УДК 530.182, 51-73

ДИАГНОСТИКА И КОРРЕКЦИЯ СИСТЕМАТИЧЕСКОЙ ОШИБКИ ПРИ ОЦЕНКЕ ЭНТРОПИИ ПЕРЕНОСА МЕТОДОМ K-БЛИЖАЙШИХ СОСЕДЕЙ

А. С. Землянников[1], *И. В. Сысоев*[1,2]

[1]Саратовский государственный университет имени Н. Г. Чернышевского

[2]Саратовский филиал Института радиотехники и электроники имени В. А. Котельникова РАН

Энтропия переноса широко используется для определения направленной связанности колебательных систем по их наблюдаемым временным рядам. При оценке энтропии переноса между связанными нелинейными системами методом K-ближайших соседей обнаружена систематическая ошибка. Предложен способ уменьшения данной ошибки: с увеличением номера соседа систематическая ошибка уменьшается. Показана возможность диагностики систематической ошибки, имея два набора измерений. Полученные результаты позволяют улучшить чувствительность и специфичность метода для нелинейных систем при малых уровнях связи.

Ключевые слова: Временные ряды, анализ связанности, энтропия переноса, нелинейные системы.

Введение

Задача определения характера связи между двумя системами по временным реализациям наблюдаемых величин возникает в различных приложениях. Сложно выявлять слабую связь и определять ее направленность, особенно в случае если объем исходных данных ограничен. Среди теоретико-информационных методов оценки связей применяется энтропия переноса [1]. Классическим подходом при оценке данной меры является её реализация с помощью разбиения фазового объема на бины – ячейки фиксированного размера [2]. Существуют также другие методы оценки энтропии переноса, например, с помощью ядерной оценки плотности [3, 4], улучшенного разбиения с помощью алгоритма Darbellay–Vajda [5], корреляционных сумм [6], энтропии Реньё [8], метода K-ближайших соседей [7, 8]. Последний подход представляется наиболее перспективным в том числе благодаря тому, что предъявляет существенно меньшие требования к объёму экспериментальной выборки. В большинстве работ в качестве эталонных моделей для исследования возможностей предполагаемых методов используют линейные системы [9–14]. Целью данной работы

является проверка определения направленности связи с помощью энтропии переноса, реализованной методом K-ближайших соседей, для эталонных нелинейных систем и выявление новых возможных проблем, которые для линейных систем не наблюдались.

1. Описание метода

Метод оценки плотностей распределения на основе определения ближайших соседей был предложен в работе [7] для расчёта взаимной информации, но может быть легко обобщён на случай энтропии переноса, как сделано в [12,13]. Пусть есть два временных ряда: ряд $\{x_i\}_{i=1}^N$ от системы X и ряд $\{y_i\}_{i=1}^N$ от системы Y, где i – дискретное время, N – длина временного ряда. Уменьшение неопределённости следующего значения y_{i+1} за счёт учёта x_i называется энтропией переноса [1] и выражается через условные энтропии Шеннона следующим образом:

$$TE_{X\to Y} = H(Y_{i+1}|Y_i) - H(Y_{i+1}|Y_i, X_i), \tag{1}$$

где ряд для сигнала Y_{i+1} получается из временного ряда $\{y_i\}$ путём сдвига на единицу вперёд в дискретном времени. Для расчётов удобнее перейти от условных энтропий к совместным

$$\begin{aligned} H(Y_{i+1}|Y_i) &= H(Y_{i+1}, Y_i) - H(Y_i), \\ H(Y_{i+1}|Y_i, X_i) &= H(Y_{i+1}, Y_i, X_i) - H(Y_i, X_i). \end{aligned} \tag{2}$$

Тогда

$$TE_{X\to Y} = H(Y_{i+1}, Y_i) - H(Y_i) - H(Y_{i+1}, Y_i, X_i) + H(Y_i, X_i). \tag{3}$$

Если ввести расстояние между трёхмерными векторами в пространстве (Y_{i+1}, Y_i, X_i) как максимум из модулей расстояний по координатам, то можно воспользоваться оценкой энтропии по Козаченко–Леоненко [15]

$$d(i,j) = \max\left(|y_{i+1} - y_{j+1}|, |y_i - y_j|, |x_i - x_j|\right). \tag{4}$$

Тогда, по аналогии с оценкой функции взаимной информации методом K-ближайших соседей [7], индивидуальные и совместные энтропии в (3) выражаются в трёхмерном случае следующим образом:

$$\begin{aligned} H(Y_i) &= \psi(N) - \langle\psi(n_{Y_i}(i)+1)\rangle_i + \langle\log\varepsilon(i)\rangle_i, \\ H(Y_{i+1}, Y_i) &= \psi(N) - \left\langle\psi(n_{Y_{i+1},Y_i}(i)+1)\right\rangle_i + 2\langle\log\varepsilon(i)\rangle_i, \\ H(Y_i, X_i) &= \psi(N) - \langle\psi(n_{Y_i,X_i}(i)+1)\rangle_i + 2\langle\log\varepsilon(i)\rangle_i, \\ H(Y_{i+1}, Y_i, X_i) &= \psi(N) - \psi(K) + 3\langle\log\varepsilon(i)\rangle_i, \end{aligned} \tag{5}$$

где $\psi(n)$ – дигамма-функция; K – номер соседа; $\varepsilon(i)/2$ – расстояние от i-й точки в трёхмерном пространстве (Y_{i+1}, Y_i, X_i) до K-го ближайшего соседа, рассчитанное по формуле (4); $n_{Y_i}(i)$ – количество элементов ряда Y_i, расстояния которых до точки y_i строго меньше $\varepsilon(i)/2$; $n_{Y_{i+1},Y_i}(i)$ – количество точек из двумерного пространства

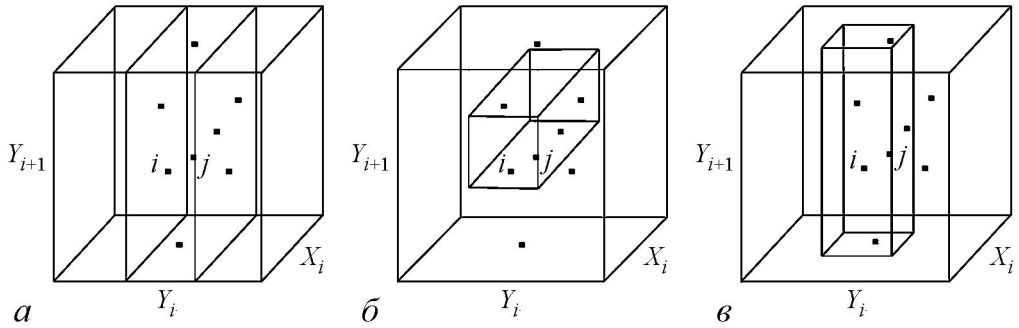


Рис. 1. Иллюстрация метода K-ближайших соседей. Изображены подпространства трёхмерного пространства (Y_{i+1}, Y_i, X_i), в которых расположены ближайшие соседи: *а* – соседи по оси Y_i, *б* – соседи по двум осям (Y_{i+1}, Y_i) и *в* – соседи по двум осям (Y_i, X_i)

(Y_{i+1}, Y_i), расстояния которых до точки (y_{i+1}, y_i) точно меньше $\varepsilon(i)/2$; аналогично для $n_{Y_i,X_i}(i)$. Расчёт числа соседей в различных сечениях пространства (Y_{i+1}, Y_i, X_i) иллюстрирует рис. 1.

Подставляя (5) в (3), получаем окончательную формулу энтропии переноса

$$TE_{X\to Y} = \psi(K) + \left\langle \psi(n_{Y_i}+1) - \psi(n_{Y_{i+1},Y_i}+1) - \psi \right\rangle_i . \tag{6}$$

2. Объект и методика исследования. Результаты

В качестве объекта исследования была выбрана нелинейная эталонная система двух однонаправленно связанных обобщённых отображений Эно (7). Логистическое отображение и его обобщения – отображение Эно и обобщённое отображение Эно – очень популярны как базовая модель нелинейной динамики, демонстрирующая сложное поведение при достаточно простом операторе эволюции [16]. Рассмотрение обобщённого отображения Эно важно тем, что варьируя m можно изменять размерность исходного объекта.

$$\begin{aligned} x_i &= 1 - \alpha_1 x_{i-1}^2 - \beta_1 x_{i-m} + \xi_i, \\ y_i &= 1 - \alpha_2 y_{i-1}^2 - \beta_2 y_{i-m} + \gamma x_i + \eta_i. \end{aligned} \tag{7}$$

Здесь γ – коэффициент связи; параметры $\alpha_1, \alpha_2, \beta_1, \beta_2$ подобраны так, чтобы в автономных системах наблюдался режим детерминированного хаоса; ξ_i, η_i – динамический шум с нулевым средним и среднеквадратичным отклонением 0.001. В численном эксперименте коэффициент связи варьировался в диапазоне $[0;\ 0.07]$ с шагом 0.01. Для каждого коэффициента связи генерировался ансамбль из 20 временных рядов $\{x_i\}_{i=1}^N$ и $\{y_i\}_{i=1}^N$, длина временного ряда составляла $N = 10000$ точек. Для каждой реализации рассчитывалась энтропия переноса методом K-ближайших соседей в заведомо верную сторону $X \to Y$. Исследование проводилось для систем с первого (логистическое отображение) по пятый порядок включительно, а также при $N = 1000$.

Зависимость оценки энтропии переноса $TE_{X\to Y}$ при номере соседа $K = 1$ от коэффициента связи γ представлена на рис. 2, *а*. Различные кривые соответствуют вариантам отображения (7) при различном m. Видно, что с ростом коэффициента связи оценка растёт, но при сравнительно малых γ наблюдается систематическая

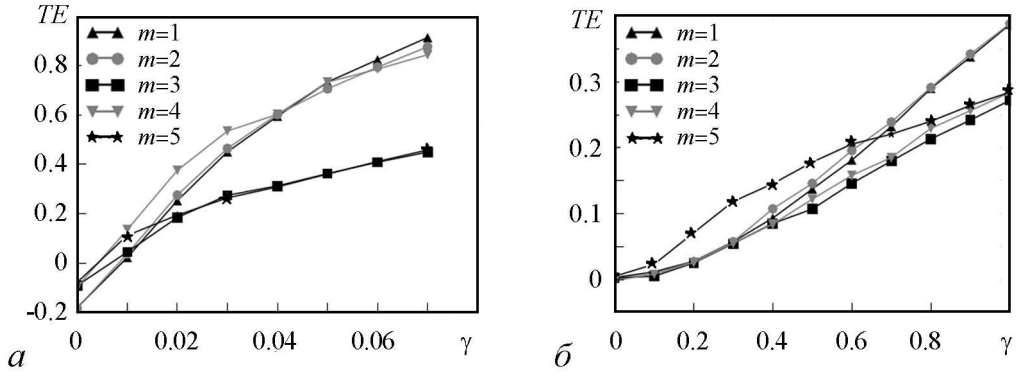


Рис. 2. *а* – график зависимости оценки энтропии переноса от коэффициента связи в заведомо верную сторону $TE_{X\to Y}(\gamma)$ для однонаправленно связанных обобщённых отображений Эно при различных m: $m = 1$ – кривая соответствует логистическому отображению; $m = 2$ – обычному отображению Эно; $m = 3$, $m = 4$ и $m = 5$ – соответствуют обобщённому отображению Эно; $N = 10000$. *б* – график зависимости оценки энтропии переноса от коэффициента связи в заведомо верную сторону $TE_{X\to Y}(\gamma)$ для однонаправленно связанных процессов авторегрессии первых пяти порядков (от $m = 1$ до $m = 5$), $N = 10000$

ошибка – полученные значения оказываются отрицательными, чего не может быть по определению энтропии переноса. Такого эффекта нет в случае линейных систем, например, для однонаправленно связанных процессов авторегрессии (рис. 2, *б*). Он также не наблюдался в работах [12, 13], что обусловлено линейностью рассмотренных там систем. Наличие систематической ошибки при оценке энтропии переноса можно исправить, увеличив длину реализации N. Однако данный способ часто неприменим на практике в силу ограниченности объёма экспериментальных данных или существенной нестационарности рассматриваемых сигналов.

Расчёт зависимости $TE_{Y\to X}$ в заведомо ложную сторону проводился во всех рассмотренных случаях. Для линейных процессов авторегрессии среднее значение $TE_{Y\to X}$ статистически (по ансамблю из 20 реализаций на уровне значимости 0.05) не отличается от нуля. Для связанных отображений Эно наблюдается систематическая ошибка – значения $TE_{Y\to X} < 0$. Отличие от нуля статистически значимо во многих случаях, то есть имеет место та же ошибка, что и при нулевой связи.

Можно предложить иной способ уменьшения ошибки, в котором вычисления проводятся при разных номерах соседа K. На рис. 3, *а* видно, что при увеличении номера соседа систематическая ошибка уменьшается и стремится к нулю. Численные эксперименты показали, что увеличение номера соседа при расчёте энтропии переноса методом K-ближайших соседей ведёт к уменьшению систематической ошибки при малых величинах γ. К сожалению, одновременно происходит занижение оценок при больших γ, что обусловлено излишним усреднением в слишком большой окрестности и может быть исправлено только увеличением длины ряда. Если целью исследования является обнаружение факта наличия связи и её направления, а не получение точных количественных мер силы связи, то недооценивание величины TE не является критическим, но тем не менее снижается чувствительность подхода. Поэтому следует искать некое компромиссное значение K, для чего необходимо уметь диагностировать имеющуюся проблему и не использовать слишком большие K в случае, когда в этом нет необходимости.

Выявить наличие нулевого сдвига и подобрать оптимальный номер соседа можно, рассчитав энтропию переноса между рядами, которые явно не связаны. Для экспериментальных данных это могут быть два набора измерений. Так, если мы име-

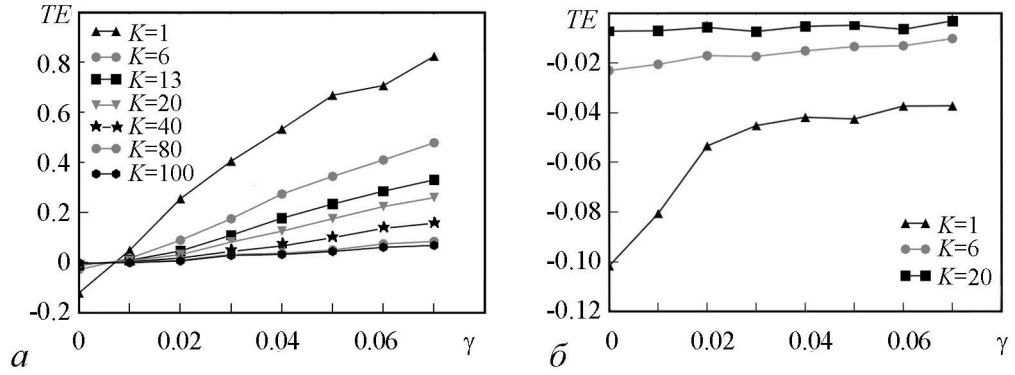


Рис. 3. *а* – график зависимости оценки энтропии переноса в заведомо верную сторону $TE_{X\to Y}(\gamma)$ от коэффициента связи однонаправленно связанных отображений Эно 4-го порядка ($m = 4$) для разных номеров соседа K; *б* – графики зависимости оценки энтропии переноса от коэффициента связи $TE_{X\to Y'}(\gamma)$ для разных номеров соседа K

ем два набора данных: первый – $\{x_i\}_{i=1}^N$ и $\{y_i\}_{i=1}^N$ и второй – $\{x'_i\}_{i=1}^N$ и $\{y'_i\}_{i=1}^N$, где в действительности $X \to Y$ и $X' \to Y'$, то оценка энтропии переноса $X \to Y'$ (и, аналогично, $Y \to X'$) должна быть равна нулю. На рис. 3, *б* представлены результаты оценивания энтропии переноса $TE_{X\to Y'}$ для двух сгенерированных с разными начальными условиями систем вида (7). Видно что при $K = 1$ имеется значительная систематической ошибка, которая уменьшается с увеличением номера соседа (кривые при $K = 6$ и $K = 20$).

Заключение

Показана применимость метода K-ближайших соседей при оценке энтропии переноса для диагностики слабой направленной связи между двумя нелинейными колебательными системами по временным рядам систем различной размерности. На примере обобщённого отображения Эно для нелинейных систем выявлена систематическая ошибка, проявляющаяся при сравнительно малых значениях коэффициента связи при оценке энтропии переноса в направлении, в котором связь действительно присутствует. Данная ошибка может вести к тому, что метод не сможет диагностировать наличие слабой связи. Предложен практический способ её уменьшения – определено, что с ростом номера соседа K систематическая ошибка уменьшается. Однако одновременно снижается чувствительность метода: начиная с некоторого значения коэффициента связи оценка энтропии переноса существенно занижается с ростом K – происходит недооценивание. Чтобы избежать возможного недооценивания или минимизировать его последствия, предложена методика диагностики наличия систематической ошибки для произвольных нелинейных систем на основе двух наборов измерений, позволяющая подобрать оптимальный компромиссный номер соседа K и тем самым минимизировать потери в чувствительности. Представленные результаты будут полезны для оценки связи по коротким временным рядам нелинейных систем, что важно в часто встречающихся на практике условиях существенной нестационарности сигналов и дефиците данных.

Работа выполнена при поддержке Российского научного фонда, грант 14-12-00291.

Библиографический список

1. *Schreiber T.* Measuring information transfer // Phys. Rev. Lett. 2000. Vol. 85, № 2. P. 461.
2. *Moddemeijer R.* On estimation of entropy and mutual information of continuous distributions // Signal Processing. 1989. Vol. 16, № 3. P. 233.
3. *Lee J., Nemati S., Silva I., Edwards B.-A., Butler J.-P., Malhotra A.* Transfer entropy estimation and directional coupling change detection in biomedical time series // BioMedical Engineering OnLine. 2012. 11:19.
4. *Silverman B.* Density estimation for statistics and data analysis. London: Chapman and Hall, 1986. 175 p.
5. *Darbellay A.G., Vajda I.* Estimation of the information by an adaptive partitioning of the observation space // IEEE Transactions on Information Theory. 1999. Vol. 45, № 4. P. 1315.
6. *Kugiumtzis D.* Transfer entropy on rank vectors // Journal of Nonlinear Systems and Applications. 2012. Vol. 3, № 2. P. 73.
7. *Kraskov A., Stögbauer H., Grassberger P.* Estimating mutual information // Phys. Rev. E. 2004. 69: 66138.
8. *Jizba P., Kleinert H., Shefaat M.* Renyi's information transfer between financial time series // Physica A. 2012. Vol. 391. P. 2971.
9. *Gomez-Herrero G., Wu W., Rutanen K., Soriano M.C., Pipa G., Vicente R.* Assessing coupling dynamics from an ensemble of time series // Arxiv preprint arXiv:1008.0539v1. 2010.
10. *Kaiser A., Schreiber T.* Information transfer in continuous process // Physica D: Nonlinear Phenomena. 2002. Vol. 166, № 1–2.
11. *Hahs D.W., Pethel S.D.* Transfer entropy for coupled autoregressive processes // Entropy. 2003. Vol. 15(3). P. 767.
12. *Lindner M., Vicente R., Priesemann V., Wibral M.* TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy // BMC Neuroscience. 2011. 12:119.
13. *Wibral M., Pampu N., Priesemann V., Siebenhuhner F., Seiwert H., Lindner M., Lizier J.T., Vicente R.* Measuring information-transfer delays // PLoS One. 2013. Vol. 8(2):e55809.
14. *Smirnov D.A.* Spurious causalities with transfer entropy // Phys. Rev. E. 2013. Vol. 87. 042917.
15. *Козаченко Л.Ф., Леоненко Н.Н.* О статистической оценке энтропии случайного вектора // Проблемы передачи информации. 1987. Т. 23:2. P. 9.
16. *Кузнецов С.П.* Динамический хаос. М.: Физматлит, 2001. 296 с.

References

1. *Schreiber T.* Measuring information transfer // Phys. Rev. Lett. 2000. Vol. 85, № 2. P. 461.
2. *Moddemeijer R.* On estimation of entropy and mutual information of continuous distributions // Signal Processing. 1989. Vol. 16, № 3. P. 233.

3. *Lee J., Nemati S., Silva I., Edwards B.-A., Butler J.-P., Malhotra A.* Transfer entropy estimation and directional coupling change detection in biomedical time series // BioMedical Engineering OnLine. 2012. 11:19.
4. *Silverman B.* Density estimation for statistics and data analysis. London: Chapman and Hall, 1986. 175 p.
5. *Darbellay A.G., Vajda I.* Estimation of the information by an adaptive partitioning of the observation space // IEEE Transactions on Information Theory. 1999. Vol. 45, № 4. P. 1315.
6. *Kugiumtzis D.* Transfer entropy on rank vectors // Journal of Nonlinear Systems and Applications. 2012. Vol. 3, № 2. P. 73.
7. *Kraskov A., Stögbauer H., Grassberger P.* Estimating mutual information // Phys. Rev. E. 2004. 69: 66138.
8. *Jizba P., Kleinert H., Shefaat M.* Renyi's information transfer between financial time series // Physica A. 2012. Vol. 391. P. 2971.
9. *Gomez-Herrero G., Wu W., Rutanen K., Soriano M.C., Pipa G., Vicente R.* Assessing coupling dynamics from an ensemble of time series // Arxiv preprint arXiv:1008.0539v1. 2010.
10. *Kaiser A., Schreiber T.* Information transfer in continuous process // Physica D: Nonlinear Phenomena. 2002. Vol. 166, № 1–2.
11. *Hahs D.W., Pethel S.D.* Transfer entropy for coupled autoregressive processes // Entropy. 2003. Vol. 15(3). P. 767.
12. *Lindner M., Vicente R., Priesemann V., Wibral M.* TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy // BMC Neuroscience. 2011. 12:119.
13. *Wibral M., Pampu N., Priesemann V., Siebenhuhner F., Seiwert H., Lindner M., Lizier J.T., Vicente R.* Measuring information-transfer delays // PLoS One. 2013. Vol. 8(2):e55809.
14. *Smirnov D.A.* Spurious causalities with transfer entropy // Phys. Rev. E. 2013. Vol. 87. 042917.
15. *Kozachenko L.F., Leonenko N.N.* // Probl. Inf. Transm. 1987. Vol. 23. P. 95.
16. *Kuznetsov S.P.* Dynamical chaos. M.: Fizmatlit, 2001. 296 s. (In Russian).

Поступила в редакцию *4.06.2015*
После доработки *3.09.2015*

DIAGNOSTICS AND CORRECTION OF SYSTEMATIC ERROR WHILE ESTIMATING TRANSFER ENTROPY WITH K-NEAREST NEIGHBOURS METHOD

A. S. Zemlyannikov[1], *I. V. Sysoev*[1,2]

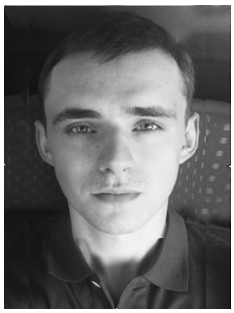
[1]Saratov State University

[2]Kotel'nikov Institute of Radio-engineering and Electronics of RAS, Saratov Branch

Transfer entropy is widely used to detect the directed coupling in oscillatory systems from their observed time series. The systematic error is detected, while estimating transfer entropy between nonlinear systems with K-nearest neighbours method. The way

to minimize this error is suggested: the error is decreasing with increase of the neighbour number. The possibility to detect the systematic error is shown using two sets of measured data. The achieved results make possible to rise the method sensitivity and specificity for weakly coupled nonlinear systems.

Keywords: Time series, coupling analysis, transfer entropy, nonlinear systems.



Землянников Андрей Сергеевич – родился в Саратове (1989), окончил Саратовский государственный университет имени Н.Г. Чернышевского. В настоящее время – аспирант кафедры динамического моделирования и биомедицинской инженерии. Участвовал в IX Всероссийской научной конференции молодых учёных «Наноэлектроника, нанофотоника и нелинейная физика». Работает инженером по медицинскому оборудованию в ГУЗ «Областной госпиталь для ветеранов войн», Саратов.

410012 Саратов, ул. Астраханская, 83
Саратовский государственный университет имени Н. Г. Чернышевского
E-mail: a89097z@yandex.ru

Сысоев Илья Вячеславович – родился в Саратове (1983), окончил факультет нелинейных процессов СГУ (2004), защитил диссертацию на соискание учёной степени кандидата физико-математических наук (2007). Доцент базовой кафедры динамического моделирования и биомедицинской инженерии, ответственный секретарь редакционной коллегии журнала «Известия вузов. ПНД». Научные интересы – исследование сигналов биологической природы методами нелинейной динамики, исследование эффективности и модернизация подходов к анализу сигналов. Автор более 40 публикаций.

410012 Саратов, Астраханская, 83
Саратовский государственный университет имени Н.Г. Чернышевского
410019 Саратов, ул. Зеленая, 38
Саратовский филиал Института радиотехники и электроники РАН
E-mail: ivssci@gmail.com