

遅延を伴う事象を分析するための 遅延相関分析の応用

Open Source Software Engineering Laboratory

B4 加藤寛瑛

遅延相関分析

- 本研究における遅延相関分析とは…
 - OSS開発におけるプロダクトメトリクスとプロセスメトリクスの時間的遅延を伴う関係进行分析すること
- 一方のシステムの進化（変化）が、一定時間後に他方のシステムの変化に影響を与えるという関係の抽出を支援する

遅延相関分析

ソースコード

- LOC（行数）
- クラス数
- 関数の数
- コミット数

不具合修正履歴

- 不具合を報告した人数
- 不具合が報告された件数
- 優先度が高い不具合件数

メール

- 総投稿数
- 新しい参加者の人数
- 平均行数

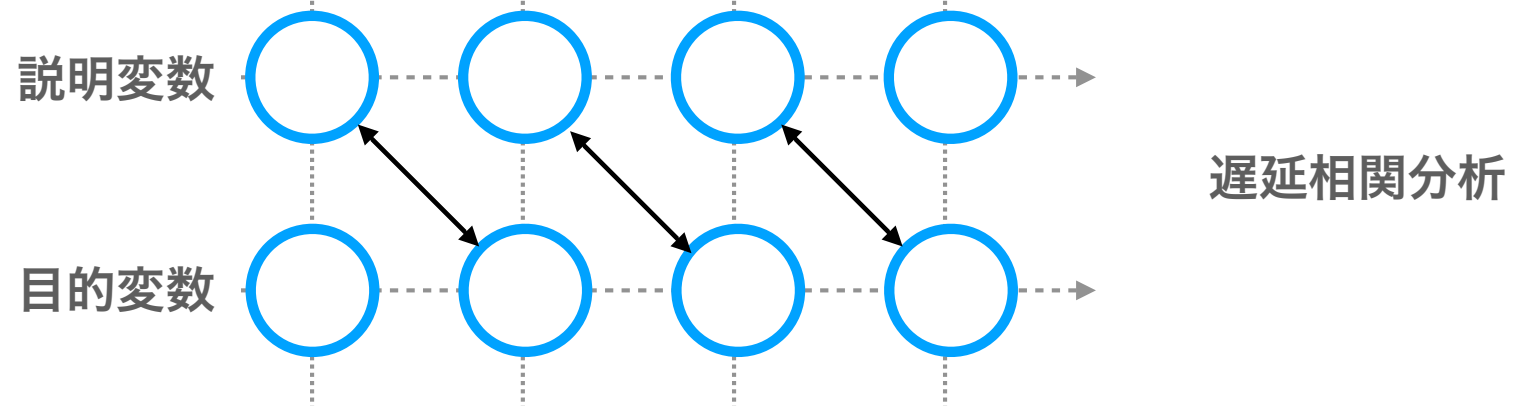
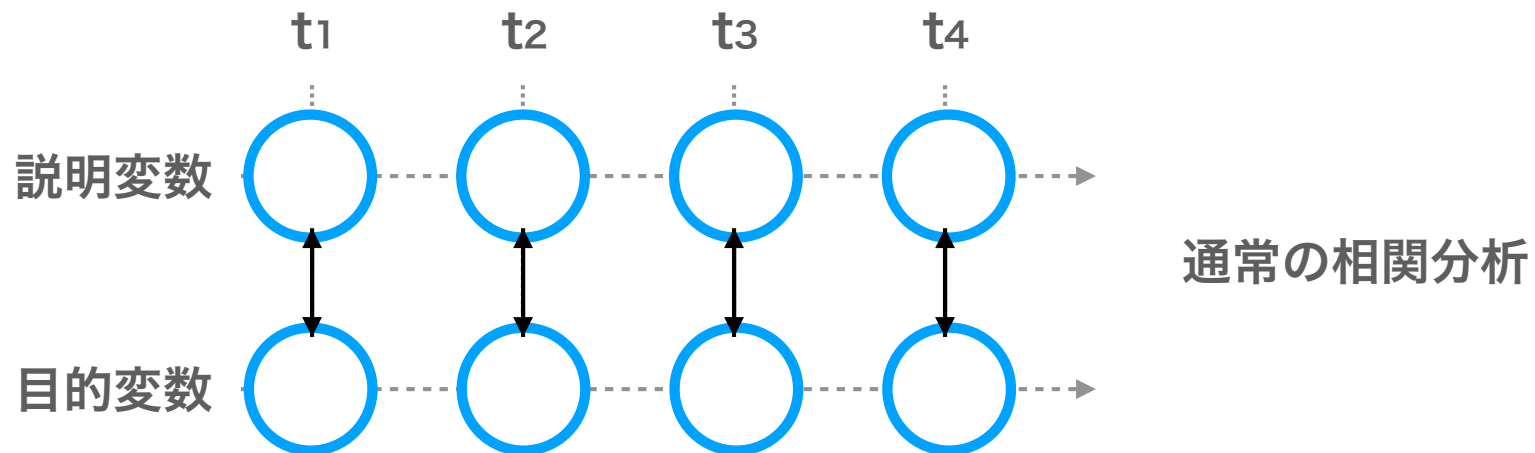


遅延相関分析に使うためのメトリクスを抽出



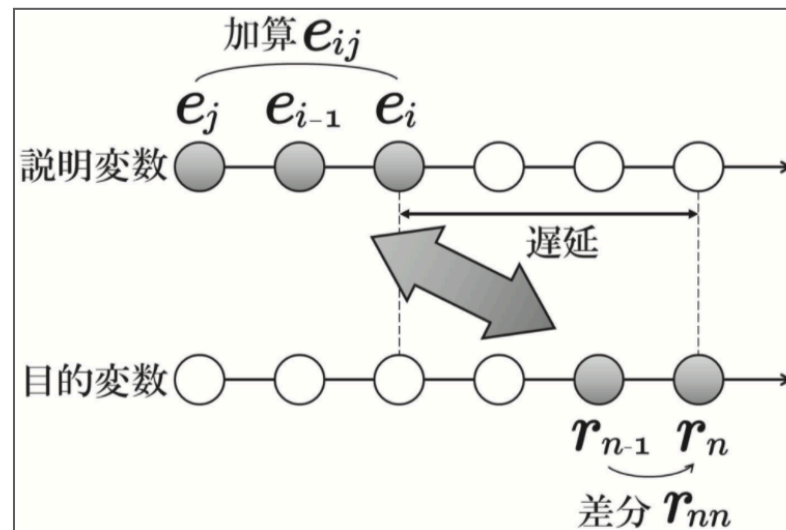
それぞれのメトリクスのデータを時系列にそって並べる

遅延相関分析



先行研究

- Eclipse Platformプロジェクトを対象に131種類のメトリクスを用いてケーススタディを行なった結果、従来の相関分析では抽出できない31件の関係を遅延相関分析により抽出できた[1]

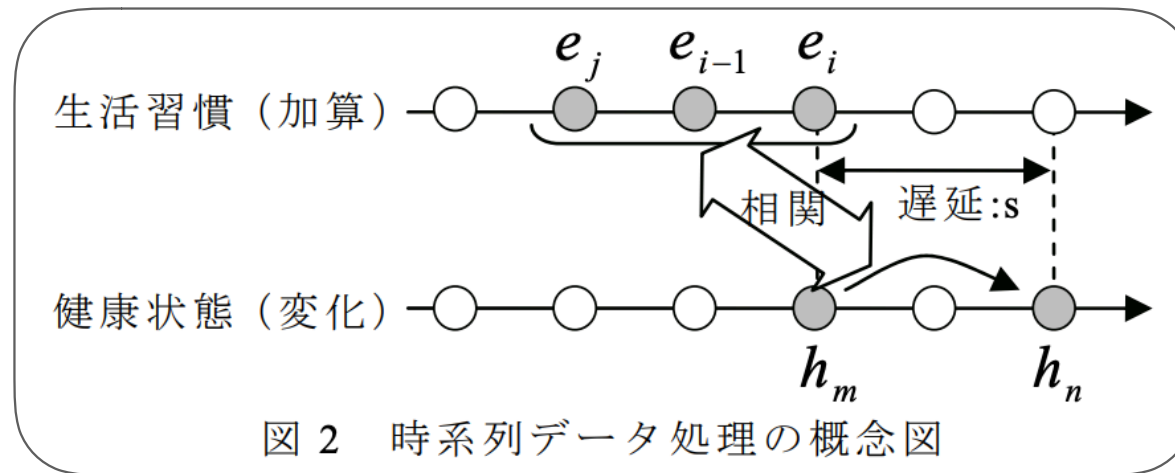


加算係数=3, 遅延係数=3, 差分係数=1
とした場合の計算例

[1]OSSシステムとコミュニティの共進化の理解を目的としたデータマイニング手法（山谷陽亮）

先行研究

- 元々は医療関係の研究を引用したもの



生活習慣に関するメトリクス（喫煙など）を説明変数として加算し、
健康状態に関するメトリクス（肺がんなど）を目的変数として差分をとる

背景・目的

- 背景

- 近年、オープンソースソフトウェア（OSS）を活用したシステム開発が一般的になっている一方で、「開発コミュニティから継続したサポートが得られるかわからない」、「緊急時にサポートが迅速に受けられない」といったデメリットは依然として改善できていない

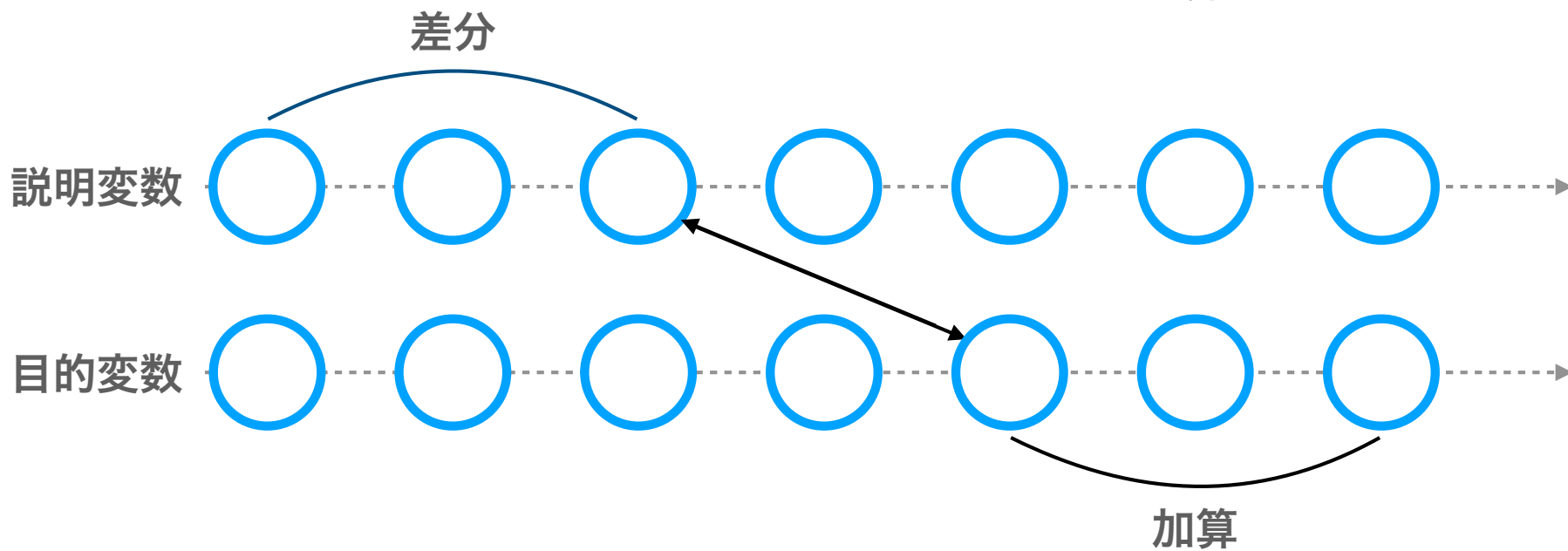
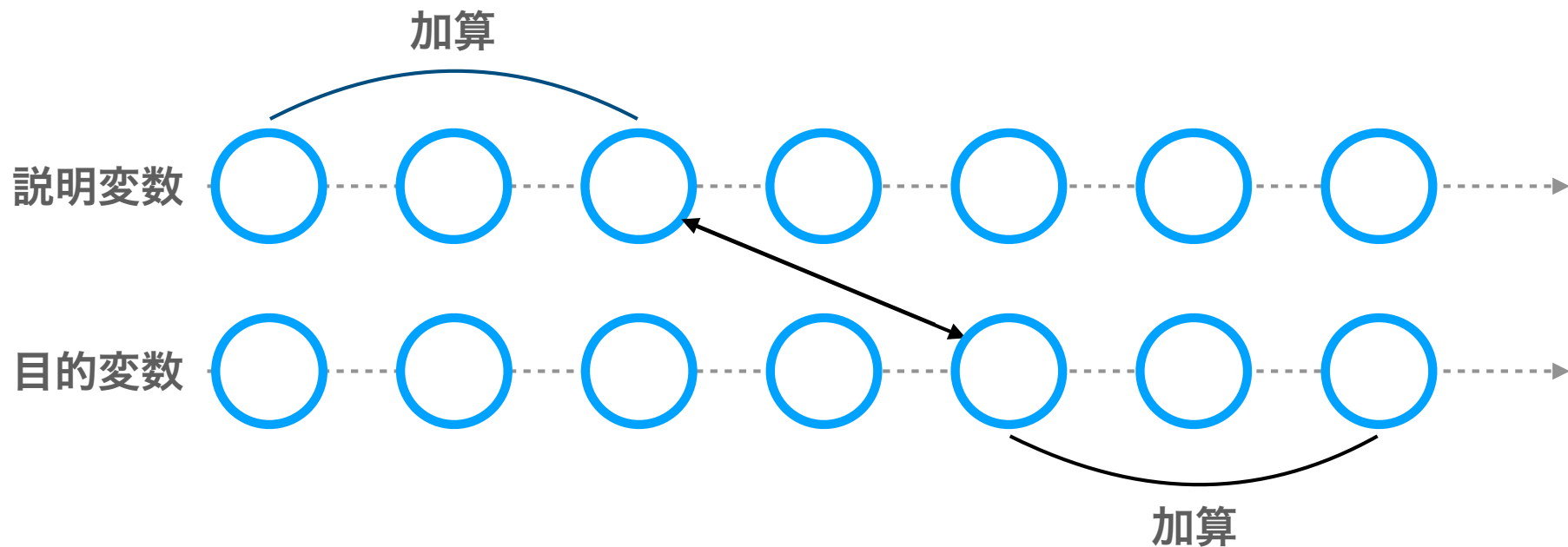
- 目的

- 山谷さんらはOSSシステムとコミュニティの共進化のプロセスを定量的に分析するために、遅延相関分析を提案した
- 上記の研究では、説明変数の加算、目的変数の差分を用いた遅延相関分析を行っていたが、**本研究では他の3通り（加算・加算、差分・加算、差分・差分）の組み合わせのデータセットを作成して分析を行う**

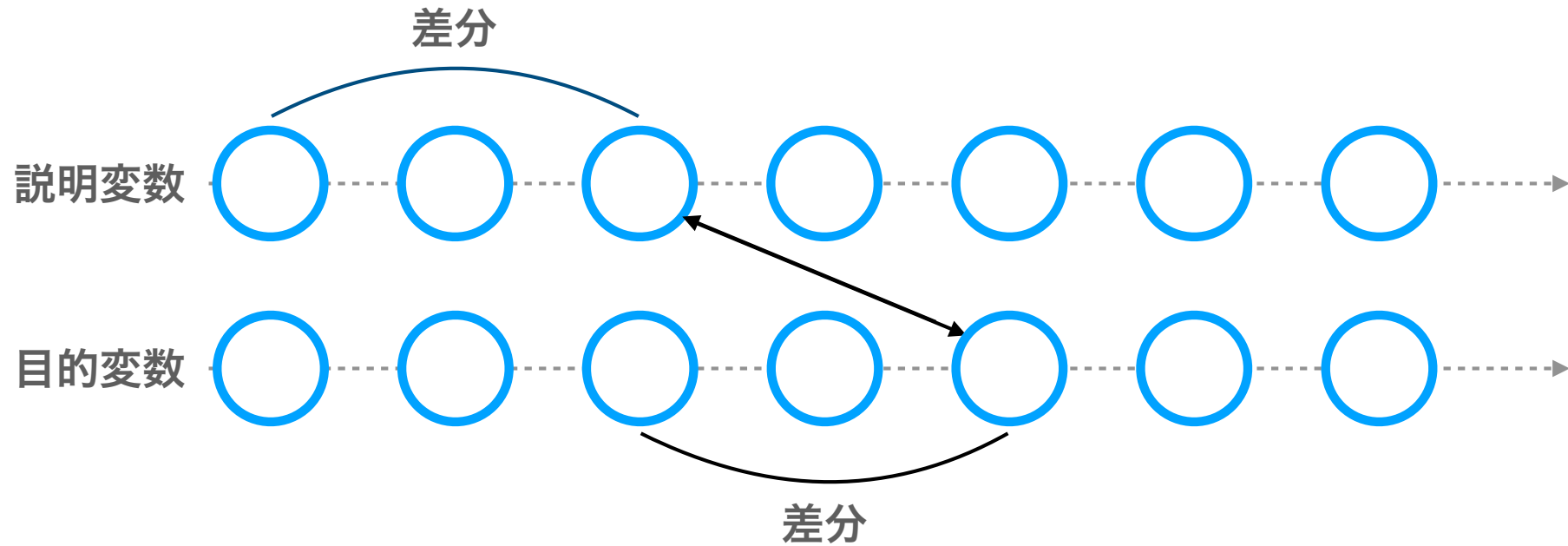
対象データ

- Apache HTTP Serverの2001年から2015年のソースコード（cファイルのみ），不具合修正履歴，Apacheのメーリングリストからそれぞれメトリクスを抽出
- ソースコード
 - Githubから抽出したソースコードを，Understand（ソースコード解析ツール）を使用しメトリクスを取得
- 不具合修正履歴
 - Bugzillaから変更履歴を取得し，メトリクス毎にデータ整形
- メールデータ
 - Apacheのメーリングリストの履歴をMbox形式で取得し，メトリクス毎にデータ整形

4通りの遅延相関分析



4通りの遅延相関分析



ToDo

- 4通りの実行した出力結果が正しいかテストする
- 目的変数の加算の取り方を考える
- 従来の相関分析と比較する
- 普遍性があるか調べる
 - Apache HTTP Serverプロジェクト内で期間を区切って分析し、有効性 (Validity) を検証する
 - (時間があれば) 他プロジェクトでも遅延相関分析をする