07

Выбор спектральных переменных и повышение точности калибровки температуры методом проекции на латентные структуры по спектрам флуоресценции $Yb^{3+}:CaF_2$

© М.А. Ходасевич[1], В.А. Асеев[2]

[1] Институт физики НАН Беларуси,
220072 Минск, Республика Беларусь
[2] Университет ИТМО,
197101 Санкт-Петербург, Россия
e-mail: m.khodasevich@ifanbel.bas-net.by

Поступила в редакцию 22.01.2018 г.

На примере спектров флуоресценции $Yb^{3+}:CaF_2$, зарегистрированных в полосе 880–1120 nm с разрешением около 0.2 nm в температурном диапазоне от 66 до 150°C с шагом 2°C, рассмотрена эффективность интервальных методов выбора переменных и дивизимного метода оптимизации интервалов с помощью генетического алгоритма с целью повышения точности калибровки температуры с помощью проекции на латентные структуры. Лучший результат по величине среднеквадратичной ошибки предсказания температуры в проверочной выборке (0.45°C) получен с помощью интервальной проекции на латентные структуры по комбинации движущихся окон. Показано, что применение методов выбора спектральных переменных позволило более чем в 2 раза повысить точность калибровки температуры.

DOI: 10.21883/OS.2018.05.45958.22-18

Решение задач калибровки или идентификации по спектральным данным с помощью классического спектроскопического подхода подразумевает использование небольшого количества спектральных отсчетов на отдельных длинах волн или в спектральных интервалах. Например, спектрометрическая идентификация химических веществ и соединений с помощью баз данных или массивов спектров ИК диапазона [1] в настоящее время проводится по интенсивности, форме и расположению небольшого количества характеристических полос. Достоинством такого подхода является четкий физический смысл спектральных отсчетов на отдельных длинах волн или в спектральных интервалах. В противоположность этому подходу многопараметрический спектральный анализ обращается к информации, скрытой в широкополосных спектрах, и представляет в малоразмерном и нефизичном базисе только существенные данные, выделенные из мультиколлинеарных спектров. Компромиссный подход [2], когда из широкополосных спектров удаляется часть, приводящая к переопределенности многопараметрических моделей, позволяет уменьшить влияние избыточности спектральных данных на качество регрессионного и дискриминантного анализов для решения задач калибровки или идентификации.

Рассмотрим некоторые методы выбора переменных на примере одного из эффективных регрессионных методов анализа многопараметрических данных — проекции на латентные структуры (PLS — projection to latent structures or partial least squares) [3]. В [4] нами было предложено использовать метод PLS для определения температуры по широкополосным спектрам апконверсионной флуоресценции, активированной ионами эрбия свинцово-фторидной наностеклокерамики, и показано [5], что этот метод позволяет получить меньшую ошибку предсказания температуры, чем измерения согласно широко применяемому методу FIR (fluorescence intensity ratio) по отношению интенсивностей двух полос флуоресценции с температурно связанных уровней [6]. В качестве объекта исследования в данной работе рассмотрен $Yb^{3+}:CaF_2$, возбуждаемый неполяризованным излучением лазерного диода ML-151 („Милон“, Россия) мощностью 1 W с максимумом спектра около 967 nm. Спектры флуоресценции регистрировались в полосе 880–1120 nm с разрешением около 0.2 nm в температурном диапазоне от 66 до 150°C с шагом 2°C. Температура образца изменялась и контролировалась с погрешностью 0.1°C с помощью печи PV10 („Coversion Ltd“, Англия) и температурного контроллера TS-200 („Thorlabs“, США). Ранее [7] на основе зарегистрированных спектров флуоресценции с помощью метода главных компонент [8] была найдена штарковская структура энергетических уровней иона иттербия, с высокой точностью совпадающая с данными, полученными традиционными методами абсорбционной и флуоресцентной спектроскопии с временным разрешением при низких температурах [9].

Спектры флуоресценции $Yb^{3+}:CaF_2$ были объединены в матрицу X размерностью 43×1024, где 43 — количество спектров, зарегистрированных при разной температуре, 1024 — количество отсчетов на разных длинах волн. Изначально количество спектральных отсчетов было равно 1154. Не несущие существенной информации края спектрального диапазона были удалены так, чтобы количество спектральных отсчетов являлось степенью числа 2. Это позволяет применить описываемые ниже

интервальные методы PLS при последовательном сужении спектральных интервалов в 2 раза и дивизимный метод оптимизации интервалов с помощью генетического алгоритма. Проведенный с помощью метода главных компонент анализ температурной зависимости счетов в первую главную компоненту, описывающую 99.95% суммарной объясненной дисперсии данных, позволил выявить выброс в экспериментальных данных при 80°C. Соответствующий спектр был удален из дальнейшего рассмотрения, и матрица X стала иметь размерность 42×1024. В методе PLS кроме матрицы X предикторов (спектры флуоресценции в нашем случае) в исходные данные входит вектор $\mathbf{Y}$ откликов (значения температуры). Метод PLS находит в пространстве предикторов маломерное пространство так называемых латентных переменных, матрицы проекций X и $\mathbf{Y}$ в которое имеют максимальную ковариацию.

Существенной особенностью метода PLS является необходимость использования обучающей и проверочной выборок. Самым надежным способом уменьшить ошибку предсказания параметра при калибровке является использование большой обучающей выборки, охватывающей весь диапазон изменения параметра [10]. В рассматриваемом случае необходимо весь набор спектров флуоресценции разделить на две непересекающиеся выборки. Отбор осуществлялся по величинам счетов в первую главную компоненту, а не по соответствующему значению температуры. Такая особенность рассмотрения позволяет целенаправленно применить методы кластерного анализа [11] при формировании проверочной выборки спектров. В силу регулярного характера изменений температуры, при которой проводятся измерения, количество спектров, отобранных в обучающую выборку, может быть относительно невелико. Из 42 спектров флуоресценции всего 7 входили в обучающую выборку: 5 целенаправленно отобранных спектров и еще 2, соответствующие крайним величинам счетов в первую главную компоненту. В [12] на рассматриваемом наборе спектров было показано, что наименьшая среднеквадратичная ошибка при калибровке температуры была достигнута путем построения обучающей выборки с помощью иерархического кластерного анализа пространства главных компонент. Таким образом, на предварительном этапе подготовки экспериментальных данных к применению метода PLS уже были применены два метода многопараметрического анализа: метод главных компонент и кластерный анализ.

После разделения исходных X и $\mathbf{Y}$ на соответствующие обучающую и проверочную выборки приступим к решению задачи исследования — оптимизации спектрального диапазона измерений с целью повышения качества калибровки температуры методом PLS.

Целый класс методов оптимизации спектрального диапазона PLS объединяется присутствием в названии термина „интервальная“ или „оконная“ [13–16]. В самой простой реализации интервальной PLS (interval PLS — IPLS) [13] весь диапазон измерений делится на заданное количество неперекрывающихся интервалов, по каждому из которых проводится отдельное моделирование. Целевыми функциями при оценке качества моделирования, как правило, служат среднеквадратичные ошибки предсказания параметра в обучающей или проверочной выборках. Спектральные интервалы могут последовательно объединяться с целью достижения минимальной величины целевой функции (forward IPLS — FIPLS), а могут по одному изыматься из полного диапазона измерений (backward IPLS — BIPLS) [14]. Интервал, который может изменять размер и сдвигаться по спектру, принято называть „окном“ [15,16]: в методах по движущемуся „окну“ (MWIPLS — moving window IPLS) [15] и по комбинации движущихся окон (SCMWIPLS — searching combination moving window IPLS) [16].

Применим указанные методы оптимизации спектрального диапазона PLS к спектрам флуоресценции $Yb^{3+}:CaF_2$. Целевой функцией калибровки будет являться среднеквадратичная ошибка $RMSEP$ (root-mean square error of prediction) предсказания температуры в проверочной выборке:

$$RMSEP = \sqrt{\frac{\sum\limits_{\text{test}}(T - T_{\text{predicted}})^2}{n}},$$

где T и $T_{\text{predicted}}$ — измеренное и предсказанное значения температуры для спектров, входящих в проверочную выборку, n — количество спектров в проверочной выборке. В рассматриваемом случае $n = 35$.

Предварительное моделирование по всему спектральному диапазону измерений флуоресценции $Yb^{3+}:CaF_2$ показало, что лучшие результаты достигаются при выборе трех латентных переменных. Этот параметр будет сохранен во всех дальнейших расчетах. Поэтому минимальное количество спектральных отсчетов в одиночном интервале не может быть меньшим 4.

Для калибровки температуры методом IPLS вначале необходимо определить зависимость минимальной величины $RMSEP$ от количества спектральных отсчетов в одиночном интервале. Количество отсчетов изменялось от 4, что давало возможность моделировать PLS по трем латентным переменным, до 1024, что соответствует всему спектральному диапазону измерения флуоресценции. Значение $RMSEP$ при моделировании по полному спектральному диапазону, т.е. методом PLS, равно 0.93°C. Минимальная $RMSEP = 0.83$°C для IPLS соответствует 128 отсчетам на интервал или разбиению диапазона измерений на 8 интервалов.

На рис. 1 представлены результаты моделирования с помощью IPLS по 8 неперекрывающимся спектральным интервалам вместе со спектром флуоресценции при $T = 66$°C. Штриховкой отмечен интервал 942.7–981.9 nm, модель IPLS по которому характеризуется минимальной среднеквадратичной ошибкой калибровки температуры. На выбранном интервале спектр флуоресценции испытывает самый быстрый рост, поэтому моделирование по этому интервалу представляется и

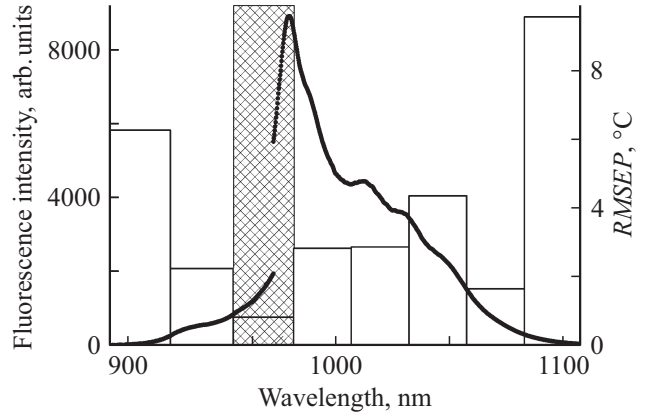

Рис. 1. Среднеквадратичная ошибка предсказания температуры в проверочной выборке методом IPLS по восьми неперекрывающимся спектральным интервалам и спектр флуоресценции $Yb^{3+}:CaF_2$ при $T = 66°C$.

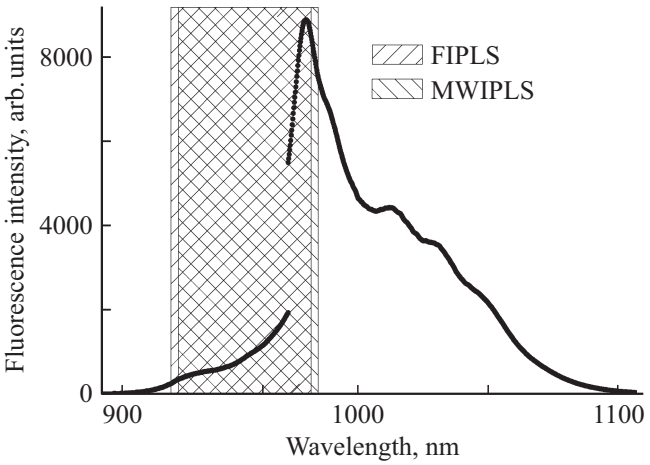

Рис. 2. Спектральные интервалы, обеспечивающие минимальные среднеквадратичные ошибки предсказания температуры в проверочной выборке методами FIPLS (917.8–981.9 nm) и MWIPLS (919.4–983.7 nm).

наиболее физичным с точки зрения учета скорости затухания флуоресценции при росте температуры. Парадоксально, что следующей по значению $RMSEP = 1.65°C$ соответствует интервал 1057.2–1082.3 nm, находящийся на длинноволновом крыле спектра. Это подтверждает, что модели, опирающиеся на применение многопараметрического анализа данных, лишь в простейших случаях могут быть сопоставлены с физическим описанием рассматриваемого явления.

По времени счета более затратными по сравнению с IPLS являются методы FIPLS и MWIPLS. На рис. 2 представлены найденные этими методами интервалы, PLS по которым характеризуются минимальными величинами *RMSEP*. Для FIPLS $RMSEP_{\min} = 0.77°C$, для MWIPLS $RMSEP_{\min} = 0.74°C$. Уменьшение среднеквадратичной ошибки при FIPLS по интервалу со 128 отсчетами по сравнению с IPLS происходит только при добавлении одного интервала, добавление еще одного приводит к ухудшению качества калибровки температуры. Видно, что возможность сдвига спектрального „окна“ в методе MWIPLS позволяет уменьшить $RMSE_{\min}$ по сравнению с FIPLS при моделировании по спектральным диапазонам одинаковой ширины.

Еще большее количество вычислений требует применение метода SCMWIPLS. Разработанная нами модификация этого алгоритма может быть описана следующим образом. Как упоминалось выше, во всех рассматриваемых методах для построения PLS были использованы 3 латентные структуры. Поэтому минимальная ширина спектрального „окна“ составляла 4 отсчета. Аналогично методу IPLS находится положение первого „окна“, и его положение фиксируется. Следующее „окно“ последовательно сдвигается в пределах всего спектрального диапазона измерений и объединяется с первым. Выбор его окончательного положения осуществляется по минимальной среднеквадратичной ошибке предсказания температуры. Так как зависимость $RMSE_{\min}$ от количества спектральных отсчетов, учитываемых в модели, может быть немонотонной, наращивание спектрального интервала моделирования необходимо продолжать до полного диапазона измерений.

На рис. 3 представлены зависимость минимальной среднеквадратичной ошибки предсказания температуры в проверочной выборке методом SCMWIPLS от количества спектральных отсчетов, учитываемых в модели, и положение интервалов, содержащих 136 отсчетов, для которых достигается минимум $RMSEP_{\min} = 0.45°C$. Видно, что наиболее информативными участками исследуемого спектра флуоресценции являются коротковолновой склон пика и длинноволновое крыло. Эти результаты подтверждают парадоксальные, с нашей точки зрения, выводы, сделанные выше при рассмотрении моделирования методом IPLS.

Описанные способы оптимизации спектрального диапазона измерений флуоресценции для калибровки температуры достаточно эффективны, но не охватывают все возможные варианты комбинаций спектральных отсчетов. Для поиска глобального минимума среднеквадратичной ошибки предсказания температуры может применяться генетический алгоритм (GA) [17] — эвристический метод поиска, использующий аналогию с такими механизмами естественного отбора в природе, как наследование, отбор наиболее приспособленных особей, мутации и кроссинговер. Генетический алгоритм как метод отбора переменных в обработке больших массивов данных используется уже на протяжении длительного времени [18,19] и до сих пор актуален в различных спектроскопических применениях [20,21].

Одной из первичных задач применения генетического алгоритма является кодирование данных [19]. Поскольку каждый спектральный интервал может либо учитываться, либо не учитываться в модели PLS, проще всего представлять их с помощью бинарных цифр. В рассматриваемом случае каждый из m спектральных интервалов

Минимальные среднеквадратичные ошибки предсказания температуры при применении различных методов оптимизации спектрального диапазона измерения флуоресценции $Yb^{3+}:CaF_2$

Метод	PLS	IPLS	FIPLS	MWIPLS	SCMWIPLS	DGA-PLS
RMSEP, °C	0.93	0.83	0.77	0.74	0.45	0.53

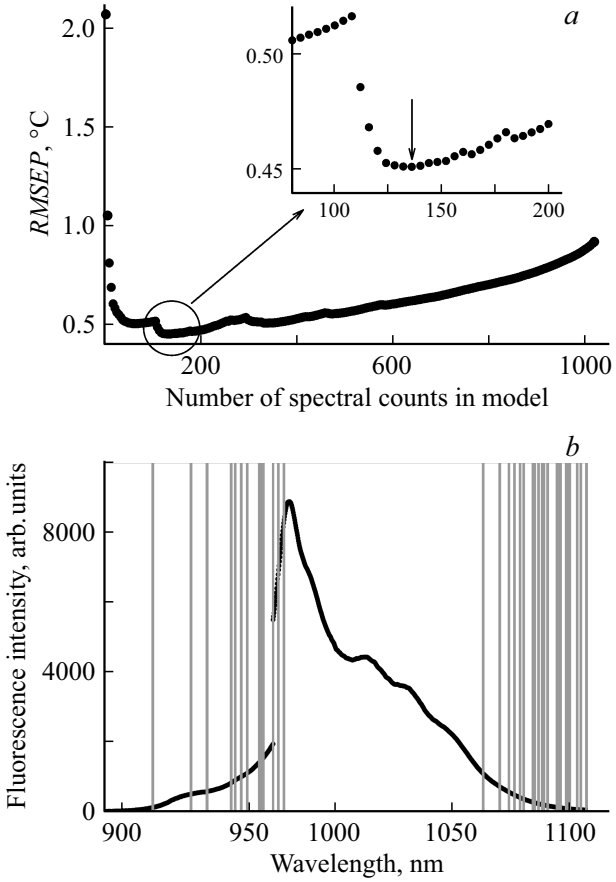

Рис. 3. Минимальная среднеквадратичная ошибка предсказания температуры в проверочной выборке методом SCMWIPLS в зависимости от количества спектральных отсчетов в модели (*a*) и положение интервалов, содержащих 136 отсчетов, для которых достигается минимум *RMSEP* (*b*).

в зависимости от использования будет кодироваться бинарными 1 (если интервал включается в модель) или 0. Строка из m бинарных 1 и 0 с помощью кода Грея преобразуется в целое число в диапазоне от 0 до $2^m - 1$. Ограничение на максимальное целое число в MatLab приводит к ограничению на m при использовании инструментария глобальной оптимизации. С учетом того, что m является степенью числа 2, максимально возможным количеством оптимизируемых интервалов для работы генетического алгоритма в MatLab является 16. Следовательно, в целочисленном диапазоне от 0 до $2^{16} - 1$ с помощью генетического алгоритма ищется минимум функции приспособленности. Далее генетический алгоритм продолжает работу со спектральным диапазоном, объединенным из отобранных на предыдущем этапе интервалов. Оптимизация учитываемых в модели интервалов с помощью такого алгоритма, основанного на делении найденного на предыдущем этапе спектрального диапазона на более мелкие интервалы, заканчивается при отыскании минимального значения *RMSEP*. По аналогии с дивизимным иерархическим кластерным анализом такую модификацию генетического алгоритма можно назвать дивизимным методом оптимизации интервалов с помощью GA или дивизимным GA (divisional GA — DGA). Поскольку генетический алгоритм является эвристическим методом, оптимизация для каждого набора спектральных интервалов проводилась 10 раз. После первого этапа применения генетического алгоритма из 1024 спектральных отсчетов были выбраны 192 ($RMSEP = 0.61°C$), после второго — 96 ($RMSEP = 0.57°C$), после третьего — 66 ($RMSEP = 0.53°C$), из которых 2 крайних были удалены для того, чтобы их количество было кратно 16. Всего 64 отсчета позволили рассмотреть 16 спектральных интервалов по 4 отсчета в каждом, что является минимумом для применения PLS по трем латентным структурам в каждом интервале. Дальнейшее уменьшение количества спектральных отсчетов в модели не привело к уменьшению среднеквадратичной ошибки предсказания температуры. На рис. 4 представлено положение интервалов, содержащих отсчеты, для которых достигается минимум *RMSEP* при оптимизации

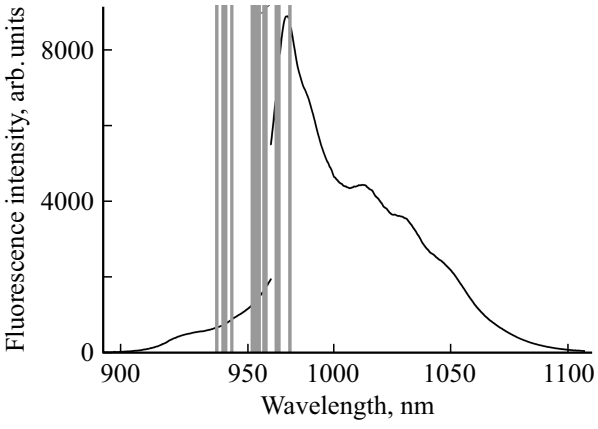

Рис. 4. Положение интервалов, содержащих отсчеты, для которых достигается минимум *RMSEP* при оптимизации спектрального диапазона методом дивизимного генетического алгоритма (937.42–938.40, 939.78–941.94, 943.30–944.28, 951.54–954.88, 956.24–957.22, 975.06–977.22, 980.94–981.92 nm).

спектрального диапазона методом DGA-PLS. Полученный результат не является глобальным минимумом, что видно из сравнения с результатами применения метода SCMWIPLS. Это обусловлено примененным нами способом кодирования наборов спектральных интервалов и ограниченностью вычислительных ресурсов.

В таблице представлены минимальные среднеквадратичные ошибки предсказания температуры при оптимизации спектрального диапазона различными примененными нами методами.

Итак, на примере флуоресценции $Yb^{3+}:CaF_2$ продемонстрированы достоинства и недостатки методов оптимизации спектрального диапазона измерения температурной зависимости флуоресценции для повышения точности калибровки температуры методом проекции на латентные структуры. Показано, что применение подхода, компромиссного между спектрометрическим и хемометрическим, позволило в исследованном случае более чем в 2 раза уменьшить среднеквадратичную ошибку предсказания температуры в проверочной выборке по сравнению с применением проекции на латентные структуры по всему измеренному диапазону спектра. Лучший результат по величине ошибки получен с помощью проведения калибровки по комбинации движущихся окон.

Список литературы

[1] *Тарасевич Б.Н.* ИК спектры основных классов органических соединений. М., 2012. 55 с.

[2] *Mehmood T., Liland K.H., Snipen L., Sæbø S.* // Chem. Intel. Lab. Sys. 2012. V. 118. P. 62. doi 10.1016/j.chemolab.2012.07.010

[3] *Geladi P., Kowalski B.* // Analyt. Chim. Acta. 1986. V. 185. P. 1.

[4] *Асеев В.А., Варакса Ю.А., Колобкова Е.В., Синицын Г.В., Ходасевич М.А.* // Опт. и спектр. 2015. Т. 118. № 5. С. 760; *Aseev V.A., Varaksa Yu.A., Kolobkova E.V., Sinitsyn G.V., Khodasevich M.A.* // Opt. Spectrosc. 2015. V. 118. N 5. P. 727. doi 10.1134/S0030400X15050033

[5] *Асеев В.А., Варакса Ю.А., Колобкова Е.В., Синицын Г.В., Ходасевич М.А., Ясюкевич А.С.* // Научно-техн. вестник инф. тех., мех. и опт. 2015. Т. 15. № 3. С. 457. doi 10.17586/2226-1494-2015-15-3-457-462

[6] *Rai V.* // Appl. Phys. B. 2007. V. 88. P. 297. doi 10.1007/s00340-007-2717-4

[7] *Khodasevich M., Varaksa Y., Sinitsyn G., Aseev V., Demesh M., Yasukevich A.* // J. Luminesc. 2017. V. 187. P. 295. doi 10.1016/j.jlumin.2017.03.014

[8] *Esbensen K.H., Geladi P.* // Compr. Chemom. 2009. V. 2. P. 211.

[9] *Petit V., Camy P., Doualan J.-L., Portier X., Moncorgé R.* // Phys. Rev. B. 2008. V. 78. P. 085131−1. doi 10.1103/PhysRevB.78.085131

[10] *Anderson R.B., Bell III J.F., Wiens R.C., Morris R.V., Clegg S.M.* // Spectrochim. Acta. B. 2012. V. 70. P. 24. doi 10.1016/j.sab.2012.04.004

[11] *Мандель И.Д.* Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.

[12] *Ходасевич М.А., Саскевич Н.А.* // Вести НАН Беларуси: сер. физ.-мат. наук. 2018. Т. 54. № 1. С. 77.

[13] *Norgaard L., Saudland A., Wagner J., Nielsen J.P., Munck L., Engelsen S.B.* // Appl. Spectr. 2000. V. 54. P. 413.

[14] *Zou X., Zhao J., Li Y.* // Vibr. Spectr. 2007. V. 44. P. 220. doi 10.1016/j.vibspec.2006.11.005

[15] *Jiang J.-H., Berry R.J., Siesler H.W., Ozaki Y.* // Anal. Chem. 2002. V. 74. P. 3555.

[16] *Du Y.P., Liang Y.Z., Jiang J.H., Berry R.J., Ozaki Y.* // Anal. Chim. Acta. 2004. V. 501. P. 183. doi 10.1016/j.aca.2003.09.041

[17] *Holland J.H.* Adaptation in Natural and Artificial Systems. MIT Press, 1992. 225 p.

[18] *Lucasius C.B., Kateman G.* // Trends in Analyt. Chem. 199l. V. 10. P. 254.

[19] *Leardi R.* // J. Chromatography A. 2007. V. 1158. P. 226. doi 10.1016/j.chroma.2007.04.025

[20] *Yang Y., Wang L., Wu Y., Liu X., Bi Y., Xiao W., Chen Y.* // Spectrochimica Acta A. 2017. V. 182. P. 73. doi 10.1016/j.saa.2017.04.004

[21] *Caredda M., Addis M., Ibba I., Leardi R., Scintu M.F., Piredda G., Sanna G.* // LWT-Food Sci. Tech. 2016. V. 65. P. 503. doi 10.1016/j.lwt.2015.08.048