4. **Демин Н. С., Жадан Л. И.** Об оптимальности процедуры исключения аномальных измерений.— Автометрия, 1983, № 4.
5. **Медич Дж.** Статистически оптимальные линейные оценки и управление.— М.: Энергия, 1973.
6. **Демин Н. С., Жадан Л. И.** Синтез и анализ оптимального алгоритма фильтрации для дискретных сигналов с аномальными помехами.— Радиотехника и электроника, 1984, т. XXIX, № 2.
7. **Алберт А.** Регрессия, псевдоинверсия и рекуррентное оценивание.— М.: Наука, 1977.
8. **Абгарян К. А.** Матричные и асимптотические методы в теории линейных систем.— М.: Наука, 1973.
9. **Маркус М., Минк Х.** Обзор по теории матриц и матричных неравенств.— М.: Наука, 1972.
10. **Ройтенберг Я. Н.** Автоматическое управление.— М.: Наука, 1978.
11. **Рао С. Р.** Линейные статистические методы и их применения.— М.: Наука, 1968.
12. **Гантмахер Ф. Р.** Теория матриц.— М.: Наука, 1966.
13. **Ланкастер Т.** Теория матриц.— М.: Наука, 1982.

УДК 519.24

Б. М. ШУМИЛОВ

(Томск)

АЛГОРИТМ СЖАТИЯ ДАННЫХ, СОДЕРЖАЩИХ ПРОТЯЖЕННЫЕ ВЫБРОСЫ

В работах [1—5] привлекалось внимание к проблеме фильтрации данных, содержащих выбросы — быстрые и значительные флуктуации сигнала, обусловленные, например, случайными продолжительными и кратковременными неполадками в устройствах аналого-цифрового преобразования данных или в чисто цифровой части автоматизированной системы обработки данных. От одиночных либо k-кратных изолированных выбросов можно избавиться, используя алгоритм скользящей медианы [6]. Для случая протяженных выбросов общие методы фильтрации не разработаны. Ниже предлагается алгоритм фильтрации и сжатия данных кусочными многочленами первой степени, который дает возможность выявить начало и конец протяженного выброса, а при необходимости и избавиться от него.

Пусть в течение интервала времени $0 \leqslant t \leqslant T$ регистрируется некоторая функциональная зависимость, причем $f(t)$ есть величина сигнала, измеренная в момент времени t. В результате аналого-цифрового преобразования в моменты времени $t_0 = 0$, $t_1, \ldots, t_i, \ldots, t_N = T$ исходная непрерывная функция отображается в последовательность отсчетов $f(t_0)$, $f(t_1), \ldots, f(t_i), \ldots, f(t_N)$. Эта последовательность и вводится в ЭВМ. Обычно моменты t_i $(i = 1, \ldots, N)$ не запоминаются, потому что их выбирают равноотстоящими, т. е. $t_i = iT/N$. Частоту аналого-цифрового преобразования $F = N/T$ выбирают довольно большой (согласно теореме Котельникова), чтобы достаточно подробно отображать развитие во времени быстропротекающих процессов.

Будем считать, что реально измеряемые значения $f(t_i)$ представляют собой сумму детерминированного полезного сигнала $u(t_i)$ и случайной помехи $\xi(t_i)$, причем помеха состоит из белого гауссова шума с нулевым средним и постоянной дисперсией σ^2 и протяженных выбросов типа ступенек. Относительно сигнала $u(t_i)$ предположим, что он имеет плавно изменяющиеся пологие участки, на которых его можно удовлетворительно описать отрезками прямых (осуществляя таким образом сжатие данных). Возьмем один такой участок $t_k, t_{k+1}, \ldots, t_{k+m}$ без точек разрыва и аппроксимируем значения $f(t_i)$, $i=k, \ldots, k+m$, по методу наименьших квадратов при помощи функции $p_m(t_i)=\alpha_m+\lambda_m t_i$, т. е.

$$\sum_{i=k}^{k+m}(f(t_i)-\alpha_m-\lambda_m t_i)^2 \to \min_{\alpha_m,\, \lambda_m}.$$

Тогда

$$\alpha_m=\frac{2}{(m+1)(m+2)}\sum_{l=0}^{m}(2m+1-3l)\,f(t_{k+l})-\frac{k}{F}\lambda_m,$$

$$\lambda_m=\frac{6F}{m(m+1)(m+2)}\sum_{l=0}^{m}(2l-m)\,f(t_{k+l}),$$

причем величина

$$\widehat{\sigma}_m^2=\frac{1}{m-1}\sum_{i=k}^{k+m}(f(t_i)-p_m(t_i))^2$$

служит оценкой для дисперсии суммы шума и отклонения сигнала $u(t_i)$ от линейной модели на рассматриваемом интервале $t_k, \ldots, t_{k+m}$. Для того чтобы величина отклонения была согласована в дискретном среднеквадратическом смысле с уровнем аддитивного белого гауссова шума с нулевым средним и постоянной дисперсией σ^2, проверяется статистическая гипотеза о том, что оценка $\widehat{\sigma}_m^2$ получена из нормального распределения $N(0, \sigma)$. Тогда отношение $\widehat{\sigma}_m^2/\sigma^2$ подчиняется распределению χ^2 с $m-1$ степенями свободы.

Алгоритм сжатия состоит в следующем. Первому участку приписывается значение $k_1=0$. Пусть найдено такое значение m_1, для которого впервые отношение $\widehat{\sigma}^2_{m_1+1}/\sigma^2$ значимо отличается от ожидаемого согласно распределению χ^2 при заданном уровне значимости. Это соответствует тому, что отношение $\widehat{\sigma}_m^2/\sigma^2$ не противоречит выдвинутой гипотезе при $2 \leqslant m \leqslant m_1$ и противоречит ей при $m=m_1+1$. Тогда точка t_{m_1+1} принимается в качестве начала следующего участка, соответствующего $k_2=m_1+1$; в результате отыскивается точка со значением $k_3=k_2+m_2+1$ и т. д. В итоге получается, вообще говоря, разрывная кусочно-линейная функция времени. Чтобы выявить разрывы на стыках двух соседних отрезков прямых, соответствующие началу либо концу протяженного выброса, для каждого значения $j=2, 3, \ldots$ вычисляются выражения

$$\widetilde{\alpha}_j=\frac{2}{(m_j+2)(m_j+3)}\sum_{l=-1}^{m_j}(2m_j-3l)\,f(t_{k_j+l})-\frac{k_j-1}{F}\widetilde{\lambda}_j,$$

$$\widetilde{\lambda}_j=\frac{6F}{(m_j+1)(m_j+2)(m_j+3)}\sum_{l=-1}^{m_j}(2l+1-m_j)\,f(t_{k_j+l}).$$

Тогда

$$\widetilde{\sigma}_j^2=\frac{1}{m_j}\sum_{i=k_j-1}^{k_j+m_j}\left(f(t_i)-\widetilde{p}_j(t_i)\right)^2,$$

где $\widetilde{p}_j(t_i)=\widetilde{\alpha}_j+\widetilde{\lambda}_j t_i$, служит оценкой для дисперсии на интервале $t_{k_j-1}, t_{k_j}, \ldots, t_{k_j+m_j}$. Если отношение $\widetilde{\sigma}_j^2/\sigma^2$ значимо отличается от ожидаемого согласно распределению χ^2 с m_j степенями свободы при заданном уровне

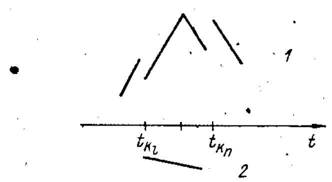

Рис. 1. Результат сжатия (*1*) сигнала, содержащего протяженный выброс, и вычитаемый отрезок прямой (*2*).

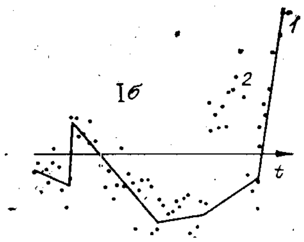

Рис. 2. Результат сжатия (*1*) реальной ЭКГ с наложенным на нее прямоугольным импульсом (*2*).

значимости, то возможна одна из следующих ситуаций: либо точка $t_{k\,j-1}$ является концом протяженного выброса, либо точка t_{k_j} — начало протяженного выброса. Если найдены две точки t_{k_l}, t_{k_n-1}, $l<n$, связанные соответственно с началом и концом протяженного выброса, то избавиться от него можно, например, вычитая отрезок прямой, интерполирующей в точках t_{k_l-1}, t_{k_n} значения $p_{m_l}(t_{k_l-1}) - p_{m_{l-1}}(t_{k_l-1})$ и $p_{m_{n-1}}(t_{k_n}) - p_{m_n}(t_{k_n})$ (рис. 1). В противном случае остается интерпретировать каждый интервал t_{k_j-1}, t_{k_j}, содержащий разрыв, как участок резкого изменения исследуемой зависимости, на котором она представима отрезком прямой, интерполирующей величины $p_{m_{j-1}}(t_{k_j-1})$, $(p_{m_j}(t_{k_j}))$. В случае когда отношение $\tilde{\sigma}_j^2/\sigma^2$ не противоречит гипотезе нормальности распределения шума, проводится склеивание двух соседних отрезков прямых по формуле локальной аппроксимации сплайнами [7], согласно которой значение аппроксимирующей ломаной в точке t_{k_j-1} принимается равным выражению

$$\frac{c_{j-1}p_{m_{j-1}}(t_{k_j-1}) + \tilde{c}_j\tilde{p}_j(t_{k_j-1}) - f(t_{k_j-1})}{c_{j-1} + \tilde{c}_j - 1},$$

$$c_{j-1} = \frac{(m_{j-1}+1)(m_{j-1}+2)}{2(2m_{j-1}+1)}, \quad \tilde{c}_j = \frac{(m_j+2)(m_j+3)}{2(2m_j+3)},$$

если в точке $t_{k_{j-1}}$ обнаружен разрыв или $j=2$, либо выражению

$$\frac{\tilde{c}_{j-1}\tilde{p}_{j-1}(t_{k_j-1}) + \tilde{c}_j\tilde{p}_j(t_{k_j-1}) - f(t_{k_j-1})}{\tilde{c}_{j-1} + \tilde{c}_j - 1}$$

(вывод последних дан в приложении).

Отметим, что истинное значение параметра σ^2 обычно неизвестно и, более того, дисперсия может меняться во времени. Тогда в качестве статистики, по которой принимается решение о соответствии оценки $\hat{\sigma}_m^2$ гипотезе нормальности закона распределения помехи на участке $[t_{k_j}, t_{k_j+m}]$, используется отношение $\hat{\sigma}_m^2/\hat{\sigma}_{m_{j-1}}^2$. В предположении квазистационарности изучаемого процесса эта величина подчиняется распределению Фишера с $(m-1,\ m_{j-1}-1)$ степенями свободы, для которого в [8] даны простые аппроксимирующие выражения. В качестве $\hat{\sigma}_{2m_0}$ можно взять оценку дисперсии на изолинии (участке, где присутствует только гауссов шум, а полезный сигнал и выбросы отсутствуют). Тогда $f(t_i) = \xi(t_i)$, $i = -m_0, \ldots, -1$, $\hat{\sigma}_{m_0}^2 = \frac{1}{m_0-1}\sum_{i=-m_0}^{-1} f(t_i)^2$. Аналогично для выявления разрыва берется отношение $\tilde{\sigma}_j^2/\hat{\sigma}_{m_{j-2}}^2$.

Данный алгоритм реализован на языке Фортран-IV на ЭВМ «Электроника 100-25». Тестовой функцией служила реальная электрокардиограмма с наложенным на нее прямоугольным импульсом (на рис. 2 она

изображена точками). Значение параметра σ^2 выбрано равным 285. В ка-
дован для практического применения. Отметим, что все вычисления могут проводиться в режиме реального времени с использованием микропроцессорной техники.

ПРИЛОЖЕНИЕ

Вывод формулы локальной аппроксимации сплайнами первой степени. Пусть среди равноудаленных точек $t_0=0,\ t_1,\ \ldots,\ t_N=T$ выбраны p узлов $t_{k_1},\ t_{k_2},\ \ldots,\ t_{k_p}$, расположенных друг от друга на расстоянии $t_{k_{j+1}}-t_{k_j}=m_jT/N$, причем $t_{k_1}=0,\ t_{k_p}=T$. Множество функций, составленных из отрезков прямых на каждом интервале $[t_{k_j},\ t_{k_{j+1}}]$ и непрерывно склеенных в точках t_{k_j}, образует пространство сплайнов первой степени $S_1(x)$. В этом пространстве можно ввести базис, состоящий из B-сплайнов $B_j(x)$, равных 1 в точке t_{k_j} и 0 вне интервала $(t_{k_{j-1}},\ t_{k_{j+1}})$ (рис. 3). Здесь предполагается, что $t_{k_0}=0,\ t_{k_{p+1}}=T$. Будем строить формулу локальной аппроксимации вида

$$S_1(x)=\sum_{j=1}^{p}\ \sum_{i=k_{j-1}}^{k_{j+1}} q_j(t_i)\,f(t_i)\,B_j(x),$$

которая точно воспроизводит любые функции $f(x)$ из пространства сплайнов первой степени. Согласно лемме де Бора — Фикса [7] для этого необходимо и достаточно, чтобы для всех $j=1,\ \ldots,\ p$ выполнялись равенства

$$\sum_{i=k_{j-1}}^{k_{j+1}} q_j(t_i)\,B_n(t_i)=\begin{cases}1, & n=j,\\ 0, & n\neq j.\end{cases}$$

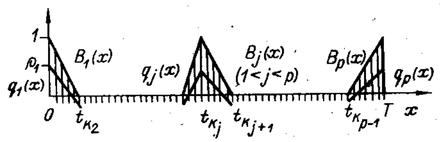

Рис. 3. B-сплайны и функции $q_j(x)$.

Для значений $n\neq j-1,\ j,\ j+1$ эти равенства соблюдаются в силу того, что B-сплайны $B_n(x)$ обращаются в нуль в точках $t_i,\ i=k_{j-1},\ \ldots,\ k_{j+1}$. Покажем, что при

$$\rho_j=\left[\frac{(m_{j-1}+1)(m_{j-1}+2)}{2(2m_{j-1}+1)}+\frac{(m_j+1)(m_j+2)}{2(2m_j+1)}-1\right]^{-1}$$

и функции $q_j(x)$, составленной из отрезков прямых и равной

$$\rho_j(1-m_{j-1})/(2m_{j-1}+1),\ \rho_j,\ \rho_j(1-m_j)/(2m_j+1)$$

соответственно в точках $t_{k_{j-1}},\ t_{k_j},\ t_{k_{j+1}}$ (см. рис. 3), выполняются и остальные равенства. Действительно, учитывая, что $t_i=iT/N$, и проводя замену $i-k_j=l$, находим

$$\sum_{i=k_{j-1}}^{k_{j+1}} q_j(t_i)\,B_{j+1}(t_i)=\sum_{i=k_j+1}^{k_j+m_j} q_j(t_i)\,B_{j+1}(t_i)=$$

$$=\rho_j\sum_{l=1}^{m_j}\left(1-\frac{3l}{2m_j+1}\right)\frac{l}{m_j}=\ldots=0.$$

Аналогично с помощью замены $k_j-i=l$ получаем

$$\sum_{j=k_{j-1}}^{k_{j+1}} q_j(t_i)\,B_{j-1}(t_i)=\rho_j\sum_{l=1}^{m_{j-1}}\left(1-\frac{3l}{2m_{j-1}+1}\right)\frac{l}{m_{j-1}}=\ldots 0.$$

Наконец, используя замены $i-k_{j-1}=l$ и $k_{j+1}-i=l'$, приходим к

$$\sum_{i=k_{j-1}}^{k_{j+1}} q_j(t_i) B_j(t_i) = \rho_j \left(\sum_{l=1}^{m_{j-1}} \frac{3l+1-m_{j-1}}{2m_{j-1}+1} \frac{l}{m_{j-1}} + \right.$$

$$\left. + \sum_{l'=1}^{m_{j-1}} \frac{3l'+1-m_j}{2m_j+1} \frac{l'}{m_j} \right) = \rho_j \left(\sum_{l=1}^{m_{j-1}} \frac{3l+1-m_{j-1}}{2m_{j-1}+1} \frac{l}{m_{j-1}} + \right.$$

$$\left. + \sum_{l'=1}^{m_j} \frac{3l'+1-m_j}{2m_j+1} \frac{l'}{m_j} - 1 \right) = \ldots = 1.$$

Теперь, учитывая, что

$$p_{m_j}(t_{k_j}) = \alpha_{m_j} + \lambda_{m_j} t_{k_j} = \frac{2}{(m_j+1)(m_j+2)} \sum_{l=0}^{m_j} (2m_j+1-3l) f(t_{k_j+l})$$

и (в силу симметрии)

$$p_{m_{j-1}}(t_{k_j}) = \frac{2}{(m_{j-1}+1)(m_{j-1}+2)} \sum_{l=0}^{m_{j-1}} (2m_{j-1}+1-3l) f(t_{k_j-l}),$$

имеем окончательный результат:

$$S_1(t_{k_j}) = \sum_{i=k_{j-1}}^{k_{j+1}} q_j(t_i) f(t_i) = \rho_j \left(\sum_{l=0}^{m_{j-1}} \left(1 - \frac{3l}{2m_{j-1}+1}\right) \times \right.$$

$$\left. \times f(t_{k_j-l}) + \sum_{l=1}^{m_j} \left(1 - \frac{3l}{2m_j+1}\right) f(t_{k_j+l}) \right) = \ldots =$$

$$= \rho_j \left(\frac{(m_{j-1}+1)(m_{j-1}+2)}{2(2m_{j-1}+1)} p_{m_{j-1}}(t_{k_j}) + \frac{(m_j+1)(m_j+2)}{2(2m_j+1)} \times \right.$$

$$\left. \times p_{m_j}(t_{k_j}) - f(t_{k_j}) \right).$$

Отсюда при соответствующем выборе значений m_j вытекают соотношения, приведенные в основной части работы.

ЛИТЕРАТУРА

1. **Шакин В. В.** Вычислительная электрокардиография.— М.: Наука, 1981.
2. **Чемоданова О. А.** Сравнение различных методов обнаружения аномальных погрешностей.— Научн. тр./Моск. лесотехн. ин-т, 1981, вып. 136, с. 168—170.
3. **Фомин А. Ф., Новоселов О. Н., Плющев А. В.** Методы и средства повышения достоверности измерений непрерывных процессов.— Измерения, контроль, автоматизация, 1981, № 4, с. 3—10.
4. **Русинов Л. А., Гуревич А. Л.** Алгоритмическое обеспечение информационно-измерительных систем с аналитическими приборами.— Измерения, контроль, автоматизация, 1982, № 3, с. 9—16.
5. **Паламарчук С. Н.** Алгоритм оценивания параметров режима и анализа грубых ошибок на основе линеаризованных выражений измеряемых величин.— В кн.: Алгоритмы обработки данных в электроэнергетике. Иркутск: СЭИ, 1982, с. 178—184.
6. **Ефимов Л. Н.** Методы порядковых статистик и рангов в задачах обработки наблюдений.— Измерения, контроль, автоматизация, 1981, № 5, с. 19—27.
7. **Шумилов Б. М.** Локальная аппроксимация сплайнами: формулы, точные на сплайнах.— Новосибирск, 1981. (Препринт/АН СССР, Сиб. отд-ние, ВЦ; 86. Семинар «Методы вычислительной и прикладной математики»/Под руководством Г. И. Марчука).
8. **Johnson E. E.** Empirical equations for approximating tabular F values.— Technometrics, 1973, vol. 15, N 2, p. 379—384.
9. **Reinsch C. H.** Smoothing by spline functions.— Numer. Mathem., 1967, vol. 10, N 4, p. 177—183.

Поступила в редакцию 19 июля 1983 г.