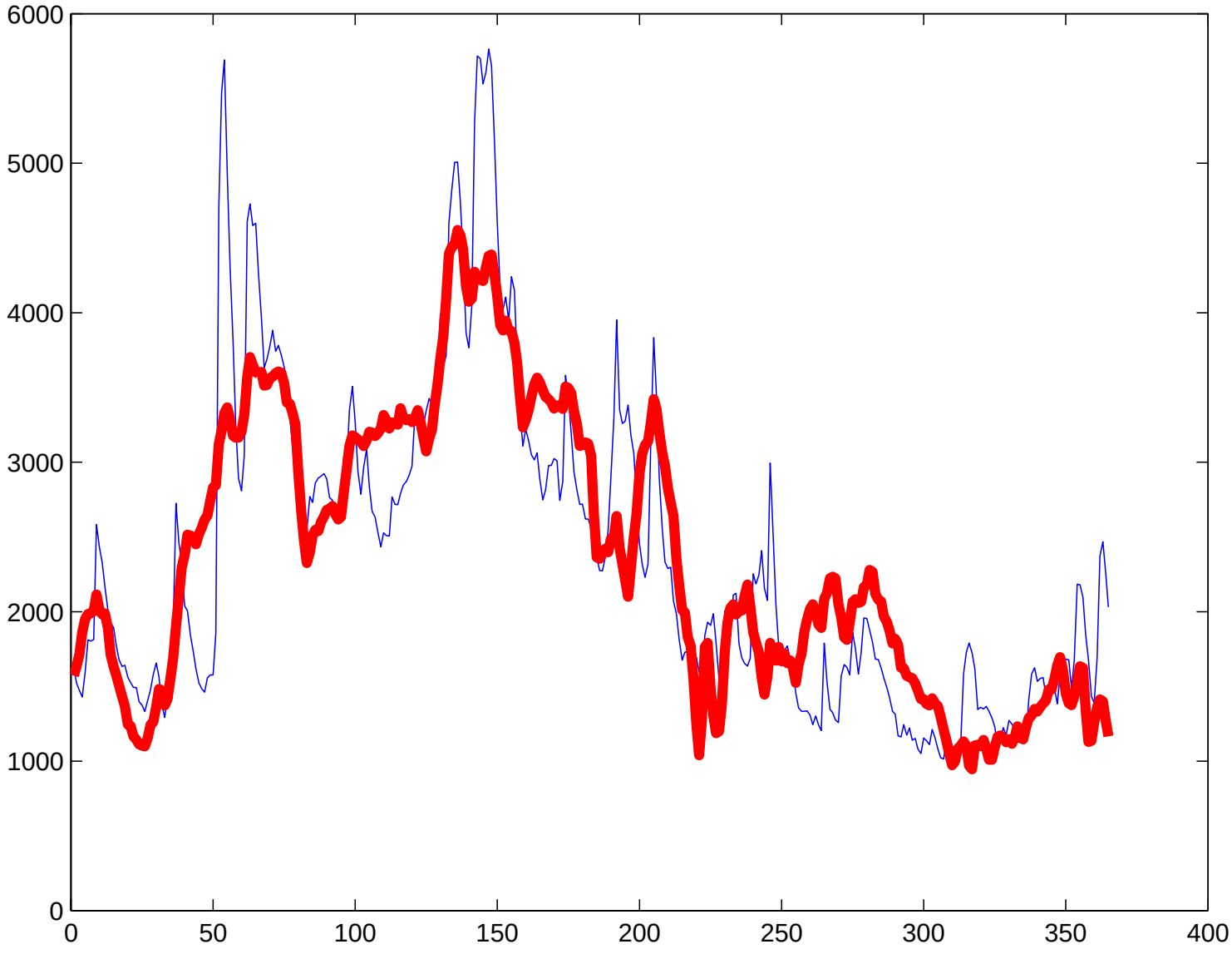
Why does it work?

We can meet two situations:

1. The large variance of the chosen linear combination is natural
- we have chosen a wrong linear combination.

2. The large variance of the chosen linear combination is caused by a shift in the mean of this linear combination - we have chosen a right linear combination.

Danube-year 1998 plus approx. with the help of 12 eigenvectors



statistic	mont. av.	6 freq.	5 freq.	12 eig.v.
tr. TNS	23.65 (38)	28.93 (32)	25.31 (22)	25.17 (37)
perm.	p.v.=0.30	p.v.=0.04	p.v.=0.12	p.v.=0.16
sim.	p.v.=0.24	p.v.=0.07	p.v.=0.09	p.v.=0.15
pen. TN	5.37 (38)	6.43 (38)	5.62 (38)	5.71 (32)
perm.	p.v.=0.14	p.v.=0.03	p.v.=0.06	p.v.= 0.08
sim.	p.v.=0.14	p.v.=0.05	p.v.=0.06	p.v.=0.08
sum MNS	16.4	17.5	16.21	17.23
perm.	p.v.=0.05	p.v.=0.02	p.v.=0.03	p.v.=0.03
sim.	p.v.=0.05	p.v.=0.05	p.v.=0.02	p.v.=0.03
sum MN	2.96	3.23	2.97	3.11
perm.	p.v.=0.05	p.v.=0.02	p.v.=0.03	p.v.=0.03
sim.	p.v.=0.05	p.v.=0.03	p.v.=0.02	p.v.=0.03

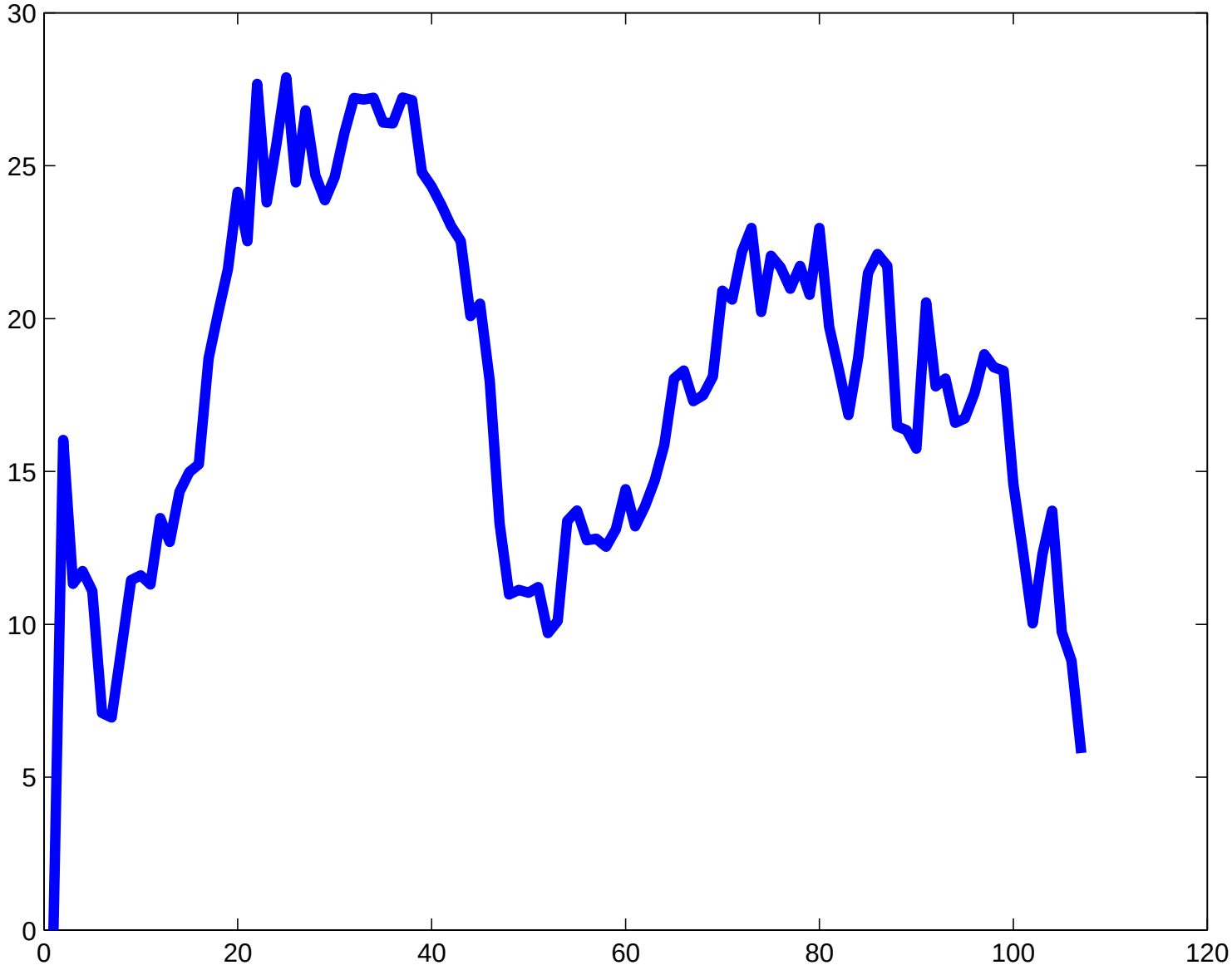
$p-$ values of sum-type test statistics are smaller than $p-$ values of max-type test statistics

Sum-type test statistics react better to more complex changes, i.e.

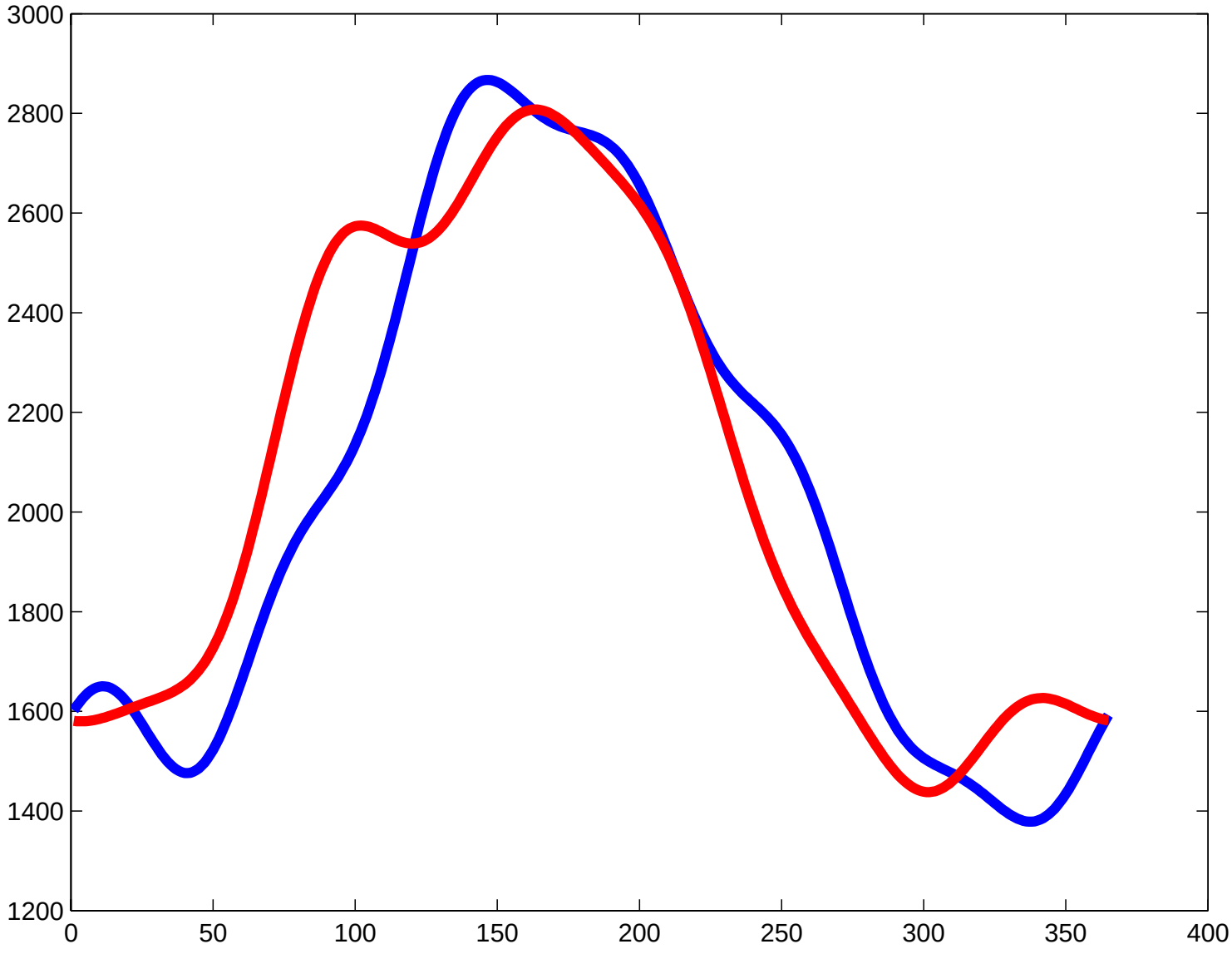
changes that have not necessarily form of a sudden shift in parameters of the stochastic model,

and to the changes where the change in different coordinates does not occur simultanously.

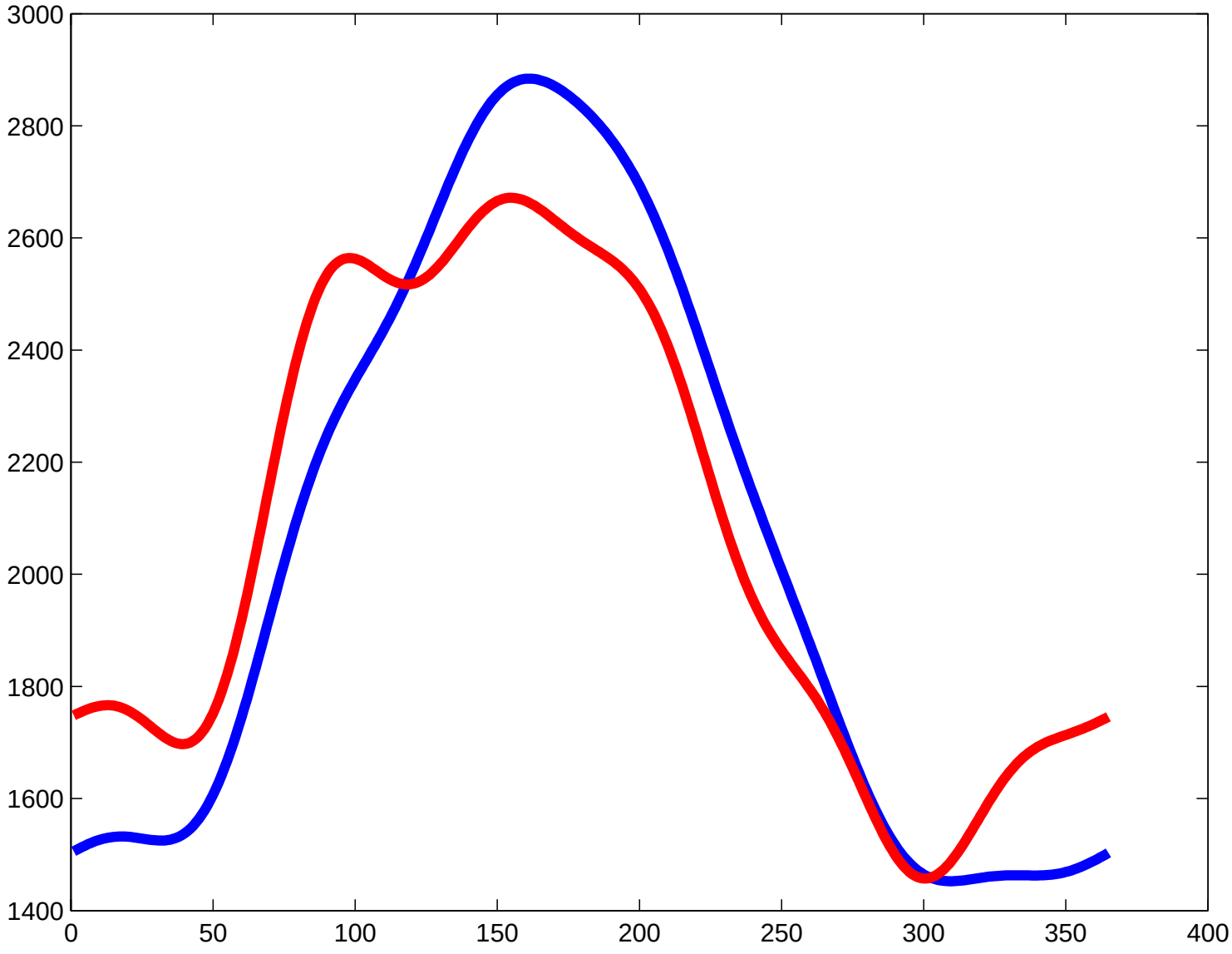
Test statistic for a change in annual cycle based on Fourier coefficients - 6 frequencies (Danube)



Approximation of annual cycle before (blue) and after (red) year 1937 using 6 Fourier frequencies



Approximation of annual cycle before (blue) and after (red) year 1971 using 6 Fourier frequencies



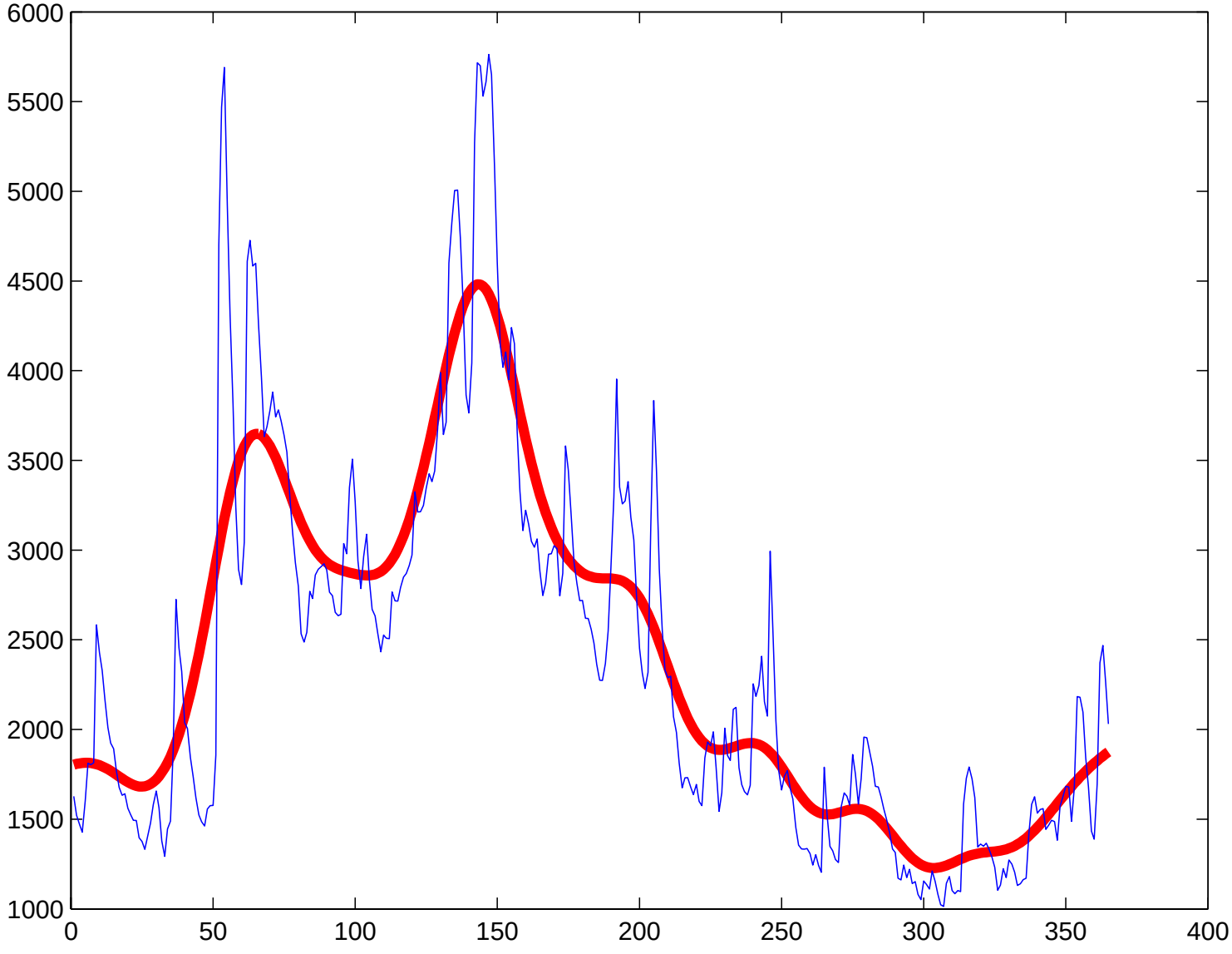
Change point detection for a change in functional data

Berkes I., Gabrys R., Horvath L., Kokoszka P. (2009): Detecting changes in the mean of functional observations. J.R. Stat. Soc. Ser. B stat. Methodol:

Suppose that the data are random functions observed at discrete time points with a noise. First, we transform by smoothing (spline smoothing, kernel smoothing) the observed vectors into smooth functions. (In this way we get rid of certain variability.) Then, we use again the principal component approach.

First,I have smoothed the daily discharges of Danube by Gaussian filter using $h_n = 30$.

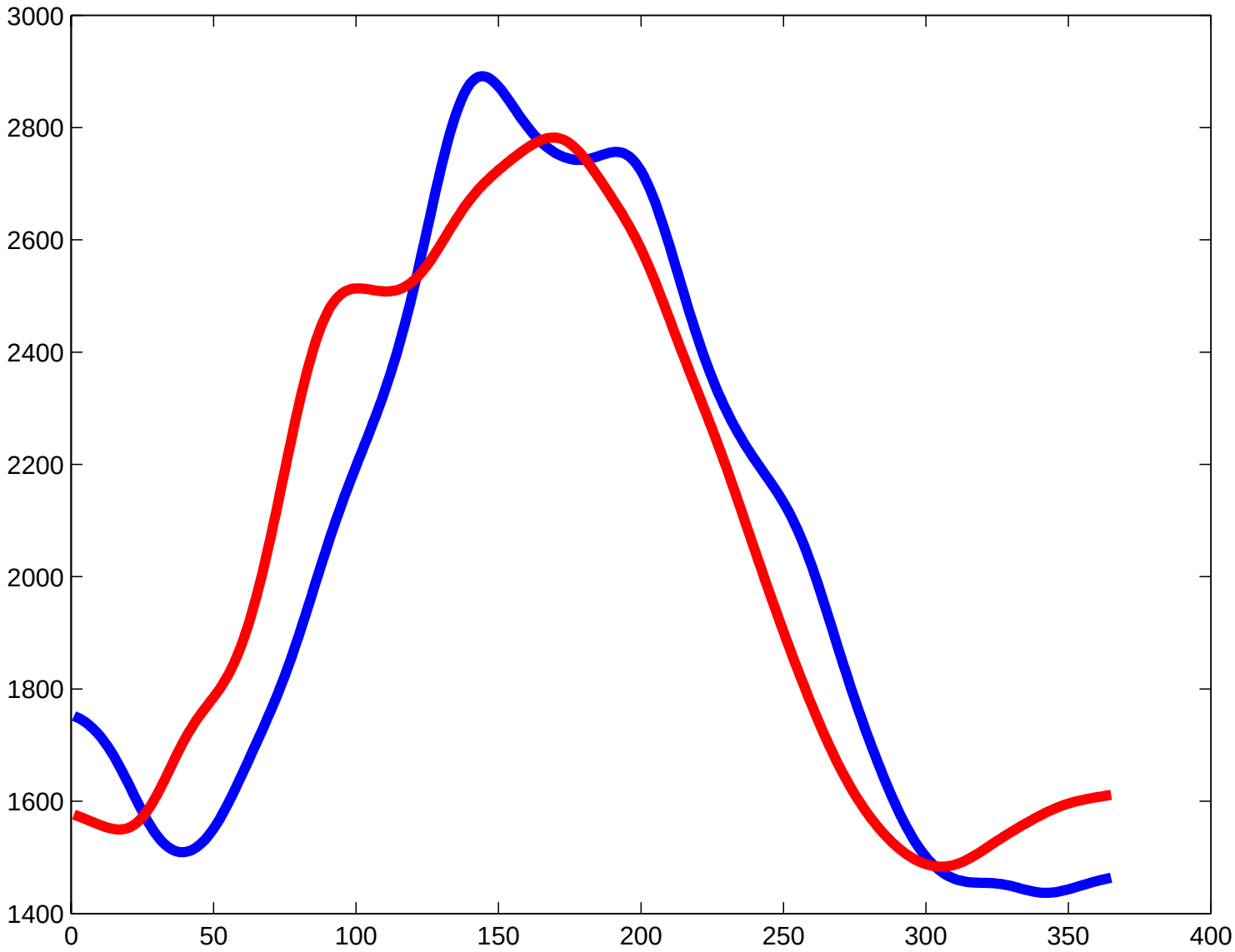
Annual cycle for the year 1998 smoothed by Gaussian filter



Then, I discretized the smoothed functions in $T = 365$ time points. After that, I applied again the method of principal components with $p = 12$. The following table shows the results for "smoothed" and "unsmoothed" vectors.

test.stat.	"smoothed"	"unsmoothed"
trim. max. TNS	26.52 p.v.=0.10	25.17 p.v.=0.16
max. pen. TN	5.98 p.v.=0.06	5.71 p.v.=0.07
sum. MNS	17.67 p.v.=0.02	17.23 p.v.=0.03
sum. pen. MN	3.27 p.v.=0.01	3.11 p.v.=0.03

Approximation of annual cycle after smoothing before (blue) and after (red) year 1937 using 12 principal components (Danube)

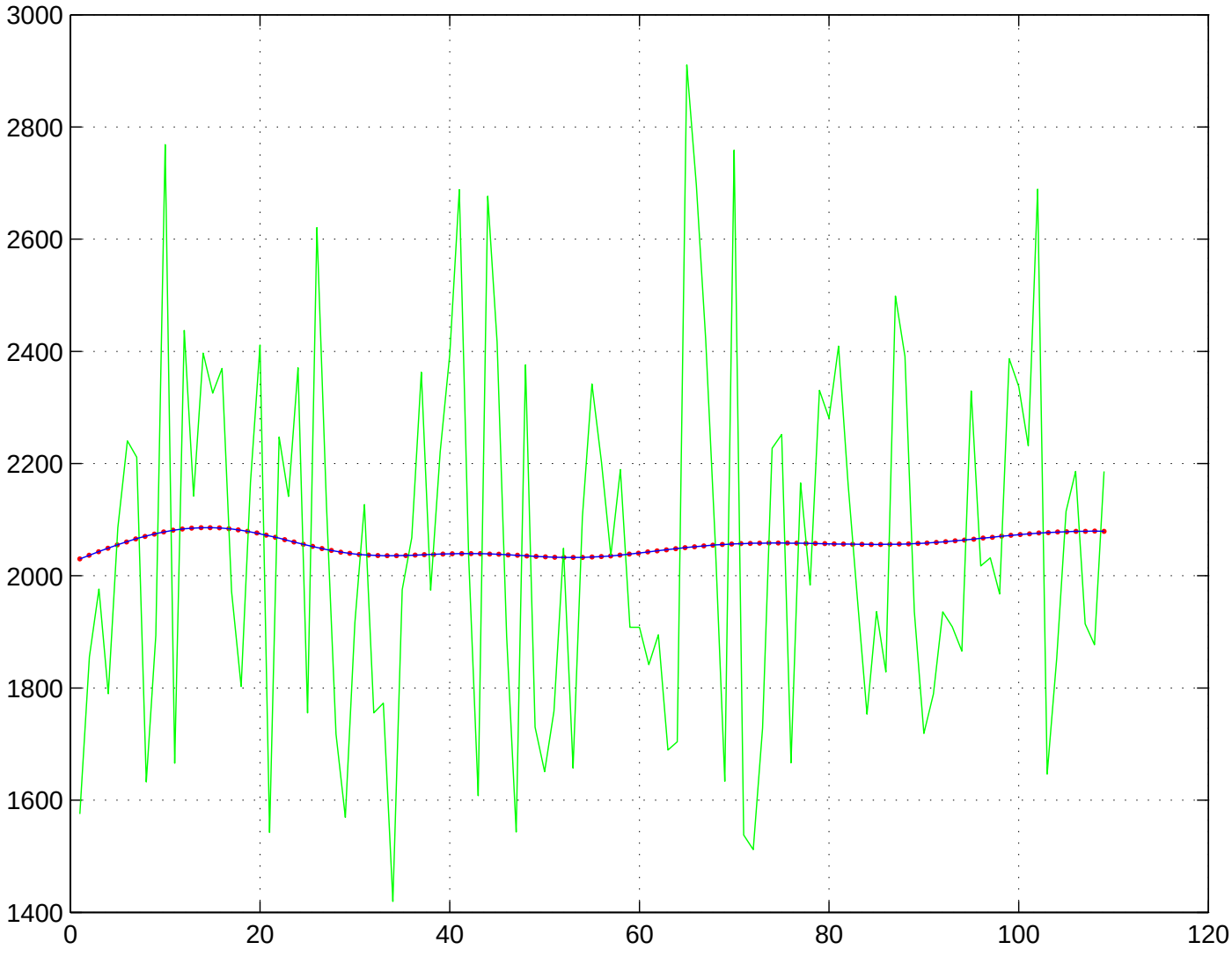


The majority of test statistics rejected the null hypothesis of no change.

Question: Is it possible that it is caused by a change (trend) in time series of annual averages?

Answer: NO

Annual averages of Danube

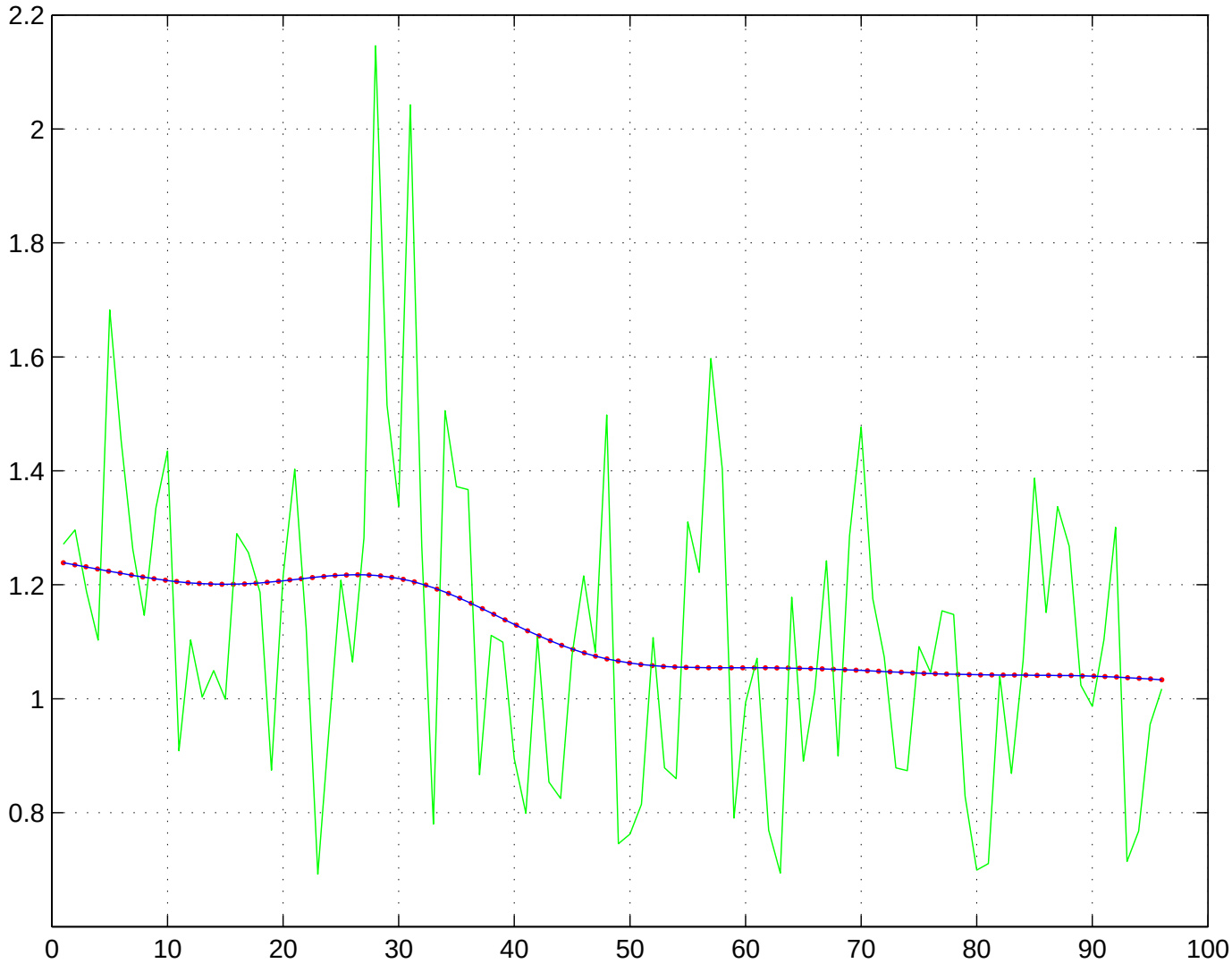


Results of tests for detecting a change(s) in annual averages (Danube)

trim. TNS	pen. TN	sum MNS	sum MN
2.53	0.24	0.286	0.034
p.v.=0.76	p.v.=0.93	p.v.=0.93	p.v.=0.97

We have also studied all months separately. The null hypothesis of no change has been rejected for March only.

Annual averages of Metuje



Null hypothesis of no change in annual averges was rejected

trim. TNS	pen. TN	sum MNS	sum MN
14.64 (36)	3.43 (36)	4.553	0.925
p.v.=0.005	p.v.=0.0008	p.v.=0.004	p.v.=0.004

May we apply the suggested procedures for

$\{Y_{i,j} - \overline{Y}_{i,\cdot}, i = 1, \ldots, n, j = 1, \ldots, 365\}$ **?**

The variance-covariance matrix of deviations of monthly averages from the annual averages

$$\mathbf{V}_i^T = (\bar{Y}_{i,1} - \bar{Y}_{i,\cdot}, \ldots, \bar{Y}_{i,12} - \bar{Y}_{i,\cdot})$$

have a singular variance-covariance matrix.

Results of tests for a change in monthly means (Metuje)

trim. TNS	pen. TN	sum MNS	sum MN
23.93 (36)	2.80 (36)	8.97	1.43
p.v.=0.23	p.v.=0.85	p.v.=0.79	p.v.=0.86

The null hypothesis of no change has been not rejected inspite of the fact that all tests have detected a change in January averages.

Conclusions

- Dimensionality of the problem should be reduced. We recommend to approximate the behaviour of daily averages of one year by few terms in the corresponding Fourier series expansion or to project the vector into the space given by p eigenvectors of the variance-covariance matrix corresponding to p largest eigenvalues (method of principal components). How the choose the dimensionality of the new problem?
- We recommend to use the sum type statistics.
- The test should be accompanied by a test for detection a change in annual means.
- For the method of principal components is it reasonable to smooth apriori the data? (Approach of functional data.)

Antoch J. and Hušková M.: *Permutation tests for a change point analysis*. Statist. Probab. Letters, **53**, 2001, 37 – 46.

Berkes I., Gabrys R., Horváth L., Kokoszka P.: Detecting changes in the mean of functional observations. J.R. Stat. Soc. Ser. B stat. Methodol. 71, 2009, 927 – 946.

Csörgő M. and Horváth L.: *Limit Theorems in Change Point Analysis*, 1997, J. Wiley, New York.

Horváth L., Kokoszka P., Steinebach J.: Testing for changes in multivariate dependent observations with an application to temperature changes. Journal of Multivariate Analysis 68, 1999, 96-119.

Jarušková D.: Asymptotic behavior of a test statistic for detection of change in mean of vectors. JSPI 140, 2010, 616-625.

Kiefer J.: K-sample analogues of the Kolmogorov-Smirnov and Cramēr - V. Mises tests, AMS 30, 1959, 420 – 447.