

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

蘆田 誠人

背景

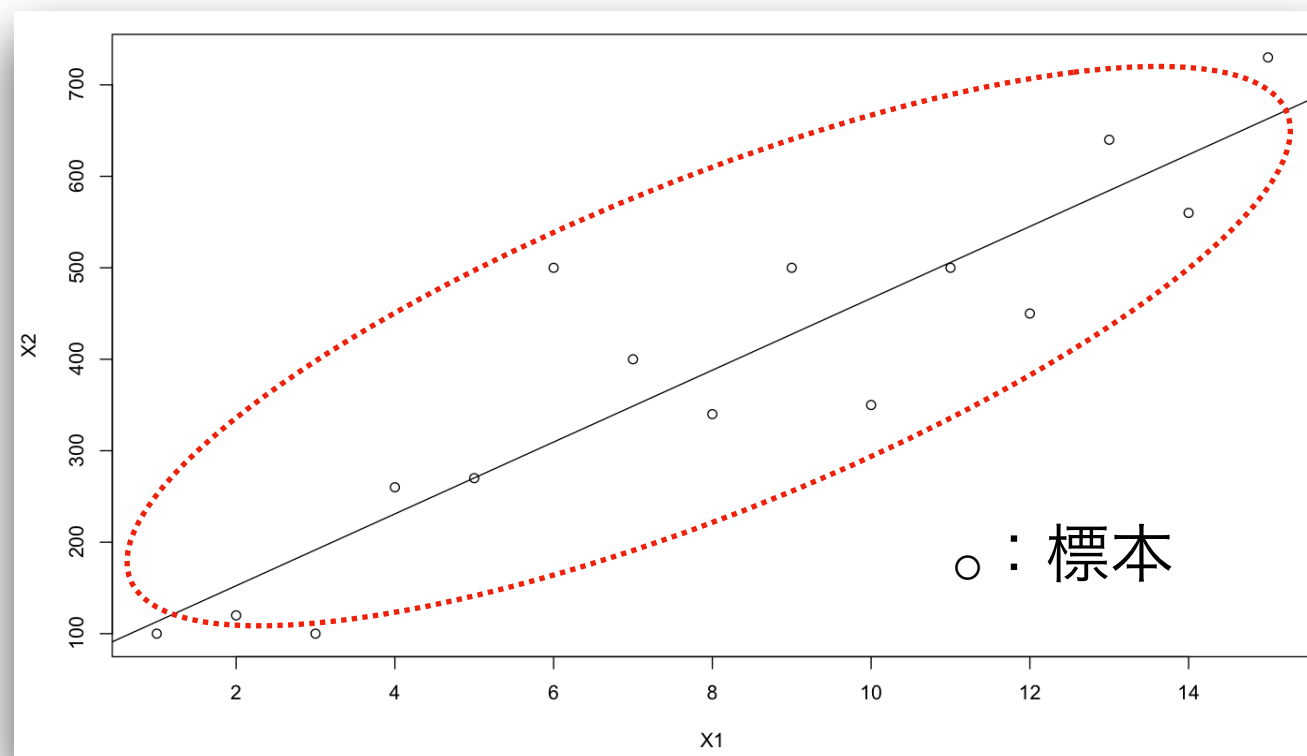
- ソフトウェア開発
 - ▶ プロジェクト内の作業を定量的に管理することが重要[1]
 - ソフトウェア開発データの計測, 分析が必要

[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/teiryou.html>

背景

- 相関分析

- ▶ 二変数間に直線的な関係があるかどうか
 - 同時点の関係の強さを相関係数で表す

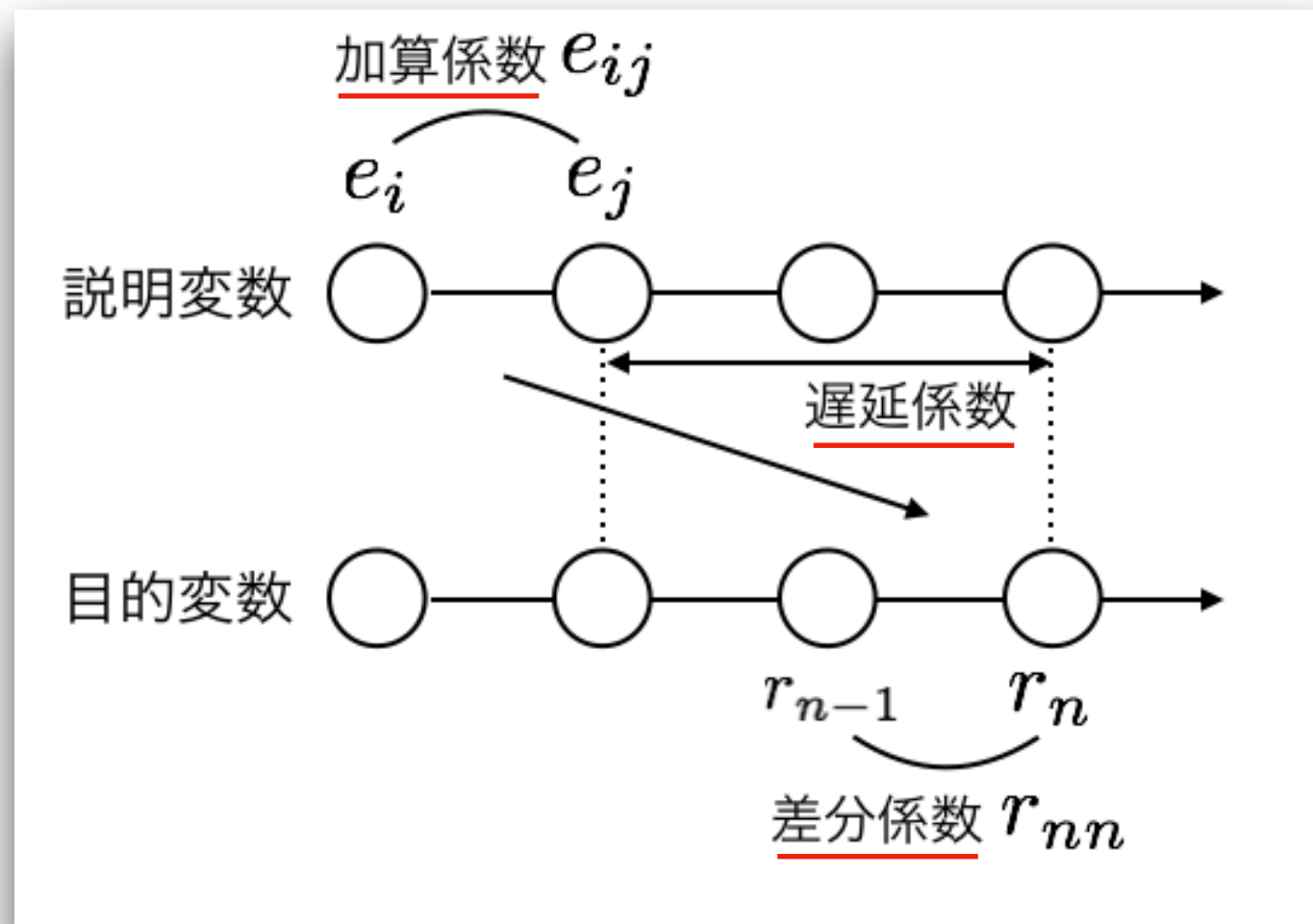


背景

- 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
 - ▶ そのような関係を相関分析で捉えるのは難しい

背景

- 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



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
 - ▶ 生活習慣と健康状態の関係が時間を遅延させたことにより抽出できた
- 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, 6 Vol. 56, No. 1, pp. 59-71, 1 2015.

時系列データに対する遅延相関分析

- 時系列データを対象とした分析はされていない
 - ▶ 時系列データに対して、遅延相関分析の有用性は明らかになっていない

目的とアプローチ

- 目的

- ▶ 時系列データを分析する際の遅延相関分析の有用性を評価する

- アプローチ

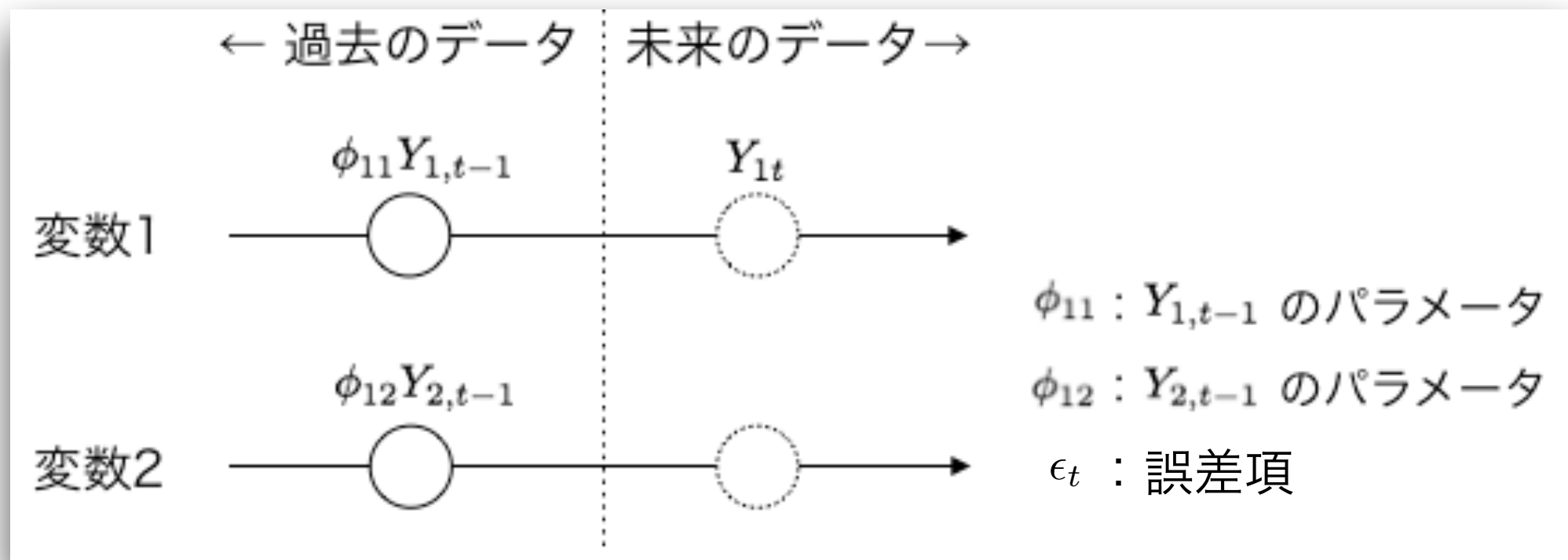
- ▶ 時系列モデルと比較する

- 1.遅延相関分析の遅延係数を用いて，回帰モデルを構築
- 2.回帰モデルと時系列モデルのそれぞれの予測誤差で比較

時系列モデル

- 過去のデータから現在のデータを予測するモデル
 - 多変量自己回帰(VAR)モデル

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



定常状態

- 分散が無限大のデータを用いてモデル適用する
 - ▶ 予測値に大きな誤差が生まれる
 - ✓ 時系列モデルを用いる際、定常な状態を保たなければならない

定常な状態のデータ例

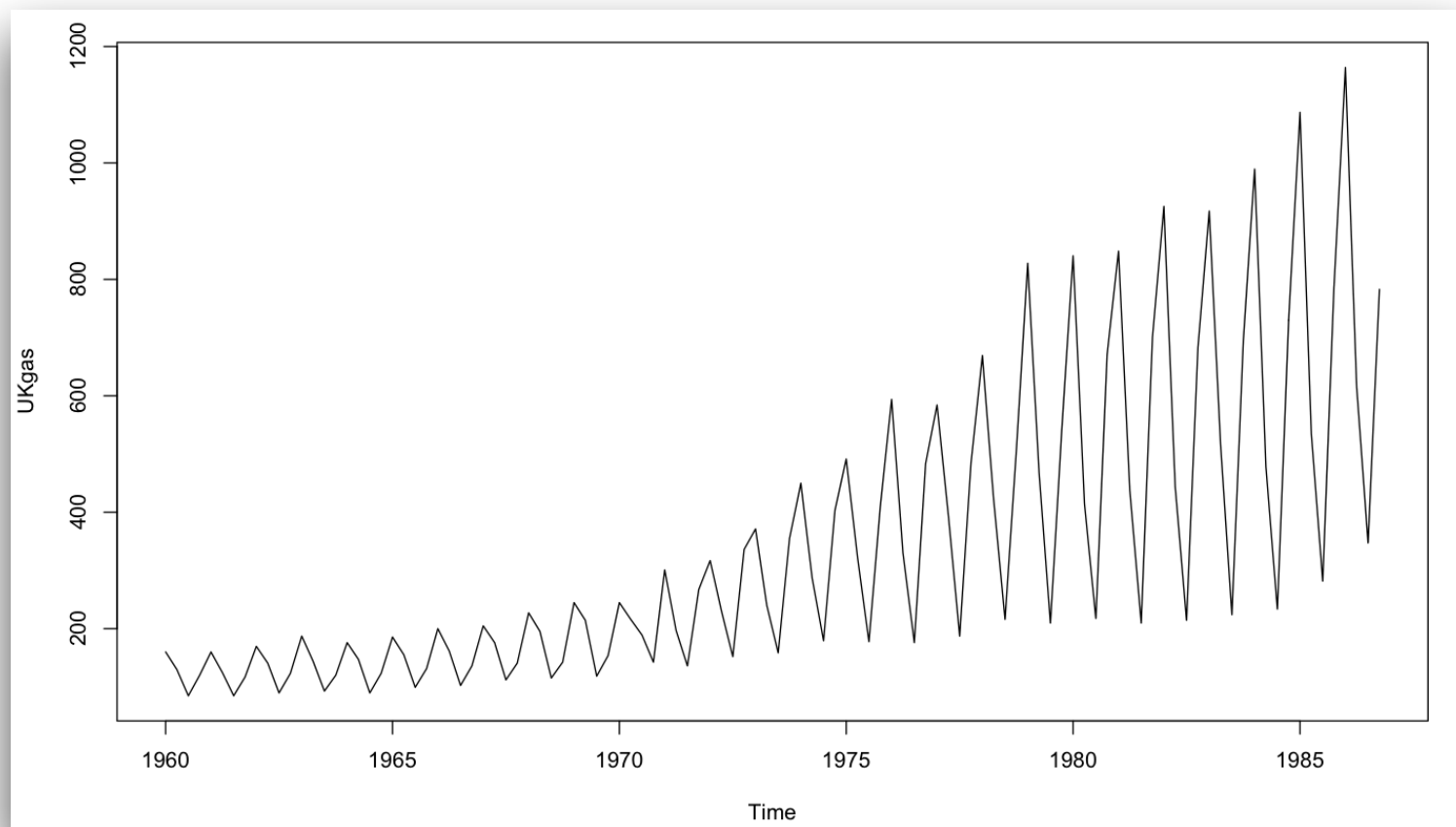
- ✓ サイコロの出目：周囲の条件が変更されても確率は不変

非定常な状態のデータ例

- ✓ 日本の外気温：季節によって朝夕の気温差は変動

定常化

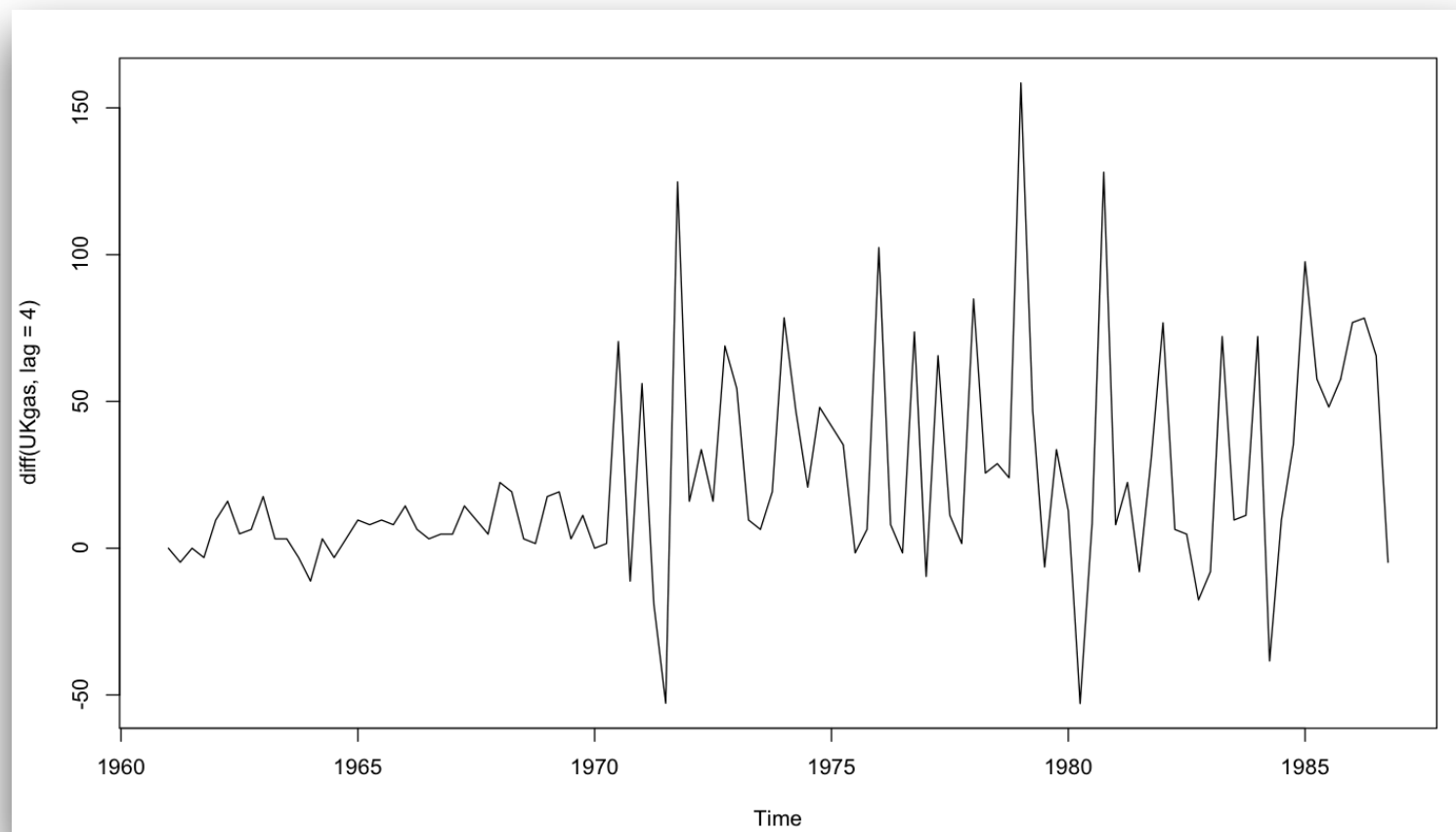
- ・ イギリスのガス消費量(1960年~1986年)
 - ▶ 生データ (縦軸：ガス消費量(サーム), 横軸：年)



- ・ 上昇トレンド
 - ・ 分散が大きくなる
- 定常条件を満たしていない**

定常化

- 上昇トレンドの除去
 - ▶ 差分を取ることでより定常な状態にする



- トレンド無し
 - 分散はほぼ一定
- 定常条件を満たした

遅延相関分析

- 三つの係数により構成される

- ▶ 加算係数(1~12)

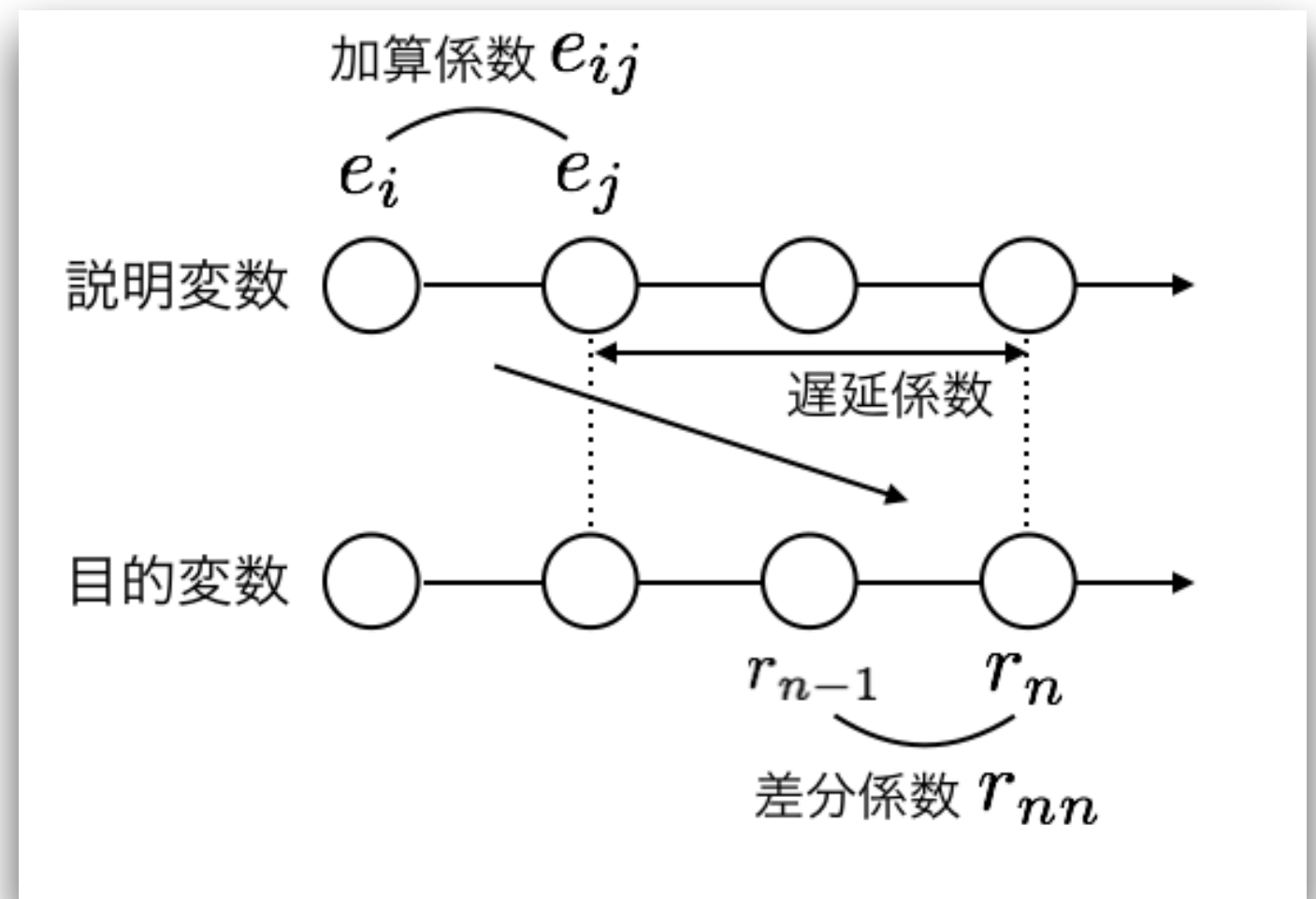
- 説明変数の累積値

- ▶ 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

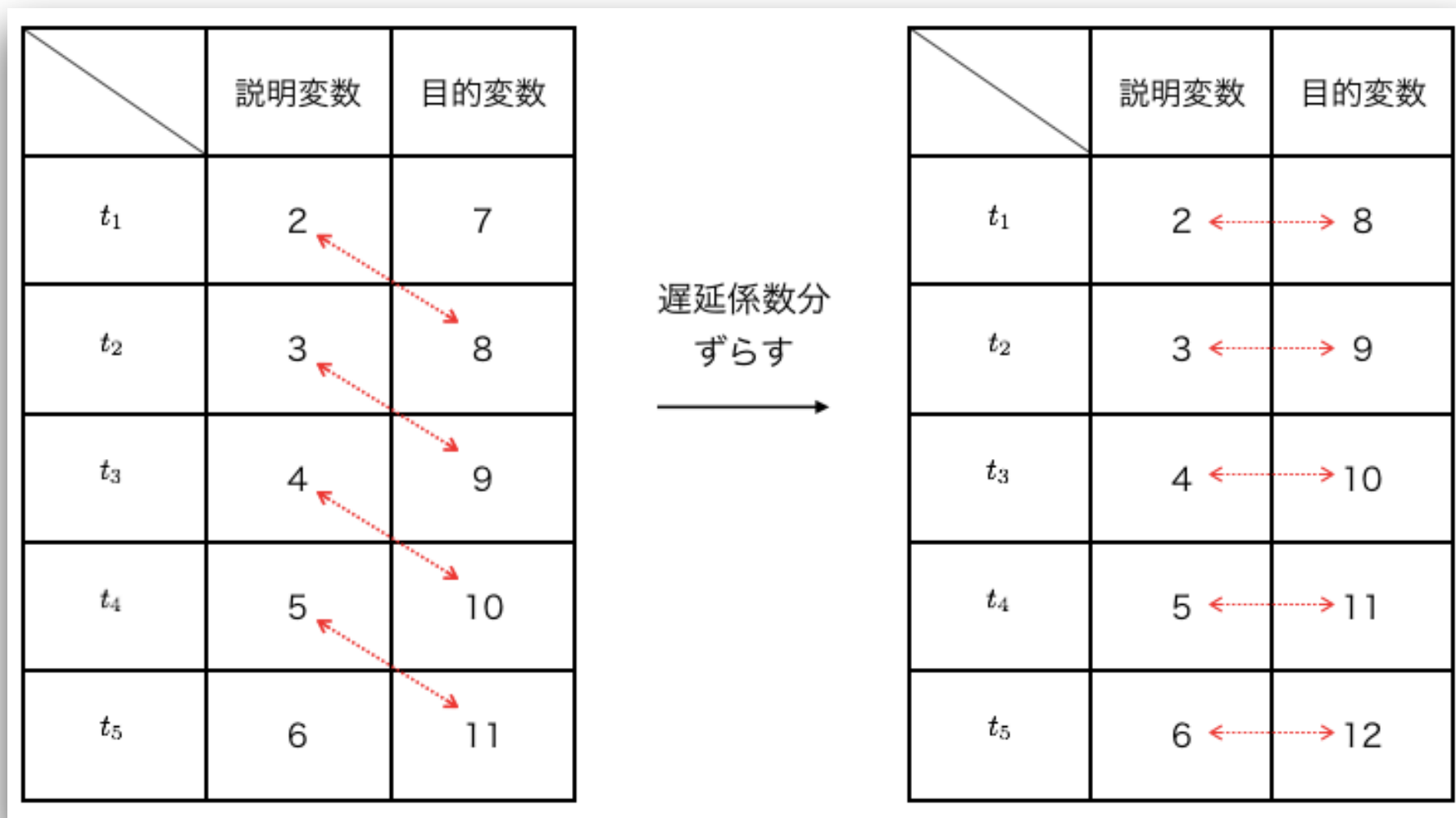
▶ 三つの係数のセットで、相関が一番高い係数のセットを記憶

第一メトリクス	第二メトリクス	相関係数	加算	差分	遅延
commitNum	countLine	0.45	2	3	1
countLine	totalFixers	0.23	10	2	5
maxCyclomatic	totalFixed	-0.13	8	2	2
totalReplies	totalReporters	0.39	4	1	10

実験方法

- 回帰モデルの構築

- ▶ 目的変数を遅延係数分ずらしたデータで回帰モデルを構築



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
ソースコード	48
不具合修正履歴	32
メーリングリスト	10

比較方法

- データセットを分割
 - ▶ 学習用データセット：2001年1月～2010年12月
 - ▶ テスト用データセット：2011年1月～2015年12月
- テスト用データと各モデルの予測値を相対誤差で比較

$$MRE = \frac{|PredictedValue - MeasuredValue|}{MeasuredValue}$$

MRE ：相対誤差 $PredictedValue$ ：予測値 $MeasuredValue$ ：実測値

実験結果

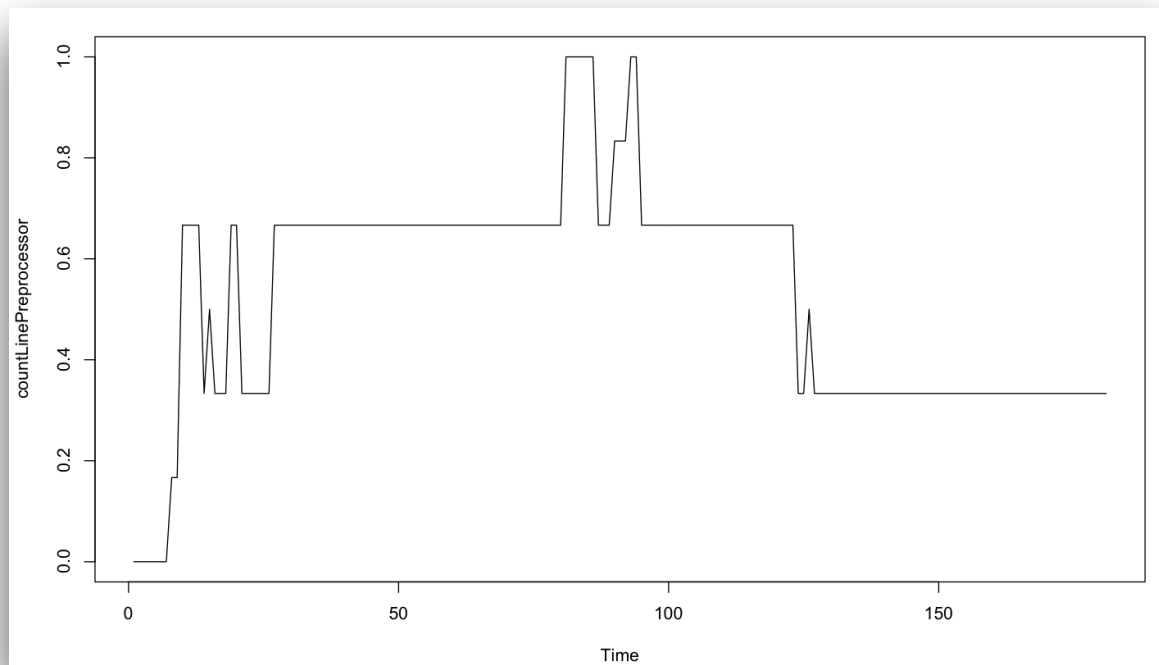
- 相関係数の絶対値が0.4以上の関係について各モデルで予測
 - ▶ 349件/8010件が抽出された

	誤差が小さかった メトリクス関係数	相対誤差平均	誤差の中央値	最大相対誤差
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
 - ▶ 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



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

考察

- 予測時期による相対誤差の変動

- ▶ 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、かつ2, 3年以上の予測をする際は回帰モデルの方が予測誤差の小さ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