# **Анализ данных с пропусками и их восстановление**

***Степанова Ю.В.***

*студент*

*Московский государственный университет имени М.В. Ломоносова, физический факультет, Москва, Россия*

*E–mail*: *stepanova.yv23@physics.msu.ru*

В эпоху развития сферы больших данных всё чаще возникают проблемы обработки неполной информации. Наличие пропусков, вызванных техническими ограничениями или внешними факторами, делает неприменимыми стандартные методы анализа, приводя к искажениям и неточным результатам. В контексте временных рядов и многомерных данных, таких как случайные векторы, задача эффективного восстановления пропущенной информации становится критически важной.

Если известны истинная ковариационная матрица и среднее значение случайного вектора, восстановление пропущенных компонентов можно осуществить, основываясь на построении условного математического ожидания 𝐸(𝑌) и условной ковариационной матрицы [1].

𝐸(𝑌) = 𝜇𝑌 + 𝐶𝑋𝑌$V\_{X}^{-1}$(X − 𝜇𝑋)

𝑉(𝑌) = 𝑉𝑌 − 𝐶𝑋Y $V\_{X}^{-1}C\_{XY}^{T}$

Здесь X – это вектор, состоящий из известных компонентов, 𝑌 - из пропущенных, 𝜇𝑋 и 𝜇𝑌 – их соответствующие средние значения, 𝐶𝑋𝑌 – матрица ковариаций между известными и неизвестными компонентами, $V\_{X}$ – ковариационная матрица известных компонентов, а 𝑉𝑌 – ковариационная матрица неизвестных. Все эти матрицы выделяются из ковариационной матрицы всего вектора. Использование условного математического ожидания позволяет оптимально оценить пропущенные значения, обеспечивая минимальные ошибки восстановления, рис. 1.



***Рис. 1.*** Пример восстановления вектора

На этом рисунке кроме восстановленных компонентов также показаны отвечающие им стандартные отклонения, полученные из условной ковариационной матрицы 𝑉(𝑌).

В реальных задачах мы, как правило, не обладаем информацией об истинной ковариационной матрице. Поэтому в качестве основы для восстановления информации используется выборочная ковариационная матрица, построенная на основе конечной выборки случайных векторов. Оценка выборочной ковариационной матрицы позволяет восстановить утерянные компоненты, основываясь на статистической структуре имеющихся данных, и применить полученные результаты в различных прикладных областях.

Традиционные методы, применяемые для построения выборочной ковариационной матрицы, предполагают полноту обучающей выборки. Однако в современных условиях больших данных обучающие выборки часто содержат пропуски. Отказ от данных с пропусками может привести к существенной потере информации, особенно если пропусков много. Поэтому остро встает потребность в разработке методов, способных полноценно анализировать данные с пропусками [2]. В данной работе рассматривается задача извлечения информации из обучающей выборки, содержащей дефекты, когда требуется оценка ковариационной структуры и восстановление утерянных компонентов случайного вектора. Рассматриваемые методы используют подходы к анализу потоков больших данных [3], в которых по мере поступления данных пополняется информация специального вида, на основании которой и строятся оценки среднего и ковариационной матрицы.

# **Литература**

1. Ширяев А.Н. Вероятность. Москва, издательство МЦНМО, 2004
2. Дещеревский А.В., Журавлев В.И., Никольский А.Н., Сидорин А.Я.

Проблемы анализа временных рядов с пропусками и методы их решения в программе WINABD // ГЕОФИЗИЧЕСКИЕ ПРОЦЕССЫ И БИОСФЕРА, 2016, T. 15, № 3, с. 5–34.

1. Голубцов П.В. Понятие информации в контексте задач обработки больших данных // НТИ. Сер. 2. Информ. Процессы и системы 2018. № 1, с. 31-36.