

多変量自己回帰モデルとの比較を通じた 遅延相関分析の有用性評価

オープンソースソフトウェア工学研究室

60190003

蘆田 誠人

背景

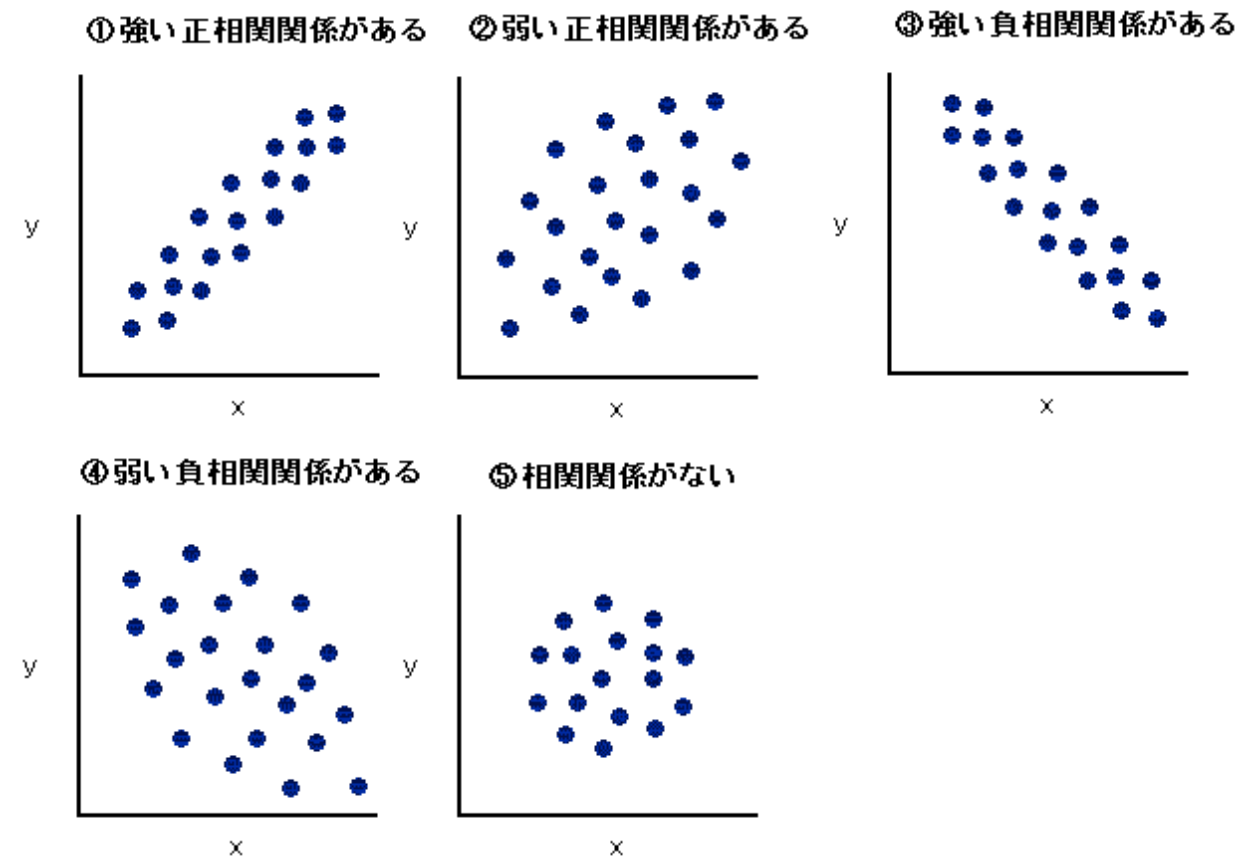
- 高品質で効率的なソフトウェアを開発するためには、プロジェクト内の作業を定量的に管理することが重要^[1]
 - ▶ 管理するためにはソフトウェア開発データの計測、分析が必要
 - 代表的なひとつとして相関分析が挙げられる

[1]定量的プロジェクト管理の推進：IPA 独立行政法人 情報処理推進機構, <https://www.ipa.go.jp/sec/system/teiryoku.html>

相関分析

- 相関分析とは

- ▶ 2変数間にどれほど直線間の関係があるか



出展：統計的品質管理学習ナビ, <http://www.sqc-works.com/qc7-04.html>

ソフトウェア開発の現状

- 実際のソフトウェア開発

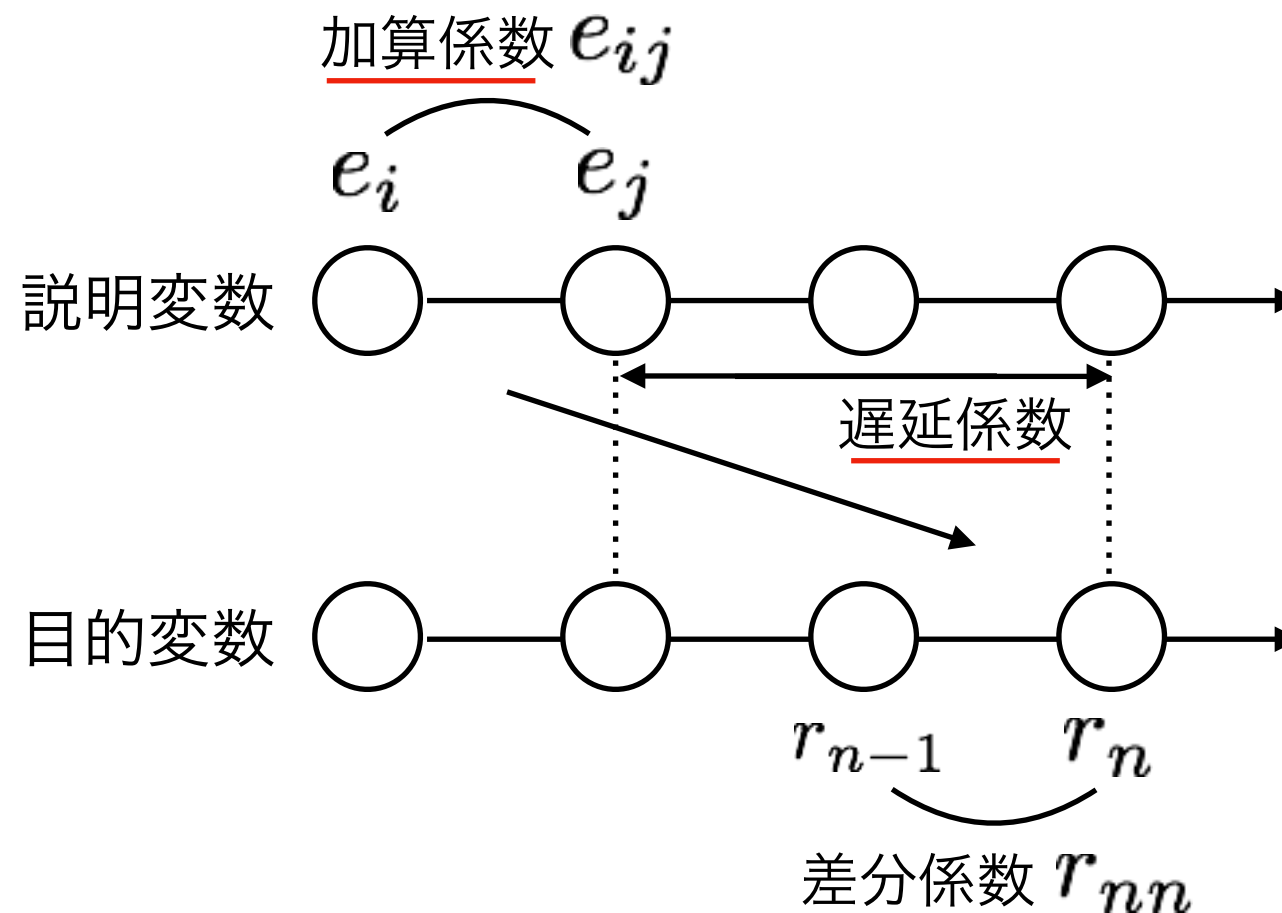
- ▶ 事象Aが発生してからある程度経過して事象Bが発生する可能性

- バグ報告件数が増えると、一定期間後バグ修正数が増加

- ✓ 時間を考慮した相関分析を行う必要がある

遅延相関分析

- 遅延相関分析
 - ▶ 時間的遅延を考慮した相関分析手法
 - ▶ 三つの係数を基に相関分析を行う



遅延相関分析

- 三つの係数

- ▶ **加算係数(1~12)**

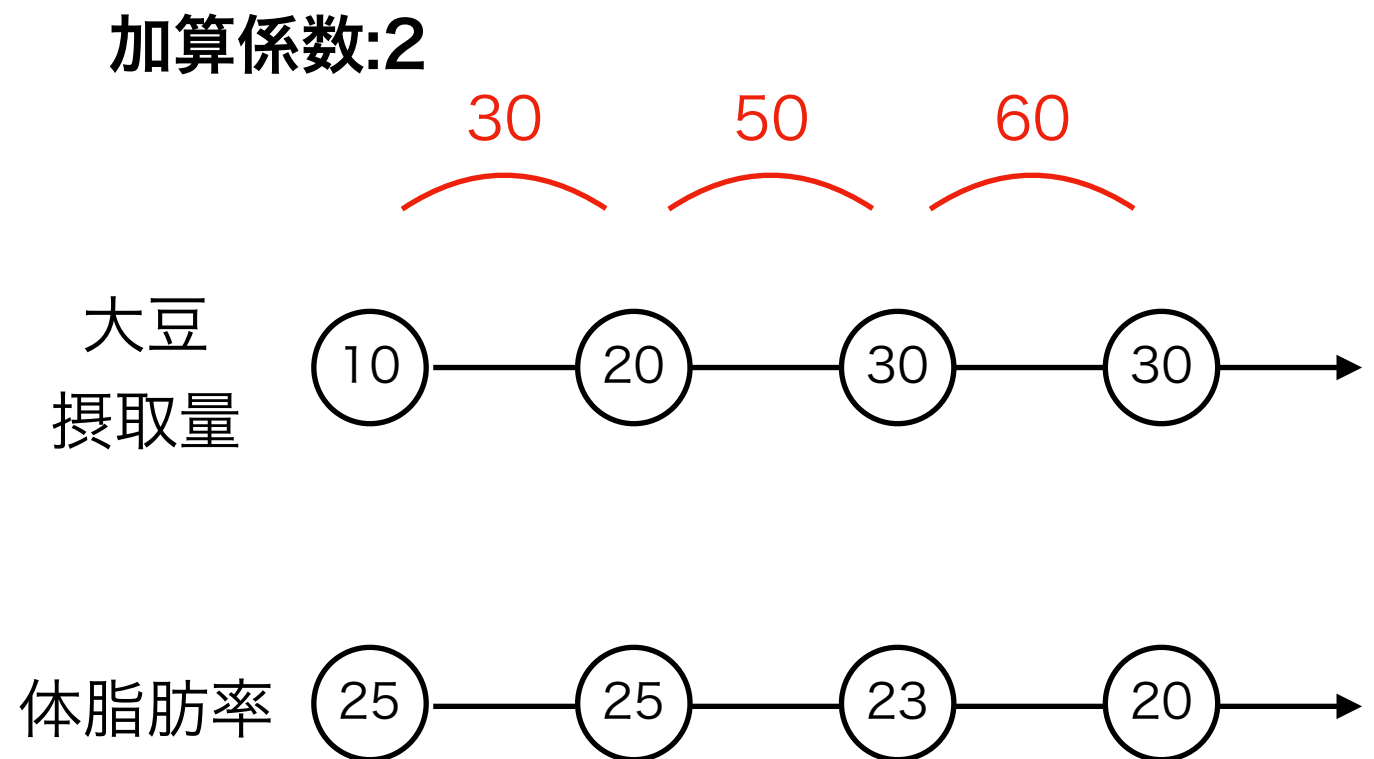
- 説明変数の累積値

- ▶ 差分係数(1~12)

- 目的変数の変化量

- ▶ 遅延係数(1~12)

- 遅延した期間



遅延相関分析

- 三つの係数

- ▶ 加算係数(1~12)

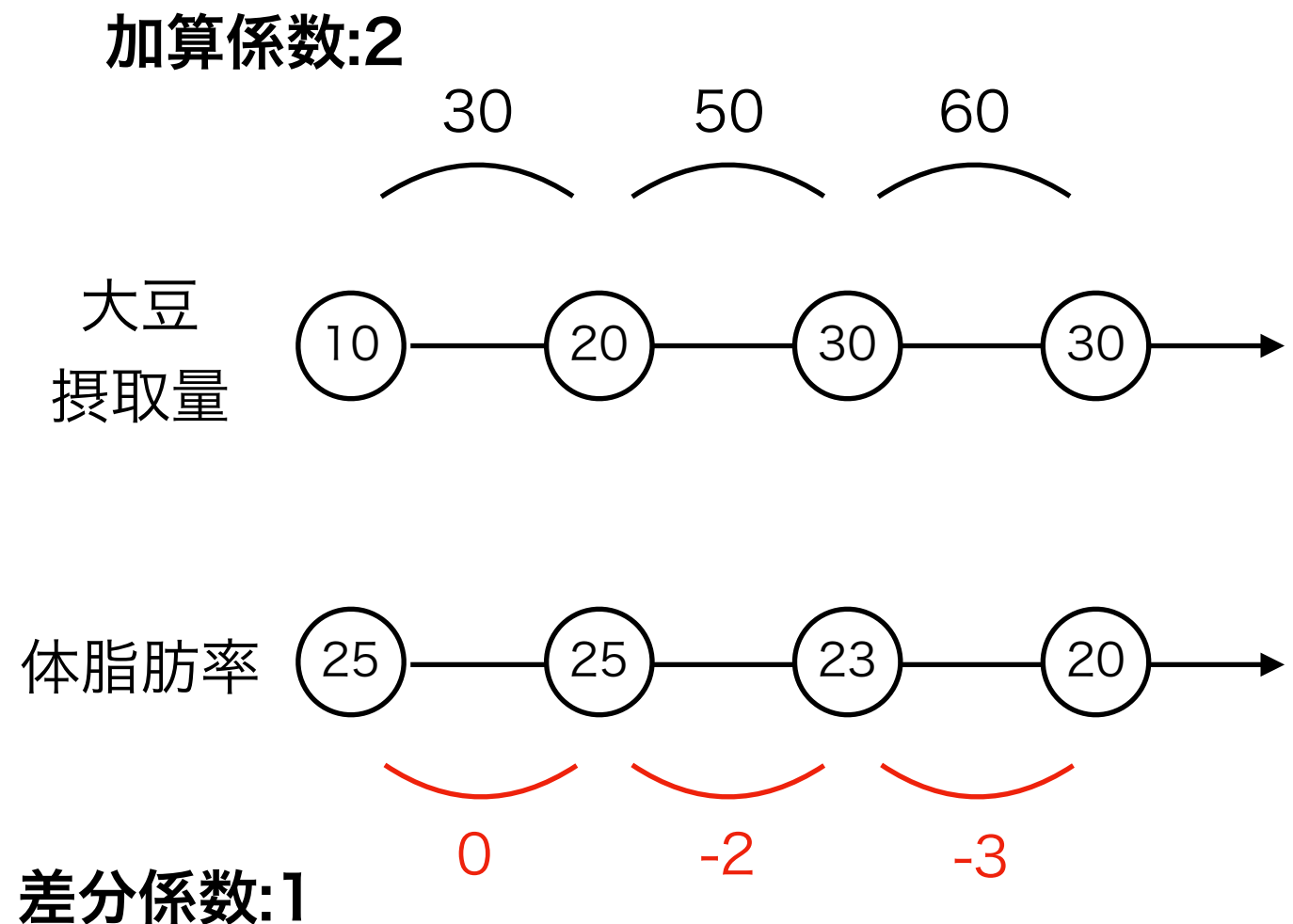
- 説明変数の累積値

- ▶ **差分係数(1~12)**

- 目的変数の変化量

- ▶ 遅延係数(1~12)

- 遅延した期間



遅延相関分析

- 三つの係数

- ▶ 加算係数(1~12)

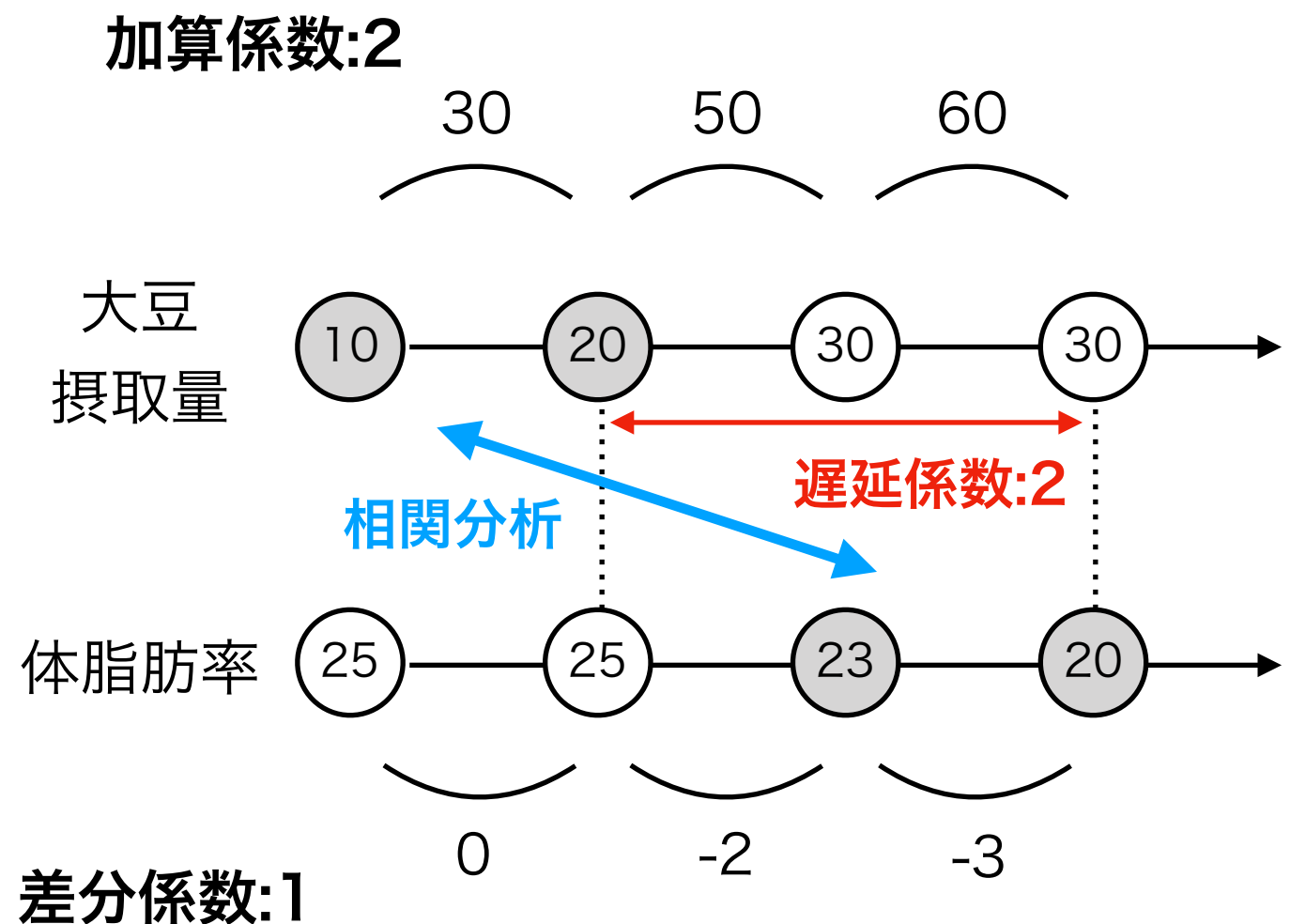
- 説明変数の累積値

- ▶ 差分係数(1~12)

- 目的変数の変化量

- ▶ 遅延係数(1~12)

- 遅延した期間



遅延相関分析

- 最大相関係数

- ▶ 三つの係数のセットで、相関係数の絶対値が一番高い係数のセットを記憶

説明変数	目的変数	加算	差分	遅延	相関係数
countLine	totalFixed	1	1	1	0.05
countLine	totalFixed	1	1	2	0.23
countLine	totalFixed	1	1	3	0.48
countLine	totalFixed	1	1	4	0.39

組み合わせを明らかに

先行研究

- 生活習慣から健康状態への影響の調査[2]
 - ▶ 生活習慣がどの程度の遅延を経て健康状態に影響を及ぼすか
 - ▶ 穀物を摂取することにより血糖値が上がる, という状態を抽出
 - 5日間運動すると2日後に体脂肪率が下がる
- OSSの共進化の調査[3]
 - ▶ システムとコミュニティの共進化を定量化
 - ▶ 従来の相関分析では見れなかった時間遅延に基づく関係が抽出できた

[2]竹内裕之, 児玉直樹. 生活習慣と健康状態に関する時系列データ解析手法の開発. In Proceedings of the 3th Forum on Data Engineering and Information Management(DEIM'08), E1-5, 2008.

[3]山谷陽亮, 大平雅雄, Passakorn Phannachitta, 伊原彰紀. 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. 情報処理学会論文誌, Vol. 56, No. 1, pp. 59-71, 1 2015.

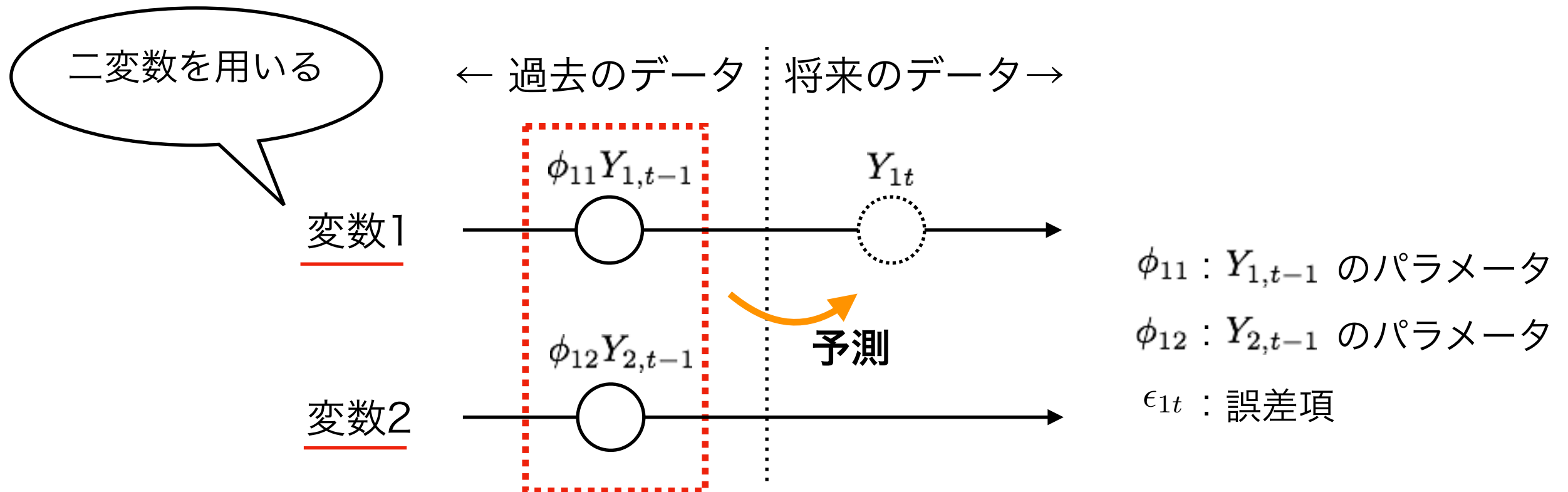
時間的遅延を伴う事象を分析する手法

- 遅延相関分析と同じく，時系列データ分析をする手法のひとつに時系列モデルがある
 - ▶ 時系列モデル：自身の過去のデータから将来のデータを予測する
時系列データ解析手法

VARモデル

- 過去のデータの値から未来のデータの値を予測するモデル
 - 多変量自己回帰(VAR)モデル（一期前のデータを用いる場合）

$$Y_{1t} = \Phi_{11}Y_{1,t-1} + \Phi_{12}Y_{2,t-1} + \epsilon_{1t}$$



遅延相関分析の問題点

- 遅延相関分析の利点が不明なのが現状
 - ▶ 時系列モデルの予測よりも精度が良ければ、遅延相関分析の有用性はある

目的とアプローチ

- 目的

- ▶ 遅延相関分析が時系列モデルよりも有用である場合があるかどうか調査し、有用性を明らかにすること

- アプローチ

- ▶ 時系列モデルの予測結果と比較する
 - 1.遅延相関分析の遅延係数に基づき、回帰モデルを構築
 - 2.回帰モデルとVARモデルのそれぞれの予測誤差で比較

回帰モデルの構築

- データ予測し，予測値を比較するため，遅延相関分析を回帰モデルに置換
- 遅延係数を考慮したデータセットを作成し，回帰モデルを構築
 - ▶ メトリクス関係の遅延係数が1の場合

時間 \	説明変数	目的変数
t_1	2	7
t_2	3	8
t_3	4	9
t_4	5	10
t_5	6	11

回帰式

$$y = ax + b$$

a : 回帰係数 b : y切片

比較方法

- テスト用データと各モデルの予測値を相対誤差で比較
 - ▶ 値が小さいほど予測精度が良い

$$MRE_t = \frac{|PredictedValue_t - MeasuredValue_t|}{MeasuredValue_t}$$

$$MMRE = \frac{\sum_{t=1}^{PredictionPeriod} MRE_t}{PredictionPeriod}$$

MRE_t : 時点tの相対誤差 $MMRE$: 予測期間内の平均相対誤差

$PredictedValue_t$: 時点tの予測値 $MeasuredValue_t$: 時点tの実測値

$PredictionPeriod$: 予測期間数

データセット

- Apache HTTP Serverプロジェクト
 - ▶ 2001年から2015年のデータからメトリクスを抽出

メトリクス	件数	例
ソースコード	48	コード行数
不具合修正履歴	32	修正されたバグ件数
メーリングリスト	10	MLに送信した件数

- ▶ 相関係数の絶対値が0.4以上のメトリクス関係
 - ✓ 349/8010件 → 予測・分析

実験結果

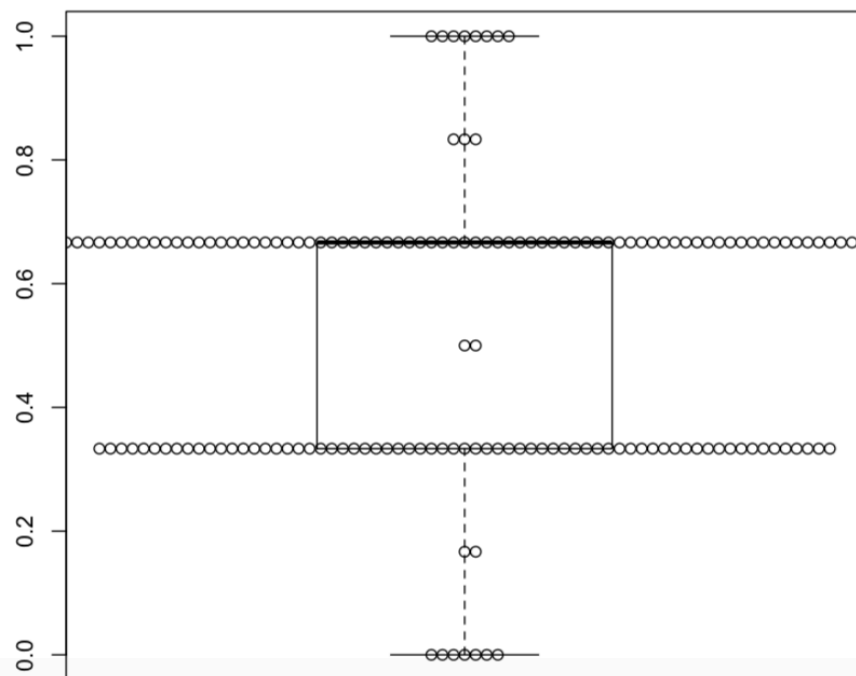
- 相関係数の絶対値が0.4以上の関係について各モデルで予測

	誤差が小さかった メトリクス関係数	平均相対誤差の 平均	平均相対誤差の 中央値	最大 平均相対誤差
VARモデル	191	0.3997	0.3212	1.8714
回帰モデル	26	0.7997	0.6155	4.8651

※132件は相対誤差が無限大となったので対象外とする

考察

- メトリクスの特徴による相対誤差の変動
 - ▶ 遅延相関分析に基づく回帰モデルの方が精度が良かった
26件について調査
 - 10件がプリプロセッサ行数に関係している
 - プリプロセッサ：ソースに対して行われる前処理を行うプログラム



値の偏りが大きいデータ

回帰モデルの方が誤差の少ない
予測ができている

プリプロセッサ行数の箱ひげ&蜂群図

考察

- 予測時期による相対誤差の変動

- ▶ 遅延相関分析に基づく回帰モデルの方が精度が良かった中から

20件調査

- 約18ヶ月先でVARモデルよりも小さくなった

- ✓ 同じく値の偏りが大きいデータであった

値の偏りが大きい, 2, 3年以上先の予測をする際は遅延相関分析に基づく回帰モデルの方が**精度の高い予測**をすることができる

まとめ

- VARモデルと比較することで、時系列データを分析する際の遅延相関分析の有用性を示した
 - ▶ 対象217件のうち、26件はVARモデルよりも予測精度が高い
- データの特徴や予測期間により使い分けができる
 - ▶ データの偏りが大きい、長期間予測：**遅延相関分析**
 - ▶ 短期間予測：**VARモデル**

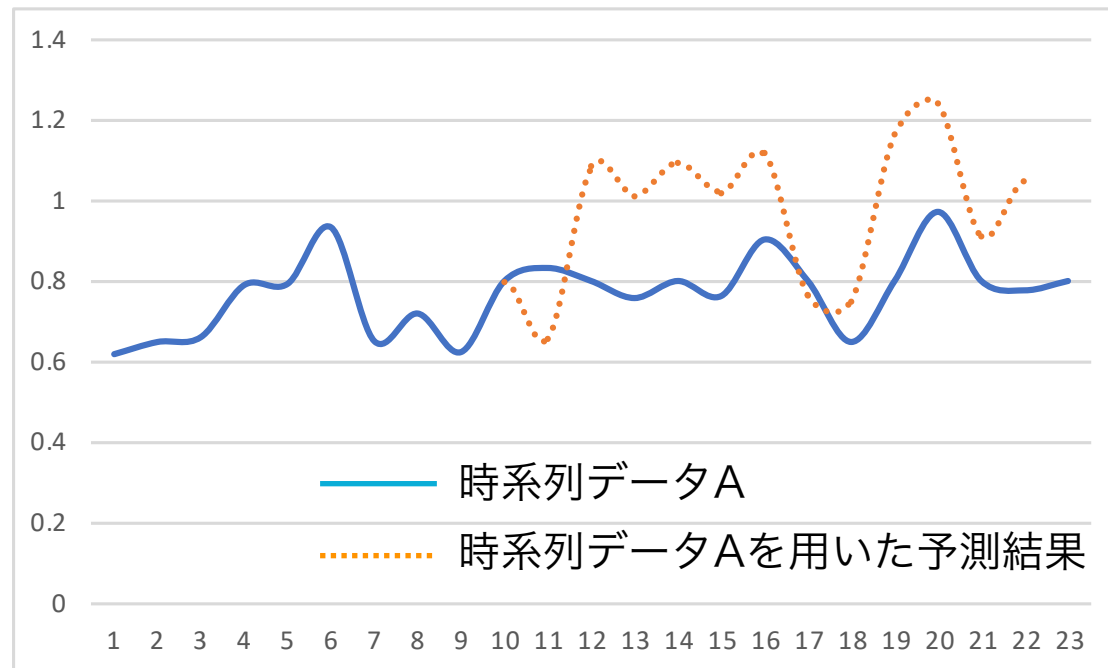
今後の課題1：メトリクス間の因果関係

- グレンジャー因果性を用いて、各メトリクス関係の因果関係を調査
 - ▶ 本研究の対象とした、相関係数が0.4以上であるメトリクス関係349件の因果関係を各方向から調査

グレンジャー因果性の定義

・グレンジャー因果性の定義

- ▶ 現在と過去の x の値を用いて予測するよりも、現在と過去の x と y を用いて予測したほうが精度が良い場合、 y から x へのグレンジャー因果性が存在する

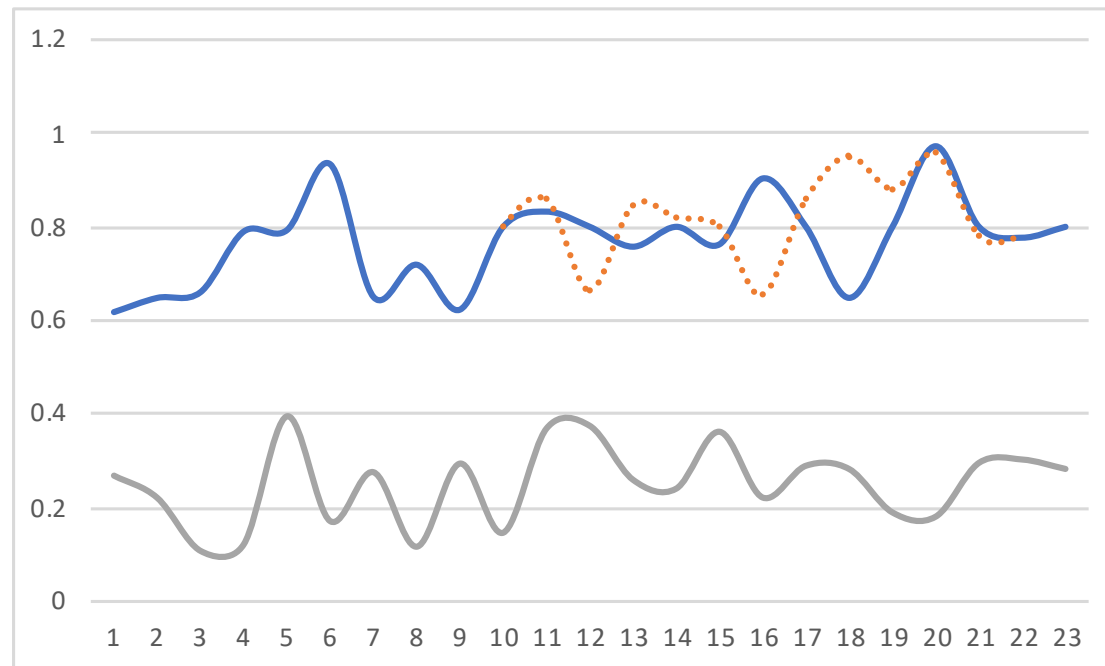


- ✓ 時系列データAのみ用いた場合
 - ・ 誤差が大きい結果となる

グレンジャー因果性の定義

・グレンジャー因果性の定義

- ▶ 現在と過去のxの値を用いて予測するよりも、現在と過去のxとyを用いて予測したほうが精度が良い場合、yからxへのグレンジャー因果性が存在する



- ✓ 時系列データA・Bを用いた場合
 - ・ 比較的誤差は小さくなる



BからAへのグレンジャー因果性が存在

グレンジャー因果性の具体例

- ・サンマの漁獲量を予測する
 - ▶ サンマの漁獲量を用いて予測：結果A
 - ▶ サンマの漁獲量とプランクトンの量を用いて予測：結果B
- 結果Bの誤差が小さい場合
 - ✓ プランクトンからサンマにかけて因果関係が**存在する**
- 結果Bの誤差が大きい場合
 - ✓ プランクトンからサンマにかけて因果関係は**存在しない**

グレンジャー因果性の応用

- ・グレンジャー因果性の応用

- ▶メトリクス間関係の因果関係がわかる

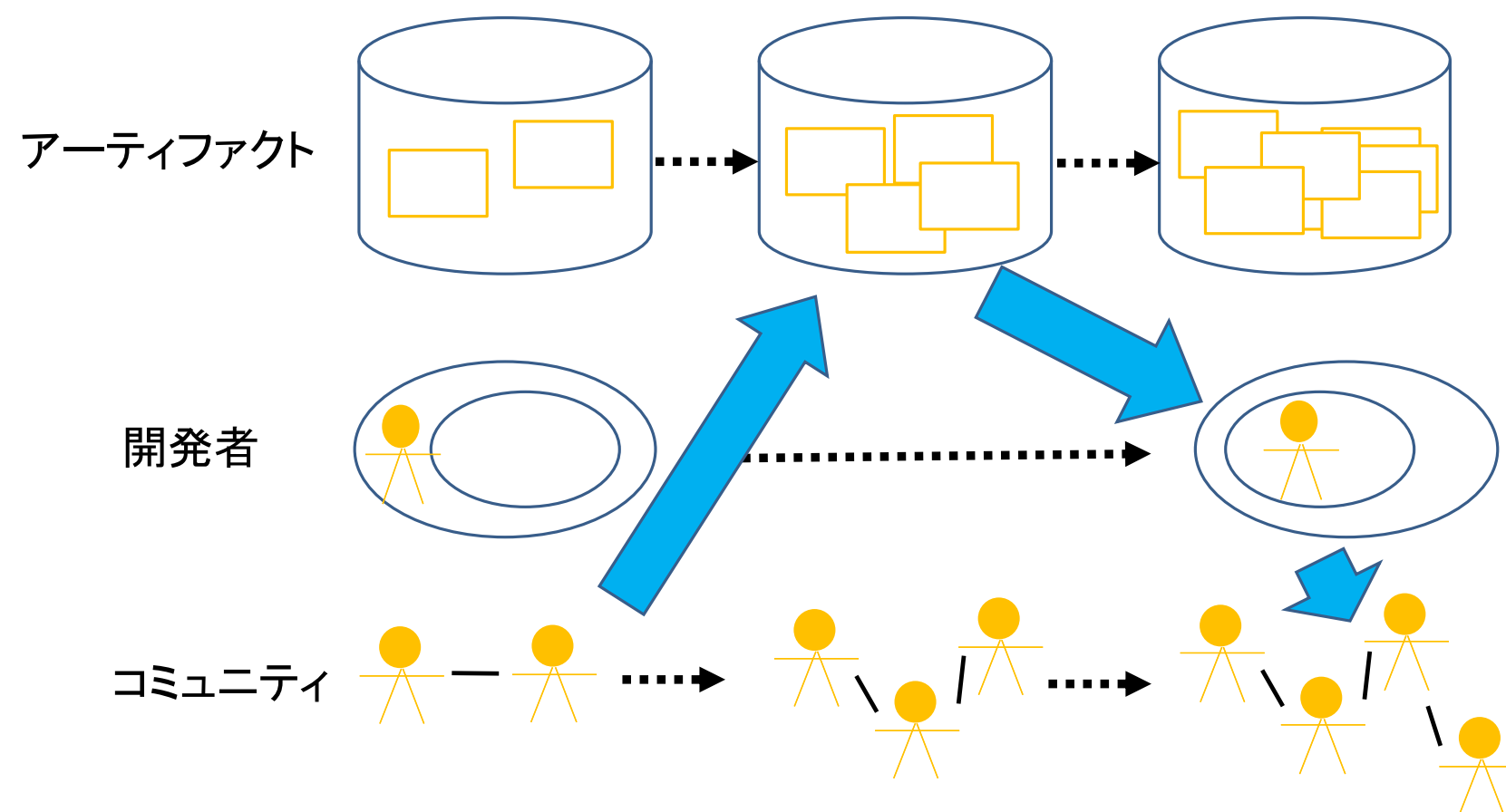
- ・各メトリクスが、どのメトリクスに対して影響を及ぼすのか、直感的に理解することができる

今後の課題2：遅延相関分析の改良

- 遅延相関分析の変数を増やす
 - ▶ 今現在は説明変数と目的変数の二つのみを扱う
 - ▶ 遅延相関分析は三つの要素から構想している[3]
 - アーティファクト
 - 開発者
 - コミュニティ

遅延相関分析の構想

- OSSはコミュニティとともに成長するもの^[4]
 - 遅延相関分析は三つの共進化を測定したい



[4] Ye,Y., Nakakoji,K., Yamamoto,Y. ,and Kishida,K. : The Co-Evolution of Systems and Open Source Software Development,Free/Open Source Software Development Koch,S.(Ed.),Idea Group Publishing, Hershey, PA chapter 3, pp.59-82 (2004)