УДК 519. 2

О НЕКОТОРЫХ РЕЗУЛЬТАТАХ ИССЛЕДОВАНИЙ ПО МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ И ЕЕ ПРИЛОЖЕНИЯМ

Ю. С. Харин

Белорусский государственный университет, 4,
220030, г. Минск, Беларусь, kharin@bsu.by

Статья содержит краткое описание полученных научных результатов исследований по математической и прикладной статистике, анализу данных и криптологии со ссылками на основные научные публикации.

Ключевые слова: *статистическая оценка; решающее правило; асимптотический анализ; прогнозирование; робастность; риск.*

1. Студенческие и аспирантские исследования в Томском государственном университете

Начало исследований инициировано моим научным руководителем профессором Геннадием Алексеевичем Медведевым [1], который в весеннем семестре моего обучения на 2-м курсе пригласил для решения задачи укрупнения состояний стохастического графа с большим числом вершин с целью нахождения стационарного распределения вероятностей. На 3-м курсе я занимался задачей оптимизации дисциплины управления m лифтами в n-этажном здании на основе стохастических графов.

Основной задачей, поставленной мне в этот период, была следующая вероятностно-статистическая задача, возникшая из прикладной НИР в области обработки и распознавания сигналов. На вероятностном пространстве (Ω, F, P) и временной области $[0,T]$ определены: узкополосный случайный процесс (сообщение) $m(t)$, случайный процесс аддитивной помехи $\xi(t)$ и зарегистрированный сигнал:

$$x(t) = f_i\left(m(t);\theta_i\right) + \xi(t),\ t \in [0,T], \qquad (1)$$

где i – номер одного из $L \geq 2$ возможных типов преобразования (кодирования) сообщения $m(t)$. Задача состоит в том, чтобы при неизвестных значениях параметров сигнала $\theta_i \in \Theta_i \subseteq R^m$ по дискретным наблюдениям (с шагом дискретизации Δ):

$$\left\{x_j = x(j\Delta):\ j = 0,1,\ldots,n\right\},\ n = \left[\frac{T}{\Delta}\right],$$

оценить истинный номер $i^0 \in \{1,\ldots,L\}$ класса сигнала, т. е. решить задачу распознавания образов. Для решения этой задачи удалось разработать метод на основе максимальных инвариантов групп преобразований, порожденных изменением параметров $\{\theta_i \in \Theta_i\}$ [2 – 4]. Результаты решения этой задачи вошли в кандидатскую диссертацию по физико-математическим наукам, защищенную досрочно в сентябре 1974 года, когда Г.А. Медведев уже переехал в г. Минск и создал в БГУ кафедру теории вероятностей и математической статистики (ТВ и МС).

2. Начало научных исследований в БГУ

Впервые я появился в Минске в феврале 1975 г. по приглашению Г.А. Медведева и познакомился с кафедрой ТВ и МС, которая тогда еще размещалась совместно с кафедрой методов оптимального управления. В весеннем семестре 1975–1976 учебного года был приглашен на повышение квалификации в БГУ, и одновременно вел учебные занятия по специальным дисциплинам на кафедре ТВ и МС. По приглашению ректората БГУ был избран по конкурсу и с 1 сентября 1976 г. начал работать в должности доцента кафедры ТВ и МС. Исследования в период 1975–1978 г.г. продолжались в области статистического распознавания образов [5 – 7], но пришлось заниматься и компьютерным моделированием автотранспортного предприятия методами теории массового обслуживания [8] в рамках хоздоговорной НИР с МАЗом.

3. Асимптотический анализ риска статистической классификации

Многие прикладные задачи математически формализуются в виде следующей задачи статистической классификации. Пусть на (Ω, F, P) определены $L \geq 2$ N-мерных случайных векторов:

$$X_i = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{iN} \end{pmatrix} \in R^N,\ i = 1,\ldots,L,$$

причем X_i имеет абсолютно непрерывное распределение вероятностей $p\left(x;\ \theta_i^0\right)$, $\theta_i^0 \in R^m$, – истинное значение вектора параметров плотности

$p(\cdot)\in\mathcal{P}$, причем среди $\left\{\theta_i^0 : i=1,\ldots,L\right\}$ нет совпадающих. Эти истинные значения параметров $\left\{\theta_i^0\right\}$, определяющие L классов случайных наблюдений, не известны и оцениваются по классифицированной обучающей выборке $A=\bigcup_{i=1}^{L}A_i$, где $A_i=\left\{X_i^{(1)},\ldots,X_i^{(n_i)}\right\}$ – обучающая случайная выборка объема n_i из i-го класса. Пусть $\hat{\theta}_i\in R^m$ – оценка максимального правдоподобия (или более общая оценка минимального контраста) параметра θ_i^0 по выборке A_i,

$$r=\mathrm{E}\left\{w\left(\nu,d_0\left(X_\nu;\left\{\hat{\theta}_i\right\}\right)\right)\right\}- \tag{2}$$

риск (математическое ожидание функции потерь) при классификации случайного наблюдения X_ν с помощью подстановочного байсовского решающего правила $d_0\left(\cdot;\left\{\hat{\theta}_i\right\}\right)$; здесь $w(\cdot)\ge 0$ – функция потерь. Удалось доказать, при мягких условиях регулярности семейства $\mathcal{P}$, что справедливо следующее асимптотическое разложение риска (2):

$$r=r_0+\sum_{i=1}^{L}\frac{\rho_i}{n_i}+\left(O\left(\min_i n_i\right)^{-1}\right), \tag{3}$$

где r_0 – байесовский риск, т. е. наименьший возможный риск классификации при известных $\left\{\theta_i^0\right\}$, $\left\{\rho_i\ge 0\right\}$ – коэффициенты, для которых получены явные выражения. Эти результаты опубликованы в [9 – 12].

4. Робастная статистическая классификация

Приблизительно в 1984 году возникла идея распространить метод асимптотических разложений риска статистической классификации на ситуации, когда подлежащие классификации наблюдения на самом деле не в точности соответствуют гипотетической модели $\left\{p_i\left(x;\,\theta_i^0\right)\right\}$: наблюдения из i-го класса имеют плотность $p_i(x)$ с отклонением от $p\left(x;\,\theta_i^0\right)$ в некоторой метрике $\rho(\cdot)$ (например, в $L_2\left[R^N\right]$):

$$\rho\left(p_i(\cdot),p_i\left(\cdot;\theta_i^0\right)\right)\le\varepsilon_i,\ i=1,\ldots,L, \tag{4}$$

где $\varepsilon_i \geq 0$ – уровень искажений для *i*-го класса. При этом удается построить обобщение асимптотического разложения (3) для функционала риска при $\{\varepsilon_i \to 0\}$:

$$r = r_0 + \sum_{i=1}^{L}\left(\frac{\rho_i}{n_i} + \alpha_i\varepsilon_i\right) + O\left(\left(\min_i n_i\right)^{-1} + \max_i \varepsilon_i\right), \qquad (5)$$

где коэффициенты разложения $\{\rho_i, \alpha_i\}$ зависят от $\{p_i(\cdot)\}$. С помощью главного члена разложения (5) строится робастное решающее правило $d = d_*(x)$:

$$\sup_{\{\rho(p_i, p_i^0) \leq \varepsilon_i\}} \sum_{i=1}^{L}\left(\frac{\rho_i}{n_i} + \alpha_i\varepsilon_i\right) \to \min_{d(\cdot)}. \qquad (6)$$

Для ряда типов искажений и метрик $\rho(\cdot)$ удалось получить разложения (5) и построить робастные решающие правила $d = d_*(x)$ согласно (6) [13 – 17].

5. Робастное статистическое прогнозирование временных рядов

Пусть на (Ω, F, P) определен случайный временной ряд $x_t \in R$, $t \in \square$, для которого имеется некоторая гипотетическая модель M_0 (например, авторегрессия порядка *s*). Задача заключается в построении прогнозирующей статистики

$$\hat{x}_{T+\tau} = f_\tau(x_1, \ldots, x_T),$$

позволяющей вычислить прогноз $\hat{x}_{T+\tau}$ для будущего значения $x_{T+\tau}$ на τ шагов вперед с минимальным риском (среднеквадратической ошибкой) прогнозирования по доступным T наблюдениям $x_1, \ldots, x_T$, используя гипотетическую модель M_0. Наиболее распространенным подходом к решению этой задачи является подстановочный подход:

$$f_\tau^0(x_1, \ldots, x_T) = \mathrm{E}_{M_0}\{x_{T+\tau} | x_1, \ldots, x_T\},$$

причем вместо истинных параметров модели M_0 используются их статистические оценки по наблюдаемой предыстории $\{x_1, \ldots, x_T\}$.

Аналогично предыдущему разделу построены асимптотические разложения риска в ситуации, когда гипотетическая модель искажается: M_ε вместо M_0 $(\varepsilon \to 0)$, и построены робастные прогнозирующие статистики [18 – 24].

6. Исследования в области криптологии

С 2000 г. в связи с созданием в БГУ НИИ прикладных проблем математики и информатики значительно интенсифицировались научные исследования по криптологии – прикладной математической науке о защите информации. Одной из задач криптологии является оценка надежности методов защиты информации на основе параметрического преобразования исходного сообщения $m = (m_1, m_2, \ldots, m_n) \in \{0,1\}^n$:

$$x = f(m;\theta),$$

где $\theta = (\theta_1, \ldots, \theta_L) \in \{0,1\}^L$, $L \le n$, – неизвестный параметр преобразования. Задача оценки надежности состоит в оценивании вычислительной сложности оценивания скрытого параметра θ с заданной точностью по наблюдаемой последовательности $(x_1, x_2, \ldots, x_N) \in \{0,1\}^N$, $N \ge n$, т. е. является задачей математической статистики. Другая статистическая задача – это статистическое тестирование криптографических генераторов случайных и псевдослучайных последовательностей на соответствие гипотетической модели равномерно распределенной случайной последовательности [25 – 29].

7. Вероятностно-статистический анализ дискретно-значных временных рядов

В связи цифровизацией экономики и общества регистрируемые наблюдения $x_t \in A$ часто являются дискретно-значными и принимают значения из некоторого конечного множества A мощности $2 \le N < +\infty$. Статистический анализ временных рядов, как известно, глубоко развит для непрерывных временных рядов x_t, когда $A \subset R^d$ – множество ненулевой меры Лебега $\mu(A) > 0$. Применение этой теории и соответствующего программного обеспечения в дискретном случае $(\mu(A) = 0)$ ведет к некорректным результатам.

Одной из универсальных моделей в этом случае является предложенная Дж. Дубом цепь Маркова порядка $s \geq 1$, определяемая обобщенным Марковским свойством:

$$\mathrm{P}\{x_t = J_0 \mid F_{t-1}\} = \mathrm{P}\{x_t = J_0 \mid x_{t-1} = J_1, \ldots, x_{t-s} = J_s\} ::= p_{J_s,\ldots,J_1,J_0}, \qquad (7)$$

где $J_0, J_1, \ldots, J_s \in A$, $t \in Z$, $F_{t-1} = \sigma\{x_{t-1}, \ldots, x_{t-s}\}$, s – «глубина памяти» процесса. Задание матрицы вероятностей одношаговых переходов $\mathrm{P} = \left(p_{J_s,\ldots,J_1,J_0}\right)$ имеет экспоненциальную по s сложность $O\left(|A|^{s+1}\right)$, что ведет к «проклятию размерности». В этой ситуации нами предлагается строить так называемые «малопараметрические» (parsimonious) модели:

$$\mathrm{P} = \mathrm{P}(\theta),\ \theta = (\theta_1, \ldots, \theta_m) \in R^m,\ m \leq |A|^{s+1}.$$

Построены, исследованы и используются в статистическом анализе следующие малопараметрические модели: ЦМ (s,r), ЦМУП(s), BCNAR(s), BiCNAR(s), SBiCNAR(s), MCSS(s) и другие [30 – 37].

Библиографические ссылки

1. *Харин Ю. С., Дудин А. Н., Хацкевич Г. А., Малинковский Ю. В.* О научной школе по теории вероятностей, случайным процессам и математической статистике // В кн.: Научные школы в воспоминаниях и размышлениях профессоров. Минск: БГУ, 2023. С. 55–65.
2. *Харин Ю. С. (совместно с Гармаш Ю.М., Горцев А.М.)* Распознавание образов // Библиографические материалы по НИР «Рубидий-Р». Томск: СФТИ, 1970. 250 с.
3. *Харин Ю. С.* Выбор признаков в задаче распознавания сигналов // Математический сборник. Вып. 1. Томск: ТГУ, 1973. С. 166–168.
4. *Харин Ю. С.* Инвариантное оценивание собственной размерности множества наблюдений // Математическая статистика и ее приложения // Труды Сиб. физ.-тех. института. Вып. 60. Томск: ТГУ, 1974. С. 31–45.
5. *Харин Ю. С.* Адаптивное формирование признаков в задаче распознавания образов. // Межвуз. сб. «Динамика систем». Вып. 8. Горький: ГГУ, 1975. С. 148–161.
6. *Харин Ю. С.* Об адаптивном формировании инвариантных признаков в задаче распознавания образов // Известия АН СССР. Тех. кибернетика, 1977, № 5. С. 155–164.
7. *Харин Ю. С.* О редукции данных при помощи оценок Парзена в задачах распознавания образов // Адаптивные системы и их приложения. Новосибирск: Наука, 1978. С. 43–50.
8. *Харин Ю. С.* Оптимизация режима функционирования замкнутой системы массового обслуживания методом статистического моделирования на ЭВМ // Оптимизация динамических систем. Минск: БГУ, 1978. 7 с.

9. *Харин Ю. С., Дучинскас К. А.* Асимптотическое разложение риска для классификатора, использующего оценки максимального правдоподобия // Статистические проблемы управления. Вып. 38. Изд-во ИМК АН ЛитССР, 1979. С. 77–93.

10. *Харин Ю. С.* О точности статистической классификации при использовании МК-оценок // Теория вероятностей и ее применения. 1981. Т. XXVI, вып. 4. С. 866– 867.

11. *Kharin Yu. S.* Asymptotic expansions for the risk of parametric and nonparametric decision functions // Transactions of the IX Prague Conf. on Inform. Theory. Statist. Decision Funct. Random Process. Prague: Academia, 1983. P. 11–16.

12. *Харин Ю. С.* Исследование риска статистических классификаторов, использующих оценки минимального контраста // Теория вероятностей и ее применения. 1983. Т. XXVIII, вып. 3. С. 592–598.

13. *Kharin Yu. S.* Robustness investigation for the decision rules by risk asymptotic expansion method // Proceedings of the Third Prague Symposium on Asymptotic Statistics. Amsterdam: Elsevier, 1984. P. 309–317.

14. *Харин Ю. С., Медведев А. Г.* Робастность решающих правил в задачах статистической классификации многомерных наблюдений // Проблемы устойчивости стохастических моделей. М.: Изд-во Всесоюзн. НИИ системных исследований, 1985. 9 с.

15. *Харин Ю. С.* Робастность в статистическом распознавании образов. Минск: «Университетское», 1992.170 с.

16. *Kharin Yu.* Robustness in Statistical Pattern Recognition. Dordrecht/ Boston/ London: Kluwer Academic Publishers, 1996. 302 p.

17. *Жук Е. Е., Харин Ю. С.* Устойчивость в кластер-анализе многомерных наблюдений. Минск: БГУ, 1998. 240 с.

18. *Kharin Yu.* Robust Forecasting of Parametric Trend of Time Series under “Outliers” // Studies in Classification and Data Analysis. Springer, 2000. Vol. 17. P. 197–206.

19. *Kharin Yu. S., Zenevich D. V.* Robustness of Statistical Forecasting by autoregression model under distortions // Theory of Stochastic Processes. 2001. Vol. 5 (21), No. 3 – 4. P. 84–91.

20. *Kharin Yu.* Robustness Analysis in Forecasting of Time Series. In: Developments in Robust Statistics. (Ed. Dutter R., Filzmoser P., Gather U., Rousseeuw P. J.). Heidelberg/ New York: Springer, 2002. P. 180–193.

21. *Kharin Yu., Huryn A.* “Plug-in” Statistical Forecasting of Vector Autoregressive Time Series with Missing Values // Austrian Journal of Statistics. 2005. Vol. 34, No. 2. P. 163–175.

22. *Харин Ю. С.* Устойчивость в статистическом прогнозировании временных рядов // Прикладная эконометрика. 2006, № 1. С. 82–93.

23. *Харин Ю. С.* Оптимальность и робастность в статистическом прогнозировании. Минск: БГУ, 2008. 263 с.

24. *Kharin Yu.* Robustness in Statistical Forecasting. Heidelberg/ New York/ Dordrecht/ London: Springer, 2013. 356 p. (ISBN 978-3-319-00839-4; DOI 10.1007/978-3-319-00840-0).

25. *Соловей О. В., Харин Ю. С.* Математические модели генераторов двоичных последовательностей с неравномерным движением регистров // Управление защитой информации. 2002. Том 6, № 2. С. 77–83.

26. *Агиевич С. В., Галинский В. А., Харин Ю. С., Микулич Н. Д.* Алгоритм блочного шифрования BelT // Управление защитой информации. 2002. Том 6, № 4. С. 407–412.

27. *Харин Ю. С., Берник В. И., Матвеев Г. В., Агиевич С. В.* Математические и компьютерные основы криптологии. – Минск: Новое Знание, 2003. 381 с.

28. *Харин Ю. С., Палуха В. Ю.* Анализ мощности статистического тестирования криптографических генераторов на основе оценок информационной энтропии // Комплексная защита информации. Витебск: ВГУ. 2019. С. 141–146.

29. *Харин Ю. С., Агиевич С. В., Васильев Д. В., Матвеев Г. В.* Криптология. Минск: БГУ, 2023. 512 с.

30. *Харин Ю. С.* Цепи Маркова с *r*-частичными связями и их статистическое оценивание // Доклады НАН Беларуси. 2004. Том 48, № 1. С. 40–44.

31. *Kharin Yu.* Statistical analysis of high-order Markov chains // Modern stochastics: theory and applications. Kiev: KSU, 2006. P. 159–160.

32. *Харин Ю. С., Петлицкий А. И.* Идентификация двоичной цепи Маркова *s*-го порядка с *r* частичными связями при наличии аддитивных искажений // Дискретная математика. 2010. Том 22, вып. 4. С. 138–155.

33. *Kharin Yu., Zhurak M.* Statistical Analysis of Spatio-Temporal Data Based on Poisson Conditional Autoregressive Model // INFORMATICA. 2015. Vol. 26, № 1. P. 67– 87.

34. *Kharin Yu. S.* Parsimonious models of high-order Markov chain for evaluation of cryptographic generators // Математические вопросы криптографии. 2017. Т. 7, № 2. С. 131–142.

35. *Kharin Yu. S., Voloshko V. A., Medved E. A.* Statistical estimation of parameters for binary conditionally nonlinear autoregressive time series // Mathematical Methods of Statistics. 2018. Vol. 26, No. 2. P. 103–118. (DOI: 10.3103/S1066530718020023).

36. *Kharin Yu., Fokianos K., Fried R., Voloshko V.* Statistical analysis of multivariate discrete-valued time series // Journal of Multivariate Analysis (квартиль Q_1, impact factor 1.5). 2022. Vol. 188, article 104805 (doi.org/10.1016/j.jmva.2021.104805). (Scopus, Web of Science).

37. *Kharin Yu., Voloshko V.* Robust estimation for Binomial conditionally nonlinear autoregressive time series based on multivariate conditional frequencies // Journal of Multivariate Analysis. 2021. Vol. 185(2). P. 11–27. (DOI: http//doi.org/10.1016/j.jmva.2021. 104777; журнал входит в 1-ю квартиль Q_1, импакт-фактор h=1.5).