

Санкт-Петербургский государственный университет  
Математико-механический факультет  
Кафедра Статистического Моделирования

«Научно-исследовательская работа» (семестр 6)

ИССЛЕДОВАНИЕ УСЛОВИЙ ДЛЯ ПОДДЕРЖИВАЮЩИХ ВРЕМЕННЫХ  
РЯДОВ В MSSA

Выполнил:

Ткаченко Егор Андреевич  
группа 19.Б04-мм

Научный руководитель:

к. ф.-м. н., доцент  
Голяндина Нина Эдуардовна

# Оглавление

<b>Введение</b>	3
1. Определения	3
2. Применение SSA и MSSA	4
3. ЛРФ	5
<b>Глава 1. Постановка задачи</b>	6
<b>Глава 2. Численные эксперименты</b>	7
2.1. Первый эксперимент: выбор компонент для MSSA	7
2.2. Второй эксперимент: сравнение ошибки прогноза методами SSA и MSSA при разных величинах шума первого ряда и отклонениях второго ряда	11
2.3. Третий эксперимент: линейные сигналы	13
<b>Заключение</b>	19
<b>Список литературы</b>	20

# Введение

Полезность умения строить прогнозы не нуждается в доказательстве. Прогноз временных рядов может использоваться в прогнозе погоды, приливов, спроса на товары и многом другом.

С помощью книги [1] был изучен базовый SSA, разложение рядов, заполнение пропусков в данных, прогноз и базовый MSSA. Для работы с временными рядами и их прогнозом использовался пакет Rssa. Проведены эксперименты с простейшими моделями сигналов для изучения связи между согласованностью сигналов и поддерживающими рядами. Исследовано, при каких отклонениях, сигналы с одинаковой структурой перестают быть согласованными.

## 1. Определения

Вещественным временным рядом длины  $N$  называется вектор

$$\mathbf{F} = (f_1, \dots, f_N), \quad f_j \in \mathbb{R}.$$

Многомерным временным рядом  $\vec{\mathbf{F}}$  называется набор  $s$  временных рядов  $\mathbf{F}^{(p)}$  длин  $N_p$ :

$$\vec{\mathbf{F}} = \{\mathbf{F}^{(p)} = (f_1^{(p)}, \dots, f_{N_p}^{(p)}), \quad p = 1, \dots, s\}.$$

$L$ -траекторная матрица (или просто траекторная матрица) ряда  $\mathbf{F}$  имеет структуру ганкелевой матрицы, а ее столбцами являются отрезки длины  $L$  ряда  $\mathbf{F}$

$$\mathcal{T}_{\text{SSA}}(\mathbf{F}) = \begin{pmatrix} f_1 & f_2 & \dots & f_K \\ f_2 & f_3 & \dots & f_{K+1} \\ \vdots & \vdots & \ddots & \vdots \\ f_L & f_{L+1} & \dots & f_N \end{pmatrix}.$$

$L$ -траекторная матрица многомерного ряда  $\vec{\mathbf{F}}$  состоит из горизонтально склеенных траекторных матриц рядов  $\mathbf{F}^{(p)} \in \vec{\mathbf{F}}$ :

$$\mathcal{T}_{\text{MSSA}}(\vec{\mathbf{F}}) = [\mathcal{T}_{\text{SSA}}(\mathbf{F}^{(1)}) : \dots : \mathcal{T}_{\text{SSA}}(\mathbf{F}^{(s)})].$$

Из траекторной матрицы можно восстановить ряд. Из любой матрицы подходящего размера можно получить траекторную матрицу проектированием на пространство ганкелевых матриц (или склеенных горизонтально ганкелевых матриц для многомерного случая).

$L$ -Ранг ряда равен рангу его  $L$ -траекторной матрицы:

$$\text{rank}_L \mathbf{F} = \text{rank } \mathcal{T}_{\text{SSA}}(\mathbf{F}), \quad \text{rank}_L \vec{\mathbf{F}} = \text{rank } \mathcal{T}_{\text{MSSA}}(\vec{\mathbf{F}}).$$

Ряд называется рядом конечного ранга  $r$ , если его  $L$ -ранг равен  $r$  для любой длины окна  $L$  и любой достаточно большой длины  $N$  [1, 2.1.2.1, стр. 34].

## 2. Применение SSA и MSSA

Алгоритмы SSA и MSSA могут быть применены для аппроксимации временного ряда рядом конечного ранга.

**Алгоритм.**

**Вход:** Ряд  $\mathbf{F}$  для SSA или многомерный ряд  $\vec{\mathbf{F}}$  для MSSA, длина окна  $L \leq N$  для SSA или  $L \leq N_p, \forall N_p$  для MSSA, ранг аппроксимирующего ряда  $r$ .

1 Вложение. Временной ряд переводится в  $L$ -траекторную матрицу  $\mathbf{X}$

$$\mathbf{X} = \mathcal{T}_{\text{SSA}}(\mathbf{F}) \text{ для SSA}, \quad \mathbf{X} = \mathcal{T}_{\text{MSSA}}(\vec{\mathbf{F}}) \text{ для MSSA}.$$

2 Сингулярное разложение. Методом SVD матрица  $\mathbf{X}$  раскладывается на сумму  $d$  матриц  $\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T$ , где  $d = \text{rank } \mathbf{X} = \text{rank } \mathbf{X}\mathbf{X}^T \leq L$ ,  $\lambda_i$  — собственные числа матрицы  $\mathbf{X}\mathbf{X}^T$  ( $\lambda_1 \geq \dots \geq \lambda_L \geq 0$ ),  $U_i$  — собственные вектора матрицы  $\mathbf{X}\mathbf{X}^T$ ,  $V_i = \mathbf{X}^T U_i / \sqrt{\lambda_i}$  — факторные вектора матрицы  $\mathbf{X}$ .

3 Группировка. Множество индексов  $\{1, \dots, d\}$  делится на  $t$  непересекающихся множеств  $I_1, \dots, I_m$ . Далее для каждого множества индексов (пусть  $I = i_1, \dots, i_t$ ) получается матрица  $\mathbf{X}_I = \mathbf{X}_{i_1} + \dots + \mathbf{X}_{i_t}$ .

Для аппроксимации рядом конечного ранга  $r$ , понадобится множество из первых  $r$  индексов  $\{1, \dots, r\}$ , соответствующую ему матрицу обозначу  $\hat{\mathbf{X}}_r = \mathbf{X}_1 + \dots + \mathbf{X}_r$ .

4 Сгруппированные матрицы  $\mathbf{X}_{I_j}$  диагональный усреднением восстанавливаются в ряды (SSA) или многомерные ряды (MSSA). Для получения аппроксимирующего ряда нужно восстановить его из матрицы  $\hat{\mathbf{X}}_r$ .

**Выход:** Аппроксимирующий ряд  $\hat{\mathbf{F}}_r$  конечного ранга  $r$ .

### 3. ЛРФ

Линейная рекуррентная формула (ЛРФ) выражает каждый член последовательности через линейную комбинацию предыдущих членов.

Ряд  $F$  длины  $N$  — управляемый ЛРФ, если существуют такие  $a_1, \dots, a_d$ , что:

$$f_{i+d} = \sum_{k=1}^d a_k f_{i+d-k}, \quad 1 \leq i \leq N-d, \quad a_d \neq 0, \quad d < N-1.$$

Важно отметить, что бесконечный ряд конечного ранга является управляемым ЛРФ [1, 2.1.2.2, стр. 35].

Вещественный временной ряд  $F$ , управляемый ЛРФ, естественным образом прогнозируется на одну точку:

$$\tilde{f}_N = \sum_{k=1}^{L-1} a_k f_{N-k}.$$

Но тогда его можно прогнозировать и на любое количество точек.

## Глава 1

### Постановка задачи

Пусть имеется временной ряд  $F^{(1)} = S^{(1)} + R^{(1)}$ , где сигнал  $S^{(1)}$  — ряд управляемый ЛРФ, шум  $R^{(1)}$  — ряд без структуры. Рассмотрим задачу прогнозирования  $S^{(1)}$ . Эта задача уже решается методом SSA, но как можно улучшить прогноз?

Пусть помимо ряда  $F^{(1)}$  имеется временной ряд  $F^{(2)} = S^{(2)} + R^{(2)}$ . Если структура сигналов  $S^{(1)}$  и  $S^{(2)}$  похожа, то использование ряда  $F^{(2)}$  может улучшить прогноз сигнала  $S^{(1)}$ , потому что второй ряд дает алгоритму больше данных, которые могут улучшить ЛРФ. Возможность такого улучшения прогноза подтверждена [1, 4.3.3.3, стр. 216]. Но второй ряд может сделать прогноз хуже, если структура его сигнала отличается от первого или шум большой.

Простейший пример похожих по структуре сигналов — гармонические колебания с одинаковыми периодом. Они даже могут быть смещены по фазе или иметь разную амплитуду.

Для объективной оценки качества прогноза будем использовать среднюю квадратичную ошибку.

$$\text{MSE}(\tilde{S}, S) = \frac{1}{N_f} \sum_{i=N+1}^{N+N_f} (\tilde{s}_i - s_i)^2,$$

где  $\tilde{S}$  — прогноз сигнала  $S$  на  $N_f$  точек,  $N$  — длина прогнозируемого сигнала  $S$ .

Ряд  $F^{(2)}$  называется поддерживающим для прогноза, если прогноз с его использованием лучше чем без него, т.е.

$$\text{MSE}(\tilde{S}_{\text{MSSA}}, S^{(1)}) < \text{MSE}(\tilde{S}_{\text{SSA}}, S^{(1)}).$$

Такое определение можно применить только в экспериментах с известным продолжением ряда. На практике не с чем сравнить прогноз, поэтому появляется вопрос. Как понять, что ряд поддерживающий не зная продолжения прогнозируемого ряда? Помочь ответить на этот вопрос может понятие согласованности.

Сигналы  $S^{(1)}$ ,  $S^{(2)}$  называются полностью согласованными, если ранг  $r_{1,2} = r_1 = r_2$  и полностью несогласованными, если  $r_{1,2} = r_1 + r_2$ , где  $r_1 = \text{rank } S^{(1)}$ ,  $r_2 = \text{rank } S^{(2)}$ ,  $r_{1,2} = \text{rank } \vec{S} = \text{rank } \{S^{(1)}, S^{(2)}\}$ .

## Глава 2

## Численные эксперименты

## 2.1. Первый эксперимент: выбор компонент для MSSA

По определению, когда сигналы похожи, их можно считать согласованными и лучше использовать (при прогнозе или восстановлении сигнала) ранг равный рангу одного сигнала. Когда сигналы отличаются, их следует считать не согласованными и использовать ранг равный сумме рангов сигналов. Исследуем, будет ли ошибка MSSA меньше при таком выборе ранга для алгоритма MSSA в восстановлении и прогнозе первого ряда. И будут ли при этом вторые ряды поддерживающими.

Выберем в качестве первого ряда простой сигнал, зависящий от параметра (например, у косинуса параметр — период) с аддитивным гауссовым шумом с дисперсией  $\sigma_1^2 = 0.2^2$ . Второй ряд будет простым сигналом того же вида, с несколько отличающимся параметром и без шума. Восстановим и спрогнозируем первый ряд с помощью SSA, MSSA считая ряды согласованными и MSSA считая ряды несогласованными.

Для устойчивости результатов, повторим это 50 раз и усредним ошибки.

Назовем относительной ошибкой прогноза (восстановления) значение

$$\frac{error_{SSA} - error_{MSSA}}{error_{SSA} + error_{MSSA}},$$

где  $error_{SSA}, error_{MSSA}$  — ошибки прогноза (восстановления) методами SSA и MSSA соответственно.

Значения относительной ошибки легко расположить на графике (она принимает значения от -1 до 1). По значению относительной ошибки легче понять, какой метод лучше (не надо сравнивать два значения ошибок, которые просто положительны и могут быть любых порядков). Но относительную ошибку нельзя считать когда  $error_{SSA} = error_{MSSA} = 0$ .

Как интерпретировать значения относительной ошибки?

- значения больше 0 значат, что что MSSA лучше SSA;
- значения меньше 0 значат, что что MSSA хуже SSA;
- значения около 0 значат что ошибки примерно равны;
- значения далеко от 0 значат, что ошибки сильно отличаются.

Все сигналы в этом и следующих экспериментах будут нормироваться, чтобы амплитуда сигнала не влияла на ошибки прогноза и восстановления. Например, для косинуса:  $s_j^{(i)} = A \cos(\frac{2\pi j}{T_i})$ , где  $A$  — такая константа, что  $\text{mean}(|s_j^{(i)}|) = 1$ .

### 2.1.1. Сигнал косинус

Сигнал  $S^{(1)}$  — косинус с периодом  $T_1 = 8$ . Сигналы  $S^{(2)}$  — косинусы с периодами  $T_2 \in \{8, 8.02, 8.04, 8.06, 8.08, 8.1, 8.15\}$ .

Ранг косинуса равен 2, поэтому для MSSA используются первые 2 или первые 4 компоненты разложения, а для SSA только 2.

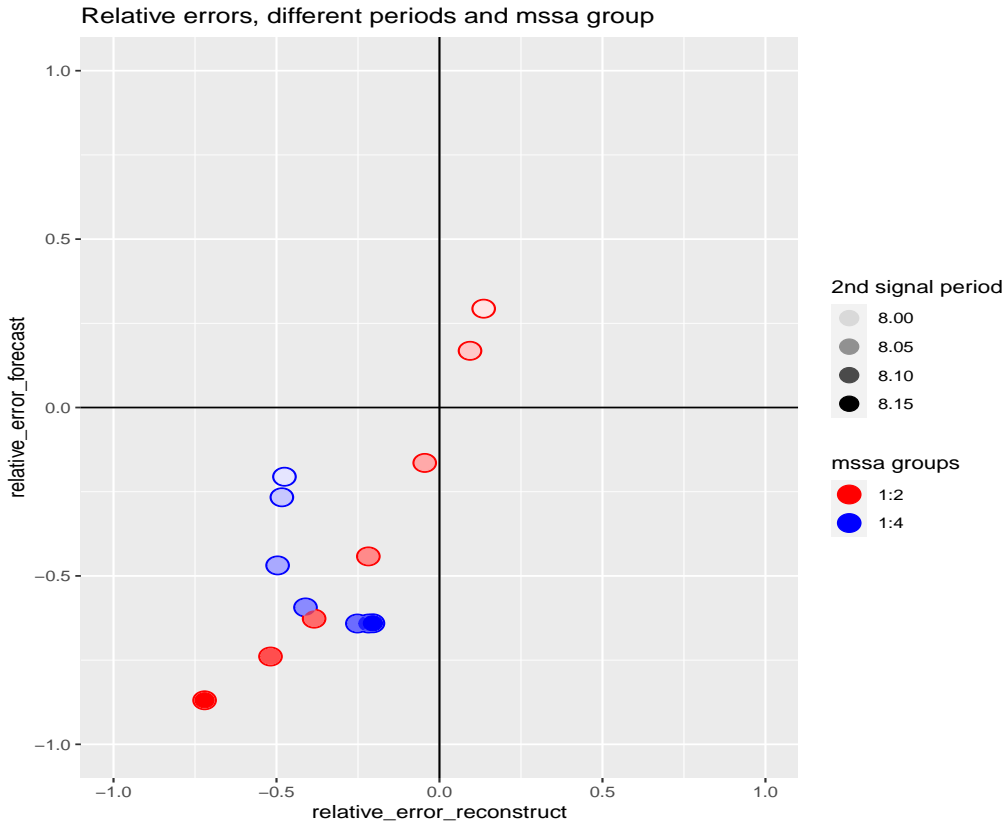


Рис. 2.1. Зависимость относительных ошибок от разницы сигналов и выбора ранга для MSSA.

На рис. 2.1 видим, что с увеличением разницы рядов использование четырех компонент становится лучше и для прогноза и для восстановления сигнала, но при этом второй ряд является поддерживающим только для случаев, когда второй сигнал совпадает с первым или очень близок к нему, а ранг для MSSA 2.

### 2.1.2. Сигнал экспонента (показательная функция)

Функция для сигналов —  $s_j^{(i)} = A \exp(j\lambda_i)$ . Сигнал  $S^{(1)}$  — нормированная показательная функция с  $\lambda_1 = 0.005$ . Сигналы  $S^{(2)}$  — нормированная показательная функция с



$\lambda_2 \in \{0.005, 0.0075, 0.01, 0.0125, 0.015, 0.02, 0.025, 0.03\}$ .

Ранг показательной функции равен 1, поэтому для **MSSA** используется первая или первые 2 компоненты разложения, а для **SSA** только первая.

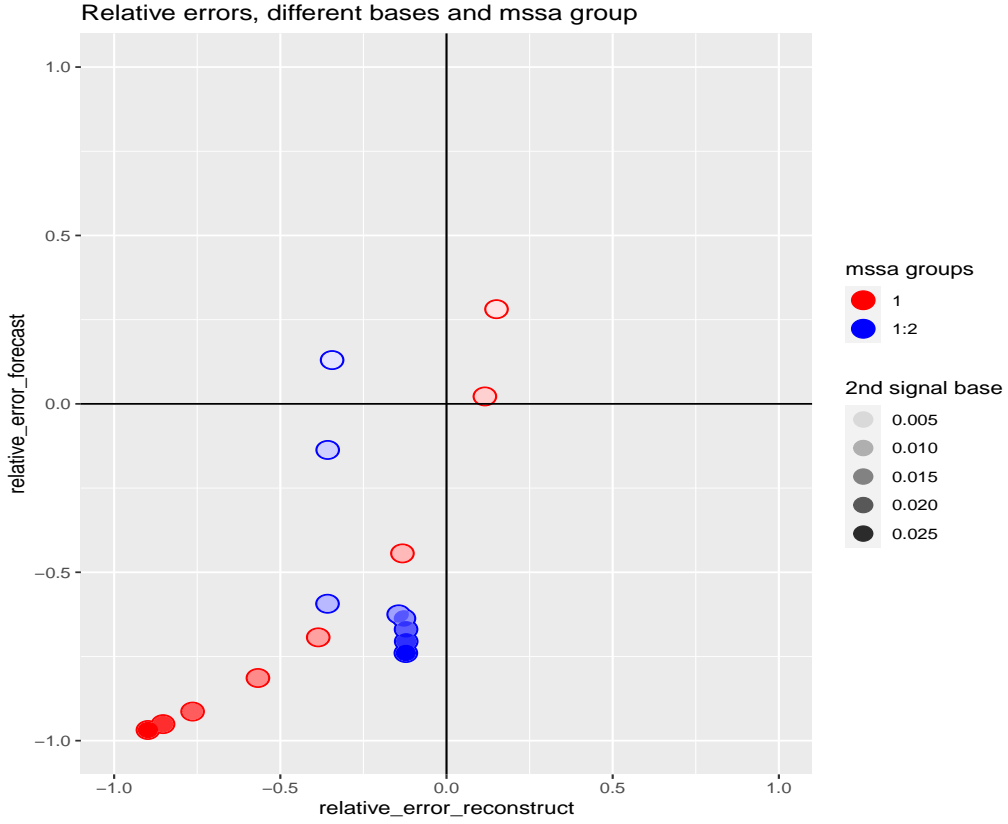


Рис. 2.2. Зависимость относительных ошибок от разницы сигналов и выбора ранга для **MSSA**.

На рис. 2.2 видим похожий результат: с увеличением разницы рядов использование двух компонент становится лучше. Второй ряд поддерживающий только для случаев, когда он равен первому или очень близок к нему, но на этот раз не только когда ранг для алгоритма **MSSA** равен рангу ряда.

### 2.1.3. Сигнал косинус с показательной модуляцией (общий период, меняющаяся модуляция)

Функция для сигналов —  $s_j^{(i)} = A \exp(j\lambda_i) \cos(\frac{2\pi j}{8})$ . Сигнал  $S^{(1)}$  — функция с  $\lambda_1 = 0.005$ . Сигналы  $S^{(2)}$  — функция с  $\lambda_2 \in \{0.005, 0.0075, 0.01, 0.0125, 0.015, 0.02, 0.025, 0.03\}$ .

Ранг косинуса с модуляцией равен 2, поэтому для **MSSA** используются первые 2 или первые 4 компоненты разложения, а для **SSA** только 2.

На рис. 2.3 видим аналогичный результат для косинусов с модуляцией при изменении модулирующей функции.

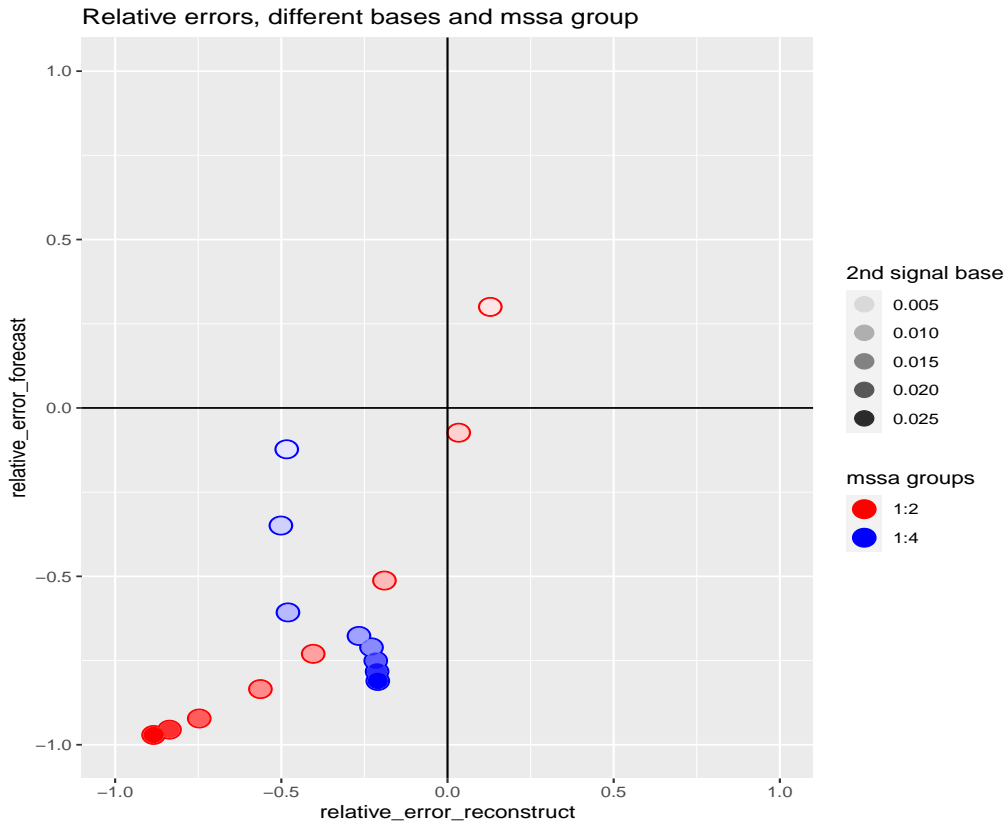


Рис. 2.3. Зависимость относительных ошибок от разницы сигналов и выбора ранга для MSSA.

#### 2.1.4. Сигнал косинус с показательной модуляцией (меняющийся период, общая модуляция)

Функция для сигналов —  $s_j^{(i)} = A \exp(0.02j) \cos(\frac{2\pi j}{T_i})$ . Сигнал  $S^{(1)}$  — функция с  $T_1 = 8$ . Сигналы  $S^{(2)}$  — функция с  $T_2 \in \{8, 8.02, 8.04, 8.06, 8.08, 8.1, 8.15\}$ .

Ранг косинуса с модуляцией равен 2, поэтому для MSSA используются первые 2 или первые 4 компоненты разложения, а для SSA только 2.

На рис. 2.4 видим аналогичный результат для косинусов с модуляцией при изменении периодов.

#### 2.1.5. Результат первого эксперимента

Для всех видов сигналов при отклонении второго сигнала от первого всегда наступал момент, когда использование удвоенного ранга дает меньшие ошибки прогноза и восстановления.

Но при этом, второй ряд редко оказывался поддерживающим, потому что большая часть наблюдений находилась в нижней левой четверти.

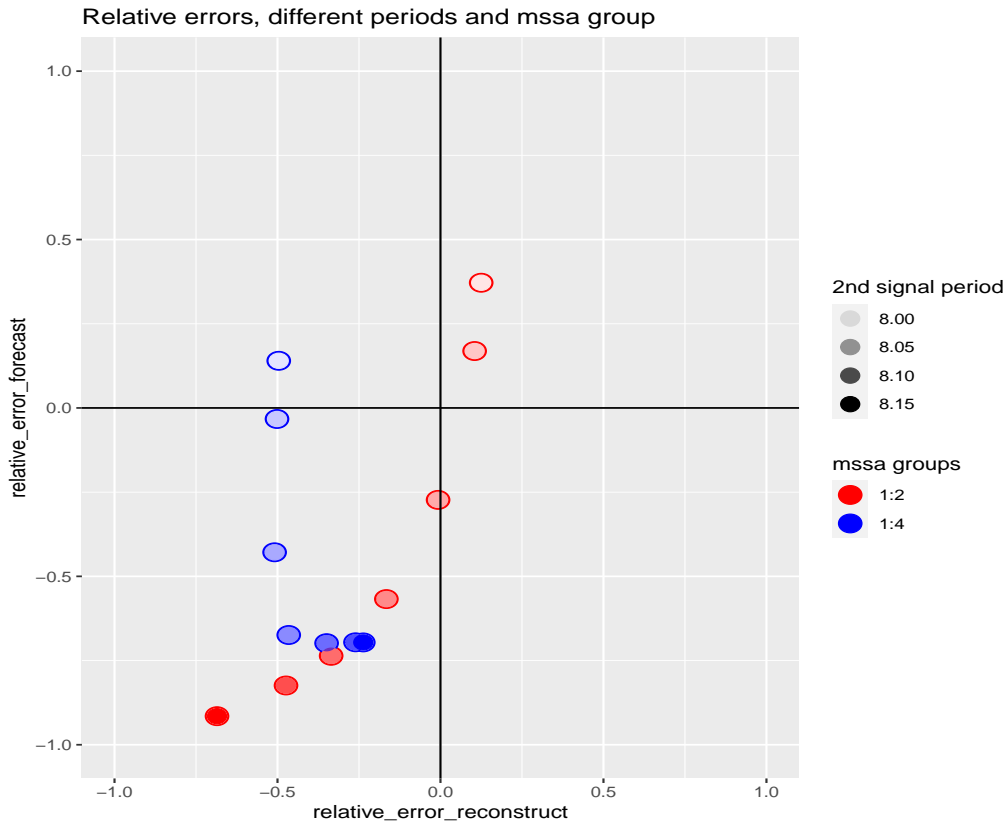


Рис. 2.4. Зависимость относительных ошибок от разницы сигналов и выбора ранга для MSSA.

## 2.2. Второй эксперимент: сравнение ошибки прогноза методами SSA и MSSA при разных величинах шума первого ряда и отклонениях второго ряда

Никита Федоров в своей выпускной квалификационной работе изучал влияние величины второго шума на результаты работы SSA, MSSA, ProjSSA. Рассмотрим влияние величины первого шума на прогноз SSA и MSSA, с не зашумленным вторым рядом.

Гипотеза: при увеличении шума первого ряда, MSSA станет лучше для любого отклонения второго ряда. Если это так, то можно найти зависимость граничного значения  $\sigma_1$  (при котором SSA становится хуже MSSA) от изменения параметра второго сигнала.

Как и в первом эксперименте, выберем в качестве первого ряда простой сигнал, зависящий от параметра с аддитивным гауссовым шумом с несколькими значениями дисперсией  $\sigma_1^2$ . Второй ряд будет простым сигналом того же вида, с несколько отличающимся параметром и без шума. Спрогнозируем первый ряд с помощью SSA и MSSA используя в алгоритме ранг равный рангу сигнала. Для устойчивости результатов, повторим это 50 раз и найдем средние ошибки. Если графики ошибок прогноза SSA и MSSA будут пересекаться, то найдем значения  $\sigma_1$  при которых это происходит, это и будут граничные

значения  $\sigma_1$ .

### 2.2.1. Сигнал косинус

Модель сигнала —  $s_j^{(i)} = A \cos(\frac{2\pi j}{T_i})$ . Параметры для сигналов:  $T_1 = 8$ ,  $T_2 \in \{8, 8.02, 8.04, 8.06, 8.08, 8.1, 8.15\}$ ,  $\sigma_1 \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1\}$ , ранг косинуса равен 2.

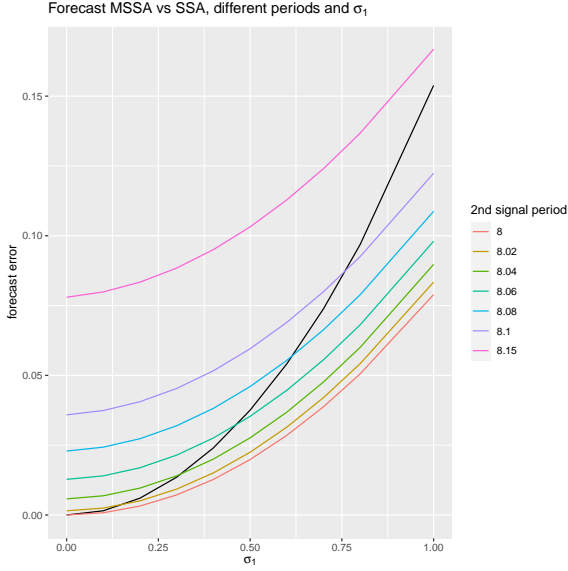


Рис. 2.5. Ошибка прогноза SSA и MSSA для косинусов.

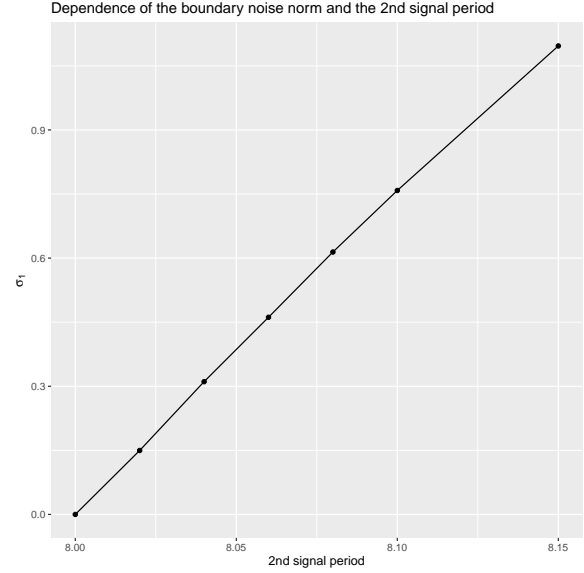


Рис. 2.6. Зависимость граничного значения  $\sigma_1$  от периода второго сигнала

На рис. 2.5 видим, что график ошибки прогноза SSA (черная линия) пересекает все графики ошибок прогноза MSSA кроме одного, но они очевидно пересекутся при большем  $\sigma_1$ . Пересечения графиков будем искать с помощью интерполяции, а для случаев, когда пересечения не было — с помощью экстраполяции.

На рис. 2.6 изображены полученные граничные значения  $\sigma_1$  для каждого второго сигнала. Видна линейная зависимость.

### 2.2.2. Сигнал экспонента (показательная функция)

Модель сигнала —  $s_j^{(i)} = A \exp(j\lambda_i)$ . Параметры для сигналов:  $\lambda_1 = 0.005$ ,  $\lambda_2 \in \{0.005, 0.01, 0.012, 0.013, 0.015, 0.0175, 0.02\}$ ,  $\sigma_1 \in \{0, 0.3, 0.6, 1, 1.5, 2\}$ , ранг косинуса равен 2.

На рис. 2.7 видим, что график ошибки прогноза SSA (черная линия) пересекает все графики ошибок прогноза MSSA.

На рис. 2.8 видна линейная зависимость, как с косинусом.

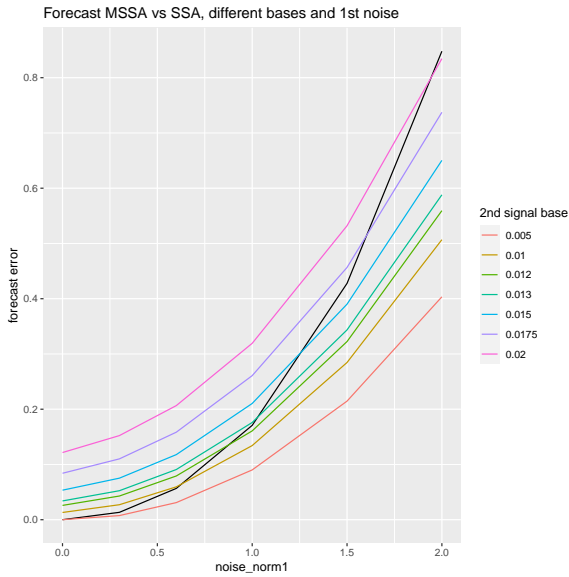


Рис. 2.7. Ошибка прогноза SSA и MSSA для показательных функций.

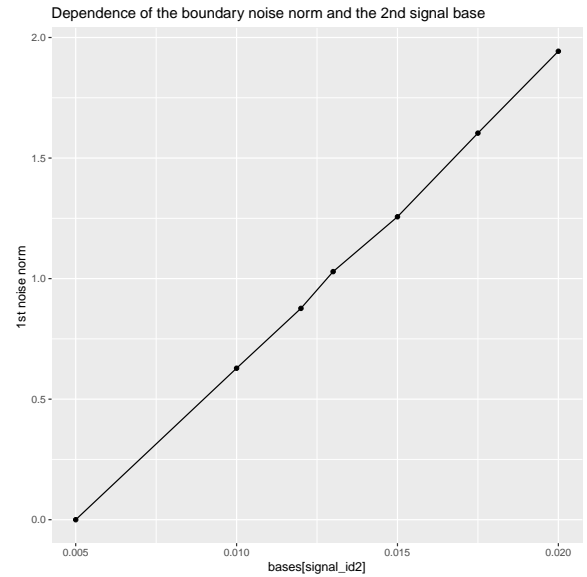


Рис. 2.8. Зависимость граничного значения  $\sigma_1$  от параметра  $\lambda_2$  второго сигнала

### 2.2.3. Результат второго эксперимента

Гипотеза подтверждена, зависимость граничных значений  $\sigma_1$  от отклонения второго сигнала линейная или очень близка к линейной.

## 2.3. Третий эксперимент: линейные сигналы

Как видно на рис. 2.9 иногда линейный сигнал можно хорошо аппроксимировать показательной функцией. Причем, качество приближения зависит от угла наклона. Из-за того, что ранг линейного сигнала равен 2, а показательного — 1, становится интересно, можно ли использовать экспоненциальный сигнал как поддерживающий для линейного?

### 2.3.1. Является ли первая компонента разложения линейного ряда показательной функцией?

Для того чтобы понять можно ли использовать экспоненциальный сигнал как поддерживающий для линейного, нужно узнать, насколько их структура похожа. Например, сравнить аппроксимацию линейного сигнала рядом ранга 1 (восстановить первую компоненту алгоритмом SSA) и аппроксимацию этого же сигнала экспоненциальной функцией.

Из рис. 2.10 видно что первая компонента разложения линейного ряда не является показательной функцией, так как показательные функции не могут дважды пересекаться.

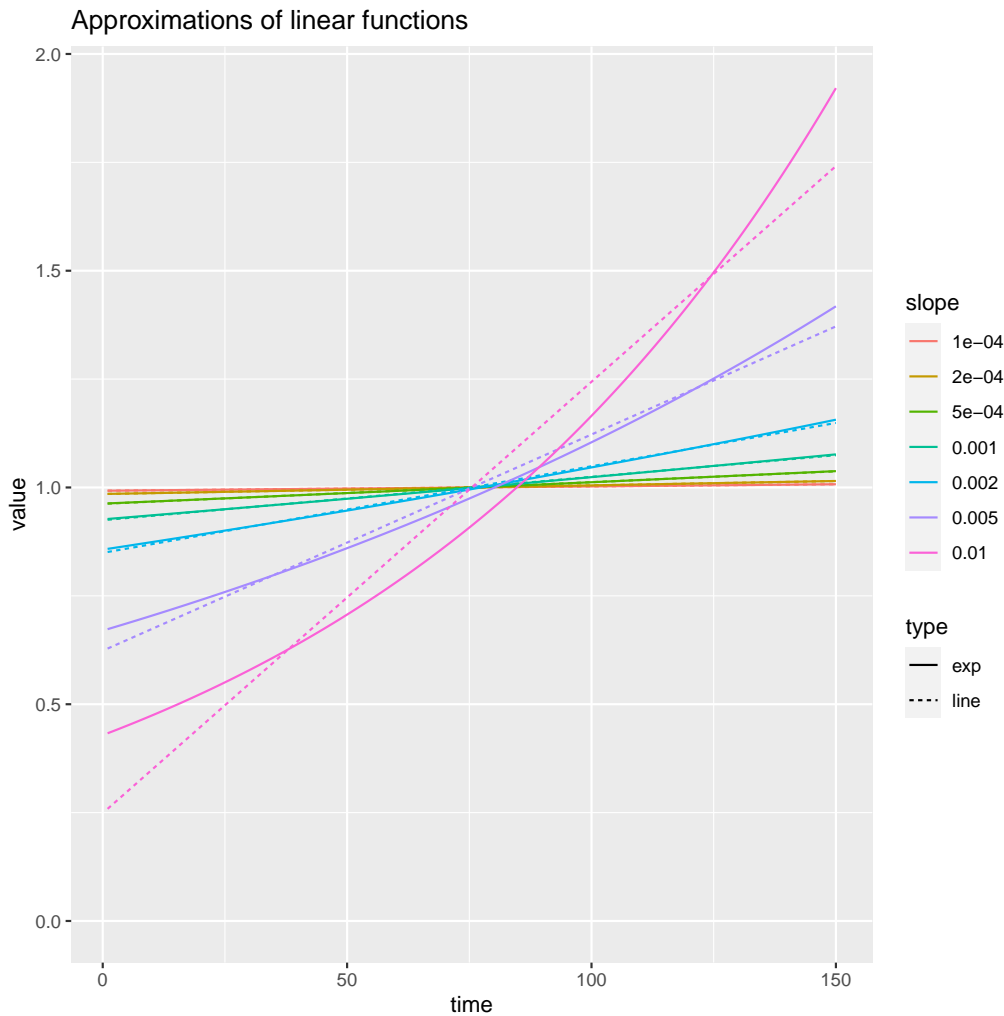


Рис. 2.9. Пример аппроксимации линейных функций показательными.

### 2.3.2. Зависимость доли второй компоненты от угла наклона

Как уже было замечено, при меньших углах наклона, аппроксимация получается лучше. Есть ли зависимость доли второй компоненты в линейном сигнале от угла наклона?

На рис. 2.11 зависимость похожа на экспоненциальную, прологарифмируем и проверим это. На рис. 2.12 видна линейная зависимость логарифма доли второй компоненты и угла наклона сигнала, поэтому доля второй компоненты действительно зависит экспоненциально от угла наклона.

### 2.3.3. При каком шуме вторая компонента теряется

На рис. 2.11 можно заметить, что доля второй компоненты мала, а значит в алгоритме SSA она может оказаться не второй и быть потеряна при достаточно большом шуме.

Как понять, что вторая компонента потерялась не изучая разложение в ручную?

Будем восстанавливать одну или две компоненты алгоритмом SSA из линейных сиг-

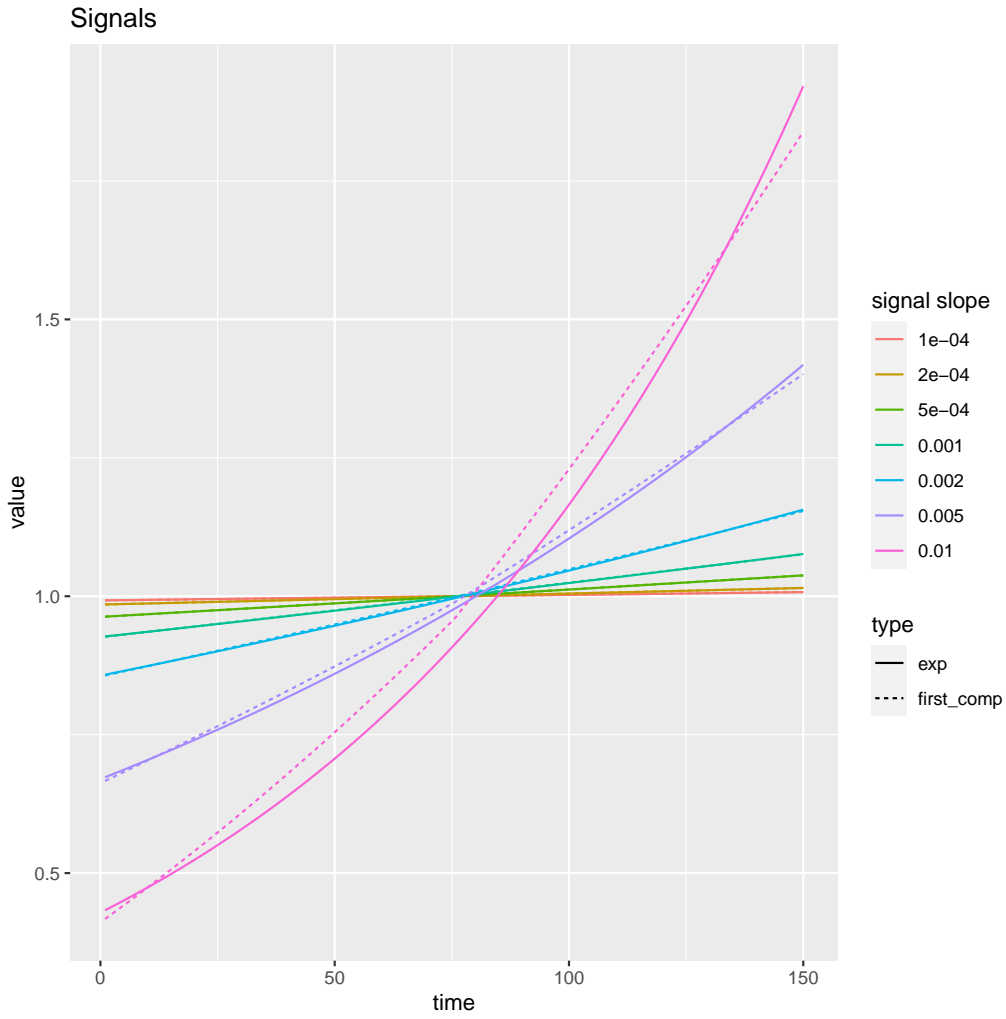


Рис. 2.10. Сравнение первых компонент сигнала и аппроксимирующих экспонент.

налов с разными углами наклона и считать ошибку восстановления. Когда вторая компонента теряется, ошибка двух компонент становится больше ошибки одной компоненты, потому что вместо нужной второй компоненты ряда берется часть шума.

На рис. 2.13 и 2.14 черная линия показывает прямую  $x = y$ , цветные линии — графики ошибок восстановления, пунктирные — двумя компонентами, сплошные — одной. Чтобы ответить на поставленный вопрос, найдем точки пересечения графиков с помощью интерполяции и построим график.

На рис. 2.15 черным пунктиром обозначена прямая  $4y = x - 7$ . Зависимость логарифмов угла наклона и дисперсии шума близка к уравнению  $4 \log_{10}(\text{slope}) = \log_{10}(\sigma^2) - 7$ , поэтому сама зависимость выражается уравнением  $\text{slope}^4 = 10^{-7} \sigma^2$ .

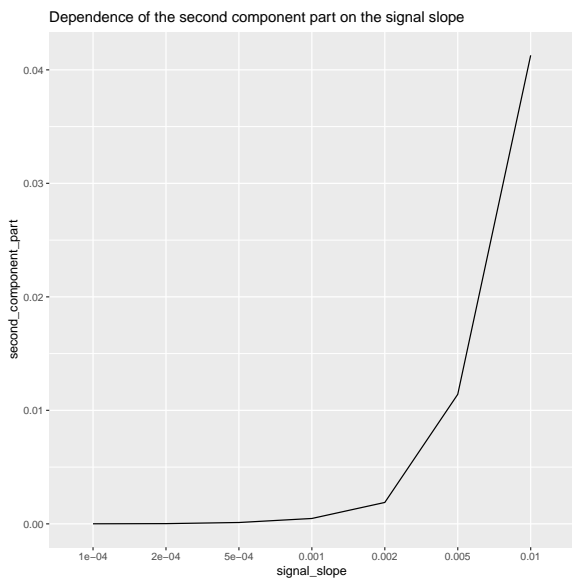


Рис. 2.11. Зависимость доли второй компоненты от угла наклона линейного сигнала.

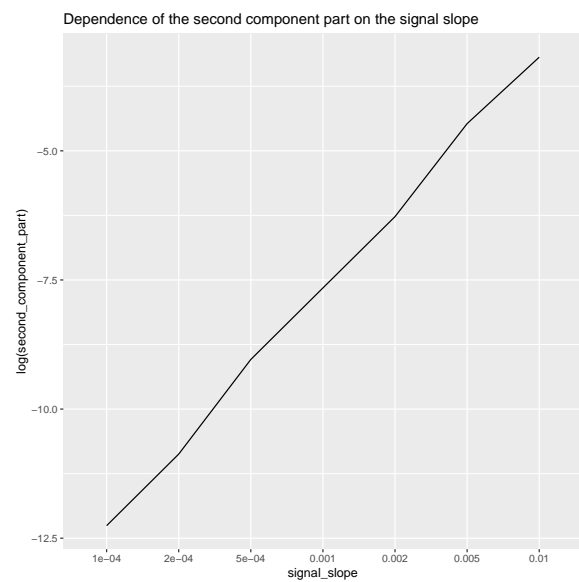


Рис. 2.12. Зависимость логарифма доли второй компоненты от угла наклона линейного сигнала.

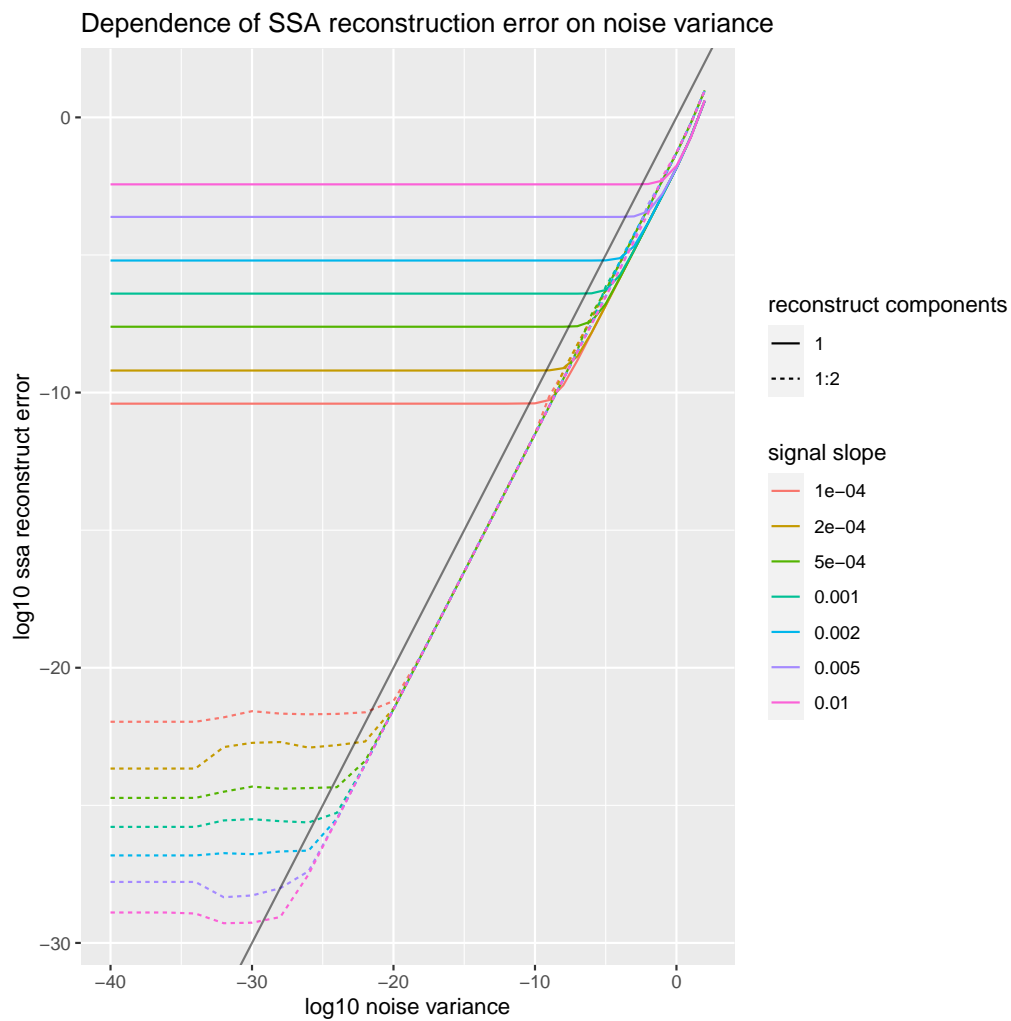


Рис. 2.13. Зависимость ошибки восстановления от  $\sigma^2$ .



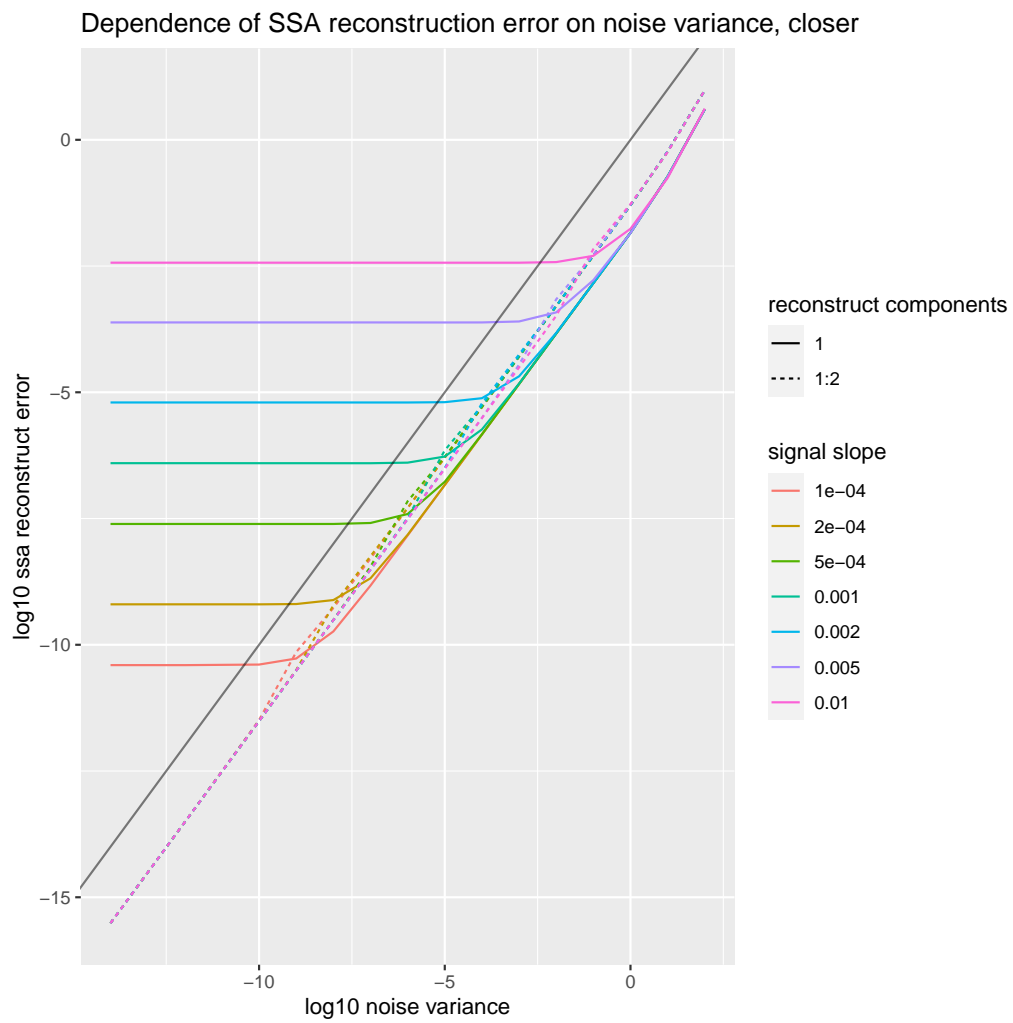


Рис. 2.14. Зависимость ошибки восстановления от  $\sigma^2$ , крупнее.

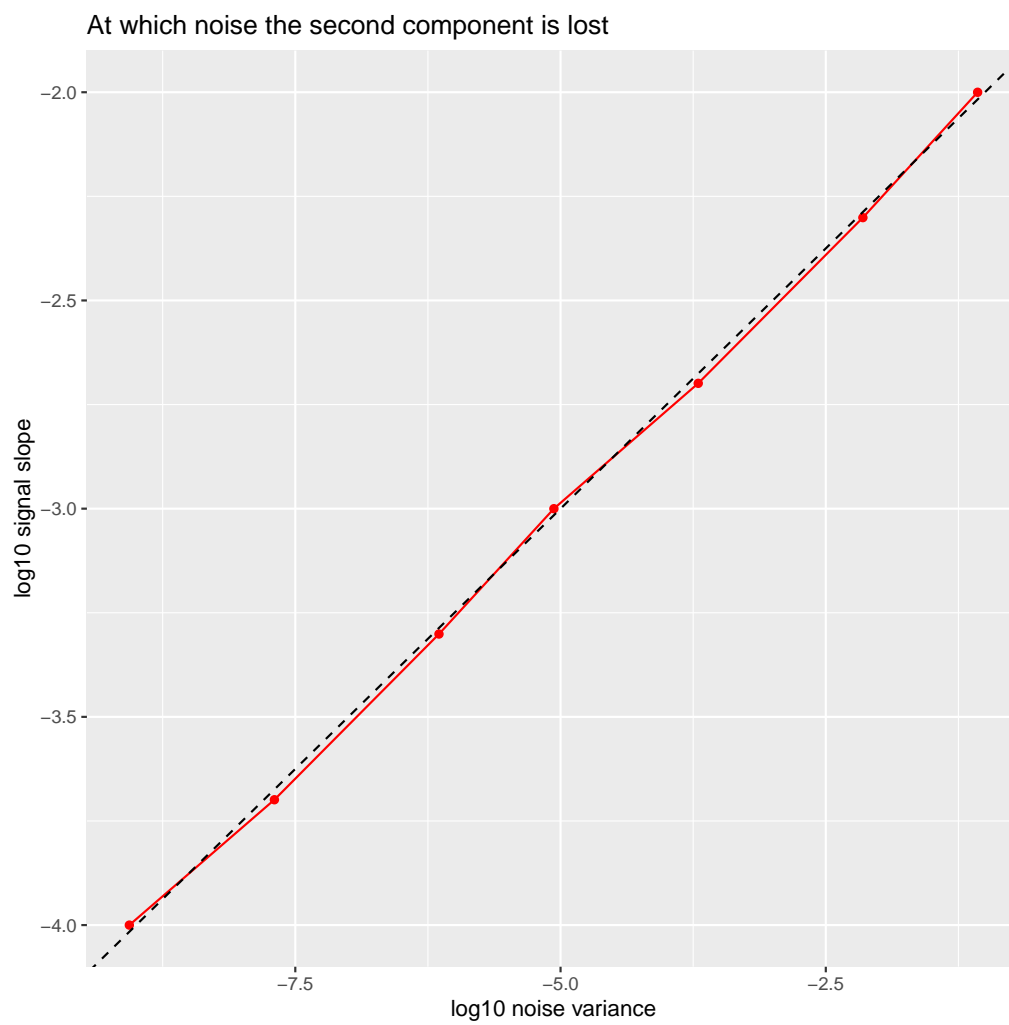


Рис. 2.15. Зависимость ошибки восстановления от  $\sigma^2$

## Заключение

Найдено много интересных зависимостей, характер и причины которых еще предстоит объяснить.

В дальнейшем планируется изучить использование экспоненциального ряда в качестве поддерживающего для линейного и наоборот.

## Список литературы

1. Golyandina N, Korobeynikov A, Zhigljavsky A. Singular Spectrum Analysis with R. — Springer, 2018. — P. 272.