Адаптация метода главных компонент к данным с пропусками

Рукавишникова Анна Александровна, гр. 422

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

Научный руководитель: к.ф.-м.н., доцент Алексеева Н. П.
Рецензент: д.ф.-м.н., профессор Кривулин Н. К.

Санкт-Петербург
2018 г.

Введение: Данные с пропусками

Таблица: Общий вид табличных данных

	y_1	y_2	$\cdots$	y_j	$\cdots$	y_n
1	x_{11}	$*$	$\cdots$	x_{1j}	$\cdots$	x_{1n}
2	$*$	x_{22}	$\cdots$	$*$	$\cdots$	x_{2n}
$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$\cdot$
i	x_{i1}	$*$	$\cdots$	x_{ij}	$\cdots$	x_{in}
$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdots$	$*$
m	x_{m1}	x_{m2}	$\cdots$	$*$	$*$	x_{mn}

Причины пропусков:

- невозможность получения или обработки данных,
- искажение или сокрытие информации,
- утеря данных.

Почему данные с пропусками являются проблемой?

Стандартные статистические методы предполагают, что все переменные в указанной модели измерены для всех случаев.

Введение: Метод главных компонент

Идея метода: редукция размерности данных при наименьшей потере информативности.

Определение

Пусть имеется k центрированных $(EX_i = 0)$ признаков $X = (X_1, \ldots, X_k)^{\mathrm{T}}$ для m индивидов, где $X_i \in \mathbb{R}^m$. Тогда j-ой **главной компонентой** называется линейная комбинация

$$Y_j = A_j^{\mathrm{T}} X = \sum_{i=1}^{k} \alpha_{ij} X_i,$$

где A_j — собственные вектора ковариационной матрицы $\Sigma = EXX^{\mathrm{T}}$, соответствующие собственным числам λ_j, упорядоченным по неубыванию: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$.

Актуальность работы

Необходимость использования ортогональных компонент в качестве вещественной и мнимой части реализаций комплексного марковского нормального стационарного (КМНС) процесса $x_j(t) = u_j(t) + iv_j(t),\ \ j = 1, \ldots, n,\ t = 1, \ldots, k$, где n — число наблюдений, k — число временных точек.

Замечание

Важным свойством метода главных компонент является ортогональность (независимость) главных компонент: $EY_m Y_n^{\mathrm{T}} = 0$ при $m \neq n$.

Структура исходных данных

Структура данных с пропусками по больным туберкулёзом лёгких:

- 30 пациентов, из которых:
  - 19 пациентов с улучшениями после лечения,
  - 11 пациентов без улучшений,
- 32 признака,
- 2 временные точки,
- 23% данных отсутствуют.

Таблица: Основные признаки (измерены в 2-х временных точках)

MMP.1 MMP.8 MMP.9	Белки, участвующие в процессах воспаления при туберкулёзе лёгких
TIMP1	Вещество, регулирующее активность MMP в тканях

Постановка задачи

- Заполнить пропуски в имеющихся данных.
- Применить метод главных компонент к полученным полным данным для каждой группы пациентов в отдельности.
- Использовать полученные главные компоненты для исследования корреляционной структуры данных и анализа динамики заболевания на основе модели KMHC процесса.

Метод заполнения пропусков Predictive Mean Matching

$\Omega = (\omega_1, \ldots, \omega_n)$ — наблюдения по неполным данным:
зависимой переменной $\mathcal{Y}(\Omega) = (y(\omega_1), \ldots, y(\omega_n))$,
независимых переменных $\mathcal{X} = (X_1(\Omega), \ldots, X_p(\Omega))$.

Модель линейной регрессии $Y = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$

- $Y = \mathcal{Y}(\Omega_1)$, $\Omega_1 \subseteq \Omega$ — полные наблюдения, $\Omega_2 = \Omega \setminus \Omega_1$,
- $X = (X_{t_1}(\Omega_1), \ldots, X_{t_k}(\Omega_1))$, $(t_1, \ldots, t_k) \subseteq (1, \ldots, p)$.
- $\hat{\beta}$, $\hat{\sigma}^2$ по МНК

Предикторы:

- $\hat{Y}(\Omega_1) = X\hat{\beta}$,
- $Y^*(\Omega_2) = X_*\beta^*$, где $X_* = (X_{t_1}(\Omega_2), \ldots, X_{t_k}(\Omega_2))$, $\beta^* \sim \mathcal{N}(\hat{\beta}, \mathrm{Cov}(\hat{\beta}))$.

$\Delta(u, v) = |\hat{Y}(u \in \Omega_1) - Y^*(v \in \Omega_2)|$

Подбор значений

$\forall v \in \Omega_2$ подбирается $y(v) = y(u) \in \mathcal{Y}(\Omega_1)$ при условии $\Delta(u, v) \leq \Delta_*$.

Заполнение пропусков в данных

- Используемая среда программирования: R.
- Используемый программный пакет: MICE (Multivariate Imputation by Chained Equations).
- Метод заполнения пропусков: Predictive Mean Matching.
- Количество полученных полных наборов данных: 5.

Применение главных компонент. КМНС процесс

Цель: для двух групп пациентов исследовать корреляционную структуру данных.

Подход к решению: оценка параметров τ и η ковариационной функции комплексного марковского нормального стационарного (КМНС) процесса (Дж.Л. Дуб, 1956)

$$\mathcal{B}(t) = \sigma^2 e^{-\eta|t| - i\tau t}, \quad \eta > 0.$$

Её выборочная оценка с m реализациями в k временных точках имеет вид:

$$\hat{\mathcal{B}}(t) = \frac{1}{(k-t)m} \sum_{l=1}^{k-t} \sum_{j=1}^{m} x_{jl} x^*_{j,l+t}.$$

Реализация

В качестве вещественной и мнимой части независимых реализаций КМНС процесса $x_j(t) = u_j(t) + iv_j(t) = x_{jt}$ берём значения первой и второй главной компоненты, соответственно.

Оценка параметров КМНС процесса

Теорема (А.Г. Барт, 2003, Н.П. Алексеева, 2012)

Пусть $x_j(t) = u_j(t) + iv_j(t) = x_{jt}$ — m независимых реализаций КМНС процесса в k временных точках,

$$A_1 = \sum_{l=1}^{k}\sum_{j=1}^{m} x_{jl}^* x_{jl}, \quad A_2 = \sum_{l=2}^{k-1}\sum_{j=1}^{m} x_{jl}^* x_{jl}.$$

Если $A_1(k-2) = A_2 k$, то ОМП $\hat{\tau}, \hat{\eta}, \hat{\sigma}$ удовлетворяют соотношениям:

$$\operatorname{tg}\hat{\tau} = \frac{\operatorname{Im}(\hat{\mathcal{B}}(1)/\hat{\mathcal{B}}(0))}{\operatorname{Re}(\hat{\mathcal{B}}(1)/\hat{\mathcal{B}}(0))}, \quad \hat{\eta} = -\ln\left|\frac{\hat{\mathcal{B}}(1)}{\hat{\mathcal{B}}(0)}\right|, \quad \hat{\sigma}^2 = \hat{\mathcal{B}}(0),$$

где $\hat{\mathcal{B}}(t)$ — выборочная оценка ковариационной функции КМНС процесса.

Результаты: устойчивость оценок КМНС процесса

Замечание

Главные компоненты определены с точностью до знака.

Утверждение

Пусть $\mathbf{u}$ и $\mathbf{v}$ — первая и вторая главные компоненты соответственно. Тогда для оценок параметров КМНС процесса $\hat{\eta} = \hat{\eta}(\mathbf{u}, \mathbf{v})$ и $\hat{\tau} = \hat{\tau}(\mathbf{u}, \mathbf{v})$ верны следующие соотношения:

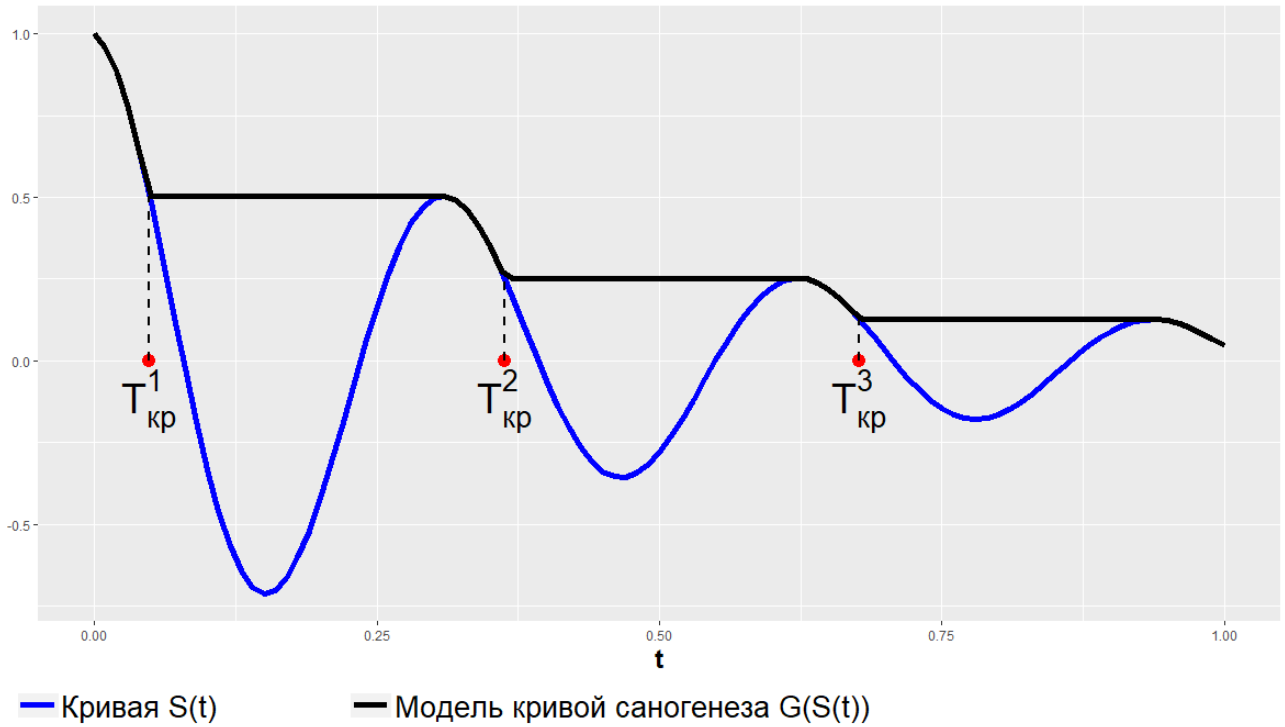
- $\hat{\eta}(\pm\mathbf{u}, \pm\mathbf{v}) = \hat{\eta}(\mathbf{u}, \mathbf{v})$,
- $\hat{\eta}(\mathbf{v}, \mathbf{u}) = \hat{\eta}(\mathbf{u}, \mathbf{v})$,
- $\hat{\tau}(-\mathbf{u}, \mathbf{v}) = \hat{\tau}(\mathbf{u}, -\mathbf{v}) = -\hat{\tau}(\mathbf{u}, \mathbf{v})$,
- $\hat{\tau}(\mathbf{v}, \mathbf{u}) = -\hat{\tau}(\mathbf{u}, \mathbf{v})$.

Параметрическая модель кривой саногенеза

Оценки $\hat{\tau}$ и $\hat{\eta}$ используются для построения вещественной части ковариационной функции КМНС процесса ($\sigma^2 = 1$)

$$S(t) = e^{-\eta t}\cos\tau t, \quad t > 0, \eta > 0.$$

Рассматриваем модель **кривой саногенеза** (А.Г. Барт, 2003) $F(t) = G(S(t))$, полученную двойным правым обращением кривой $S(t)$, а именно: $F(t) = (S_{11}^-)_{11}^-(t)$, где $S_{11}^-(t) = \sup\{x : S(x) \geq t\}$.



Критические точки

Важную роль в течении болезни играют **критические точки** $T^j_{\text{кр}}$ — моменты времени, в которые возможно изменение характера течения болезни.

Задача

Найти первую критическую точку для двух групп пациентов и выяснить, у кого она раньше.

Решение

По имеющимся оценкам параметров КМНС процесса $\hat{\tau}$ и $\hat{\eta}$ находим первые критические точки по формуле (А.Г. Барт, 2003)

$$T^1_{\text{кр}} = \frac{2\pi - \phi}{\tau} + Q,$$

где $\operatorname{tg}\phi = \eta/\tau, \quad e^{2\pi\eta/\tau}\left(\cos\tau Q + \frac{\eta}{\tau}\sin\tau Q\right) = e^{\eta Q}.$

Результаты: корреляционная структура

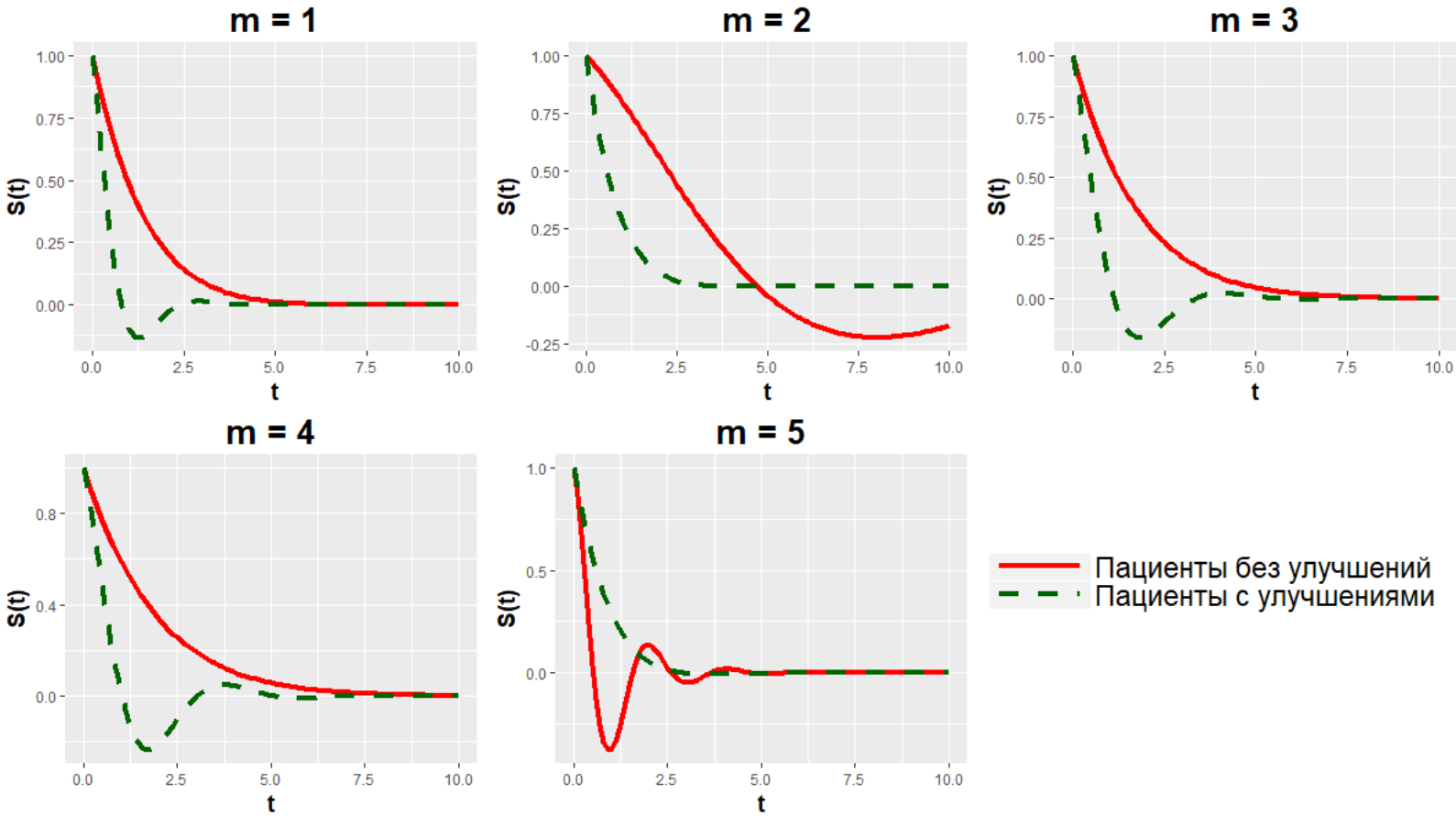


Рис.: $S(t) = e^{-\eta t} \cos \tau t$, **m** – номер заполнения.

Результаты по 50 заполнениям: в 33 случаях распад корреляционых связей для пациентов с улучшениями больше. По критерию знаков p-значение $= 0.016 < 0.05$.

Результаты: модели кривых саногенеза

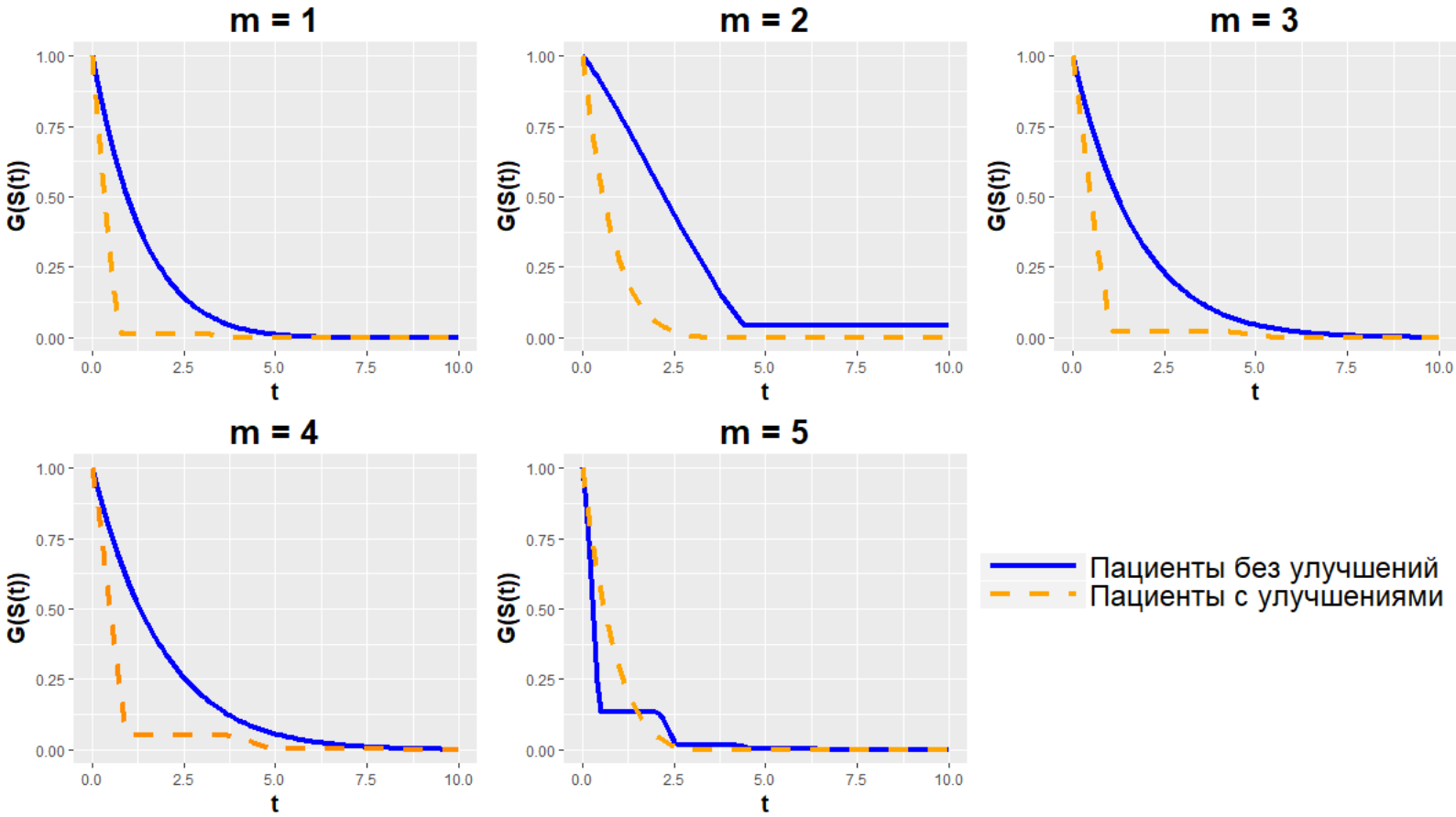


Рис.: Модель кривой саногенеза для обеих групп пациентов при разных заполнениях пропусков (**m** — номер заполнения).

Вывод: для пациентов без улучшений модель кривой саногенеза выше.

Результаты: критические точки

m	$T^1_{\text{кр}}$ для пациентов без улучшений	$T^1_{\text{кр}}$ для пациентов с улучшениями
1	6.57	0.77
2	4.41	3.40
3	10.96	1.08
4	11.18	0.94
5	0.45	2.84

Таблица: Первая критическая точка при разных заполнениях пропусков (**m** — номер заполнения).

Результаты по 50 заполнениям: в 38 случаях первая критическая точка для пациентов с улучшениями раньше. По критерию знаков p-значение $= 0.0002 < 0.05$.

Заключение

- Реализован алгоритм заполнения пропусков в реальных медицинских данных на основе метода PMM.
- Осуществлен анализ корреляционной структуры данных на основе модели KMHC процесса.
- Доказана устойчивость оценок основных параметров KMHC процесса к знакопеременной структуре главных компонент и их перестановке.
- Показана статистическая значимость порядка критических точек для двух групп пациентов.
- Создана программа на языке программирования R, позволяющая производить вычисления от заполнения пропусков до оценки критических точек.