ПАРАЛЛЕЛЬНАЯ РЕАЛИЗАЦИЯ ВЫЧИСЛИТЕЛЬНО ТРУДОЕМКИХ РОБАСТНЫХ АЛГОРИТМОВ ОПРЕДЕЛЕНИЯ ПАРАМЕТРОВ МОДЕЛЕЙ

В.А. Фурсов, Е.В. Гошин

В работе рассматриваются робастные алгоритмы решения вычислительно трудоемких задач определения линейных по параметрам моделей. Предполагается, что наблюдения могут содержать большое число грубых ошибок типа сбоев. Исследуются два типа устойчивых к грубым ошибкам методов, имеющих высокую вычислительную сложность: семейство алгоритмов RANSAC и метод согласованной идентификации. Приводятся результаты сравнительных экспериментальных исследований точности, надежности и эффективности параллельной реализации на кластере СГАУ «Сергей Королев».

1. Введение

В работе рассматриваются вычислительные аспекты решения задач идентификации моделей в условиях априорной неопределенности. Если априорные вероятностные модели отсутствуют, обычно применяют метод наименьших квадратов (МНК). Известно, что МНК-оценки являются оптимальными, когда ошибки измерений имеют нормальное распределение [1,2], но весьма чувствительны к нарушениям условий оптимальности. В частности, при наличии в исходных данных грубых ошибок типа сбоев МНК теряет работоспособность.

Резкой критике методы оценивания, основанные на использовании стандартной (нормальной) априорной статистической гипотезы, подверг Р. Калман [3]. В указанной работе была высказана идея улучшения метода наименьших квадратов путем поиска подсистемы наиболее свободной от шума. Похожие идеи идентификации, основанные на иных предположениях, приводились в работах [4,5]. Однако, возможно в связи с отсутствием на тот момент необходимых вычислительных мощностей, оба этих подхода остались на уровне теоретических идей.

В последние годы так называемые робастные методы оценивания, в которых устойчивость к грубым ошибкам типа сбоев достигается ценой высокой вычислительной сложности алгоритмов приобретает все большую популярность. Наиболее широко известным алгоритмом этого класса является RANSAC [6,7]. Другой подход – согласованная идентификация предложен в работах [8,9]. Наиболее полная версия этого метода описана в работе [10].

Оба эти подхода для получения оценок за приемлемое время требуют применения высокопроизводительных многопроцессорных систем, т.к. в основе своей являются переборными. Большие вычислительные затраты в данном случае неизбежная плата за недостаток «хороших» данных. Возможность применения простых (а следовательно относительно дешевых) статистических схем обработки, как правило, является следствием значительных затрат на получение достаточно большого числа точных измерений. Тем не менее, применение указанных подходов оправдано в ситуациях, когда получение таких измерений, по каким-либо причинам, невозможно или нецелесообразно.

В настоящей работе рассматриваются параллельные схемы реализации указанных алгоритмов, приводятся результаты их сравнительных исследований по критериям точности, надежности и эффективности.

2. Описание алгоритмов

Рассматривается задача идентификации линейной по параметрам модели вида

$$y(t_i)=\sum_{j=1}^{M} c_j x_j(t_i)+\xi(t_i),\quad i=\overline{1,N}\quad , \qquad (1)$$

где $x_j(t_j)$, $y_j(t_j)$ – входной и выходной сигналы, а $\xi(t_i)$ – ошибка измерений, аппроксимации и др.

Если за время, необходимое для проведения N наблюдений, изменения параметров c_j, $j=\overline{1,M}$ несущественны, систему N соотношений (1) можно переписать в матричном виде:

$$y=Xc+\xi \quad , (2)$$

где $N\times M$-матрица **X**, $N\times 1$векторы **y**, ξ и $M\times 1$-вектор **c** определяются как

$$X=\begin{bmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_M(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_M(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_N) & x_2(t_N) & \cdots & x_M(t_N) \end{bmatrix}, \quad y=\begin{bmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_N) \end{bmatrix}, \quad \xi=\begin{bmatrix} \xi(t_1) \\ \xi(t_2) \\ \vdots \\ \xi(t_N) \end{bmatrix}, \quad c=\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_M \end{bmatrix} .$$

Задача идентификации заключается в том, чтобы по NxM-матрице X и Nx1-вектору y (N>M) построить оценку $\hat{c}$ вектора параметров c при неизвестном Nx1-векторе ошибок ξ . Предполагается, что матрица X и все возможные квадратные матрицы, составленные из ее строк неособенные. Если это не так, может быть предусмотрена процедура предварительного контроля обусловленности и отбора таких подсистем. Построение таких процедур рассматривалось в работе [1].

Приведем краткое описание алгоритмов решения описанной задачи с использованием методов RANSAC [6][7] и согласованной идентификации [8]. Идея метода RANSAC (*RANdom SAmple Consensus*) состоит в том, что по имеющимся данным последовательно строятся модели («гипотезы»), при этом все исходные данные разделяют на два типа: удовлетворяющие модели, «не-выбросы» или «инлаеры» (*inlier*) и ложные точки, шумы — случайные включения в исходные данные, «выбросы» или «аутлаеры» (*outlier*). Качество гипотезы определяется количеством элементов исходных данных ей удовлетворяющих. Степень согласия гипотезы и исходных данных вычисляется следующим образом:

$$R(\hat{c}_k)=\sum_i p(r_i(\hat{c}_k))$$

$$p(r_i)=\begin{cases}0, & r_i<T,\\ 1, & r_i>T,\end{cases}$$

где $r_i(\hat{c}_k)$ - невязка оценки $\hat{c}_k$ для i-й строки матрицы X, а T — фиксированный порог, задаваемый из некоторых общих соображений или свойств исходных данных. Принимается гипотеза с наименьшей величиной $R(\hat{c}_k)$.

В методе согласованной идентификации из исходной системы (2) формируется множество так называемых подсистем нижнего уровня с помощью весовых матриц диагонального вида – $G_k=diag(g_{k,1},\dots,g_{k,N})$, k=1,2,… Элементы этих матриц могут быть только нулями и единицами. Ненулевые элементы задаются для различных сочетаний из S_k индексов, так что $rank\, G_k=S_k$ для всех k. В результате получаем множество подсистем вида:

$$y_k=X_k c_k+\xi_k \quad , k=1,2,\dots, (3)$$

где

$$y_k=G_k y \quad , \quad X_k=G_k X \quad , \quad \xi_k=G_k y \quad ,$$

$$dim(R(X_k))=rank\, G_k=\sum_{i=1}^{N} g_{k,i}=\|g_k\|_2=S_k \quad ,$$

где $R(X_k)$ – пространство столбцов матрицы X_k , а g_k – N×1-вектор: $G_k=Eg_k$.

Если размерность подсистем нижнего уровня фиксирована, т.е. $S_k = S$, то число подсистем нижнего уровня равно C_N^S (S<N). Вычисляя для каждой из построенных таким образом подсистем МНК-оценку:

$$\hat{c}_k=[X^T G_k X]^{-1} X^T G_k y \quad , (4)$$

получаем множество Ξ всех возможных оценок на подсистемах нижнего уровня размерности S:

$$\Xi=\{\hat{c}_k=[X^T G_k X]^{-1} X^T G_k y \mid \forall G_k:\ \|g_k\|=S\} \quad , \quad |\Xi|=C_N^S \quad .(7)$$

Аналогичным образом (из нулей и единиц) строится множество диагональных $N\times N$ *весовых матриц* H_l:

$$H_l=diag(h_{l,1},\dots,h_{l,N}) \quad , \quad rank\, H_l=P \quad (S<P<N) \quad ,$$

с использованием которых формируются так называемые подсистемы *верхнего уровня*:

$$\tilde{y}_l=\tilde{X}_l c_l+\tilde{\xi}_l \quad , (5)$$

где

$$\tilde{X}_l=H_l X \quad , \quad \tilde{y}_l=H_l y \quad , \quad \tilde{\xi}_l=H_l y \quad , \quad l=\overline{1,L} \quad , \quad L=C_N^P \quad .$$

Ясно, что каждой такой подсистеме будет соответствовать свое подмножество подсистем нижнего уровня и, соответственно, *подмножество промежуточных оценок*:

$$\Theta_l=\{\hat{c}_k\in\Xi \mid \forall k:\ g_k^T h_l=\|g_k\|=S\} \quad , \quad |\Theta_l|=C_P^S, \quad \forall l=\overline{1,C_N^P} \quad , (6)$$

где h_l – N×1-вектор: $H_l=E\, h_l$.

Для характеристики множеств Θ_l вводится *критерий взаимной близости*:

$$W(\Theta_l)=W(\hat{c}_{l,1},\dots,\hat{c}_{l,K}), \quad c_{l,k}\in\Theta_l, k=\overline{1,K} \quad , \quad l=\overline{1,L} \quad . (7)$$

Множество Θ_l с минимальным значением $W(\Theta_l)$ обозначим $\hat{\Theta}$ и будем называть *наиболее согласованным множеством оценок*. Задача заключается в отыскании этого множества и построении на нем точечной оценки. Нахождение наиболее согласованного множества по существу сводится к отысканию индекса $\hat{l}$:

$$W(\hat{l}) = \min_{l} W(\Theta_l) \quad .$$

В рамках описанной общей постановки согласованной идентификации могут использоваться разные схемы формирования подсистем и критерии взаимной близости. Далее в работе используется критерий взаимной парной близости вида:

$$W[\Theta_l] = \sum_{i,j=1}^{K} (\hat{c}_{l,i} - \hat{c}_{l,j})^2 \quad . (8)$$

3. Параллельные схемы реализации робастных алгоритмов

Алгоритмы вычисления решения методами RANSAC и согласованной идентификации — параллельные по обрабатываемым подсистемам нижнего уровня. Для формирования подсистемы нижнего уровня используется следующий алгоритм генерации сочетаний.

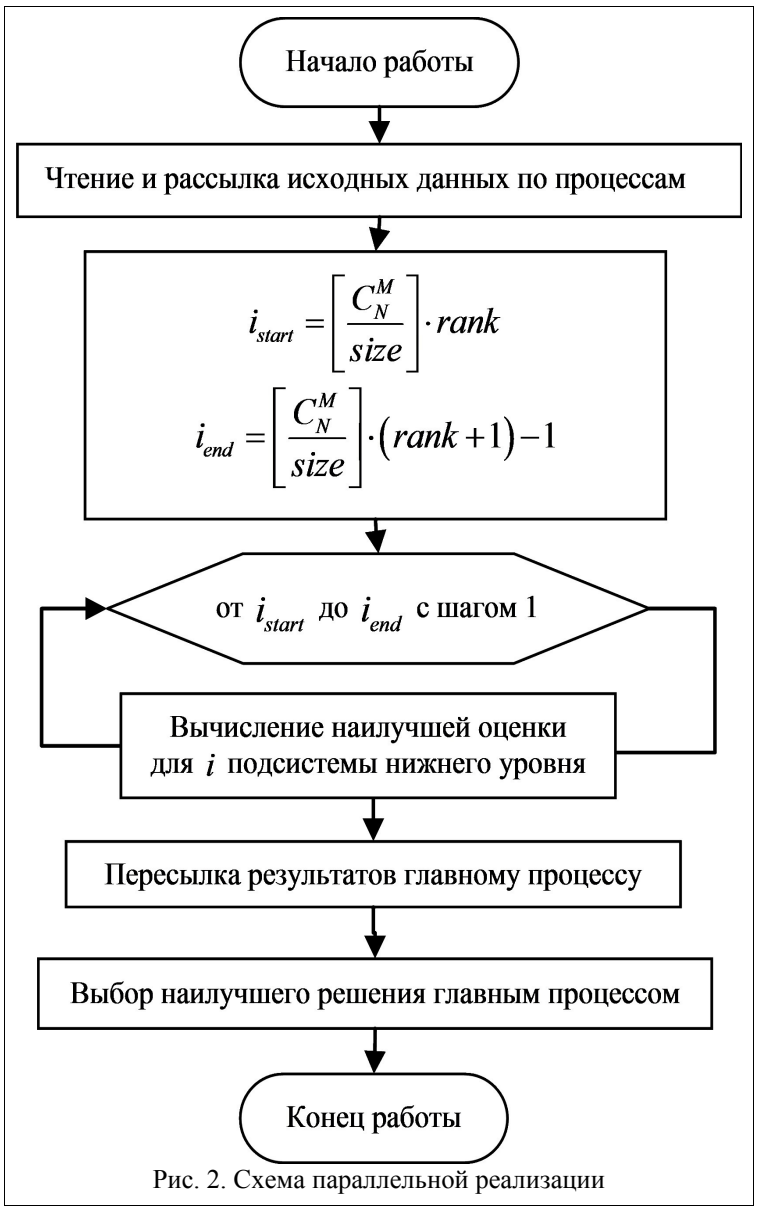

Рис. 2. Схема параллельной реализации

Общая схема параллельной реализации алгоритмов, рассматриваемых в данной работе, приведена на рисунке 2.Вычисление оценки в каждом процессе производится следующим образом. Среди всех подсистем нижнего уровня выбирается диапазон подсистем для обработки в этом процессе. Затем для всех подсистем нижнего уровня из выбранного диапазона производится вычисление решения и нахождение числа инлаеров для этого решения среди строк матрицы, не вошедших в оцениваемую подсистему нижнего уровня.

В методе RANSAC в качестве критерия для отнесения строки к инлаерам используется следующее выражение:

$$X_i - \begin{cases} inlier, & if \dfrac{\sum\limits_j X_{ij}\hat{c}_j - y_j}{\sum\limits_j X_{ij}^2} < t, \\ outlier, & if \dfrac{\sum\limits_j X_{ij}\hat{c}_j - y_j}{\sum\limits_j X_{ij}^2} > t, \end{cases}$$

где X_i - строка матрицы, $\hat{c}_j$ - оценка на подсистеме нижнего уровня, t - порог.

Следует заметить, что обычно при реализации алгоритма RANSAC процесс останавливают на некотором шаге, если есть основания полагать, что достаточно «хорошая» модель найдена. В данном случае число инлаеров рассчитывается для всех решений, а не для некоторого случайного подмножества. Такая схема

реализации принята для обеспечения достижения алгоритмом RANSAC гарантированно наилучшего результата, т.к. одна из задач настоящей работы состоит в сравнении его точностных характеристик с точностью метода согласованной идентификации.

При реализации описанного выше метода согласованной идентификации на многопроцессорной вычислительной системе основная проблема связана с необходимостью хранить на каждом узле большое число оценок, вычисляемых на всех возможных вариантах подсистем нижнего уровня. Общее их число равно C_M^N, где N — число наблюдений (строк матрицы **X**), M — число параметров модели (столбцов матрицы **X**). Для хранения одной оценки необходим массив из M элементов. Если для каждого элемента выделяется 8 байт (тип double), для хранения всех оценок потребуется $8C_N^M M$ байт. Далее рассматривается модификация метода согласованной идентификации, в которой в отличие от алгоритмов, описанных в работах [9], [10], имеется возможность последовательного просмотра формируемых моделей.

Общая схема параллельного алгоритма реализации метода согласованной идентификации также соответствует приведенной на рисунке 2 общей схеме. Отличие от RANSAC имеет место лишь на шаге вычисления наилучшей оценки. Это связано со способом вычисления этой оценки. Для каждой обрабатываемой подсистемы нижнего уровня формируется подмножество подсистем, на котором затем вычисляется оценка согласованности. Выбор подмножества подсистем для заданной квадратной подсистемы осуществляется следующим образом. Для получения подсистемы, входящей в это подмножество, в рассматриваемой подсистеме одна строка заменяется строкой исходной системы, не входящей в эту подсистему, Поскольку любая из M строк рассматриваемой подсистемы может быть заменена на одну из (N - M) строк, не входящих в эту подсистему, в результате формируется $M\times(N-M)$ подсистем.

Описанный метод формирования оценок избавляет от необходимости хранения решений всех квадратных подсистем нижнего уровня и уменьшает число сравнений, однако при этом требуется каждый раз заново вычислять эти решения. В результате последовательной обработки всех подсистем нижнего уровня исходной системы и сравнения полученных для каждой подсистемы оценок согласованности находится наиболее согласованное решение.

4. Анализ степени параллелизма алгоритмов

Для сравнительной оценки эффективности параллельной реализации описанных выше алгоритмов представляет интерес оценить относительный вес накладных расходов на организацию взаимодействия процессоров, синхронизацию параллельных вычислений и т.п. Известно [11], что эффективность параллельной реализации на однородной вычислительной системе определяется следующим соотношением:

$$E=\frac{T_1}{sT_s}=\frac{T_1}{T_1+T_0}=\frac{1}{1+\frac{T_0}{T_1}},$$

где T_0 – временные затраты на накладные расходы, T_1 – время решения задачи при ее последовательной реализации, T_s – время решения задачи при параллельной реализации.

Рассмотрим, из чего составляется время выполнения последовательной и параллельной программы. Для этого введем следующие обозначения для временных затрат на реализацию отдельных работ. Пересылка и получение данных – T_{exch}, расчет начальной и конечной подсистемы на каждом узле – T_{util}, вычисление решения и оценки этого решения – T_{solve}, выбор наилучшего решения – T_{choose}. Для простоты будем полагать, что число обрабатываемых подсистем нижнего уровня L нацело делится на число процессов s, тогда:

$$T_1=LT_{solve},$$

$$T_s=T_{exch}+T_{util}+\frac{L}{s}T_{solve}+T_{choose},$$

$$T_0=sT_s-T_1=s\left(T_{exch}+T_{util}+T_{choose}\right),$$

а эффективность определится как

$$E=\frac{1}{1+\frac{s\left(T_{exch}+T_{util}+T_{choose}\right)}{LT_{solve}}}.$$

Если $T_{exch}+T_{util}+T_{choose}\ll T_{solve}$, то даже для предельного случая $s=L$ эффективность параллельного алгоритма близка к единице. Это связано с тем, что в процессе вычислений отсутствует взаимодействие между узлами на каждой итерации. Данные передаются только перед началом вычислений однократно и после расчета предварительных результатов на каждом узле. Поскольку число L подсистем

нижнего уровня велико (например, для матрицы размерности 20×7 $L=77520$, а s обычно ограничено, оба описанных алгоритма имеют высокую степень масштабируемости.

5. Экспериментальные исследования

Сравнительные экспериментальные исследования метода согласованной идентификации и алгоритма RANSAC проводились с целью сопоставления их точности и вычислительной сложности в одинаковых условиях функционирования. Эксперименты проводились на наборах данных, которые моделировались следующим образом.

Строилась система вида (2) с числом оцениваемых параметров M равным 4, 5 и 6 и числом наблюдений N равным, соответственно, 12, 15 и 18. Таким образом, для каждой реализации системы (2) генерировалась матрица $\boldsymbol{X}$ размерности 12x4 (15x5, 18x6) и вектор y размерности 12x1 (15x1, 18x1). Компоненты вектора параметров с задавались в виде равномерно распределенных случайных чисел (РРСЧ) в диапазоне от 1 до 10. Наблюдения (компоненты N×1-векторов, из которых составлена матрица **X**) моделировались как случайные последовательности с фиксированными дисперсиями для каждой реализации. Шум формировался таким образом, чтобы для нормальных ошибок отношение сигнал/шум находилось в пределах 40-60 дБ. Для аномальных ошибок отношение сигнал/шум задавалось в пределах 0-10 дБ.

В большинстве известных модельных примеров, используемых в публикациях, посвященных исследованию эффективности метод RANSAC, число наблюдений N значительно (на 2-3 порядка) превышает число оцениваемых параметров M. Если при этом интенсивность помех находится в разумных пределах, традиционые статистические схемы обработки дают хороший результат. Поэтому при большом числе степеней свободы цель большинства работ состоит в том, чтобы показать преимущества метода RANSAC в условиях, когда число аномальных ошибок достигает 80-90% от общего числа наблюдений.

В настоящей работе исследуется случай, когда число наблюдений N – одного порядка с M (в 3-4 раза больше). При этом статистические схемы обработки принципиально непригодны. В то же время интенивность аномальных ошибок задавалась более реалистичной – 50-60% от числа степеней свободы N-M полученной системы.

Для сравнительной оценки точности и надежности алгоритмов используются следующие показатели: идентификация считается верной, если отношение нормы вектора погрешности к норме вектора параметров не превосходит 0.3.

На рисунке 3 приведены графики средних значений погрешности верных идентификаций для различных значений числа оцениваемых параметров (светлый – для метода RANSAC, темный – для метода согласованной идентификации). На рисунке 4 для тех же значений числа оцениваемых параметров приведены графики числа ложных идентификаций на 100 экспериментов. Нетрудно заметить, что на всех реализациях метод согласованной идентификации показывал лучшие результаты, как по надежности, так и по точности результатов.

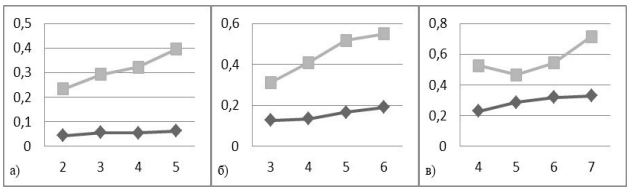

Рис. 3. Средняя погрешность верных идентификаций: а) N=12, M=4; б) N=15, M=5; в) N=18, M=6.

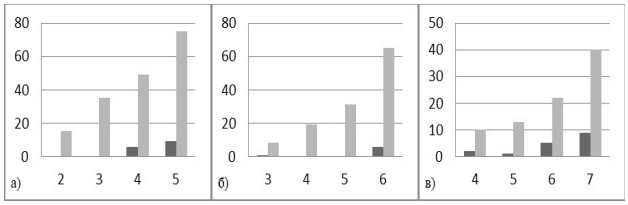

Рис. 4. Процент ложных идентификаций: а) N=12, M=4; б) N=15, M=5; в) N=18, M=6

Для сравнения вычислительной сложности и оценки степени параллелизма алгоритмов проведены эксперименты по оценке фактически достижимых ускорения и эффективности. На рисунках 5,6 приведены полученные в эксперименте графики ускорения и эффективности параллельной реализации для RANSAC и метода согласованной идентификации для различных размерностей задачи.

Заметим, что решение описанной задачи на многопроцессорной системе, например, при N = 60, M = 20 с использованием 20-ти процессоров занимает около 5 мин. В то же время, решение на однопроцессорной системе с использованием процессора того же типа превышает 1 час как для метода RANSAC, так и для метода согласованной идентификации.

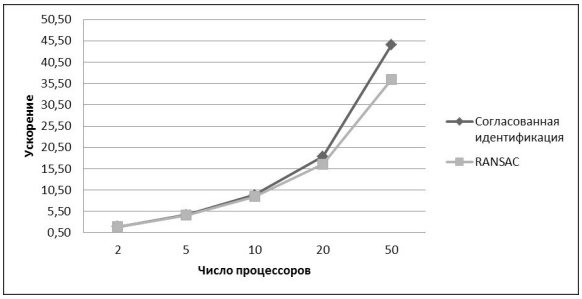

Рис. 5. Ускорение вычислений

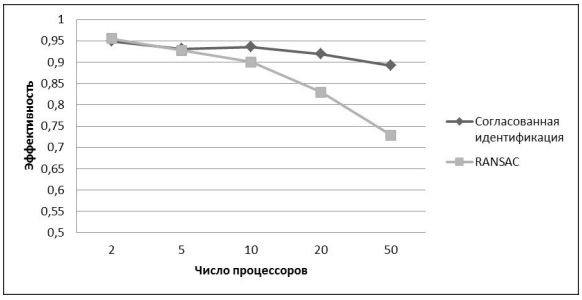

Рис. 6. Эффективность параллельной реализации

6. Заключение

Таким образом, показана возможность достижения существенно более высокой точности и надежности оценок методом согласованной идентификации. Вместе с тем, вычислительная сложность рассмотренной модификации метода согласованной идентификации пока еще остается более высокой по сравнению с методом RANSAC. Конечно, высокая вычислительная сложность метода вызывает неудовлетворенность, однако на практике возникают ситуации, когда задачу определения параметров модели необходимо решить с максимально возможной точностью по одному, возможно малому и притом сильно зашумленному, набору данных. К счастью, как показано в настоящей работе, при реализации на многопроцессорных системах это не является серьезной проблемой, т.к. алгоритм обладает высокой степенью параллелизма.

7. Благодарности

Работа выполнена при поддержке РФФИ (проекты № 11-07-12051-офи-м, № 12-07-00581-а).

ЛИТЕРАТУРА:

1. D. Graupe. “Identification of systems”. Colorado State University Fort Collins. Robert E. Kriger Publishing Company Huntington, New York, 1978.
2. L.Ljung.“System Identification”. Theory for the User. University of Linkoping, Sweden Prentice - Hall, Inc., 1987.
3. Р.Е. Калман, Идентификация систем с шумами. Успехи математических наук, т. 40, вып. 4(244) 1985.
4. Allen D.M. The prediction sum of squares as a criterion for selecting predictor variables. –University of Kentucky, Technical report, 1971.
5. Вапник В.Н., Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979.
6. Martin A. Fischler and Robert C. Bolles. "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography", Comm. Of the ACM 24, 1981, pp. 381–395.
7. P.H.S. Torr and D.W. Murray. "The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix". International Journal of Computer Vision 24, 1997, pp. 271-300.
8. Фурсов В.А. Проблемы вычисления оценок по малому числу наблюдений. Лекция в тр. молодежной школы "Математическое моделирование 2001", Самара, 13-16 июня 2001, с. 56-63.
9. Fursov V.A. Estimates Conformity Principle in the Problems of Identification. Computional Science – ICCS 2003. International Conference Melbourne, Australia and St-Petersburg, Russia. June 2003, Proceedings, Part II, pp. 463-470.
10. Фурсов В.А. Согласованная идентификация управляемого объекта по малому числу наблюдений. Теоретический и прикладной научно-технический журнал "Мехатроника, автоматизация, управление."Москва. Новые технологии. 2010. 3(108). 2-8 с.
11. Гергель В.П., Фурсов В.А. Лекции по параллельным вычислениям: учеб. пособие / Самара: Изд-во Самар. гос. аэрокосм. ун-та, 2009. – 164 с.