

УДК 681.51; 621.391; 519.21

© В. М. Малыхин, А. В. Меркушева

МЕТОДЫ И АЛГОРИТМЫ РАЗДЕЛЕНИЯ СМЕСИ СИГНАЛОВ. II. ПРИМЕНЕНИЕ М-ГРАДИЕНТА К АНАЛИЗУ НЕЗАВИСИМЫХ КОМПОНЕНТ

Анализируется задача разделения смеси сигналов (с восстановлением вида ее компонент) при отсутствии информации о пропорциях и типе смешивания. Метод основан на использовании информационного критерия и адаптивного алгоритма для обучающейся нейронной сети. Рассмотрены несколько видов распределения первичных сигналов, поступающих на сенсоры информационно-измерительной сети, и соответствующие им изменения функции преобразования нейронов. Подход к разделению смеси сигналов включает применение модифицированного градиента (м-градиента) в схеме анализа независимых компонент. (Статью I цикла см. "Научное приборостроение", 2009, т. 19, № 2).

Кл. сл.: смесь сигналов, методы разделения, адаптивный алгоритм, нейронная сеть, м-градиент, анализ независимых компонент

ИСПОЛЬЗОВАНИЕ М-ГРАДИЕНТА И ОБУЧЕНИЕ НЕЙРОННОЙ СЕТИ (НС)

Простая схема линейного смешивания сигналов (рис. 1) в векторно-матричной форме описывается соотношением (1):

$$\mathbf{x}(k) = \mathbf{H}\mathbf{s}(k) + \mathbf{v}(k), \quad (1)$$

где $\mathbf{x}(k) = [x_1(k), \dots, x_m(k)]^T$ — зашумленный сигнал на сенсорах; $\mathbf{s}(k) = [s_1(k), \dots, s_n(k)]^T$ — первичный сигнал (ПС); $\mathbf{v}(k) = [v_1(k), \dots, v_m(k)]^T$ — вектор шума; $\mathbf{H}$ — неизвестная смешивающая матрица размера $m \times n$ и полного ранга. При этом предполагается, что доступным является только вектор-сигнал $\mathbf{x}(k)$, регистрируемый на сенсорах информационно-измерительной сети (ИИС).

Цель — разделение сигналов смеси (РСС) — достигается построением НС с прямым распространением сигнала или с обратными связями (так называемой рекуррентной НС) и соответствующего этим сетям адаптивного обучающего алгоритма, который позволяет оценивать компоненты первичного сигнала $\mathbf{s}(k)$, идентифицировать смешивающую матрицу (СМ) $\mathbf{H}$ или разделяющую матрицу $\mathbf{W}$.¹ Свойство адаптивности алгоритма состоит в возможности отслеживать изменяющуюся во времени систему сигналов.

Сначала выполнен анализ обучающих алгоритмов для простого (мгновенного) смешивания сигналов при известном и одинаковом числе n ПС и

сенсоров (m). При этом $\mathbf{H}$ и $\mathbf{W}$ — несингулярные матрицы размера $n \times n$. Предполагается, что первичные сигналы $s_i(k)$ взаимно независимы и имеют нулевые средние значения (т. е. $E\{s_i(k)\} = 0$), а аддитивный шум $\mathbf{v}(k)$ пренебрежимо мал или становится малым после предварительной обработки. На следующих этапах эти ограничения будут постепенно сняты.

Критерий Кульбака—Лейблера как мера статистической независимости

Для получения хорошей оценки $\mathbf{y} = \mathbf{W}\mathbf{x}$ ПС $\mathbf{s}$ используется "целевая функция" $\rho(\mathbf{y}, \mathbf{W})$. Ее математическое ожидание, называемое функцией риска $R(\mathbf{W}) = E\{\rho(\mathbf{y}, \mathbf{W})\}$, представляет меру взаимной независимости выходного сигнала $\mathbf{y}(k)$. Функция риска $R(\mathbf{W})$ достигает минимума при условии, что компоненты сигнала $\mathbf{y}$ становятся независимыми, а матрица $\mathbf{W}$ (с точностью до масштабирования и перестановки) становится равной обратной СМ $\mathbf{H}^{-1}$. Для достижения такого минимума используется критерий Кульбака—Лейблера [2], который служит мерой независимости. Если $p_{\mathbf{y}}(\mathbf{y}; \mathbf{W})$ — плотность вероятности (ПВ) стохастической ("случайной") переменной $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{H}\mathbf{s}$ и $q(\mathbf{y})$ — ПВ переменной $\mathbf{y}$, у которой все компоненты y_i статистически независимы и, следовательно,

$q(\mathbf{y}) = \prod_{i=1}^n q_i(y_i)$ (т. е. представима

произведением ПВ компонент $q_i(y_i)$), то $q(\mathbf{y})$ — служит в критерии плотностью для сопоставления.

¹ Более полное описание задачи разделения смеси сигналов содержится в [1].

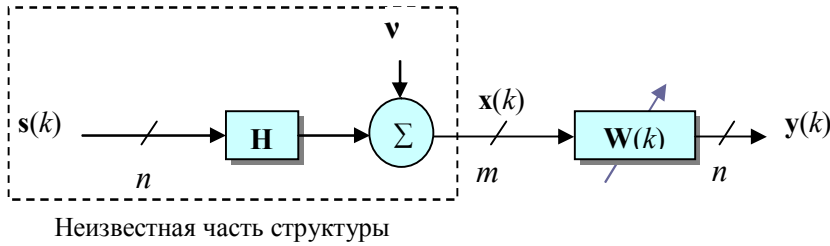


Рис. 1. Блок-схема, иллюстрирующая анализ независимых компонент и разделение сигналов смеси

Критерий Кульбака—Лейблера (ККЛ) дает меру отличия ПВ $p_y(\mathbf{y}; \mathbf{W})$ (т. е. ПВ, где $\mathbf{y}$ получено с действительной величиной $\mathbf{W}$) от ПВ $q(\mathbf{y})$, предполагающей независимость компонент.

Таким образом, функция риска $R(\mathbf{W}) = E\{\rho(\mathbf{y}, \mathbf{W})\}$ представима в виде ККЛ $K_{pq}(\mathbf{W})$:

$$R(\mathbf{W}) = E\{\rho(\mathbf{y}, \mathbf{W})\} = K_{pq}(\mathbf{W}) = K \left[p_y(\mathbf{y}; \mathbf{W}) \right]_{q(\mathbf{y})} = \int p_y(\mathbf{y}, \mathbf{W}) \log \frac{p(\mathbf{y}, \mathbf{W})}{q(\mathbf{y})} d\mathbf{y}. \quad (2)$$

Следовательно, $R(\mathbf{W}) = K_{pq}(\mathbf{W})$ показывает, насколько ПВ $p_y(\mathbf{y}; \mathbf{W})$ далека от сопоставляемой ПВ $q(\mathbf{y})$. Если ПВ $q(\mathbf{y})$ равна истинному распределению $p_s(\mathbf{s})$ первичных сигналов, $p_y(\mathbf{y}; \mathbf{W}) = p_s(\mathbf{s})$ при $\mathbf{W} = \mathbf{H}^{-1}$, поскольку в этом случае $\mathbf{y} = \mathbf{W} \mathbf{H} \mathbf{s} = \mathbf{s}$.²

Для задачи анализа независимых компонент принимается, что $q(\mathbf{y})$ является произведением ПВ независимых переменных y_i : $q(\mathbf{y}) = \tilde{p}(\mathbf{y}) = \prod_{i=1}^n p_i(y_i)$, где $p_i(y_i)$ — ПВ y_i ($i = 1, \dots, n$).³ Поэтому мера независимости принимает форму выражения (3):

² Использование $K_{pq}(\mathbf{W})$ в качестве функции риска $R(\mathbf{W})$ имеет дополнительное преимущество. Даже если $q(\mathbf{y})$ не равно p_s , минимизация $R(\mathbf{W})$ все равно определяет $\mathbf{W} = \mathbf{H}^{-1}$. Кроме того, большую часть обучающих алгоритмов, не использующих ККЛ явно, можно трактовать на основе функции риска в виде $K_{pq}(\mathbf{W})$ (отличие состоит только в выборе ПВ для сопоставления $q(\mathbf{y})$). Среди них — алгоритмы на основе максимизации энтропии [3], анализа независимых компонент [4] и анализа максимального правдоподобия [5].

³ Строго говоря, $p_i(y_i)$ — это маргинальная ПВ для y_i : $p_i(y_i) = \int_{-\infty}^{+\infty} p_y(\mathbf{y}) d\tilde{\mathbf{y}}^i$, где интегрирование осуществляется по $\tilde{\mathbf{y}}^i[y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n]^T$ (т. е. из $\mathbf{y}$ удаляется переменная y_i).

$$K_{pq} = E\{\rho(\mathbf{y}, \mathbf{W})\} = \int p_y(\mathbf{y}) \log \frac{p(\mathbf{y})}{\prod_i q_i(y_i)} d\mathbf{y}, \quad (3)$$

где $\rho(\mathbf{y}, \mathbf{W}) = \log(p_y(\mathbf{y}) / q(\mathbf{y}))$.

ККЛ имеет также трактовку взаимной информации:

$$K_{pq} = -H(\mathbf{y}) - \sum_{i=1}^n \int_{-\infty}^{+\infty} p_y(y_i) \log q_i(y_i) dy_i, \quad (4)$$

где дифференциальная энтропия выходного сигнала $\mathbf{y} = \mathbf{W} \mathbf{x}$ определяется соотношением

$$H(\mathbf{y}) = - \int_{-\infty}^{+\infty} p_y(\mathbf{y}) \log p_y(\mathbf{y}) d\mathbf{y}. \quad (5)$$

Когда $q_i(y_i) = \tilde{p}_i(y_i)$, то (учитывая, что $d\mathbf{y} = d\tilde{\mathbf{y}}^i dy_i$) второе слагаемое в (4) может быть выражено через маргинальную энтропию $H_i(y_i)$:

$$\begin{aligned} \int_{-\infty}^{+\infty} p_y(\mathbf{y}) \log \tilde{p}_i(y_i) d\mathbf{y} &= \int_{-\infty}^{+\infty} \log \tilde{p}_i(y_i) \int_{-\infty}^{+\infty} p_y(\mathbf{y}) d\tilde{\mathbf{y}}^i dy_i = \\ &= \int_{-\infty}^{+\infty} \tilde{p}_i(y_i) \log \tilde{p}_i(y_i) dy_i = E\{\log(\tilde{p}_i(y_i))\} = -H_i(y_i). \end{aligned} \quad (6)$$

Следовательно, критерий КЛ может быть выражен разностью между энтропией $H(\mathbf{y})$ и маргинальной энтропией $H_i(y_i)$:

$$K_{pq} = E\{\rho(\mathbf{y}, \mathbf{W})\} = -H(\mathbf{y}) + \sum_{i=1}^n H_i(y_i). \quad (7)$$

Для $\mathbf{y} = \mathbf{W} \mathbf{x}$ дифференциальная энтропия выражается соотношением⁴

$$H(\mathbf{y}) = H(\mathbf{x}) + \log |\det(\mathbf{W})|,$$

⁴ На основании использования правила соответствия ПВ линейному преобразованию $\mathbf{W}$.

где $H(\mathbf{x}) = -\int_{-\infty}^{+\infty} p_x(\mathbf{x}) \log p_x(\mathbf{x}) d\mathbf{x}$ не зависит от матрицы $\mathbf{W}$. С использованием этих соотношений функция риска для общего вида ПВ q приобретает форму

$$R(\mathbf{W}) = K_{pq} = E\{\rho(\mathbf{y}, \mathbf{W})\} = -H(\mathbf{x}) - \log|\det(\mathbf{W})| - \sum_{i=1}^n E\{\log q_i(y_i)\}; \quad (8)$$

причем $H(\mathbf{x})$ можно опустить, т. к. это слагаемое не включает зависимости от $\mathbf{W}$.

Основные правила обучения, основанные на использовании м-градиента

Для получения матрицы $\mathbf{W}$ (обратной относительно СМ $\mathbf{H}^{-1}$) требуется минимизация функции риска (8). Поэтому обучающий алгоритм реального времени (РВ) основывается на методе градиентного спуска:

$$\Delta \mathbf{W}(k) = \mathbf{W}(k+1) - \mathbf{W}(k) = -\eta(k) \frac{\partial \rho[\mathbf{y}(k), \mathbf{W}]}{\partial \mathbf{W}},$$

где $\eta(k)$ — скорость обучения, $\partial \rho / \partial \mathbf{W}$ — $(n \times n)$ -матрица градиента с элементами $\partial \rho / \partial w_{ij}$.

Вычисление матрицы градиента (с учетом (8)) приводит к соотношению

$$\Delta \mathbf{W}(k) = \eta(k) [\mathbf{W}^{-T}(k) - \mathbf{f}[\mathbf{y}(k)] \mathbf{x}^T(k)],$$

где $\mathbf{f}(\mathbf{y}) = [f_1(y_1), f_2(y_2), \dots, f_n(y_n)]^T$ — вектор-столбец с i -й компонентой, которая равна

$$f_i(y_i) = -\frac{d \log g_i(y_i)}{dy_i} = -\frac{d g_i(y_i) / dy_i}{g_i(y_i)} = -\frac{g'_i(y_i)}{g_i(y_i)},$$

где $g_i(y_i)$, $i=1, \dots, n$ — приближенные (обычно параметрические) модели ПВ первичных сигналов $\{s_i\}$. Обычно $-\partial \rho / \partial \mathbf{W}$ представляет направление максимального уменьшения функции ρ , но если пространством оптимизируемых параметров является множество матриц $\{\mathbf{W}\}$, то, как показано Амари, Кардосо, Ченом, Сичоки (Amari, Cardoso, Chen, Cichocki) [6, 7], такое направление определяет модифицированный градиент (**м-градиент**), который представляется соотношением⁵

⁵ Строго говоря, множество матриц $\{\mathbf{W}\}$ является мультипликативной группой (с единицей $\mathbf{I}$). В работах [6] и [7] для такого пространства матриц $\{\mathbf{W}\}$ введена "естественная" метрика, которая порождает м-градиент. Иногда м-градиент называют также естественным.

$$-\frac{\partial \rho(\mathbf{y}, \mathbf{W})}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = [\mathbf{I} - \mathbf{f}(\mathbf{y}) \mathbf{y}^T] \mathbf{W}.$$

Следовательно, алгоритм обучения представляется в следующей форме:

$$\begin{aligned} \Delta \mathbf{W}(k) &= -\eta(k) \frac{\partial \rho[\mathbf{y}(k), \mathbf{W}]}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = \\ &= \eta(k) [\mathbf{I} - \mathbf{f}[\mathbf{y}(k)] \mathbf{y}^T(k)] \mathbf{W}(k). \end{aligned} \quad (9)$$

Другой подход к вычислению м-градиента для $\rho(\mathbf{y}, \mathbf{W})$, выраженного соотношением (8), состоит в использовании дифференциала ρ по $\mathbf{W}$:

$$d\rho = \rho(\mathbf{y}, \mathbf{W} + d\mathbf{W}) - \rho(\mathbf{y}, \mathbf{W}) = \sum_{i,j} \frac{\partial \rho}{\partial w_{ij}} dw_{ij}.$$

Это выражение представляется в виде $d\rho = -\text{tr}(d\mathbf{W} \mathbf{W}^{-1}) + \mathbf{f}^T(\mathbf{y}) d\mathbf{y}$, где tr — след матрицы; $\mathbf{f}(\mathbf{y})$ — вектор-столбец с компонентами $f_i(y_i) = -g'_i(y_i) / g_i(y_i)$ [7]. Для оценки ПС $\mathbf{y} = \mathbf{W} \mathbf{x}$ справедливо соотношение $d\mathbf{y} = d\mathbf{W} \cdot \mathbf{x} = d\mathbf{W} \cdot \mathbf{W}^{-1} \mathbf{y}$. Следовательно, можно обозначить $d\mathbf{X} = d\mathbf{W} \cdot \mathbf{W}^{-1}$, и компоненты dx_{ij} являются линейными комбинациями базиса $\{dw_{ij}\}$.

В базисе $d\mathbf{X}$ м-градиент представлен теперь выражением $d\rho = -\text{tr}(d\mathbf{X}) + \mathbf{f}^T(\mathbf{y}) d\mathbf{X} \cdot \mathbf{y}$, которое приводит к алгоритму обучения

$$\Delta \mathbf{X}(k) = -\eta(k) \frac{d\rho}{d\mathbf{X}} = \eta(k) [\mathbf{I} - \mathbf{f}(\mathbf{y}(k)) \mathbf{y}^T(k)]. \quad (10)$$

С учетом $\Delta \mathbf{X}(k) = \Delta \mathbf{W}(k) \cdot \mathbf{W}^{-1}(k)$ (10) преобразуется в алгоритм обучения для матрицы $\mathbf{W}$:

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \eta(k) [\mathbf{I} - \mathbf{f}[\mathbf{y}(k)] \mathbf{y}^T(k)] \mathbf{W}(k). \quad (11)$$

Применение м-градиента в соотношении (8) приводит к алгоритму обучения с усреднением (символ $\langle \rangle$):

$$\Delta \mathbf{W}(k) = -\eta(k) \frac{\partial K_{pq}}{\partial \mathbf{W}} \mathbf{W}^T \mathbf{W} = \eta(k) [\mathbf{I} - \langle \mathbf{f}(\mathbf{y}) \mathbf{y}^T \rangle] \mathbf{W},$$

а представление его в РВ (опуская операцию усреднения) показывает эквивалентность этого алгоритма с полученным ранее выражением (9).

ОБОБЩЕНИЕ АЛГОРИТМА, ОСНОВАННОГО НА М-ГРАДИЕНТЕ

Чтобы уменьшить элемент неопределенности масштаба (свойственной задаче РСС), в алгоритме

Правило обучения, представленное выражением (9), введено в [8], а м-градиент использован несколько ранее в [9].

обучения, основанном на м-градиенте, используются ограничения на величину восстановленных сигналов смеси: $E\{f_i(y_i)\}=1$. Однако если первичные сигналы (ПС) нестационарные и их средние значения изменяются достаточно быстро, эти ограничения требуют столь же быстрого изменения значения "разделяющей" матрицы $\mathbf{W}$. Избежать этой трудности позволяет использование особой формы ограничений: направление изменения матрицы $\mathbf{W}$ должно быть ортогонально классу эквивалентности матриц, образованному за счет масштабирования. Такой алгоритм, определенный соотношением (12), адаптируется к быстрым изменениям величины ПС:

$$\Delta \mathbf{W}(k) = \eta(k) [\Lambda(k) - \mathbf{f}[\mathbf{y}(k)] \mathbf{y}^T(k)] \mathbf{W}(k), \quad (12)$$

где $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ — положительно определенная матрица и $\lambda_i = f(y_i) y_i$.

В другой модификации алгоритм обучения имеет вид

$$\Delta \mathbf{W}(k) = \eta(k) [\Lambda(k) - \mathbf{y}(k) \mathbf{g}^T(\mathbf{y}(k))] \mathbf{W}(k), \quad (13)$$

где использована видоизмененная Атиком и Редлихом (Atick, Redlich) [10] форма м-алгоритма. В этом случае нелинейность $g_i(y_i)$ является обратной относительно функции $f_i(y_i)$: $g_i = f_i^{-1}$.

Еще более общая форма алгоритма обучения определяется соотношением (14) [11]:

$$\Delta \mathbf{W}(k) = \eta(k) \cdot [\Lambda(k) - \alpha \cdot \langle \mathbf{f}[\mathbf{y}(k)] \mathbf{y}^T(k) \rangle + \beta \cdot \langle \mathbf{y}(k) \mathbf{g}^T[\mathbf{y}(k)] \mathbf{y}^T(k) \rangle] \mathbf{W}(k), \quad (14)$$

где Λ — диагональная матрица, которая предназначена для нейтрализации диагональных элементов всей матрицы в правой части соотношения (12), а α и β — адаптивно определяемые параметры [7].

М-градиент при ограничениях на ортогональность

Можно считать, что над вектором-сигналом $\mathbf{x}(k) = \mathbf{Q} \mathbf{H} \mathbf{s}(k) = \mathbf{A} \mathbf{s}(k)$ уже выполнена предварительная обработка отбеливания и сигнал нормализован [1], так что

$$\begin{aligned} \mathbf{R}_{xx} &= E\{\mathbf{x}(k) \mathbf{x}^T(k)\} = \mathbf{I}_m; \\ \mathbf{R}_{ss} &= E\{\mathbf{s}(k) \mathbf{s}^T(k)\} = \mathbf{I}_n. \end{aligned} \quad (15)$$

При равном числе ПС и сенсоров ($n = m$) из (15) следует, что $\mathbf{A} \mathbf{A}^T = \mathbf{I}_n$ и СМ $\mathbf{A} = \mathbf{Q} \mathbf{H} \in \mathbb{R}^{n \times n}$ является ортогональной. В этом случае для восстановления ПС $\mathbf{y} = \mathbf{W} \mathbf{x}$ можно рассматривать только ортогональные матрицы $\mathbf{W}$ (такие, что $\mathbf{W}^{-1} = \mathbf{W}^T$).

В рассмотренном выше соотношении $d\mathbf{X} = d\mathbf{W} \cdot \mathbf{W}^{-1} = d\mathbf{W} \cdot \mathbf{W}^T$ базисные функции $d\mathbf{W}$ являются косо-симметричными (т. е. $d\mathbf{X} = -d\mathbf{X}^T$). Это видно из цепочки равенств

$$\mathbf{0} = d\mathbf{I} = d(\mathbf{W} \mathbf{W}^T) = d\mathbf{W} \cdot \mathbf{W}^T + \mathbf{W} \cdot d\mathbf{W}^T = d\mathbf{X} + d\mathbf{X}^T.$$

Отсюда следует, что м-градиент $\partial \rho / \partial \mathbf{X} = (\partial \rho / \partial \mathbf{W}) \cdot \mathbf{W}^T$ также косо-симметричный, и поэтому

$$(\partial \rho / \partial \mathbf{X}) \cdot \mathbf{W}^T = \mathbf{f}(\mathbf{y}) \mathbf{y}^T - \mathbf{y} \mathbf{f}^T(\mathbf{y}).$$

Алгоритм обучения для разделяющей матрицы $\mathbf{W} \in \mathbb{R}^{n \times n}$, построенный с использованием м-градиента, в случае отбеленных данных определяется выражением

$$\begin{aligned} \mathbf{W}(k+1) &= \\ &= \mathbf{W}(k) - \eta(k) \cdot \tilde{\nabla} \rho(\mathbf{y}, \mathbf{W}) = \mathbf{W}(k) - \eta(k) \cdot \frac{\partial \rho}{\partial \mathbf{X}} \mathbf{W}. \end{aligned} \quad (16)$$

Матрица $\mathbf{W}$ является (приблизительно)⁶ ортогональной на каждом шаге итерации. В этом случае $\mathbf{x} = \mathbf{W}^T \mathbf{y}$, и алгоритм (14) приобретает более простую форму:

$$\begin{aligned} \mathbf{W}(k+1) &= \mathbf{W}(k) - \eta(k) \times \\ &\times [\mathbf{f}(\mathbf{y}(k)) \cdot \mathbf{y}^T(k) - \mathbf{y}(k) \cdot \mathbf{f}^T(\mathbf{y}(k))] \cdot \mathbf{W}(k). \end{aligned} \quad (17)$$

Практически благодаря косо-симметричности члена $[\mathbf{f}(\mathbf{y}(k)) \cdot \mathbf{y}^T(k) - \mathbf{y}(k) \cdot \mathbf{f}^T(\mathbf{y}(k))]$ в (17) выполнение обработки смеси сигналов, связанной с отбеливанием (декорреляцией), может производиться одновременно с РСС⁷, а сам алгоритм принимает вид

$$\begin{aligned} \Delta \mathbf{W}(k) &= \eta(k) \times \\ &\times [\mathbf{I} - \mathbf{y}(k) \mathbf{y}^T(k) - \mathbf{f}(\mathbf{y}(k)) \cdot \mathbf{y}^T(k) + \mathbf{y}(k) \cdot \mathbf{f}^T(\mathbf{y}(k))] \times \\ &\times \mathbf{W}(k). \end{aligned} \quad (18)$$

Алгоритм (18) основан на вероятностном (стохастическом) приближении и соответствует выполнению в реальном времени. Вариант этого алгоритма, предусматривающий процедуру усреднения по времени, имеет вид

⁶ Эта матрица точно ортогональная для алгоритма с непрерывным временем. Для дискретного времени реализуется почти точная ортогональность в том случае, если параметр скорости обучения η будет иметь достаточно малое значение.

⁷ В связи с этим алгоритм (17) условно назван "легким" [9].

$$\Delta \mathbf{W}(k) = \eta(k) \times \left[\mathbf{I} - \langle \mathbf{y}(k) \mathbf{y}^T(k) \rangle - \langle \mathbf{f}(\mathbf{y}(k)) \cdot \mathbf{y}^T(k) \rangle + \langle \mathbf{y}(k) \cdot \mathbf{f}^T(\mathbf{y}(k)) \rangle \right] \cdot \mathbf{W}(k). \quad (19)$$

ВЫДЕЛЕНИЕ ГРУППЫ ПЕРВИЧНЫХ СИГНАЛОВ ИЗ ИХ СМЕСЕЙ, РЕГИСТРИРУЕМЫХ СЕНСОРАМИ ИИС

Алгоритмы, использующие метод м-градиента вместо "простого" последовательного РСС, позволяют осуществлять выделение произвольной группы ПС. Положим, что число сенсоров m больше (часто неизвестного) количества n первичных сигналов, сигналы сенсоров отбелены и требуется выделить группу определенного количества e первичных сигналов. Тогда сигналы сенсоров удовлетворяют условию $\mathbf{R}_{\bar{\mathbf{x}}\bar{\mathbf{x}}} = E\{\bar{\mathbf{x}}\bar{\mathbf{x}}^T\} = \mathbf{I}_n$ (где $\bar{\mathbf{x}} = \mathbf{x}_1 = \mathbf{Q}\mathbf{x}$); общая смешивающая $(n \times n)$ -матрица $\mathbf{A} = \mathbf{Q}\mathbf{H}$ является ортогональной (т. е. $\mathbf{A}\mathbf{A}^T = \mathbf{A}^T\mathbf{A} = \mathbf{I}_n$), и для РСС при $e = n$ разделяющая $(n \times n)$ -матрица имеет вид $\mathbf{W}_* = \mathbf{A}^{-1} = \mathbf{A}^T$.

Для выделения e первичных сигналов используется подходящая форма ККЛ:

$$K_{pq}(\mathbf{W}_e) = K \left[p_y(\mathbf{y}, \mathbf{W}_e) \Big|_{q(\mathbf{y})} \right] = \int p_y(\mathbf{y}, \mathbf{W}_e) \log \frac{p(\mathbf{y}, \mathbf{W}_e)}{q(\mathbf{y})} d\mathbf{y}, \quad (20)$$

где $q(\mathbf{y}) = \prod_{i=1}^e q_i(y_i)$ представляет ПВ группы соответствующих независимых сигналов на выходе структуры РСС. Следовательно, с учетом (8) минимизируемая "функция стоимости" имеет вид $\rho(\mathbf{y}, \mathbf{W}_e) = -\sum_{i=1}^e \log q_i(y_i)$ при ограничении $\mathbf{W}_e \mathbf{W}_e^T = \mathbf{I}_e$ (или эквивалентно: $\mathbf{w}_i \mathbf{w}_j^T = \delta_{ij}$), где $\mathbf{W}_e \in \mathbb{R}^{e \times n}$ — разделяющая матрица с $e \leq n$, $\mathbf{w}_i$ — i -я вектор-строка матрицы $\mathbf{W}_e$.⁸ Для того чтобы эти ограничения выполнялись в процессе обучения, форма м-градиента в алгоритме выделения группы ПС должна использоваться в следующем виде:

$$\Delta \mathbf{W}_e(k) = \mathbf{W}_e(k+1) - \mathbf{W}_e(k) =$$

$$= -\eta(k) \cdot \left(\frac{\partial \rho(\mathbf{y}, \mathbf{W}_e)}{\partial \mathbf{W}_e} - \mathbf{W}_e \left(\frac{\partial \rho(\mathbf{y}, \mathbf{W}_e)}{\partial \mathbf{W}_e} \right)^T \mathbf{W}_e \right). \quad (21)$$

При этом если в начале вычислений выполнено ограничение $\mathbf{W}_e(0) \cdot \mathbf{W}_e^T(0) = \mathbf{I}_e$, то оно сохраняется на каждом шаге итерации, т. е. [11]

$$\mathbf{W}_e(k) \cdot \mathbf{W}_e^T(k) = \mathbf{I}_e.$$

Соотношение (21) соответствует правилу обучения в виде (22) с начальным значением $\mathbf{W}_e(0)$, удовлетворяющим условию $\mathbf{W}_e(0) \cdot \mathbf{W}_e^T(0) = \mathbf{I}_e$.⁹

$$\mathbf{W}_e(k+1) = \mathbf{W}_e(k) - \eta(k) \times \{ \mathbf{f}[\mathbf{y}(k)] \cdot \bar{\mathbf{x}}^T(k) - \mathbf{y}(k) \cdot \mathbf{f}^T[\mathbf{y}(k)] \cdot \mathbf{W}_e(k) \}. \quad (22)$$

МОДЕЛЬ ОБОБЩЕННОГО ГАУССОВА РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ ДЛЯ ПС И РАЗДЕЛЕНИЕ СМЕСИ СИГНАЛОВ ПО СХЕМЕ АНАЛИЗА НЕЗАВИСИМЫХ КОМПОНЕНТ (АНК)

Оптимальная нелинейная функция $\mathbf{f}(\mathbf{y})$, входящая в обучающий алгоритм (9) и определенная ранее выражением $\mathbf{f}(\mathbf{y}) = [f_1(y_1), f_2(y_2), \dots, f_n(y_n)]^T$, где $f_i(y_i) = -g'_i(y_i)/g_i(y_i)$, требует знания распределения ПВ первичных сигналов, которое обычно бывает неизвестным. Каждая модель ПВ соответствует некоторому виду функции $\mathbf{f}(\mathbf{y})$, которая в терминах нейронной сети, реализующей алгоритм обучения, называется активационной функцией. Для некоторых типов моделей ПВ выражение для активационной функции приведено в таблице. Например, для косинус-гиперболического распределения (имеющего "надгауссову" форму ПВ)¹⁰ активационная функция имеет вид $f_i(y_i) = \text{tg } h(\gamma_i y_i)$, где γ_i обычно имеет значение $1/\sigma_{y_i}^2$.

⁹ При $e = n$ (т. е. при "полном" РСС) разделяющая матрица $\mathbf{W}_* = \mathbf{W}$ является ортогональной, обучающее правило упрощается и сводится к алгоритму (17), предложенному Кардосо и Лахельдом (Cardoso, Laheld) [9]: $\mathbf{W}(k+1) = \mathbf{W}(k) - \eta(k) \times$

$$\times \{ \mathbf{f}[\mathbf{y}(k)] \cdot \mathbf{y}^T(k) - \mathbf{y}(k) \cdot \mathbf{f}^T[\mathbf{y}(k)] \} \cdot \mathbf{W}(k).$$

¹⁰ Термин "надгауссова" ПВ используется для распределения с эксцессом, большим чем у гауссовой ПВ, поскольку вершина такого распределения выше гауссова пика. В таком же смысле далее говорится о "подгауссовой" ПВ, если величина ее эксцесса меньше, чем у гауссовой ПВ, и, следовательно, вершина этого распределения ниже, чем у гауссовой ПВ.

⁸ Вид ограничения связан с тем, что смешивающая матрица $\mathbf{A} = \mathbf{Q}\mathbf{H}$ является квадратной ортогональной, а разделяющая ("размешивающая") матрица $\mathbf{W}_e$ после выделения e первичных сигналов должна удовлетворять соотношению $\mathbf{W}_e \mathbf{A} = [\mathbf{I}_e, \mathbf{0}_{n-e}]$.

Типовые распределения ПВ $q(y)$ и соответствующие нормализованные активационные функции $f(y) = -\text{dlog} q(y)/dy$

Название ПВ	ПВ-функция $q(y)$	Активационная функция $f(y)$
Гаусса	$(1/\sqrt{2\pi}\sigma) \cdot \exp(- y ^2/2\sigma^2)$	y/σ^2
Лапласа	$(1/2\sigma) \cdot \exp(- y /\sigma)$	$\text{sign}(y)/\sigma$
Коши	$(1/\pi\sigma) \cdot \frac{1}{1+(y/\sigma)^2}$	$\frac{2y}{\sigma^2 + y^2}$
Треугольное	$\frac{1}{\sigma(1- y /\sigma)}$	$\frac{\text{sign}(y)}{\sigma(1- y /\sigma)}$
Косинус-гиперболическое	$1/(\pi \cosh(y/\sigma^2))$	$\text{tgh}(y/\sigma^2)$
Экспоненциально-степенное (или обобщенное Гаусса)	$\frac{r}{2\sigma \Gamma(1/r)} \cdot \exp\left(-\frac{1}{r} \left \frac{y}{\sigma}\right ^r\right), \quad r \geq 0$	$\frac{ y ^{r-1}}{\sigma^r} \cdot \text{sign}(y)$

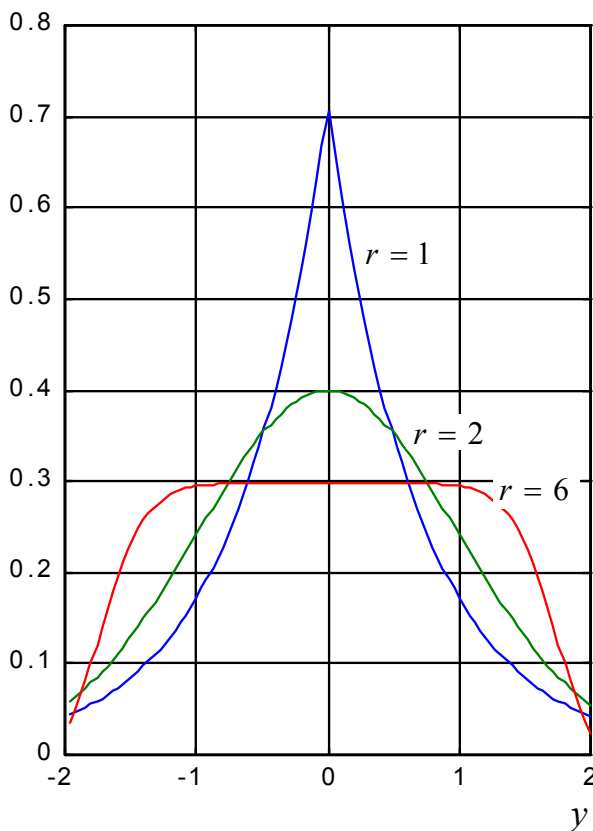


Рис. 2. Экспоненциально-степенная ПВ при различных значениях параметра r

Экспоненциально-степенная ПВ (при различных соотношениях величины ее параметров σ и r) представляет целое семейство распределений, отличающихся формой и включающих распределения Лапласа (при $r = 1$), Гаусса ($r = 2$) и равномерное (при $r > 5 \div 6$) [12], [13]. Таким образом, семейство экспоненциально-степенных ПВ (ЭС_ПВ) включает над- и подгауссовы формы распределения (рис. 2).

Обучающий алгоритм (9), включающий нелинейные функции $f_i(y_i) = -g'_i(y_i)/g_i(y_i)$, при использовании ЭС_ПВ для ПС адаптируется к форме распределения и определяется его параметрами. Для ЭС_ПВ в виде

$$q_i(y_i) = \frac{r_i}{2\sigma_i \Gamma(1/r_i)} \exp\left(-\frac{1}{r_i} \left|\frac{y_i}{\sigma_i}\right|^{r_i}\right), \quad (23)$$

где r_i — переменный параметр, $\Gamma(x) = \int_0^\infty y^{x-1} \exp(-y) dy$ — гамма-функция, $\sigma_i^r = E\{|y_i|^r\}$ — обобщенная мера дисперсии ЭС_ПВ, параметр r_i принимает значение в диапазоне $(0 \div \infty)$.

Оптимальные нормированные нелинейные активационные функции выражаются соотношением

$$f_i(y_i) = -\text{dlog}(q_i(y_i))/dy_i = |y_i|^{r_i-1} \text{sign}(y_i),$$

и т. к. $\text{sign}(y_i) = y_i / |y_i|$, то $f_i(y_i) = y_i / |y_i|^{2-r_i}$.¹¹

Для оценки значений $f_i(y_i) = y_i / |y_i|^{2-r_i} = y_i(k) / \hat{\sigma}_i^{(2-r_i)}(k)$ может использоваться также скользящее среднее, т. е. в качестве $\hat{\sigma}_i^{(2-r_i)} = \hat{m}_{2-r_i}$ берется оценка

$$\hat{\sigma}_i^{(2-r_i)}(k+1) = (1-\eta_0)\hat{\sigma}_i^{(2-r_i)}(k) + \eta_0 |y_i(k)|^{(2-r_i)}.$$

Параметр r_i может быть фиксированным (если доступно знание статистики ПС) или получен в процессе обучения нейронной сети, реализующей РСС. Правило подстройки параметра r_i , основанное на градиенте, имеет вид:

$$\begin{aligned} \Delta r_i(k) &= -\eta_i(k) \partial \rho / \partial r_i = \eta_i(k) \cdot \partial \log(q_i(y_i)) / \partial r_i \cong \\ &\cong \eta_i(k) \cdot \frac{0.1 r_i(k) + |y_i(k)|^{r_i(k)} \left(1 - \log(|y_i(k)|^{r_i(k)})\right)}{r_i^2(k)}. \end{aligned}$$

В случае предельно обостренной ПВ (т. е. для ПС в виде так называемых спайк-сигналов) и $r_i \cong 0$ в качестве оптимальной получается функция $f_i(y_i) = y_i / \langle |y_i|^2 \rangle = \frac{y_i(k)}{\hat{\sigma}_i^2(k)}$, предложенная ранее Матсуокой (Matsuoka) [14]. Аналогично для ПС с плотностью распределения Лапласа (т. е. для $r_i = 1$) $f_i(y_i) = y_i / \langle |y_i| \rangle = y_i(k) / \hat{\sigma}_i(k)$, а для больших значений r_i ($r_i = 5 \div 10$) и (почти) равномерной ПВ — $f_i(y_i) = y_i(k) \langle |y_i(k)|^8 \rangle = y_i(k) \cdot \hat{\sigma}_i^8(k)$.

В общем случае для смеси ПС с величиной эксцесса как выше, так и ниже гауссова и при наличии шума в виде спайк-сигналов использование подхода, основанного на м-градиенте, приводит к алгоритму обучения НС в форме

$$\Delta \mathbf{W}(k) = \eta \cdot \mathbf{F}(\mathbf{y}) \mathbf{W}(k), \quad (24)$$

где элементы матрицы $\mathbf{F}(\mathbf{y})$ определены соотношениями

$$f_{ij} = \begin{cases} -f_i(y_i) y_j & \text{для } \gamma_{ij} < 1, \\ -y_i f_j(y_j) & \text{для других случаев,} \end{cases}$$

где

$$\gamma_{ij} = \sigma_i^2 \sigma_j^2 E \{f'_i(y_i)\} E \{f'_j(y_j)\} \text{ и}$$

¹¹ В особом случае импульсивных (или "спайк-") сигналов параметр r может принимать значение между нулем и единицей. В этом случае целесообразно принимать несколько модифицированную активационную функцию в виде $f_i(y_i) = y_i / [|y_i|^{2-r_i} + \varepsilon]$, $0 < r_i < 1$, где ε — малое положительное число ($\sim 10^{-3}$).

$$f_i(y_i) = y_i / (|y_i|^{2-r_i} + \varepsilon)$$

при значении r_i между нулем и единицей.

В качестве другого подхода после некоторых упрощений можно использовать правило обучения (24) с диагональной матрицей $\Lambda(k) = \text{diag}\{\mathbf{F}(\mathbf{y}(k)) \cdot \mathbf{g}^T(\mathbf{y}(k))\}$. Активационные функции этой матрицы адаптивно меняются со временем, устойчивы к спайк-шуму и имеют вид:

$$\begin{aligned} f_i(y_i) &= \begin{cases} y_i / \hat{\sigma}_i^{2-r_i} & \text{для } k_4(y_i) > \delta_0, \\ y_i & \text{для других случаев;} \end{cases} \\ g_i(y_i) &= \begin{cases} y_i & \text{для } k_4(y_i) > -\delta_0, \\ y_i / \hat{\sigma}_i^{2-r_i} & \text{для других случаев,} \end{cases} \end{aligned} \quad (25)$$

где $k_4(y_i) = E\{y_i^4\} / E^2\{y_i^2\} - 3$ — это нормализованное (относительно гауссова) значение эксцесса распределения плотности вероятности ПС, δ_0 — малое пороговое значение.

Алгоритм обучения (24) и (25) наблюдает и оценивает статистики каждого выходного сигнала, зависящие от знака или величины нормализованного эксцесса (который является мерой отличия ПВ сигнала от гауссовой ПВ). Затем алгоритм автоматически отбирает (переключает) подходящие нелинейные активационные функции так, чтобы успешно разделять все негауссовы ПС.

Подобные методы применимы к другим ПВ. Так, для обобщенного распределения Коши, определяемого тремя параметрами $r_i > 0$, $v_i > 0$ и σ_i^2 , плотность вероятности имеет вид:

$$q_i(y_i) = \frac{B(r_i, v_i)}{\left\{1 + (1/v_i) \left(\frac{y_i}{A(r_i)}\right)^{r_i}\right\}^{v_i+1/r_i}}, \quad (26)$$

где

$$A(r_i) = [\sigma_i^2 \Gamma(1/r_i) / \Gamma(3/r_i)]^{1/2};$$

$$B(r_i, v_i) = r_i v_i^{-1/r_i} \Gamma(r_i + 1/r_i) / 2 A(r_i) \Gamma(v_i) \Gamma(1/r_i).$$

При этом активационная функция выражается соотношением

$$f_i(y_i) = \frac{(r_i v_i + 1)}{(v_i |A(r_i)|^{r_i} + |y_i|^{r_i})} |y_i|^{r_i-1} \cdot \text{sign}(y_i). \quad (27)$$

Статистические моменты и эксцесс ЭС_ПВ

Более полное представление об ЭС_ПВ дают статистические моменты этого распределения, особенно второй и четвертый, которые определяют оценку эксцесса. Общее выражение n -го мо-

мента ЭС_ПВ — $m_n = \int_{-\infty}^{\infty} y^n p(y; r) dy$, и в силу симметрии ПВ при нечетном n $m_n = 0$ (в том числе среднее $m_1 = 0$). Вычисление четных моментов, которые полностью характеризуют ЭС_ПВ, основано на формуле [12], [15]:

$$\int_0^{\infty} y^{n-1} e^{-\mu y^n} dy = \frac{1}{a} \mu^{-1/n} \Gamma\left(\frac{n}{a}\right).$$

Второй момент ЭС_ПВ определяется выражением

$$m_2 = \int_{-\infty}^{\infty} y^2 p(y, r) dy = 2 \int_0^{\infty} y^2 \frac{r}{2\sigma \Gamma(1/r)} e^{-\left|\frac{y}{\sigma}\right|^r} dy =$$

$$= \frac{r}{\sigma \Gamma(1/r)} \int_0^{\infty} y^2 e^{-\left(\frac{y}{\sigma}\right)^r} dy,$$

которое подстановкой $y = z \cdot \sigma$ приводится к вы-

$$\text{ражению } m_2 = \frac{r\sigma^2}{\Gamma(1/r)} \int_0^{\infty} z^2 e^{-z^r} dz = \sigma^2 \frac{\Gamma(3/r)}{\Gamma(1/r)}.$$

Подобным путем определяется момент четвертого порядка $m_4 = \sigma^4 \frac{\Gamma(5/r)}{\Gamma(1/r)}$ и момент $2k$ -й степе-

$$\text{ни } m_{2k} = \sigma^{2k} \frac{\Gamma((2k+1)/r)}{\Gamma(1/r)}.$$

Безразмерная (в отличие от моментов) величина эксцесса является показателем уплощенности или обостренности ПВ. "Нормализованный" эксцесс $k(y)$ показывает его превышение над гауссовым и определяется выражением: $k(y) = m_4 / (m_2)^2 - 3$.

Метод РСС на основе анализа независимых компонент (АНК) при критерии эксцесса ЭС_ПВ ПС

Модели ЭС_ПВ (с учетом данных таблицы и комментария к формуле (22)) соответствует нелинейная функция алгоритма (18), имеющая вид $f_i(y_i) = -d \log q_i(y_i) / d y_i = |y_i|^{r_i-1} \text{sign}(y_i)$. При $r_i = 4$ $f_i(y_i)$ становится кубической функцией, которая хорошо соответствует ПВ с положительным нормированным эксцессом (т. е. с надгауссовой ПВ).

Для выбора подходящей величины r_i оценивается эксцесс k -го выходного сигнала y_i , после чего r_i определяется с помощью графической зависимости $k(r)$. При этом, хотя величина эксцесса, оцененная по выходу восстановленного сигнала, не вполне соответствует величине $k(r)$ для ПС, процедура РСС не нарушается [11], поэтому в анализируемом методе достаточно использовать всего

несколько видов нелинейных функций (активационных функций системы РСС, реализуемой на основе нейронной сети). Кроме того, при низких значениях $k(r)$ (для подгауссовой ПВ) его величина очень слабо зависит от r , она практически постоянна для r в пределах от 4 до 12. Поэтому если оценка для $k(r)$ отрицательна, то можно использовать $r_i = 4$.

При положительном значении $k(r)$ зависимость от параметра r достаточно значительна, так что можно предполагать несколько возможных значений этого параметра. Однако опыт применения метода показывает, что для успешности РСС бывает достаточно использовать 2–3 значения параметра r для всего практического диапазона положительных значений эксцесса (т. е. у сигналов с надгауссовой ПВ).

Из числа применяемых вариантов нелинейных функций для алгоритма РСС по методу АНК [15] целесообразно проанализировать устойчивость алгоритма (18) для различных случаев: 1) $r_i = 4$ при $k_i < 0$; 2) $r_i = 1$; 3) $r_i = 0.8$ для $k_i > 0$ [15].

1) $r_i = 4$ — **подгауссовы распределения**. Выбор $r_i = 4$ предполагает у ПВ ПС наличие отрицательного эксцесса ($k_i < 0$) и соответствует кубической нелинейной (активационной) функции, т. е. $f_i(y_i) = |y_i|^2 y_i$.

2) $r_i = 1$ — **распределение Лапласа**. При выборе $r_i = 1$ ЭС_ПВ переходит в распределение Лапласа, т. е. $q_i(y_i) = (1/2\sigma_i) \exp(-|y_i|/\sigma_i)$, которое соответствует "ступенчатой" нелинейной функции $f_i(y_i) = \text{sign}(y_i) = y_i/|y_i|$ (активационная функция с жестким ограничением).

3) $r_i < 1$ — **распределение с отчетливо выраженным пиком**. При резко выраженном пике ПВ (большом значении эксцесса, $k \gg 1$) выбирается $r_i < 1$. Это соответствует невозрастающей нелинейной функции с сингулярностью в окрестности нуля. Для практических приложений $f_i(y_i)$ принимается постоянной (и ограниченной) величиной в пределах интервала $[-\varepsilon, \varepsilon]$, который определяется малым числом ε .

Дисперсия y_i для ЭС_ПВ определяется соотношением $E\{y_i^2\} = \sigma_i^2 \cdot \Gamma(3/r_i) / \Gamma(1/r_i)$, и при ограничении нормализацией $E\{y_i^2\} = 1$ дисперсия имеет величину $\sigma_i = [\Gamma(1/r_i) / \Gamma(3/r_i)]^{1/2}$. Условием устойчивости алгоритмов обучения (11) и (18), предложенным Амари, Ченом и Сичоки [7] и Кардосо, Лахэльдом [9], является проверка условия $\chi_i + \chi_j > 0$, где $\chi_i = E\{f_i'(y_i)\} - E\{f_i(y_i)y_i\}$. Для каждой отдельной компоненты y_i восстановленного ПС достаточным условием устойчивости является $\chi_i > 0$. Вне области $y_i \in [-\varepsilon, \varepsilon]$ вычис-

ление $E\{f'_i(y_i)\}$ и $E\{f_i(y_i)y_i\}$ проводится по соотношениям:¹²

$$\begin{aligned} E\{f'_i(y_i)\} &= \\ &= \int_{-\infty}^{+\infty} (r_i - 2) |y_i|^{r_i-2} \cdot \frac{r_i}{2\sigma_i \Gamma(1/r_i)} \exp(-(|y_i|^{r_i}/\sigma_i^{r_i})) \cdot dy_i = \\ &= \frac{(r_i - 2)\sigma_i^{r_i-2}}{\Gamma(1/r_i)} \Gamma[(r_i - 1)/r_i]; \\ E\{f_i(y_i)y_i\} &= \\ &= \int_{-\infty}^{+\infty} y_i |y_i|^{r_i-1} \text{sign}(y_i) \cdot \frac{r_i}{2\sigma_i \Gamma(1/r_i)} \exp(-(|y_i|^{r_i}/\sigma_i^{r_i})) \cdot dy_i = \\ &= \frac{r_i \sigma_i^{r_i+1}}{\Gamma(1/r_i)} \Gamma[(r_i + 1)/r_i]. \end{aligned}$$

Модель ПВ Пирсона

Рассмотренные выше плотности вероятности, Коши и ЭС_ПВ, принадлежат к семейству симметричных распределений, но алгоритмы, основанные на м-градиенте, могут функционировать достаточно плохо, если ПВ сигналов имеет значительную асимметрию или при негауссовой форме у них величина эксцесса близка к нулю.

Широкий класс ПВ (симметричных и несимметричных) может моделироваться системой распределений Пирсона, которая описывается дифференциальным уравнением

$$q'(y) = \frac{(y-a)q(y)}{b_0 + b_1 y + b_2 y^2},$$

где a, b_0, b_1, b_2 — параметры, зависящие от ПВ оцененных ПС. Оптимальная активационная функция для системы ПВ Пирсона имеет вид $f(y) = -q'(y)/q(y) = (a-y)/(b_0 + b_1 y + b_2 y^2)$. К разновидностям форм модели Пирсона относятся широко используемые ПВ, в том числе бета, гамма, нормальное (Гаусса), Стюдента и некоторые другие специальные виды распределений. Параметры a, b_0, b_1, b_2 оценивают непосредственно методом моментов:

$$\begin{aligned} a = b_1 &= -\frac{m_3(m_4 + 3m_2^2)}{d}, \quad b_0 = -\frac{m_2(4m_2 m_4 - 3m_3^2)}{d}, \\ b_2 &= -\frac{(2m_2 m_4 - 3m_3^2 - 6m_2^3)}{d}, \end{aligned}$$

¹² Поскольку гамма-функция $\Gamma(x)$ имеет много сингулярных точек (обращаясь в ∞), особенно при $x < 0$, это важно учитывать при выборе $r_i < 1$.

где $d = 10m_2 m_4 - 18m_2^3 - 12m_3^2$; m_2, m_3, m_4 — второй, третий и четвертый статистические моменты у. Преимущество в использовании системы ПВ Пирсона в том, что она позволяет эффективно реализовать РСС при несимметричных распределениях.

АЛГОРИТМЫ НА ОСНОВЕ М-ГРАДИЕНТА ДЛЯ РСС НЕСТАЦИОНАРНЫХ ПЕРВИЧНЫХ СИГНАЛОВ

Во многих приложениях (РСС для нескольких речевых или биомедицинских сигналов) ПС, подлежащие восстановлению, являются нестационарными и имеют изменяющуюся со временем дисперсию.

Основным предположением при РСС является статистическая независимость ПС. При этом если ПС взаимно независимы, но каждый из них имеет негауссову ПВ независимо и одинаково распределенных временных отсчетов, то для РСС требуется использование статистик высокого порядка [16]. Тогда процедура разделения смеси связана с методом анализа независимых компонент (АНК), в котором многомерные данные разлагаются в линейную сумму неортогональных базисных векторов, где коэффициенты статистически независимы. Напротив, когда первичные сигналы нестационарные, то для РСС достаточно статистик второго порядка.

Основные исходные положения модели

В качестве основы моделирования алгоритмов обучения РСС использованы 3 допущения, которые являются незначительными ограничениями и не уменьшают общность метода реализации разделения ПС с применением нейросетевых алгоритмов.

1) Смешивающая матрица **Н** имеет полный ранг по столбцам.

2) ПС $\{s_i(k)\}$ являются статистически независимыми и имеют нулевое среднее.

3) При условии, что $i, j = 1, \dots, n$ и $i \neq j$, отношение $\hat{\sigma}_{s_i}^2(k) / \hat{\sigma}_{s_j}^2(k)$ при изменении времени (аргумента k) не является постоянной величиной.¹³

¹³ Оценка дисперсии сигнала $\hat{\sigma}_{s_i}^2(k)$ осуществляется методом скользящего среднего:

$$\hat{\sigma}_{s_i}^2(k) = (1 - \eta_0) \hat{\sigma}_{s_i}^2(k-1) + \eta_0 y_i^2(k).$$

Первые два допущения обычны для большинства подходов к РСС, третье — существенно для анализируемых ниже алгоритмов, т. к. именно оно позволяет осуществлять РСС с использованием только статистик второго порядка (СВТ).

Критерий на основе СВТ для формирования алгоритмов РСС

Критерий для РСС по методу АНК (при приведенных выше допущениях) основан на взаимной информации. Формирование этой характеристики требует знания ПВ восстановленных сигналов, но поскольку эти ПВ в начале процедуры РСС неизвестны, то алгоритмы, связанные с методом АНК, основываются на гипотетических распределениях. При этом явно или неявно в схему получения алгоритмов РСС приходится включать и статистики более высокого порядка.

Достаточно эффективным является критерий Матсуоки [14], который позволяет не использовать кросс-корреляционную матрицу $E\{y_i(k)y_j(k)\}$ (вычисления которой достаточно трудоемки):

$$J(\mathbf{W}, k) = (1/2) \sum_{i=1}^n \log(\hat{\sigma}_{y_i}^2(k)) - \log \det \hat{\mathbf{R}}_{yy}^{(k)}, \quad (28)$$

где $\hat{\sigma}_{y_i}^2(k) = \langle y_i^2(k) \rangle$ — оценка переменной во времени дисперсии в реальном времени (РВ); $\det(\cdot)$ — обозначает определитель матрицы; $\hat{\mathbf{R}}_{yy}^{(k)} = \langle \mathbf{y}(k)\mathbf{y}^T(k) \rangle$ — автокорреляционная матрица сигналов $\mathbf{y}(k)$, которая оценивается тоже методом скользящего среднего в РВ:

$$\hat{\mathbf{R}}_{yy}^{(k)} = (1 - \eta_0) \hat{\mathbf{R}}_{yy}^{(k-1)} + \eta_0 \mathbf{y}(k)\mathbf{y}^T(k).$$

Критерий (28) представляет неотрицательную функцию, достигающую минимума, только если выполняется условие $E\{y_i(k)y_j(k)\} = 0$ ($i, j = 1, \dots, n$ и $i \neq j$), т. е. если корреляционная матрица сигналов является диагональной.¹⁴

¹⁴ Это следует из так называемого неравенства Адамара $\det(\mathbf{R}) \leq \prod_{i=1}^n r_{ii}$, где $\mathbf{R} = [r_{ij}]$ — неотрицательно определенная ($n \times n$)-матрица, и равенство достигается только при условии, что для $i \neq j$ $r_{ij} = 0$. Поэтому определитель принимает максимальное (а $(-\det)$ — минимальное) значение как раз при условии, что корреляционная матрица сигналов $\hat{\mathbf{R}}_{yy}^{(k)}$ — диагональная, т. е. $E\{y_i(k)y_j(k)\} = 0$ ($i, j = 1, \dots, n$ и $i \neq j$).

Алгоритм обучения с использованием м-градиента

Алгоритмы обучения ориентированы на два типа структуры нейронной сети, обеспечивающей РСС (рис. 3): 1) нейронная сеть (НС) с прямым распространением сигналов, 2) полносвязная рекуррентная нейронная сеть.

1) Выходной сигнал $\mathbf{y}(k)$ нейронной сети с прямым распространением (рис. 3, а) определяется выражением $\mathbf{y}(k) = \mathbf{W} \mathbf{x}(k)$. Поэтому полный дифференциал $dJ(\mathbf{W})$ критерия оптимизации параметров НС (матрицы $\mathbf{W}$ ее синаптических весов) определяется соотношением:

$$\begin{aligned} dJ(\mathbf{W}) &= J(\mathbf{W} + d\mathbf{W}) - J(\mathbf{W}) = \\ &= \frac{1}{2} d \left\{ \sum_{i=1}^n \log \langle y_i^2(k) \rangle \right\} - \\ &\quad - \frac{1}{2} d \left\{ \log \det \langle \mathbf{y}(k)\mathbf{y}^T(k) \rangle \right\}. \end{aligned} \quad (29)$$

Для второго слагаемого в (29) справедлива оценка:

$$\begin{aligned} d \{ \log \det \langle \mathbf{y}(k)\mathbf{y}^T(k) \rangle \} &= d \{ \log \det (\mathbf{W} \mathbf{R}_{xx} \mathbf{W}^T) \} = \\ &= 2 d \{ \log \det (\mathbf{W}^{-1}) \} + d \{ \log \det \hat{\mathbf{R}}_{xx} \} = \\ &= 2 \operatorname{tr} \{ d\mathbf{W} \mathbf{W}^{-1} \} + d \{ \log \det \hat{\mathbf{R}}_{xx} \}. \end{aligned}$$

При этом можно опустить $\hat{\mathbf{R}}_{xx}$, т. к. у него нет зависимости от матрицы весов $\mathbf{W}$. Поэтому, принимая матричный дифференциал $d\mathbf{X}$ (при $m = n$) в виде $d\mathbf{X} = d\mathbf{W} \mathbf{W}^{-1}$, получаем оценку

$$d \{ \log \det \langle \mathbf{y}(k)\mathbf{y}^T(k) \rangle \} = 2 \operatorname{tr} \{ d\mathbf{X} \}.$$

Подобным образом оценивается и первое слагаемое в (29):

$$\begin{aligned} d \left\{ \sum_{i=1}^n \log \langle y_i^2(k) \rangle \right\} &= \sum_{i=1}^n \frac{2 \langle y_i(k) dy_i(k) \rangle}{\langle y_i^2(k) \rangle} = \\ &= 2 E \{ \mathbf{y}^T(k) \Lambda^{-1}(k) d\mathbf{y}(k) \} = 2 E \{ \mathbf{y}^T(k) \Lambda^{-1}(k) d\mathbf{X} \mathbf{y}(k) \}, \end{aligned}$$

где $\Lambda(k)$ — диагональная матрица, у которой элемент i на диагонали $\lambda_i = \hat{\sigma}_{y_i}^2 = \langle y_i^2(k) \rangle$. Использование матричного дифференциала $d\mathbf{X}$ позволяет представить критерий обучения НС в виде соотношения $dJ(\mathbf{W})/d = E \{ \Lambda^{-1}(k) \mathbf{y}(k)\mathbf{y}^T(k) \} - \mathbf{I}$. С учетом того, что $\Delta \mathbf{W}(k) = \Delta \mathbf{X}(k) \mathbf{W}(k)$, применение градиентного метода непосредственно дает алгоритм обучения в РВ: $\Delta \mathbf{W}(k) = -\eta(k) \times \times [dJ(\mathbf{W})/d\mathbf{X}] \mathbf{W}(k) = \eta(k) [\mathbf{I} - \Lambda^{-1}(k) \mathbf{y}(k)\mathbf{y}^T(k)] \mathbf{W}(k)$, который можно также представить в виде

$$\mathbf{W}(k+1) = \mathbf{W}(k) + \tilde{\eta}(k) [\Lambda(k) - \hat{\mathbf{R}}_{yy}^{(k)}] \mathbf{W}(k), \quad (30)$$

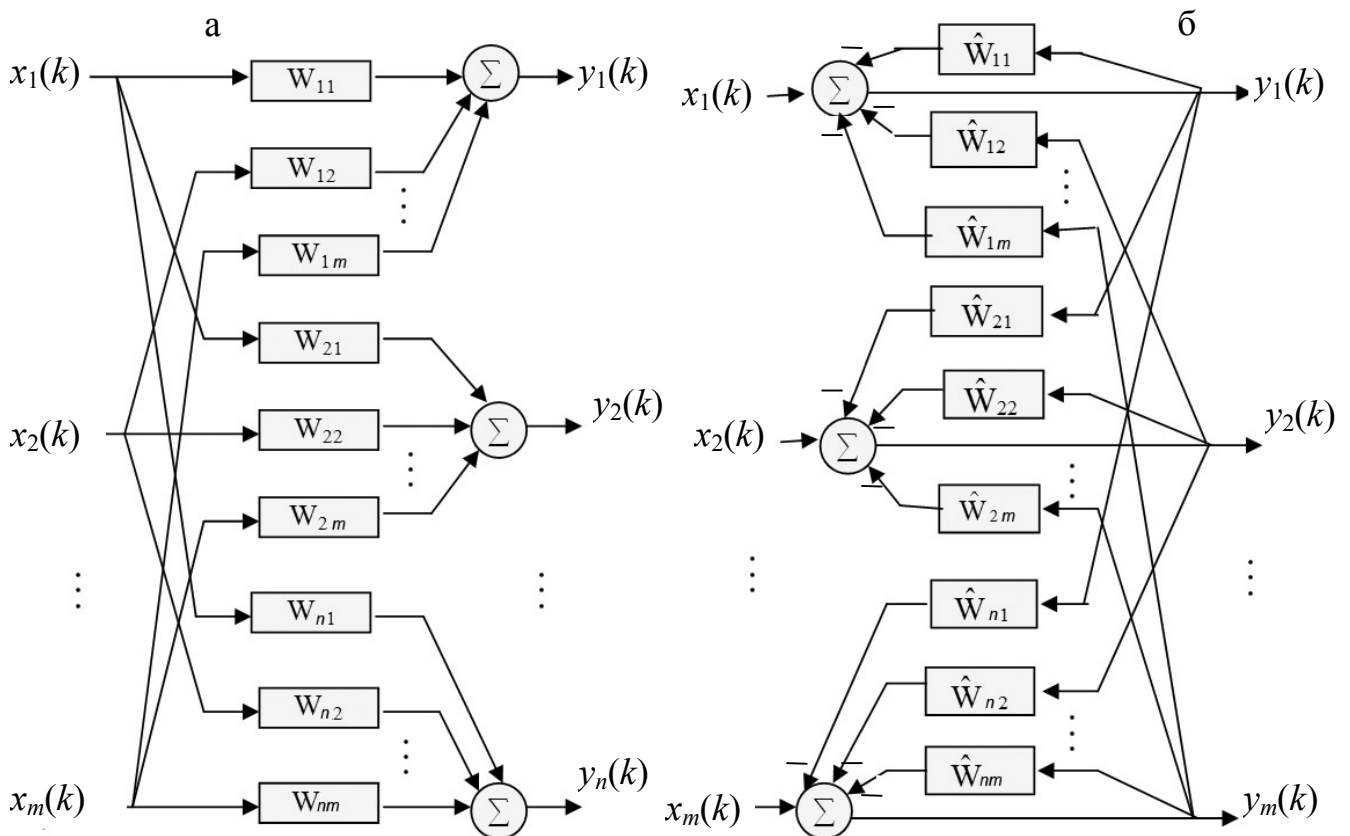


Рис. 3. Две структуры нейронных сетей.

а — с прямым распространением сигнала; б — полносвязная рекуррентная нейронная сеть

где Λ — диагональная матрица с элементом i на диагонали $\lambda_i \approx E\{y_i^2\}$, который можно оценивать в РВ: $\lambda_i(k) = (1 - \eta_0)\lambda_i(k-1) + \eta_0 y_i^2(k)$. Так же оценивается корреляционная матрица выходных сигналов НС: $\hat{\mathbf{R}}_{yy}^{(k)} = (1 - \eta_0)\hat{\mathbf{R}}_{yy}^{(k-1)} + \eta_0 \mathbf{y}(k)\mathbf{y}^T(k)$.

Алгоритм (30) отражает одну из форм метода АНК и приспособлен к реализации РСС при нестационарных ПС [7], он устойчив и практически не зависит от вида распределения ПС [17].

2) Структурой полносвязной рекуррентной нейронной сети (рис. 3, б), предназначенной для реализации РСС, выполняются преобразования, в соответствии с которыми выходной вектор-сигнал $\mathbf{y}(k)$ определяется соотношением $\mathbf{y}(k) = \mathbf{x}(k) - \hat{\mathbf{W}}\mathbf{y}(k) = (\mathbf{I} + \hat{\mathbf{W}})^{-1}\mathbf{x}(k)$ (при естественном предположении, что матрица $(\mathbf{I} + \hat{\mathbf{W}})$ не сингулярная). Путем анализа, аналогичного проведенному для НС с прямым распространением сигнала, получается алгоритм обучения для полносвязной рекуррентной нейронной сети. Матрица $\hat{\mathbf{W}}$ весовых коэффициентов НС, которая одновременно дает пре-

образование, обратное к смешивающей матрице, удовлетворяет соотношению

$$\Delta \hat{\mathbf{W}}(k) = -\tilde{\eta}(k)[\mathbf{I} + \hat{\mathbf{W}}(k)][\Lambda(k) - \hat{\mathbf{R}}_{yy}^{(k)}]. \quad (31)$$

ЗАКЛЮЧЕНИЕ

Рассмотрены методы разделения сигналов смеси и восстановления их формы. Проанализировано влияние формы распределения первичных сигналов, которые поступают на сенсоры ИИС и создают компоненты смешанных сигналов. Разделение сигналов реализуется на основе адаптивного алгоритма для нейронной сети (НС). Вид функции преобразования нейронов этой сети находится в некоторой зависимости от типа плотности распределения первичных сигналов. Для разделения сигналов смеси использованы метод анализа независимых компонент (АНК) и модифицированная форма градиента (м-градиент). Адекватность метода АНК подтверждается проверкой с помощью информационного критерия.

— Адаптивные алгоритмы обучения НС по-

строены с использованием м-градиента, который эффективен в случае оптимизации на множестве матриц.

– Даны критерий взаимной информации и критерий Кульбака—Лейблера, определяющие меру различия распределений.

– Рассмотрены схемы разделения сигналов при различных неизвестных распределениях первичных сигналов (ПС) и количестве ПС.

– Оптимальный выбор нелинейных функций преобразования нейронов увязан с различными видами распределения первичных сигналов, образующими смесь.

– Активационные функции для нейронных структур выведены вполне корректно для распределения ПС в виде ЭС_ПВ.

– Рассмотрены модели адаптивной фильтрации и связанные с ними адаптивные алгоритмы обучения НС, ориентированные как на выполнение в реальном времени, так и с обучением на микровыборках сигналов с предполагаемой для ПС статистикой.

– Процедуры РСС приспособлены к случаям, когда ПС нестационарные или имеют характер коротких импульсов (спайк-сигналов). Алгоритмы применимы также к смесям неизвестного количества ПС, имеющих неопределенный вид ПВ.

Построению алгоритмов с использованием м-градиента посвящены работы Амари и Кардосо. Основные подходы к РСС нестационарных сигналов исследованы Матсуокой, Охиа и Кавамотой.

СПИСОК ЛИТЕРАТУРЫ

1. Меркушева А.В., Малыхина Г.Ф. Методы и алгоритмы разделения смеси сигналов. I. Применение декорреляции и статистик второго порядка // Научное приборостроение. 2009. Т. 19, № 2. С. 90–103.
2. Кульбак С. Теория информации и статистика. М.: Изд. Наука, 1967. 408 с.
3. Bell A.J., Sejnowski T.J. An Information Maximization Approach to Signal Mixture Separation and Deconvolution // Neural Computation. 1995. V. 7, N 6. P. 1129–1139.
4. Comon P. Independent Component Analysis, a New Conception? // Signal Processing. 1994. N 3. April. P. 287–314.
5. Moulines E., Duhamel P., Cardoso J.-F., Mayrargue S. Subspace Method for the Identification of Multichannel FIR Filters // IEEE Transaction on Signal Processing. 1995. V. 43, N 2. P. 516–525.

6. Amari S., Cardoso J.-F. Source Separation – Semi-Parametric Statistical Approach // IEEE Transaction on Signal Processing. 1997. V. 45, N 11. P. 2692–2700.
7. Amari S., Chen T.-P., Cichocki A. Stability Analysis of Adaptive Blind Source Separation // Neural Networks. 1997. V. 10, N 8. P. 1345–1351.
8. Cichocki A., Unbehauen R. Robustneural Networks on-Line Learning for Blind Identification and Separation of Sources // IEEE Transaction on Circuits and Systems I: Fundamental Theory and Applications. 1996. V. 43, N 11. P. 894–906.
9. Cardoso J.-F., Laheld B.H. Equivariant Adaptive Source Separation // IEEE Transaction on Signal Processing. 1996. V. 14, N 12. P. 3017–3030.
10. Atick J.J., Redlich A.N. Convergent Algorithm for Sensory Receptive Field Development // Neural Computation. 1993. V. 5, N 1. P. 45–60.
11. Amari S. Natural Gradient Works Efficiently in Learning // Neural Computation. 1998. V. 10. P. 271–276.
12. Малыхина Г.Ф., Меркушева А.В. Детектирование речевого сигнала и фильтрация с адаптивным порогом // Микропроцессорные средства измерений. Сб. трудов ФТК, № 2. Изд-во СПб ГПУ, 2001. С. 26–35.
13. Меркушева А.В. Фильтрация нестационарного сигнала (речи) в вейвлет-области с адаптацией к виду и динамике шума // Научное приборостроение. 2003. Т. 13, № 2. С. 73–87.
14. Matsuoka K., Ohia M., Kawamoto M. A Neural Net for Separation of non-Stationary Signals // Neural Networks. 1995. V. 8, N 2. P. 411–419.
15. Choi S., Cichocki A., Amari S. Flexible Independent Component Analysis // Journal of VLSI Signal Processing. 2000. T. 26, N 1 / 2. P 25–38.
16. Малыхина Г.Ф., Меркушева А.В. Сеть с симметричной функцией преобразования нейронов на основе статистик высокого порядка (СВП) для снижения шума нестационарного сигнала // Научное приборостроение. 2008. Т. 18, № 4. С. 129–136.
17. Choi S., Cichocki A., Amari S. Equivariant Unstationary Source Separation // Neural Networks. 2002. V. 15, N 1.

Санкт-Петербург

Материал поступил в редакцию 23.06.2009.

SIGNAL MIXTURE DECOMPOSITION METHODS AND ALGORITHMS. II. APPLICATION OF M-GRADIENT TO ANALYSIS OF INDEPENDENT COMPONENTS

V. M. Malykhin, A. V. Merkusheva

Saint-Petersburg

The problem of signal mixture decomposition (with reconstruction of its component forms) is analyzed without any information concerning the proportions and type of mixing. The method is based on using information criteria and adaptive algorithm for learning neural network. Several forms of distributions were considered for original signals reaching the sensors of information-measurement system. Variations of neural transfer functions corresponding to several types of distributions are given. The approach to signal mixture separation includes application of m-gradient in the analysis scheme for independent component analysis.

Keywords: signals mixture, separation methods, adaptive algorithms, neural network, m-gradient, independent component analysis