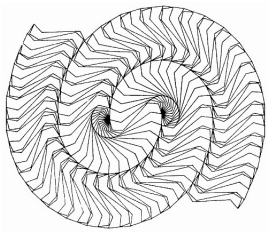
ERNST MORITZ ARNDT
UNIVERSITÄT GREIFSWALD

Wissen
lockt.
Seit 1456

Change-Point Detection and U-Statistics

Martin Wendler
joint with: B. Bucchia, C. Gerstenberger, H.Dehling, O.Sh. Sharipov,
J. Tewes, D. Vogel, D. Wied

3rd Workshop on Goodness-of-fit and Change-Point Problems



Institut für
Mathematik und Informatik

Outline

CUSUM-Statistic

test for change of expectation $\mu_i := E[X_i]$

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_n$$
$$\text{gegen } H_1 : \exists k \in 1, \ldots, n-1 : \ \mu_1 = \ldots = \mu_k \neq \mu_{k+1}, \ldots, \mu_n.$$

test statistic:

$$T_n = \max_{k=1,\ldots,n} \frac{k(n-k)}{\sqrt{\hat{\sigma}^2 n^3}} \left| \frac{1}{k} \sum_{i=1}^{k} X_i - \frac{1}{n-k} \sum_{i=k+1}^{n} X_i \right|$$
$$= \frac{1}{\sqrt{n\hat{\sigma}^2}} \max_{k=1,\ldots,n} k \left| \frac{1}{k} \sum_{i=1}^{k} X_i - \frac{1}{n} \sum_{i=1}^{n} X_i \right|$$

with $\hat{\sigma}^2_{\bar{X}}$: estimator of variance

Generalization of CUSUM Statistic

change of location: compare mean of first k observations with mean of all observations

$$T_n = \frac{1}{\sqrt{n\hat{\sigma}^2}} \max_{k=1,\ldots,n} k \left| \frac{1}{k} \sum_{i=1}^{k} X_i - \frac{1}{n} \sum_{i=1}^{n} X_i \right|$$

change of scale: compare scale estimator s_k for first k observations with scale estimator s_n for all observations

$$V_n := \frac{1}{\sqrt{n\hat{\sigma}^2_{s_n}}} \max_{k=1,\ldots,n} k \left| s_k - s_n \right|$$

Scale Estimators

- sample standard deviation: $\hat{\sigma}_n := \left(\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2 \right)^{1/2}$
- mean absolute deviation: $d_n := \frac{1}{n-1} \sum_{i=1}^{n} |X_i - \text{med}(X)|$
- Gini's mean difference: $g_n := \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} |X_i - X_j|$
- median of absolute deviations:
  $\text{mad} := \text{med}\left(|X_i - \text{med}(X)|,\ i = 1, \ldots, n \right)$
- Q_n-estimator: Q_n^{α} defined as α-quantile of $|X_i - X_j|,\ 1 \le i < j \le n$

U-Statistics

$h : \mathbb{R}^2 \to \mathbb{R}$ measureable, symmetric function

Definition

$$U_n(h) := \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} h(X_i, X_j)$$

called ***U*-statistic** with kernel h

examples:

- $h(x, y) = |x - y|$: $U_n(h) = g_n$ (Gini's mean difference)
- $h(x, y) = \frac{1}{2}(x - y)^2$: $U_n(h) = \hat{\sigma}_n^2$ (sample variance)

U-Quantiles

empirical quantile:

- sample $x_1, \ldots, x_n$
- ordered sample: $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ with $\{x_1, \ldots, x_n\} = \{x_{(1)}, \ldots, x_{(n)}\}$
- empirical α-quantile: $x_{(\lceil \alpha n \rceil)}$

empirical U-quantile

- $h : \mathbb{R}^2 \to \mathbb{R}$ measureable, symmetric function
- α-U-Quantil $U_n^{-1}(\alpha)$: empirical α-quantile of $h(x_i, x_j)$, $1 \leq i < j \leq n$

Examples:

- $h(x, y) = |x - y|$: Q_n^α
- $h(x, y) = (x + y)/2$: Hodges-Lehmann-estimator

Functional CLT for U-Statistics

Assumptions:

- short range dependence
- mild continuity assumption
- moment assumption

Theorem (Dehling, Vogel, Wendler, Wied 2017+)

$$\left(\frac{[nt]}{\sqrt{n}} \left(U_{[nt]}(h) - U \right) \right)_{t \in [0,1]} \Rightarrow (\sigma_\infty W(t))_{t \in [0,1]}$$

- W Brownian motion
- $U = E[h(X_i, \tilde{X}_i)]$, with independent copy $\tilde{X}_i$
- σ_∞^2: long run variance (needs to be estimated)

Technical conditions

assume

- there are constant M and $\delta > 0$, such that for all $l, n \in \mathbb{N}$ $E\left|h\left(f_l(Z_{-l}, \ldots, Z_l), f_l(Z_{n-l}, \ldots, Z_{n+l})\right)\right|^{2+\delta} \leq M$
- $(X_n)_{n\in\mathbb{N}}$ is P-NED of an absolutely regular sequence $(Z_n)_{n\in\mathbb{N}}$ such that $a_l\phi(l^{-6}) \leq C\left(l^{-6\frac{2+\delta}{\delta}}\right)$ and $\sum_{k=1}^{\infty} k\beta(k)^{\frac{\delta}{2+\delta}} < \infty$
- continuity condition

$$E\left(\sup_{|x-X|\leq\epsilon, |y-Y|\leq\epsilon} |h(x,y) - h(X,Y)|\right)^2 \leq L\epsilon,$$

Near Epoch Dependence in Probability

Definition

$(X_n)_{n\in\mathbb{N}}$ called P *-near epoch dependent* (NED) on a process $(Z_n)_{n\in\mathbb{Z}}$, if there exist functions f_k, non-negative numbers $a_k \to 0$ and a function $\Phi : (0,\infty) \to (0,\infty)$ such that

$$P\left(|X_0 - f_k(Z_{-k},\ldots,Z_k)| > \varepsilon\right) \; < \; \Phi(\varepsilon)a_k$$

for all $\varepsilon > 0$ and $k \in \mathbb{N}$.

- similar concepts: stochastic stability (Bierens, 1981), S-mixing (Berkes, Hörmann, Schauer, 2009)
- no moments needed
- satisfied for many time series models (e.g. linear process with heavy tailed innovations)

Estimation of Long Run Variance

- $U = E[h(X_i, \tilde{X}_i)]$
- $h_1(x) = E[h(x, X_i)] - U$

long run variance:

$$\sigma_\infty^2 = 4 \sum_{k \in \mathbb{Z}} \mathrm{Cov}(h_1(X_1), h_1(X_{1+|k|}))$$

consistent estimator: bandwidth b, kernel K

- $\hat{h}_1(x) = \frac{1}{n} \sum_{i=1}^n h(x, X_i) - U_n(h)$
- $\widehat{\mathrm{Cov}}(X_1, X_{1+|k|}) = \frac{1}{n} \sum_{j=1}^{n-|k|} \hat{h}_1(X_j) \hat{h}_1(X_{j+|k|})$
-

$$\hat{\sigma}_\infty^2 = 4 \sum_{k \in \mathbb{Z}} K(\frac{k}{b}) \widehat{\mathrm{Cov}}(X_1, X_{1+|k|})$$

Functional CLT for U-Quantiles

assumptions:

- short range dependence
- continuity condition
- differentiability $t \mapsto P(h(X_i, \tilde{X}_i) \leq t)$
- no moment condition!

Theorem (Vogel, Wendler 2017)

$$\left(\frac{[nt]}{\sqrt{n}} \left(U^{-1}_{[nt]}(\alpha) - E[U^{-1}_{[nt]}(\alpha)] \right) \right)_{t \in [0,1]} \Rightarrow (\sigma_{\alpha,\infty} W(t))_{t \in [0,1]}$$

- long run variance: $\sigma^2_{\alpha,\infty}$
- consistent estimator $\hat{\sigma}^2_{\alpha,\infty}$ exists

Application to Change-Point Tests

scale estimators s_n:

- $\hat{\sigma}_n^2$, g_n: U-statistics
- Q_n^α: U-quantiles

$$V_n = \frac{1}{\sqrt{n\hat{\sigma}_{s_n}^2}} \max_{k=1,\dots,n} k\left|s_k - s_n\right| = \max_{k=1,\dots,n} \left|\frac{k}{\sqrt{n\hat{\sigma}_{s_n}^2}} s_k - \frac{k}{n}\frac{n}{\sqrt{n\hat{\sigma}_{s_n}^2}} s_n\right|$$

$$\Rightarrow \sup_{t\in[0,1]} |W(t) - tW(1)| =: \sup_{t\in[0,1]} |B(t)|$$

- continuous mapping theorem
- process B with $B(t) = W(t) - tW(1)$: Brownian Bridge
- Kolmogorov-Smirnov-distribution

Simulation Results

- independent random variables
- normal-, t_5- or t_3-distributed
- sample size $n = 120$
- bandwidth $b_n = 10$, quartic kernel K
- case 1: hypothesis (stationarity)
- case 2: jump of variance (factor $\lambda = 1.5$) at observation 60
- case 3: jump of variance (factor $\lambda = 2$) at observation 60

Simulation Results II

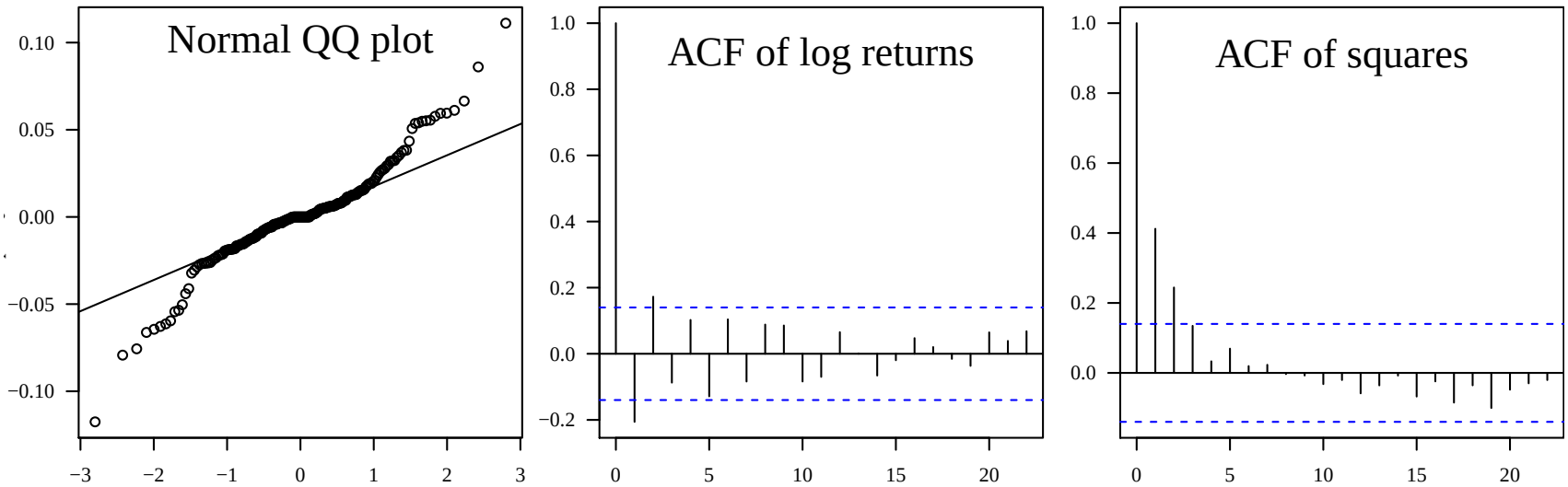
emirical rejection frequency, asymptotical size 0.05

	hypothesis	alternative $\lambda = 1.5$	alternative $\lambda = 2$
N(0,1)	0.03 / 0.03 / 0.04	0.43 / 0.57 / 0.49	0.79 / 0.94 / 0.84
t_5	0.02 / 0.02 / 0.04	0.18 / 0.34 / 0.28	0.43 / 0.75 / 0.63
t_3	0.04 / 0.02 / 0.08	0.08 / 0.22 / 0.20	0.18 / 0.51 / 0.45

$$\hat{\sigma}_n^2 \; / \; g_n \; / \; Q_n^{0.8}$$

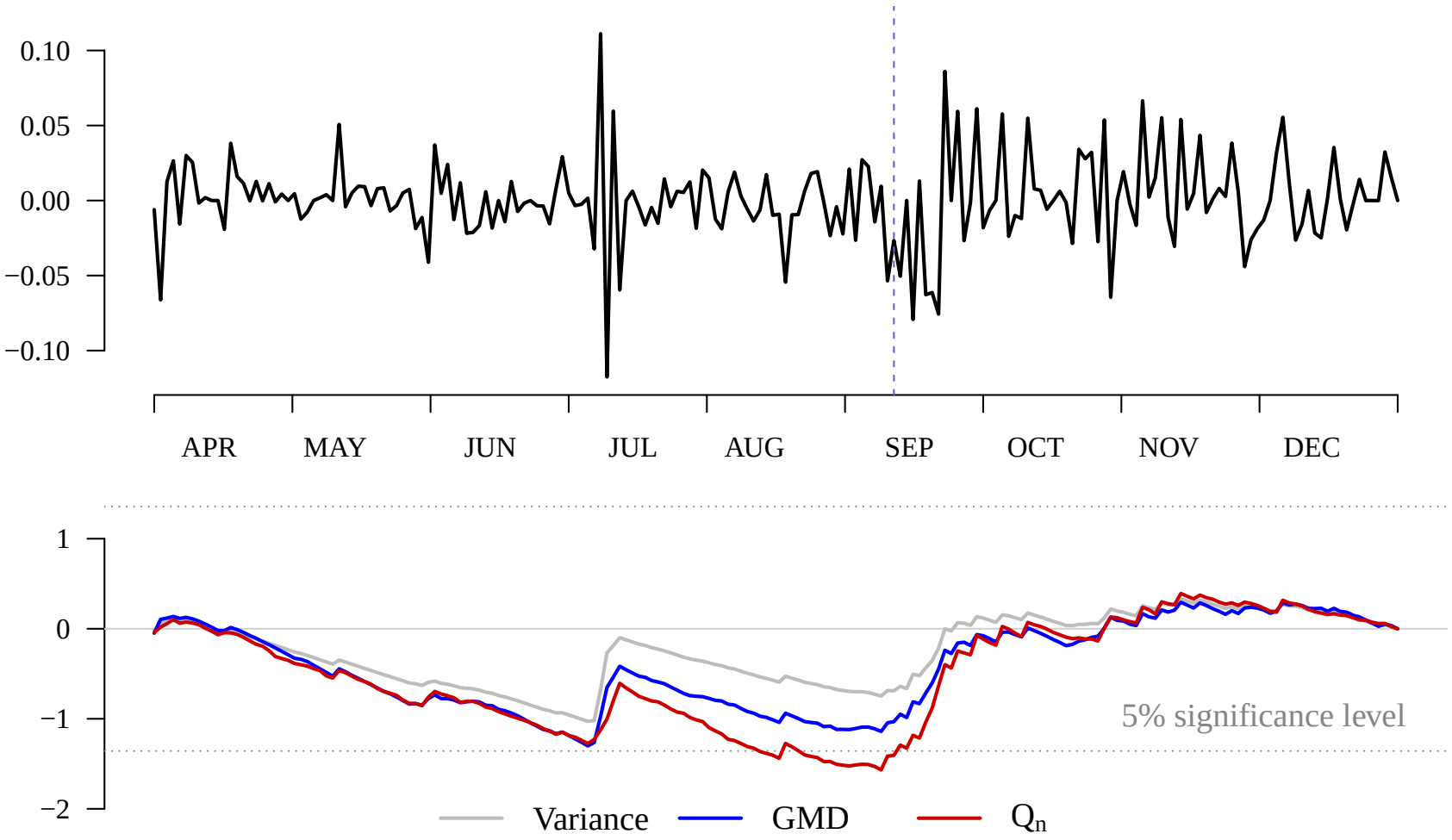
Data Example

- closing prices Volkswagen stock
- log returns
- april to december 2001



- dependence (GARCH-effect?)
- heavy tailed

Data Example II



Change in linear dependence

- bivariate time series $(X_n, Y_n)_{n \in \mathbb{N}}$
- hypothesis: stationarity
- alternative: $\exists k \in 1, \ldots, n-1 : \ \mathrm{Cov}(X_1, Y_1) = \ldots = \mathrm{Cov}(X_k, Y_k) \neq \mathrm{Cov}(X_{k+1}, Y_{k+1}), \ldots, \mathrm{Cov}(X_n, Y_n)$

non-robust test (Wied, Krämer, Dehling, 2012):
Pearson's correlation coefficient

$$\hat{\rho}_n = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\hat{\sigma}_X^2 \hat{\sigma}_Y^2}}$$

test statistic

$$T_n := \max_{1 \leq k \leq n} \frac{k}{\sqrt{n}} |\hat{\rho}_k - \hat{\rho}_n|$$

Robust Test: Kendall's τ

- Kendall's-τ (U-statistic)

$$\tau_n = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} \mathbb{1}_{\{(X_j - X_i)(Y_j - Y_i) > 0\}}$$

- estimates probability of concordance

$$\tau = P\left((X - X')(Y - Y') > 0\right)$$

Theorem (Dehling, Vogel, Wendler, Wied, 2017+)

$F_{X,Y}$ continuous, $(X_n, Y_n)_{n \in \mathbb{N}}$ P-NED, then

$$\max_{1<k<n} \frac{k}{\sqrt{n} 2 \hat{D}} |\tau_k - \tau_n| \to \sup_{0 \le t \le 1} |B(t)|$$

Simulation Results I

- i.i.d. data
- sample size $n = 500$
- at observation 250, the correlation coefficient jumps from 0.4 to ρ
- "correlation after jump": $\rho = 0.4$ (no jump), $\rho = 0.6,\ 0.8$
- different marginal distributions: normal, t_ν with degrees of freedom $\nu = 20, 5, 3$
- bandwidth and kernel for the variance estimation: $b_n = [2n^{1/3}]$, $\kappa(x) = (1 - x^2)^2 \mathbb{1}_{[0,1]}(x)$ (quartic kernel)
- 1000 samples for each parameter setting evaluated

Simulation Results II

empirical rejection frequency, asymptotical level 0.05

	ρ after jump:		
Verteilung	0.4	0.6	0.8
normal	0.04 / 0.05	0.70 / 0.65	1.00 / 1.00
t_{20}	0.04 / 0.04	0.65 / 0.63	1.00 / 1.00
t_5	0.04 / 0.04	0.41 / 0.55	0.95 / 1.00
t_3	0.06 / 0.03	0.25 / 0.52	0.69 / 1.00

Pearson's correlation coefficient / Kendall's τ

Change in Distribution

- $(Y_n)_{n\in\mathbb{N}}$ time series
- hypothesis: stationarity
- alternative: $\exists k: \; Y_1, \ldots, Y_k \sim F,\ Y_{k+1}, \ldots, Y_n \sim G$
- distribution functions $F \neq G$
  with $F(t) := P(Y_1 \leq t)$, $G(t) := P(Y_n \leq t)$

estimator: empirical distribution function

$$F_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{Y_i \leq t\}}$$

idea:

- F_n: mean of functions $\left(\mathbb{1}_{\{X_i \leq t\}}\right)_{t\in\mathbb{R}}$
- use CUSUM-statistic $T_n = \frac{1}{\sqrt{n}} \max\limits_{k=1,\ldots,n} k \left\| F_k - F_n \right\|$
- what norm $\|\cdot\|$?

Test Statistic

- Hilbert space H with inner product $\langle f, g \rangle := \int f(t)g(t)w(t)dt$
- norm: $\|f\| := \sqrt{\langle f, f \rangle}$
- F_n: mean of H-valued observations $X_i := \mathbb{1}_{\{Y_i \leq \cdot\}}$

Crámer-von Mises-type test statistic:

$$T_n^2 := \max_{k=1,\ldots,n} \frac{k^2}{n} \int \left(F_k(t) - F_n(t)\right)^2 w(t)dt$$

problem: unknown infinite-dimensional covariance parameter, so critical values unkown
solution: bootstrap

Simulation Results I

- sample size $n = 50, 100, 200$
  block length $p = 7, 8, 12$
- AR(1)-process: $X_t = 0.5X_{t-1} + \epsilon_t$
  $(\epsilon_t)_{t\in\mathbb{Z}}$ iid normal
- 1. case: hypotheses (stationarity)
- 2. case: at middle of sample: mean changes from $\mu = 0$ to $\mu = 1$
- 3. case: first half: $X_t^2 + X_t'^2$, second half: $4 - X_t^2 - X_t'^2$ (change in skewness)
- 1000 iterations for each scenario
- 1000 bootstrap replicates for each iteration

Simulation Results II

empirical rejections frequency, asymptotic size 0.05

observations/ block length	hypotheses	change of location	change of skewness
$n = 50, p = 7$	0.06/ 0.08	0.43 / 0.31	0.06 / 0.24
$n = 100, p = 8$	0.04/ 0.06	0.71 / 0.70	0.05 / 0.46
$n = 200, p = 12$	0.05/ 0.06	0.95 / 0.94	0.05 / 0.85

CUSUM-test / Crámer-von Mises-test

Summary

- test for changes in scale, correlation, distribution
- functional limit theorems for U-statistics and U-quantiles
- long run variance estimation or bootstrap

questions? remarks?

References

- B. Bucchia, M. Wendler (2017): Change-Point Detection and Bootstrap for Hilbert Space Valued Random Fields, Journal of Multivariate Analysis 155, 344-368.
- H. Dehling, O.Sh. Sharipov, M. Wendler (2015): Bootstrap for dependent Hilbert space-valued random variables with application to von Mises statistics, Journal of Multivariate Analysis 133, 200-215.
- H. Dehling, D. Vogel, M. Wendler, D. Wied (2017+): Testing for changes in Kendall's tau, to appear in Econometric Theory arXiv: 1203.4871.
- C. Gerstenberger, D. Vogel, M. Wendler (2016): Tests for scale changes based on pairwise differences. preprint arXiv:1611.04158.
- O.Sh. Sharipov, J. Tewes, M. Wendler (2016): Sequential block bootstrap in a Hilbert space with application to change point analysis, Canadian Journal of Statistics 44(3), 300-322.
- D. Vogel, M. Wendler (2017): Studentized sequential U-quantiles under dependence with applications to change-point analysis, Bernoulli 23(4b), 3114-3144.