SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 000

Analiza vremenskih nizova zasnovana na kompleksnim mrežama

Lovre Mrčela

Zagreb, lipanj 2017.

Umjesto ove stranice umetnite izvornik Vašeg rada.
Da bi ste uklonili ovu stranicu obrišite naredbu `\izvornik`*.*

SADRŽAJ

1. Uvod

Klasične metode statističke arbitraže uzimaju u obzir parove vrijednosnica čije cijene se ponašaju slično tijekom određenom vremenskog perioda. Sličnost se mjeri kointegracijom, korelacijom, ili nekom drugom mjerom, s ciljem pronalaska trenutka kada te cijene izlaze van statistički utvrđenog intervala visoke pouzdanosti. Takve prilike mogu se iskoristiti predviđanjem da će se cijene u idućem trenutku ponovno vratiti unutar intervala visoke pouzdanosti, te se u skladu s tim predviđanjem može provesti trgovanje.

In this paper, we propose a new method based on those predictions that are obtained by the statistical arbitrage method, using statistical measures as a proxy for describing the preference relations between pairs of assets. Next, a graph is formed based on those relations, so that mutual interaction of assets might be analyzed. This graph imposes a preference relation among the assets that are included in it. Finally, assets are sorted by preference and included into the portfolio. The idea of this method is to create a generalization of statistical arbitrage methods that is more robust and performs better when working with a larger number of assets by trying to take into account mutual interaction of assets.

2. Metode i koncepti

2.1. Statistička arbitraža zasnovana na standardnoj devijaciji

2.2. Relacija preferencije i funkcija korisnosti

Neka je Ω skup općenitih dobara. Relacija preferencije, označena sa $\succ$ i definirana nad $\Omega \times \Omega$, je strogi slabi uređaj koji odgovara načinu na koji ljudi preferiraju jednu vrijednosnicu u odnosu na drugu. Između dvaju dobara $a, b \in \Omega$ relacija može, ali i ne mora postojati. Primjerice, a može biti više preferirano u odnosu na b, ili b u odnosu na a, ali moguća je situacija gdje su oba dobra podjednako preferirane. U tom slučaju radi se o indiferentnosti između a i b, i to se označava kao $a \sim b$.

Ova relacija specifična je po tome što je $(\forall x, y, z \in \Omega)$:

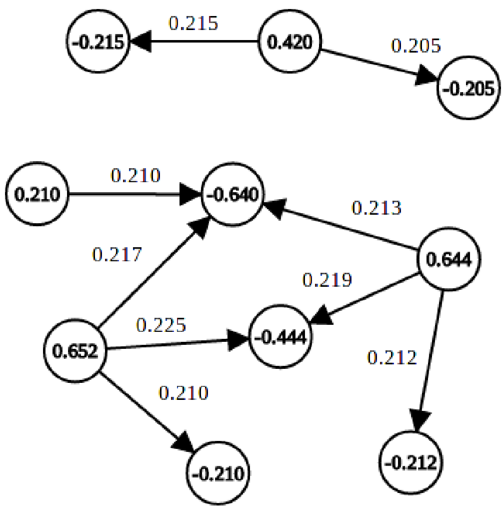
- *irrefleksivna*: $\neg(x \succ x)$ — za nijedno dobro ne vrijedi da je više preferirano od samog sebe,
- *asimetrična*: $x \succ y \Rightarrow \neg(y \succ x)$ — ako je x više preferirano od y, onda y nije više preferirano od x,
- *tranzitivna*: $x \succ y \wedge y \succ z \Rightarrow x \succ z$ — ako je x više preferirano od y, te y više preferirano od z, tada je i x više preferirano od z,
- *tranzitivna po indiferentnosti*: $x \sim y \wedge y \sim z \Rightarrow x \sim z$ — ako je x podjednako preferirano kao i y, te y podjednako preferirano kao i z, tada je i x podjednako preferirano kao i z.

Ovakva vrsta relacije prirodno opisuje odnose među različitim vrijenosnicama, npr. dionica A u nekom trenutku može biti više preferirana od dionice B. Razlog tome je što je za čovjeka lakše ocijeniti odnos (više, manje, jednako preferirano) između svakog para vrijednosnica, nego pridijeliti svakoj vrijednosnici individualnu mjeru preferencije, pogotovo ako se radi o velikom broju vrijednosnica. No ipak, u svrhu konstruiranja portfelja korisnije je posjedovati individualnu mjeru preferencije za svaku vrijednosnicu. Stoga je poželjno pronaći način da se iz relacije preferencije dobiju individualne mjere preferencije.

Funkcija korisnosti $U\colon \Omega \to \mathbb{R}$ je preslikavanje iz skupa dobara u skup realnih brojeva, na način da poredak preslikanih realnih brojeva odgovara poretku dobara prema individualnoj preferenciji, tj. vrijedi $\forall x, y \in \Omega, U(x) > U(y) \Leftrightarrow x \succ y$. Jedno takvo preslikavanje ostvaruje se korištenjem metode potencijala koja je opisana u poglavlju 2.4. Uz to što određuje poredak dobara prema individualnoj preferenciji, funkcija korisnosti također unosi i mjeru intenziteta preferencije prema nekoj vrijednosnici, što znači da nosi više informacija od same relacije.

2.3. Graf toka preferencija

Graf toka preferencija je težinski usmjereni graf, bez višestrukih bridova i petlji. Njegovi čvorovi predstavljaju vrijednosnice, usmjereni bridovi preferenciju jedne vrijednosnice nad drugom, a težine bridova odgovaraju jačini preferencije. Ukoliko između dva čvora nedostaje brid, smatra se da su pripadne vrijednosnice podjednako preferirane (indiferentnost u odlučivanju). Graf kao cjelina opisuje tok preferencija među vrijednosnicama. Primjer grafa prikazan je na slici 2.1.



Slika 2.1: Primjer grafa toka preferencija. Na bridovima su prikazane jačine preferencija jedne vrijednosnice u odnosu na drugu, a u čvorove su upisane izračunate individualne mjere preferencija za pripadnu vrijednosnicu.

Konstrukcija grafa temelji se na metodi statističke arbitraže. Brid koji ide od čvora i do čvora j s težinom $w_{i,j}$ je prisutan u grafu ako i samo ako su se dvije pripadne vrijednosnice i i j ponašale slično tijekom proteklog vremenskog perioda, ali su prema određenim statističkim mjerama trenutačno razdvojile. Težina $w_{i,j}$ opisuje magnitudu ovog razdvajanja. Detaljan opis korištenih statističkih mjera dan je u odjeljku 3.1.

Veze u ovom grafu na jedan način nameću relaciju preferencije među vrijednosnicama koje su opisane grafom, na način da brid koji ide iz čvora i u čvor j pokazuje da je vrijednosnica i više preferirana od vrijednosnice j. Kako ne graf sadrži petlje, tj. nijedna vrijednosnica nije u relaciji sama sa sobom, ispunjeno je svojstvo irrefleksivnosti; a kako nema višestrukih bridova među čvorovima vrijedi i svojstvo asimetričnosti. Ipak, problemi se pojavljuju kod prethodno spomenutih svojstava tranzitivnosti, i tranzitivnosti po indiferentnosti, koja ne moraju uvijek vrijediti za proizvoljno konstruiran graf. Poželjno bi bilo da graf nameće relaciju preferencije koja zadovoljava sva četiri prethodno navedena svojstva, no kada se radi o većem broju vrijednosnica, konstruiranje grafa koji posjeduje takva svojstva postaje nepraktično. Umjesto postizanja konzistentnosti sa svojstvima relacije preferencije, upotrijebljena je mjera konzistentnosti koja opisuje koliko je neki graf sličan sa svojom najbližom konzistentnom rekonstrukcijom, te se ona koristi kao dodatan parametar pri konstruiranju portfelja. Opis mjere konzistentnosti te način dobivanja najbliže konzistentne rekonstrukcije grafa dan je u odjeljku 2.4.

2.4. Metoda potencijala

From previously obtained graph it is possible to tell which pair of assets has the highest preference flow. However, it is not yet possible to directly tell which are the most or least preferrable assets, or obtain the measure of preference for individual assets. To calculate preferences for each node in the graph, we use the potential method[**?**]. The potential of a node corresponds to difference in amount of flow going in and out of the node.

For the observed graph $\mathcal{G}$, let there be a total of N nodes, and maximum of $E = \binom{N}{2}$ edges, in case of a complete graph. If $\mathcal{G}$ is not complete, we complete it by adding edges to it with weight 0 (direction doesn't matter), thus forming a complete graph $\mathcal{G}$ with $\binom{N}{2}$ edges.

Let $\boldsymbol{B}$ be the $E \times N$ incidence matrix of $\mathcal{G}$. Let $\boldsymbol{f}$ be $E \times 1$ vector that contains edge weights (i.e. preference flows). Order of the edges must be the same as order of the edges in $\boldsymbol{B}$. As mentioned before, in place of missing edges we simply put 0. Let $\boldsymbol{\phi}$ be $N \times 1$ vector that contains potentials of each node, in order that is the same as order of the nodes in $\boldsymbol{B}$.

Now, if $\mathcal{G}$ was consistent, then $\boldsymbol{B}$, $\boldsymbol{\phi}$, and $\boldsymbol{f}$ would satisfy the equation

$$\boldsymbol{B}\boldsymbol{\phi} = \boldsymbol{f}. \tag{2.1}$$

Equation (2.1) states that the difference between potential of any two nodes should result in weight of the edge between them. This is possible only for consistent graphs, and most of the time our graphs will be inconsistent. In that case we try to find an approximate solution $\boldsymbol{\phi}^*$ that minimizes the square error:

$$\begin{gathered}
\boldsymbol{\phi}^* = \arg\min_{\boldsymbol{\phi}} \left\{ \|\boldsymbol{B}\boldsymbol{\phi} - \boldsymbol{f}\|^2 \right\} \\
\Downarrow \\
\frac{\partial \left\|\boldsymbol{B}\boldsymbol{\phi}^* - \boldsymbol{f}\right\|^2}{\partial \boldsymbol{\phi}^*} = \boldsymbol{0}.
\end{gathered} \tag{2.2}$$

Solving (2.2) via commonly used techniques of matrix calculus brings us to the following equation:

$$\begin{aligned}
2\boldsymbol{B}^\top \left[\boldsymbol{B}\boldsymbol{\phi}^* - \boldsymbol{f}\right] &= \boldsymbol{0} \\
\boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{\phi}^* &= \boldsymbol{B}^\top \boldsymbol{f}.
\end{aligned} \tag{2.3}$$

Equation (2.4) determines $\boldsymbol{\phi}^*$ up to a constant (i.e. solution has one degree of freedom), so the following constraint is also included:

$$\boldsymbol{j}^\top \boldsymbol{\phi}^* = 0 \tag{2.4}$$

where $\boldsymbol{j}$ is vector of ones with same dimension as $\boldsymbol{\phi}^*$. This ensures an unique solution for which total amounts of positive and negative potential will be equal.

Joining the previous two equations together by adding (2.4) to each row in (2.4) results in:

$$\begin{aligned}
\boldsymbol{B}^\top \boldsymbol{B}\boldsymbol{\phi}^* + \boldsymbol{J}\boldsymbol{\phi}^* &= \boldsymbol{B}^\top \boldsymbol{f} \\
\left[\boldsymbol{B}^\top \boldsymbol{B} + \boldsymbol{J}\right] \boldsymbol{\phi}^* &= \boldsymbol{B}^\top \boldsymbol{f},
\end{aligned} \tag{2.5}$$

where $\boldsymbol{J}$ is a matrix of ones with same dimension as $\boldsymbol{B}^\top\boldsymbol{B}$. Finally, solving (2.5) for $\boldsymbol{\phi}^*$ gives us:

$$\boldsymbol{\phi}^* = \left[\boldsymbol{B}^\top \boldsymbol{B} + \boldsymbol{J}\right]^{-1} \boldsymbol{B}^\top \boldsymbol{f}. \tag{2.6}$$

Furthermore, the term $\left[\boldsymbol{B}^\top \boldsymbol{B} + \boldsymbol{J}\right]^{-1}$ in (2.6) can be simplified to $\frac{1}{N}\boldsymbol{I}$ due to $\boldsymbol{B}^\top \boldsymbol{B}$ being the Laplace matrix of a complete graph; thus, we can simplify (2.6) some more:

$$\boldsymbol{\phi}^* = \frac{1}{N}\boldsymbol{B}^\top \boldsymbol{f}, \tag{2.7}$$

to get as computationally optimal expression as possible.

Afterwards, we can calculate the consistent reconstruction $\boldsymbol{f}^*$ of preference flow by simply plugging back $\boldsymbol{\phi}^*$ into (2.1):

$$\boldsymbol{f}^* = \boldsymbol{B}\boldsymbol{\phi}^*. \tag{2.8}$$

The reconstructed preference flow $\boldsymbol{f}^*$ compared to the original preference flow $\boldsymbol{f}$ may even contain some new and/or lose some old edges. In addition, $\boldsymbol{B}$, $\boldsymbol{\phi}^*$, and $\boldsymbol{f}^*$ now describe a consistent graph $\mathcal{G}^*$. It is now possible to define a consistency measure κ as follows:

$$\kappa = \frac{\|\boldsymbol{f}^*\|}{\|\boldsymbol{f}\|}. \tag{2.9}$$

Equation (2.9) represents the cosine of the angle between $\boldsymbol{f}$ and $\boldsymbol{f}^*$ in the column space of matrix $\boldsymbol{B}$. Consistency measure κ tells us how consistent graph $\mathcal{G}$ was, compared to the $\mathcal{G}^*$. κ ranges from 0 to 1, with 0 meaning full inconsistency (virtually unreachable), and 1 meaning full consistency.

3. Algoritam

Let there be total of N assets in D days. Let price of asset i at the time step t be $a_i^{(t)}$, for $i \in [1, 2, \ldots, N]$ and $t \in [0, 1, \ldots, D-1]$. The log prices $b_i^{(t)}$, and log price differences $c_{i,j}^{(t)}$ between assets i and j are obtained as follows:

$$b_i^{(t)} = \log\left(a_i^{(t)}\right) \tag{3.1}$$

$$c_{i,j}^{(t)} = b_i^{(t)} - b_j^{(t)}, \tag{3.2}$$

and rolling means $m_{i,j}^{(t)}$ and standard deviations $d_{i,j}^{(t)}$ of log price differences over the past time window of size T are obtained as follows:

$$m_{i,j}^{(t)} = \frac{1}{T}\sum_{\tau=t-T}^{t-1} c_{i,j}^{(\tau)} \tag{3.3}$$

$$d_{i,j}^{(t)} = \sqrt{\frac{1}{T}\sum_{\tau=t-T}^{t-1}\left(c_{i,j}^{(\tau)} - m_{i,j}^{(t)}\right)^2}. \tag{3.4}$$

Note that in summation used in (3.3, 3.4) time step t was intentionally excluded, therefore summation goes only to $t-1$. We use these calculations as basis for creating the portfolio.

Note that period over which means and standard deviations of log price differences are calculated does not include time step t. Also note that calculating them separately for each time step t is rather computationally inefficient when dealing with rolling windows of data. Therefore, it is advisable to use a rolling algorithm as described in the appendix. On that note, $c_{i,j}^{(t)}$, $m_{i,j}^{(t)}$, and $d_{i,j}^{(t)}$ may be more efficiently stored if stored contiguously in memory as a matrix, using following coding scheme: a pair (i, j), where $i < j$, should be encoded to k as:

$$k = N \cdot (i-1) + j - 1 - i \cdot (i+1)/2\ , \tag{3.5}$$

and decoded from k as:

$$i = \left\lfloor N + 1/2 - \sqrt{(N+1/2)^2 - 2(N+k)}\right\rfloor, \tag{3.6}$$

$$j = k + i \cdot (i+1)/2 - N \cdot (i-1) + 1. \tag{3.7}$$

An example of proposed coding is shown on figure 3.1.

i/j	1	2	3	4	5
1	·	0	1	2	3
2	·	·	4	5	6
3	·	·	·	7	8
4	·	·	·	·	9
5	·	·	·	·	·

i	1	1	1	1	2	2	2	3	3	4
j	2	3	4	5	3	4	5	4	5	5
k	0	1	2	3	4	5	6	7	8	9

Slika 3.1: Example of the proposed coding scheme, for $N = 5$. A dot $(\cdot)$ indicates that that combination is not used.

3.1. Konstrukcija grafa toka preferencija

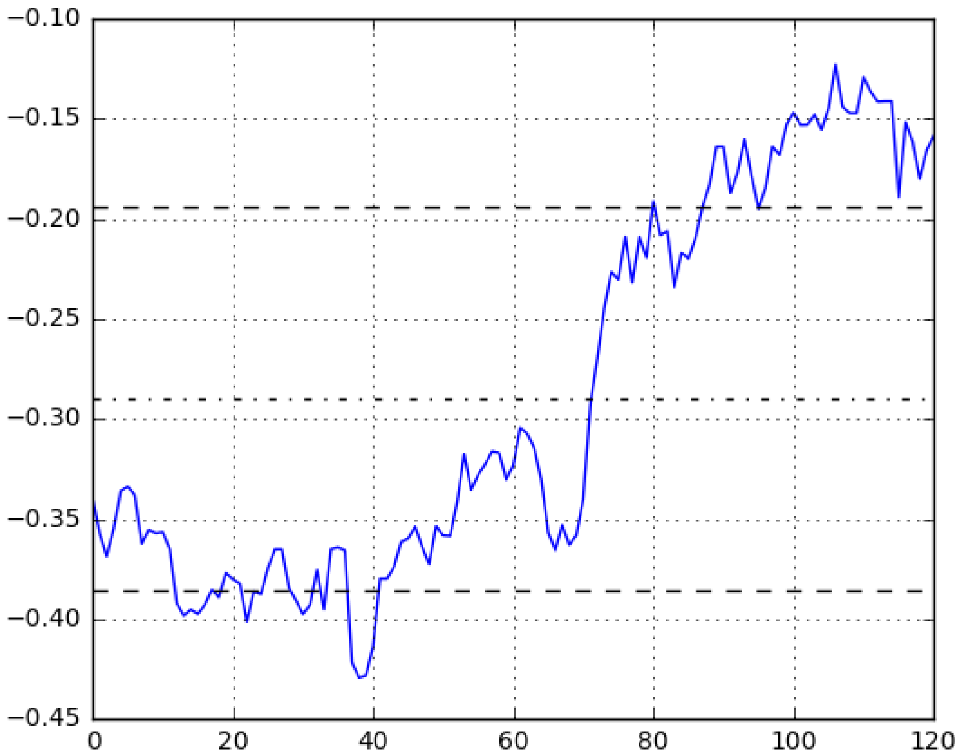
Using the obtained $c_{i,j}^{(t)}$, $m_{i,j}^{(t)}$, and $d_{i,j}^{(t)}$ it is now possible to create a graph of preference flow among assets for each time step t. Considering one time step t, we find all such pairs of assets (i, j) that satisfy:

$$\left| c_{i,j}^{(t)} - m_{i,j}^{(t)} \right| > \alpha \cdot d_{i,j}^{(t)}, \tag{3.8}$$

i.e. current log price difference is at least α deviations distant from mean value of the past time window. An illustration is shown on Fig. 3.2. Afterwards, for each observed pair (i, j) that exceeds the threshold we add into graph vertices i and j, with a weighed edge of weight $w_{i,j}^{(t)}$ going from i to j. Weight $w_{i,j}^{(t)}$ is obtained as:

$$w_{i,j}^{(t)} = \left(c_{i,j}^{(t)} - m_{i,j}^{(t)} \right) \Big/ d_{i,j}^{(t)} \, . \tag{3.9}$$

Thus it is possible to create a graph of preference flow for each time step $t \in [T, T+1, \ldots, D-1]$. At some time steps it is possible that the graph could be empty, if it is the case that no pair (i, j) satisfies (3.8). Setting lower values for parameter α yields denser graphs, and setting $\alpha = 0$ always yields complete graphs.



Slika 3.2: Log price difference between a pair of assets (i, j) during a period of $T + 1$ time steps, where $T = 120$. Dashed and dotted line represents mean value, and region between two dashed lines represents α standard deviation range from mean value, here $\alpha = 1$; both are calculated in the first T time steps. During the time step $T + 1$, log price difference goes over α standard deviations above mean value of past period of size T. This would mean that two of the assets i and j would be added to the graph of preference flow at the time step $T + 1$. Weight $w_{i,j}$ describes current deviation from mean value.

3.2. Konstrukcija portfelja prema grafu toka preferencija

We obtain preference for each asset via the potential method, as described earlier in ??. By obtaining the measure of preference for each asset it is possible to pick assets for the portfolio. The most preferred assets should be bought while the least preferred should be short-sold if possible.

Let $\boldsymbol{\phi}^{(t)} = \begin{bmatrix} \phi_1^{(t)} & \phi_2^{(t)} & \dots & \phi_N^{(t)} \end{bmatrix}$ denote vector of preferences of assets at time step t and $\phi_i^{(t)}$ denote the preference for asset i at time step t. When picking the assets for the portfolio we take into consideration the consistency measure κ as well. Lower values of κ suggest that we should diversify our portfolio by including some more assets in the order of preference, while higher values suggest that it is safe to do trading with smaller number of assets. Portfolio diversification might be seen as a strategy for protection from fundamental risks, e.g. risk of asset default.

The bound on the assets which will be taken into portfolio is proportional to the consistency measure κ. Depending on the nature of assets we may tune the consistency measure κ to be more or less inclined to diversification by transforming it to κ':

$$\kappa' = a + (1-a)\kappa^b, \tag{3.10}$$

where $a \in [0,1], b \in \mathbb{R}^+$. For default values of $a = 0$, $b = 1$, κ' equals κ.

For determining the assets that should be held in the portfolio at time step t, we find such assets i for which holds:

$$\phi_i^{(t)} \geq \kappa' \cdot \Phi, \tag{3.11}$$

where Φ is $\max_j \{|\phi_j|\}$. Likewise, for short-selling we choose those assets i for which holds:

$$\phi_i^{(t)} \leq -\kappa' \cdot \Phi. \tag{3.12}$$

For $a = 0$ diversification completely depends on consistency κ, while for $a = 1$ only the most preferred asset is held in the portfolio (no diversification). On the other hand, when $0 < b < 1$, algorithm is less inclined to diversification even when consistency is low, and when $b > 1$, algorithm is more inclined to diversification even when consistency is high.

4. Rezultati

Results were obtained by testing on a set of 203 stocks that were contiguously included in S&P 500 index from Jan 1st, 1980 thru Dec 31st, 2003, which includes 6261 trading days. A total of 20503 pairs of assets were probed for statistical arbitrage at each time step. Summary of results for various parameters is shown in the table 4.1. Best profit and Sharpe ratio has been achieved when using $\alpha = 0$.

Tablica 4.1: Results for $T = 60, \alpha = 0$.

Parameter:							
a	0.0			0.5			1.0
b	0.5	1.0	2.0	0.5	1.0	2.0	/
Average return (year)	0.95339	0.88967	0.84463	0.98336	0.95663	0.89704	1.00223
Volatility (year)	0.77042	0.76595	0.74077	0.77905	0.77054	0.76660	0.78363
Sharpe ratio (year)	1.23750	1.16152	1.14020	1.26225	1.24150	1.17015	1.27896
Profit:							
positive only	89.27624	88.89440	88.32548	89.58775	89.29840	89.04414	89.55020
negative only	-58.37779	-59.03715	-58.41396	-58.24852	-58.32385	-59.02220	-58.05316
total	30.89846	29.85725	29.91152	31.33923	30.97456	30.02195	31.49704
positive to negative ratio	1.52928	1.50574	1.51206	1.53803	1.53108	1.50866	1.54256
Average accuracy	0.36485	0.39276	0.43413	0.34902	0.36458	0.39145	0.33241
Average turnover ratio	0.59976	0.64224	0.73597	0.57585	0.59947	0.64089	0.55112
Actual profit, if transaction cost were 0.1%	23.46019	21.89215	20.78402	24.19757	23.53996	22.07361	24.66204

5. Zaključak

Algorithm works on pairs of assets, looking for those deviations which are uncommon, so generally it is expected to perform better where there is larger number of assets as more deviations will be discovered. It adapts to the inconsistence of preferences by picking variable number of assets into the portfolio.

Literatura

[1] T. Oetiket, H. Partl, Hyna, and E. Schlegl. *The not-so-short introduction to Latex*. URL `http://www.ctan.org/tex-archive/info/lshort/english/lshort.pdf`.

Analiza vremenskih nizova zasnovana na kompleksnim mrežama

Sažetak

Sažetak na hrvatskom jeziku.

Ključne riječi: Ključne riječi, odvojene zarezima.

Title

Abstract

Abstract.

Keywords: Keywords.