Модификации метода анализа сингулярного спектра для анализа временных рядов

Погребников Николай Вадимович, гр. 21.Б04-мм

Санкт-Петербургский государственный университет
Прикладная математика и информатика
Вычислительная стохастика и статистические модели

Научный руководитель — доктор физико-математических наук, Н. Э. Голяндина
Рецензент — лектор, университет Кардиффа А. Н. Пепелышев

Санкт-Петербург, 2025

Введение

$\mathsf{X} = (x_1, \ldots, x_N)$ – временной ряд длины N, $x_i \in \mathbb{R}$.
$\mathsf{X} = \mathsf{T} + \mathsf{P} + \mathcal{E}$, где:

- T – тренд, медленно меняющаяся компонента;
- P – сумма периодических компонент;
- $\mathcal{E}$ – шум, случайная составляющая.

Базовый метод SSA:
Анализ сингулярного спектра (Singular Spectrum Analysis, SSA) – метод, целью которого является разложение исходного ряда на сумму интерпретируемых компонент [Golyandina, Nekrutkin и Zhigljavsky 2001. *Analysis of Time Series Structure: SSA and Related Techniques*].

Постановка задачи

Модификации SSA:

- Generalized Singular Spectrum Analysis (GSSA) – модификация SSA с добавлением весов к исходному ряду [Gu и др. 2024];
- Circulant Singular Spectrum Analysis (CiSSA) – модификация SSA с фиксированным базисом [Bogalo, Poncela и Senra 2021];
- Functional Singular Spectrum Analysis (FSSA) – многомерная модификация SSA с одним непрерывным параметром [Haghbin и др. 2021].

Задача: Описание и анализ модификаций в контексте теории SSA, сравнение алгоритмов, реализация их на языке R.

$\mathsf{X} = \mathsf{S} + \mathcal{E} = \mathsf{S}^{(1)} + \mathsf{S}^{(2)} + \mathcal{E}$ — временной ряд длины N.
S – сигнал, $\mathsf{S}^{(1)}$ и $\mathsf{S}^{(2)}$ – компоненты сигнала.
L – длина окна. I_1 и I_2 – множества для группировки.
Алгоритм SSA [Golyandina, Nekrutkin и Zhigljavsky 2001]:

1. **Построение траекторной матрицы:**
   $\mathbf{X} = \mathcal{T}(\mathsf{X}) = \mathcal{T}_{\mathtt{SSA}}(\mathsf{X}) = [X_1 : \ldots : X_K]$, $K = N - L + 1$,
   $X_i = (x_i, \ldots, x_{i+L-1})^\top$, $1 \le i \le K$.
2. **Сингулярное разложение (SVD)** траекторной матрицы:
   $$\mathbf{X} = \sum_{i=1}^{d} \sqrt{\lambda_i} U_i V_i^T = \sum_{i=1}^{d} \mathbf{X}_i,\ d = \operatorname{rank}(\mathbf{X}).$$
3. **Группировка**:
   $$\mathbf{S}^{(1)} = \sum_{i \in I_1} \mathbf{X}_i,\ \mathbf{S}^{(2)} = \sum_{i \in I_2} \mathbf{X}_i,\ \mathbf{E} = \sum_{i \notin I_1 \cup I_2} \mathbf{X}_i.$$
   $\mathbf{X} = \mathbf{S}^{(1)} + \mathbf{S}^{(2)} + \mathbf{E}$.
4. **Восстановление** ряда: $\mathsf{X} = \tilde{\mathsf{S}}^{(1)} + \tilde{\mathsf{S}}^{(2)} + \tilde{\mathcal{E}}$.

Два подхода к выбору I_1 и I_2:

1. **По компонентам SVD:**
   $I_1 = \{1, \ldots, r_1\}, \quad I_2 = \{r_1 + 1, \ldots, r_1 + r_2\}$.
   Восстановление по $I = I_1 \cup I_2$ (первым $r = r_1 + r_2$ компонентам): $\tilde{\mathsf{S}} = \tilde{\mathsf{S}}^{(1)} + \tilde{\mathsf{S}}^{(2)}$ – оценка сигнала.

2. **По частотному критерию:**
   $I_j = \left\{i : P_{\Omega_j}(U_i) > T_0\right\}, \quad j = 1, 2,$
   где:
   - $\Omega_j = \bigcup_{k=1}^{n_j} [\omega_k^-, \omega_k^+]$ – диапазоны частот,
   - $P_{\Omega_j}(U_i)$ – вклад диапазонов частот Ω_j для i-й компоненты,
   - T_0 - пороговое значение.

Многомерные варианты SSA

Нужно переопределить входные данные и $\mathcal{T}$.

`MSSA`. Параметры: $1 \le L \le \max(N_p)$.

- $\mathsf{X} = \{\mathsf{X}^{(p)} = \left(x_j^{(p)}\right)_{j=1}^{N_p}, \ p = 1, \ldots, s\}$,
- $\mathcal{T}_{\mathsf{MSSA}}(\mathsf{X}) = \mathbf{X} = [\mathbf{X}^{(1)} : \ldots : \mathbf{X}^{(s)}]$, $\mathbf{X}^{(p)} = \mathcal{T}_{\mathtt{SSA}}(\mathsf{X}^{(p)})$.

`2D-SSA`. Параметры: (L_x, L_y), где $1 \le L_x \le N_x$, $1 \le L_y \le N_y$.

- $\mathsf{X} = (x_{ij})_{i,j=1}^{N_x, N_y}$, где $N_x \times N_y$ — размер массива.
- Из X выделяются все возможные подматрицы $\mathsf{X}_{k,l}^{(L_x, L_y)}$.
- Каждая подматрица $\mathsf{X}_{k,l}^{(L_x, L_y)}$ преобразуется в столбец: $X_{k+(l-1)K_x} = \mathrm{vec}(\mathsf{X}_{k,l}^{(L_x, L_y)})$.
- $\mathcal{T}_{\mathtt{2D\text{-}SSA}}(\mathsf{X}) = \mathbf{X} = [X_1 : \ldots : X_{K_x K_y}]$.

Вложенный вариант SSA

$\mathsf{X} = \mathsf{S} + \mathcal{E} = \mathsf{S}^{(1)} + \mathsf{S}^{(2)} + \mathcal{E}.$

Определение 1 (Golyandina и Shlemov 2015)

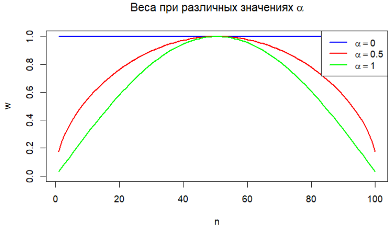
Вложенный вариант SSA — двухэтапный метод:

1. $\tilde{\mathsf{S}}$ – восстановление SSA по $I = \{1, \ldots, r\}$.
2. Применение другого метода к $\tilde{\mathsf{S}}$ для улучшения разделимости: $\tilde{\mathsf{S}} = \tilde{\mathsf{S}}^{(1)} + \tilde{\mathsf{S}}^{(2)}$.

SSA EOSSA [Golyandina, Dudnik и Shlemov 2023] – вложенный вариант SSA, улучшающий разделимость и позволяющий производить группировку по частотам.

Метод GSSA. Алгоритм

$\mathsf{X} = \mathsf{S} + \mathcal{E} = \mathsf{S}^{(1)} + \mathsf{S}^{(2)} + \mathcal{E}$,
параметры L, $\alpha \geq 0$, I_1 и I_2.
Весовая функция:
$\boldsymbol{w}^{(\alpha)} = \left(\left|\sin\left(\frac{\pi n}{L+1}\right)\right|^{\alpha}\right)$,
$n = 1, \ldots, L$.



Шаг 1 алгорима `GSSA` [Gu и др. 2024]:
$\mathbf{X}^{(\alpha)} = \mathcal{T}^{(\alpha)}(\mathsf{X}) = [X_1^{(\alpha)} : \ldots : X_K^{(\alpha)}]$,
$X_i^{(\alpha)} = (w_1 x_{i-1}, \ldots, w_L x_{i+L-2})^{\mathrm{T}}$, $1 \leq i \leq K$.
Шаги 2-4: аналогичны `SSA`.

Замечание 1

При $\alpha = 0$, `GSSA` — в точности базовый алгоритм `SSA`.

Сравнение SSA и GSSA. Линейные фильтры

$\mathsf{X} = (x_1, \ldots, x_N)$. Восстановление $\widetilde{\mathsf{X}} = (\tilde{x}_1, \ldots, \tilde{x}_N)$:
для SSA по компоненте $\sqrt{\lambda} U V^\top$, $U = (u_1, \ldots, u_L)^\top$;
для GSSA по $\sqrt{\lambda^{(\alpha)}} U^{(\alpha)} {V^{(\alpha)}}^\top$, $U^{(\alpha)} = (u_1^{(\alpha)}, \ldots, u_L^{(\alpha)})^\top$.

Запись SSA через линейный фильтр для средних точек:

$$\widetilde{x}_n = \sum_{j=-(L-1)}^{L-1} \left(\sum_{k=1}^{L-|j|} u_k u_{k+|j|}/L \right) x_{n-j}, \quad L \le n \le K.$$

Аналогичное представление для GSSA:

$$\widetilde{x}_n = \sum_{j=-(L-1)}^{L-1} \left(\sum_{k=1}^{L-|j|} u_k^{(\alpha)} u_{k+|j|}^{(\alpha)} w_k \Big/ \sum_{i=1}^{L} w_i \right) x_{n-j}, \quad L \le n \le K.$$

Замечание 1

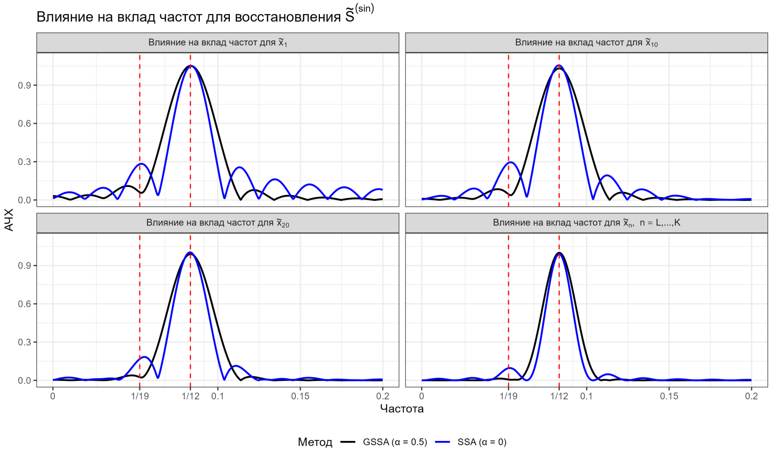
Представление $\tilde{x}_n$ для $n < L$ или $n > K$ зависит от n.

Сравнение SSA и GSSA. Пример

$$\mathsf{X} = \mathsf{S} + \mathcal{E} = \mathsf{S}^{(\sin)} + \mathsf{S}^{(\cos)} + \mathcal{E} =$$

$$\left\{\sin\left(\frac{2\pi}{12}n\right) + \frac{1}{2}\cos\left(\frac{2\pi}{19}n\right) + \varepsilon_n,\ n = 1:191\right\},\ \varepsilon_n \sim \mathrm{N}(0, 0.1^2).$$

Группировка: $I_{\sin} = \{1, 2\}, I_{\cos} = \{3,\ 4\}.\ L = 48,\ K = 144.$



GSSA. Вывод: вложенный вариант SSA + GSSA

Таблица 1: $\mathsf{X} = \mathsf{S}^{(\sin)} + \mathsf{S}^{(\cos)} + \mathcal{E}$, MSE оценок.

Метод/Ошибка	$\mathsf{S}^{(\sin)}$	$\mathsf{S}^{(\cos)}$	S
SSA	5.68e-03	5.44e-03	**7.48e-04**
GSSA, $\alpha = 0.5$	**1.21e-03**	**1.25e-03**	1.04e-03
SSA $(r = 4)$ + GSSA, $\alpha = 0.5$	**1.06e-03**	**1.12e-03**	**7.48e-04**

Применяя SSA для выделения сигнала, затем GSSA для разделения компонент, получается наилучший результат.

Метод CiSSA. Алгоритм

$\mathsf{X} = \mathsf{S} + \mathcal{E} = \mathsf{S}^{(1)} + \mathsf{S}^{(2)} + \mathcal{E}$. L, $\Omega_j = \bigcup_{k=1}^{n_j} [\omega_k^-, \omega_k^+]$, $j = 1, 2$.

Алгоритм `CiSSA` [Bogalo, Poncela и Senra 2021]:

1. **Построение траекторной матрицы:** как в `SSA`.
2. $l = 1 : L$,
   $U_l = L^{-1/2}(u_{l,1}, \ldots, u_{l,L})$, $u_{l,j} = \exp\left(-\mathrm{i}2\pi(j-1)\frac{l-1}{L}\right)$.
   Элементарное разложение: $\omega_k = \frac{k-1}{L}$, $k = 1 : \lfloor \frac{L+1}{2} \rfloor$
   $\mathbf{X}_{\omega_1} = U_1 U_1^{\mathrm{H}} \mathbf{X}, \quad \mathbf{X}_{\omega_k} = U_k U_k^{\mathrm{H}} \mathbf{X} + U_{L+2-k} U_{L+2-k}^{\mathrm{H}} \mathbf{X};$
   $\mathbf{X}_{\omega_{\frac{L}{2}+1}} = U_{\frac{L}{2}+1} U_{\frac{L}{2}+1}^{\mathrm{H}} \mathbf{X}$, если $\quad L \bmod 2 = 0$,
   Разложение: $\mathbf{X} = \sum_{k=1}^{d} \mathbf{X}_{\omega_k}$, $d = \lfloor \frac{L+1}{2} \rfloor$ (или $\frac{L}{2} + 1$).
3. **Группировка**:
   $\mathbf{S}^{(1)} = \sum_{\omega \in \Omega_1} \mathbf{X}_\omega, \ \mathbf{S}^{(2)} = \sum_{\omega \in \Omega_2} \mathbf{X}_\omega, \ \mathbf{E} = \sum_{\omega \notin \Omega_1 \cup \Omega_2} \mathbf{X}_\omega.$
   $\mathbf{X} = \mathbf{S}^{(1)} + \mathbf{S}^{(2)} + \mathbf{E}$.
4. **Восстановление ряда:** как в `SSA`.

Метод CiSSA. Особенности

1. Базис фиксированный, зависит от L, N (SSA: базис адаптивный, зависит от X, L, N).
2. Автоматическая группировка по заданному набору частот.
3. Сложность при подборе L для попадания в сетку частот.
4. Сложность при подборе Ω_j.

Примеры:

1. $\mathsf{X} = \mathsf{S}^{(\sin)} + \mathsf{S}^{(\cos)} + \mathcal{E} =$
   $= \left\{\sin(\frac{2\pi}{12}n) + \frac{1}{2}\cos(\frac{2\pi}{3}n) + \varepsilon_n,\ n = 1:191\right\}.$
2. $\mathsf{X} = \mathsf{S}^{(e\cdot\sin)} + \mathsf{S}^{(e\cdot\cos)} + \mathcal{E} =$
   $= \left\{e^{\frac{n}{200}}\sin(\frac{2\pi}{12}n) + e^{\frac{n}{100}}\cos(\frac{2\pi}{3}n) + \varepsilon_n,\ n = 1:191\right\}.$

Параметры:

- $\varepsilon_n \sim N(0, 0.1^2)$, $\delta = 1/191$.
- Группировки по частотам:
  1. $\Omega_{\sin} = \left[\frac{1}{12} \pm 6\delta\right]$, $\Omega_{\cos} = \left[\frac{1}{3} \pm 6\delta\right]$;
  2. $\Omega_{e\cdot\sin} = \left[\frac{1}{12} \pm 6\delta\right]$, $\Omega_{e\cdot\cos} = \left[\frac{1}{3} \pm 6\delta\right]$.
- `CiSSA`: $L = 12, 48, 96, 97$.
- `SSA EOSSA`: $L = 96, 97$, $r = 4$, $T_0 = 0.5$.

Пример 1: $\mathsf{X} = \mathsf{S}^{(\sin)} + \mathsf{S}^{(\cos)} + \mathcal{E}$.

Метод	**Параметры**	MSE $(\mathsf{S}^{(\sin)})$	MSE $(\mathsf{S}^{(\cos)})$	MSE (S)
`CiSSA`	$L = 12$	1.1e-03	1.1e-03	2.3e-03
`CiSSA`	$L = 48$	**9.9e-04**	1.0e-03	2.0e-03
`CiSSA`	$L = 96$	**8.7e-04**	**9.2e-04**	1.8e-03
`CiSSA`	$L = 97$	3.5e-03	7.6e-03	9.6e-03
`SSA EOSSA`	$L = 96$	**4.2e-04**	**4.4e-04**	**8.6e-04**
`SSA EOSSA`	$L = 97$	**4.3e-04**	**4.7e-04**	**9.1e-04**

- `SSA EOSSA` более точно отделяет сигнал от шума и разделяет его компоненты по сравнению с `CiSSA`.

Пример 2: $\mathsf{X} = \mathsf{S}^{(e \cdot \sin)} + \mathsf{S}^{(e \cdot \cos)} + \mathcal{E}$.

Метод	**Параметры**	$\mathrm{MSE}\left(\mathsf{S}^{(e \cdot \sin)}\right)$	$\mathrm{MSE}\left(\mathsf{S}^{(e \cdot \cos)}\right)$	$\mathrm{MSE}\left(\mathsf{S}\right)$
`CiSSA`	$L = 97$	1.4e-02	2.7e-01	2.5e-01
`CiSSA`	$L = 12$	1.2e-03	2.2e-03	3.3e-03
`CiSSA`	$L = 48$	2.1e-03	2.3e-02	2.3e-02
`CiSSA`	$L = 96$	6.3e-03	8.4e-02	8.4e-02
`SSA EOSSA`	$L = 96$	**4.0e-04**	**4.6e-04**	**8.6e-04**
`SSA EOSSA`	$L = 97$	**3.9e-04**	**3.6e-04**	**7.6e-04**

- При экспоненциальной модуляции точность оценивания методом `CiSSA` значительно ухудшается.

Метод CiSSA. Выводы

- SSA EOSSA с группировкой по частотам обеспечивает лучшую точность в сравнении с CiSSA.
- В отличие от SSA EOSSA, метод CiSSA не требует задания r для выделения сигнала и порога T_0 при автоматической группировке по частотам.

Метод FSSA. Алгоритм

$\mathsf{X} = \{x_n(\tau_k)\}_{n=1,\ldots,N}^{k=1,\ldots,M}$ – 2D объект, $\tau_k \in [0,1]$. L – длина окна.
I_1, I_2 – группировка. $\{\nu_i(\tau) \in \mathbb{H}\}_{i=1}^d$ – ЛНЗ функции.
Алгоритм `FSSA` [Haghbin и др. 2021]:

1. **Аппроксимация с непрерывным параметром:**
$$\mathsf{Y} = \left\{y_n(\tau) = \sum_{i=1}^{d} a_{i,n}\nu_i(\tau)\right\}_{n=1,\ldots,N}^{\tau\in[0,1]},\ a_{i,n} = f(\mathsf{X}, \{\nu_i(\tau)\}_{i=1}^d).$$
2. **Построение траекторной матрицы как функции от τ:**
$\mathbf{Y}(\tau) = [\boldsymbol{y}_1(\tau), \ldots, \boldsymbol{y}_K(\tau)]$, $K = N - L + 1$,
$\boldsymbol{y}_i(\tau) = (y_i(\tau), \ldots, y_{i+L-1}(\tau))^\top, \quad i = 1, \ldots, K.$
3. **Разложение SVD:** $\mathbf{Y}(\tau) = \sum\limits_{i=1}^{Ld} \mathbf{Y}_i(\tau) = \sum\limits_{i=1}^{Ld} \sqrt{\lambda_i} v_i \otimes \boldsymbol{\psi}_i(\tau)$.
4. **Группировка**: $\mathbf{S}^{(1)}(\tau) = \sum\limits_{i\in I_1} \mathbf{Y}_i(\tau)$, $\mathbf{S}^{(2)}(\tau) = \sum\limits_{i\in I_2} \mathbf{Y}_i(\tau)$.
5. **Восстановление**: $\tilde{\mathsf{S}}^{(j)} = \left\{s_n^{(j)}(\tau)\right\}_{n=1,\ldots,N}^{\tau\in[0,1]}$, $j = 1, 2$.
6. **Дискретизация**: $\left\{s_n^{(j)}(\tau_k)\right\}_{n=1,\ldots,N}^{k=1,\ldots,M}$, $j = 1, 2$.

$x_n(\tau_i) = s_n(\tau_i) + \varepsilon_{n,i},\ \varepsilon_{n,i} \sim N(0, 0.1^2),\ i = 1:M,\ n = 1:N.$
$s_n(\tau_i) = e^{\tau_i^2} \cos(2\pi\omega n) + \cos(4\pi\tau_i)\sin(2\pi\omega n),\ \tau_i = \frac{i-1}{M-1}.$

$\{\nu_i(\tau)\}_{i=1}^d$ – первые $d = 15$ функций cubic B-spline базиса.
MSSA и FSSA: $I = \{1, 2\}$. 2D-SSA: $I = \{1, \ldots, 8\}$, $L_y = 50$.

ω	N	$L_x = 20$			$L_x = 40$		
		2D-SSA	MSSA	FSSA	2D-SSA	MSSA	FSSA
0.25	50	0.009	0.029	0.010	0.010	0.021	0.014
	100	0.008	0.027	0.006	0.006	0.020	0.007
	150	0.007	0.026	0.005	0.006	0.020	0.006
	200	0.008	0.026	0.005	0.006	0.019	0.005

Таблица показывает RMSE выделения сигнала. Результаты 2D-SSA и FSSA сравнимы по точности, тогда как MSSA значительно хуже.

Результаты

GSSA введением весов улучшает разделимость в отсутствии шума. Это свойство объяснено с помощью рассмотрения применения SSA как системы линейных адаптивных фильтров. Предложен вложенный вариант, объединяющий преимущества SSA и GSSA.

CiSSA заменяет адаптивный базис SSA фиксированным, упрощая выделение компонент в заданных частотах. Однако наличие шума, проблемы с выбором окна и группировкой снижают эффективность CiSSA. В результате, SSA EOSSA с группировкой по частотам выигрывает по точности.

FSSA было выявлено, что это метод для анализа двумерных данных, поэтому сравнение с MSSA некорректное. Схож по точности с 2D-SSA.

Методы CiSSA и GSSA реализованы на языке R с использованием пакета Rssa. Для метода FSSA реализовано восстановление компонент с использованием пакета Rfssa.

- Bogalo, J., P. Poncela и E. Senra (2021). «Circulant singular spectrum analysis: A new automated procedure for signal extraction». В: *Signal Processing* 179. 107824. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2020.107824.
- Golyandina, N., P. Dudnik и A. Shlemov (2023). «Intelligent identification of trend components in singular spectrum analysis». В: *Algorithms* 16.7. 353. DOI: 10.3390/a16070353.
- Golyandina, N., V. Nekrutkin и A. Zhigljavsky (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. Chapman и Hall/CRC.
- Golyandina, N. и A. Shlemov (2015). «Variations of singular spectrum analysis for separability improvement: non-orthogonal decompositions of time series». В: *Statistics and Its Interface* 8.3, с. 277—294. ISSN: 1938-7997. DOI: 10.4310/sii.2015.v8.n3.a3.

Gu, J. и др. (2024). «Generalized singular spectrum analysis for the decomposition and analysis of non-stationary signals». В: *Journal of the Franklin Institute* 361.6. 106696. ISSN: 0016-0032. DOI: https://doi.org/10.1016/j.jfranklin.2024.106696.

Haghbin, H. и др. (2021). «Functional singular spectrum analysis». В: *Stat* 10.1. e330. DOI: https://doi.org/10.1002/sta4.330.

Определения, использующиеся для FSSA:

- Функциональным временным рядом будем называть $\boldsymbol{x} \in \mathbb{H}^k$: $\boldsymbol{x}(s) = \big(x_1(s), \ldots, x_N(s)\big)^\top$, $x_i \in \mathbb{H} = \mathcal{L}^2([0,1])$, $s \in [0,1]$.
- Скалярное произведение: $\langle \boldsymbol{x}, \boldsymbol{y} \rangle_{\mathbb{H}^k} = \sum_{i=1}^k \langle x_i, y_i \rangle$, $\langle x, y \rangle = \int_0^1 x(s)y(s)ds$.
- Тензорное произведение: для $x \in \mathbb{H}_1$, $y \in \mathbb{H}_2$, оператор $x \otimes y : \mathbb{H}_1 \to \mathbb{H}_2$, $(x \otimes y)h = \langle x, h \rangle y$, $h \in \mathbb{H}_1$.