

IBM Research

異常検知についての最近の話題: イベント 系列からの因果発見問題を中心に

IBM ワトソン研究所 井手剛

Tsuyoshi Idé, T.J. Watson Research Center, IBM Research

Slides → https://ide-research.net/papers/2021_Idé_Causal_discovery_from_events.pdf

自己紹介 (詳しくは <http://ide-research.net>)

- IBMワトソン研究所の研究員
 - “Trusted AI” というグループに所属
- 専門分野
 - 機械学習の応用研究。特に、異常検知など(本→)
- 履歴
 - 博士論文は物性理論(強相関電子系の光物性)
 - ✓ 東大大学院(物理学専攻)
 - ← 東北大学工学部(機械工学科)
 - ← 苫小牧工業高等専門学校(機械工学科)
 - 2000年 IBM東京基礎研究所 (Tokyo, Japan)
 - 2013年 IBMワトソン研究所 (New York, USA)



Agenda

■ 基礎編

- 機械学習の問題設定
- 異常検知の基礎

■ 最近の話題

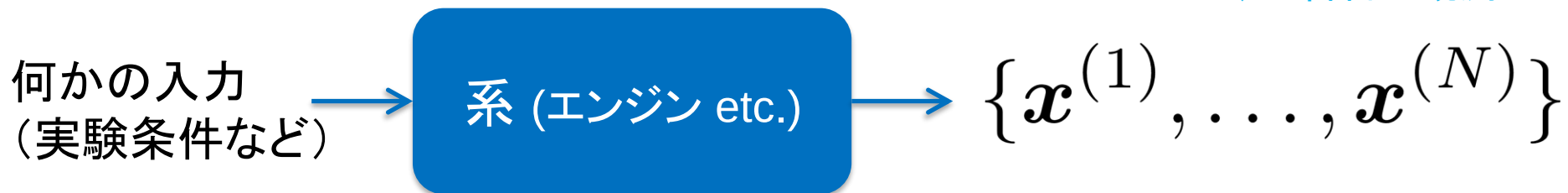
- 異常検知における解釈可能性 [Idé et al., AAAI 21]
- 確率的イベントに対する因果分析 [Idé et al., NeurIPS 21]

機械学習の基本的な問題設定

- データの要約を確率分布の形で作り、それにより、未来への予言能力を得ること(「汎化能力」)

- 観測を繰り返してN個の標本(ベクトル)を得る

$x^{(1)}$ は、「1番目に観測したベクトル」



- 系には何かの内部構造があり、その内部構造は何かのパラメーターで表現できると信ずる

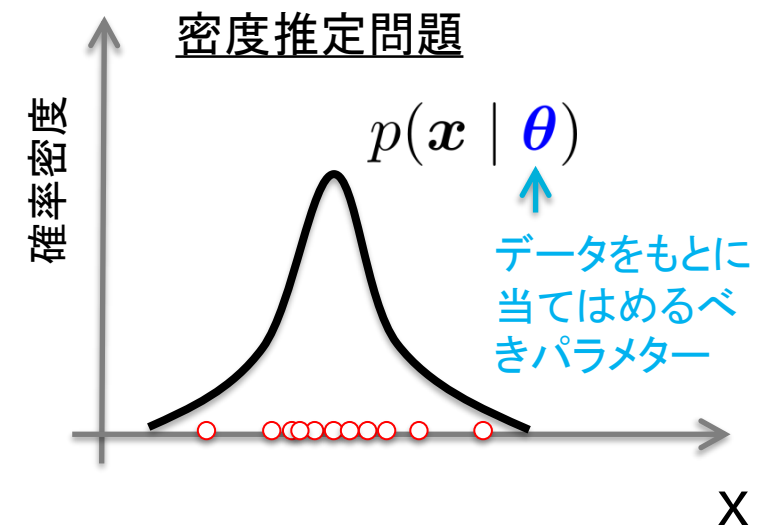
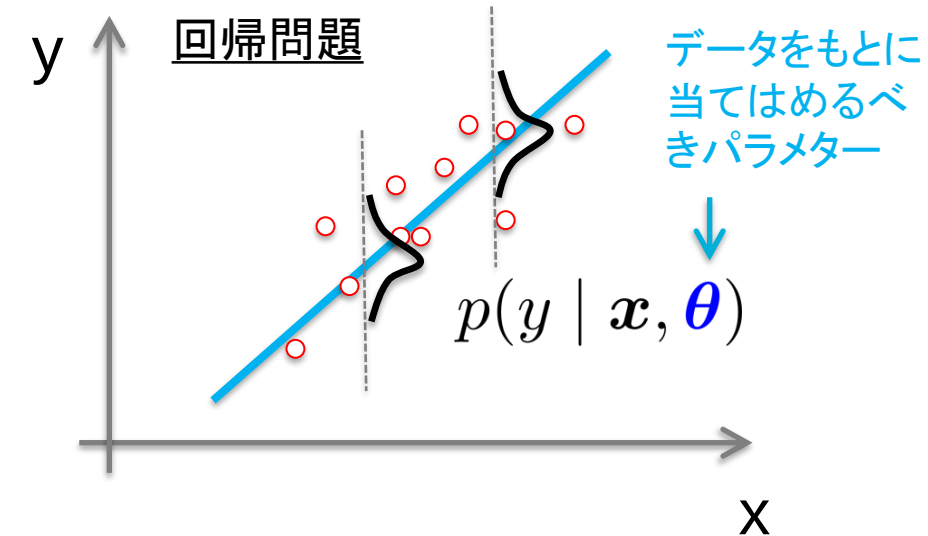
- 「N+1回目の(未知)のベクトルは、どういう値をとるだろうか？」

- 未来への予言能力を追求するという意味で、物理学と類似

機械学習の問題の多くは、回帰、分類、密度推定に大別できる。

- 回帰問題(「教師つき学習」とも)
 - 入力 x と出力 y のペアが多数与えられる
 - y は実数。それらのペアは独立と仮定
 - i.i.d.なデータを想定した手法(大半)
- 分類問題(「教師つき学習」とも)
 - 入力 x と出力 y のペアが多数与えられる
 - y はカテゴリカル変数
 - i.i.d.=identically and independently distributed
 - 観測値がその都度独立に(履歴を忘れて)系から出てくると想定。
- 密度推定(「教師なし学習」とも)
 - 出力 x のみが多数与えられる
 - データから、 x の分布を学習
- 時系列推定(システム同定)
 - 観測値
 - 未来の観測値の分布を学習する。

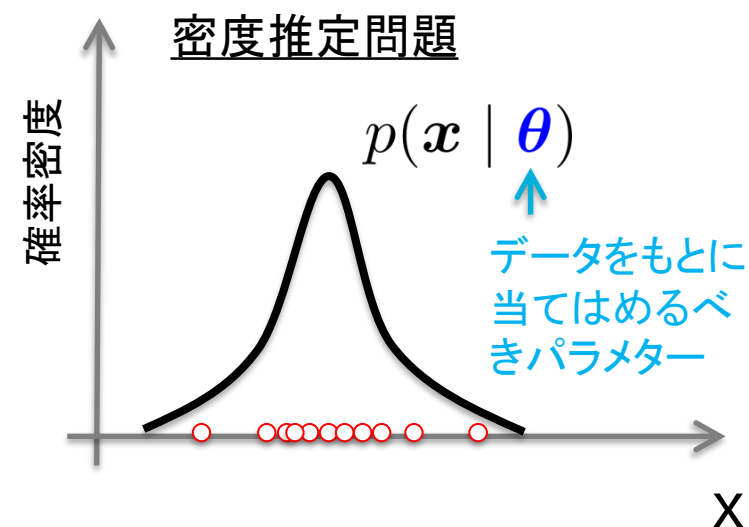
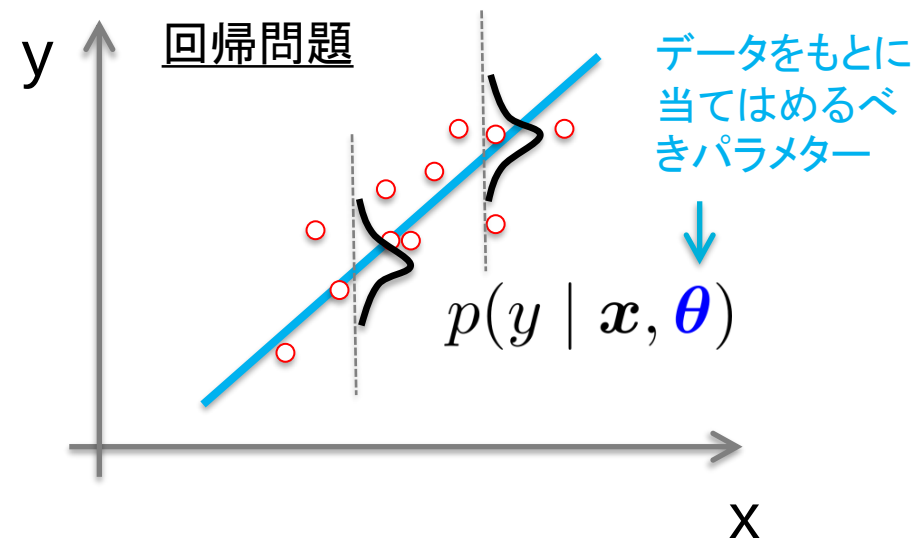
独立標本の仮定を置かない手法



機械学習の問題の多くは、回帰、分類、密度推定に大別できる。 異常検知で重要なのは密度推定。

- 回帰問題（「教師つき学習」とも）
 - 入力 x と出力 y のペアが多数与えられる
 - ✓ y は、実数。それらのペアは独立と仮定
 - データから、 x を与えたときの y の条件付き分布を学習
- 分類問題（「教師つき学習」とも）
 - 入力 x と出力 y のペアが多数与えられる
 - ✓ y は、カテゴリカル変数
 - データから、 x を与えたときの y の条件付き分布を学習
- 密度推定（「教師なし学習」とも）
 - 出力 x のみが多数与えられる
 - データから、 x の分布を学習
- 時系列推定（システム同定）
 - 観測値が時系列的に与えられる。
 - 未来の観測値の分布を学習する。

- 密度推定の種類
- いわゆる強化学習も含む
- 独立標本という仮定を置かないため、難度が高い



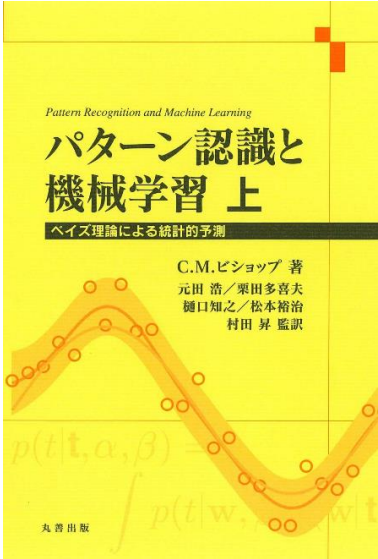
ベイズ公式 = 機械学習の基礎方程式 $\equiv$ 自由エネルギー最小原理

- ベイズ公式: $\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}$ を観測した後の未知パラメーター θ の事後分布
 - $q(\theta | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) \propto P_{\text{prior}}(\theta) \times \prod_{n=1}^N P_{\text{obs.}}(\mathbf{x}^{(n)} | \theta)$
 - ✓ $P_{\text{prior}}(\theta)$: データがないときに想定する θ の分布 (事前分布)
 - ✓ $P_{\text{obs.}}(\mathbf{x} | \theta)$: $\mathbf{x}$ を観測する時の測定ノイズの分布 (多くの場合正規分布)
- $q(\theta | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$ を明示的に求めるのはむづかしいので、事後確率を最大にする θ の値を代表の値 (推定値) として採用することも多い
 - $L(\theta) \equiv \ln q(\theta | \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)})$ において、最大値を与える θ を求める。
 - これをMAP (Maximum a posteriori) 推定法という
- 負の対数尤度 $-L$ を自由エネルギーと見なせば、自由エネルギー最小原理と形式上同じ
 - だが、究極には予測分布 (predictive distribution) を求めることが機械学習のゴール

機械学習 ≡ データの熱力学

- 自由エネルギーから、熱平衡状態を表現するモデルパラメーターを学習する問題。
 - 「熱平衡」とは、モデルがデータに一番よく「なじんだ」状態
- あるいは、ベイズ公式を運動方程式のようなものだと思ってもよい。
 - 力学の問題を解くときは、運動方程式をまず立てる。その過程で未知量、観測量すべき量、解を得るための技術的問題点などが頭の中で整理される
 - 機械学習を使ったモデリングでも、ベイズ公式を基礎方程式と見て、同様の手順に従う(べき)
- 機械学習の多くの手法はベイズ理論の近似として統一的に理解できる

手法の名前	近似の方針
最尤推定	事前分布を定数と仮定。事後分布をデルタ関数で近似。
リッジ回帰	事前分布は正規分布。事後分布をデルタ関数で近似。
Lasso回帰	事前分布はラプラス分布。事後分布をデルタ関数で近似。
変分推定法	事前分布は共役分布に選ぶ。事後分布を変数ごとの積の形に仮定。



ビショップの教科書はベイズ原理を普遍的原理として扱っているので、自然科学者には読みやすいはず

Agenda

■ 基礎編

- 機械学習の問題設定
- 異常検知の基礎

■ 最近の話題

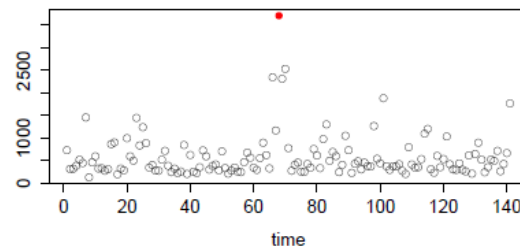
- 異常検知における解釈可能性 [Idé et al., AAAI 21]
- 確率的イベントに対する因果分析 [Idé et al., NeurIPS 21]

異常検知の代表的なタスクは、外れ値検出と変化点検出。 “one-size-fits-all”な手法はない

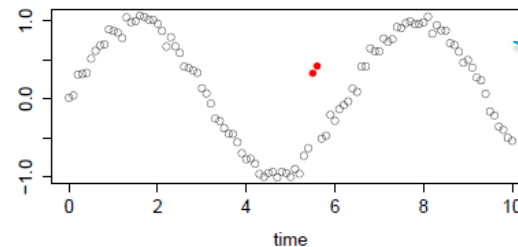
- 異常の普遍的な定義はない。しいて言えば「正常じゃないもの」
 - “Happy families are all alike; every unhappy family is unhappy in its own way.”
- 『アンナ・カレーニナ』, トルストイ

異常の例

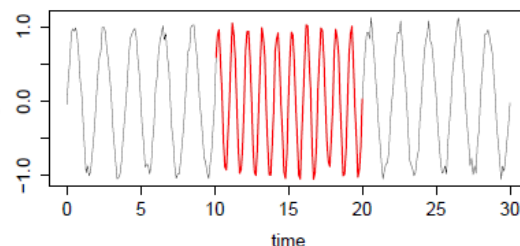
外れ値 (独立とみなせるデータ)



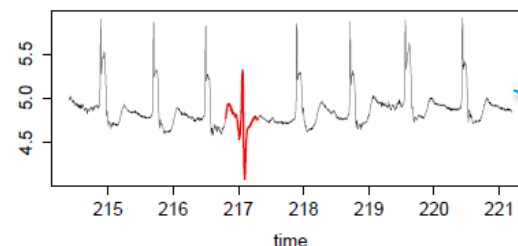
外れ値 (時系列相関の強いデータの強いデータ)



変化点



異常部位



井手、杉山、異常検知と変化検知、講談社、2015.

異常を定義するためには、正常状態におけるデータのばらつきの情報が必須。

■ 例:「平均から外れているので外れ値」

- データの背後に分布を仮定して、「平均値」という統計量に着目している
- 統計量が今観測したデータをどれだけ説明するかに注目している
- あるしきい値を置いて、それを超えたものを異常と判定している

分布の例: 多次元正規分布

- μ は平均ベクトル
- Σ は共分散行列

$$\mathcal{N}(x \mid \mu, \Sigma) = \frac{|\Sigma|^{-\frac{1}{2}}}{(2\pi)^{\frac{M}{2}}} \exp \left\{ -\frac{1}{2} (x - \mu)^{\top} \Sigma^{-1} (x - \mu) \right\}$$

- 観測値は常に揺らいでいるので、単に期待値からのずれだけを見ても何も言えない。ずれが過去の経験に比べて有意に大きいかが問題。

異常検知モデリングにおける3つの問題

■ 分布推定問題

- 正常状態の確率分布としてどういうモデルを使うべきか
- そのモデルに含まれる未知パラメータをどうデータから求めるか

問題ごとに
工夫が必要

■ 異常度設計問題

- 異常度を確率分布からどのように計算するか
- 異常診断に直結するような情報をどう引き出すか

やり方はほぼ
決まっている

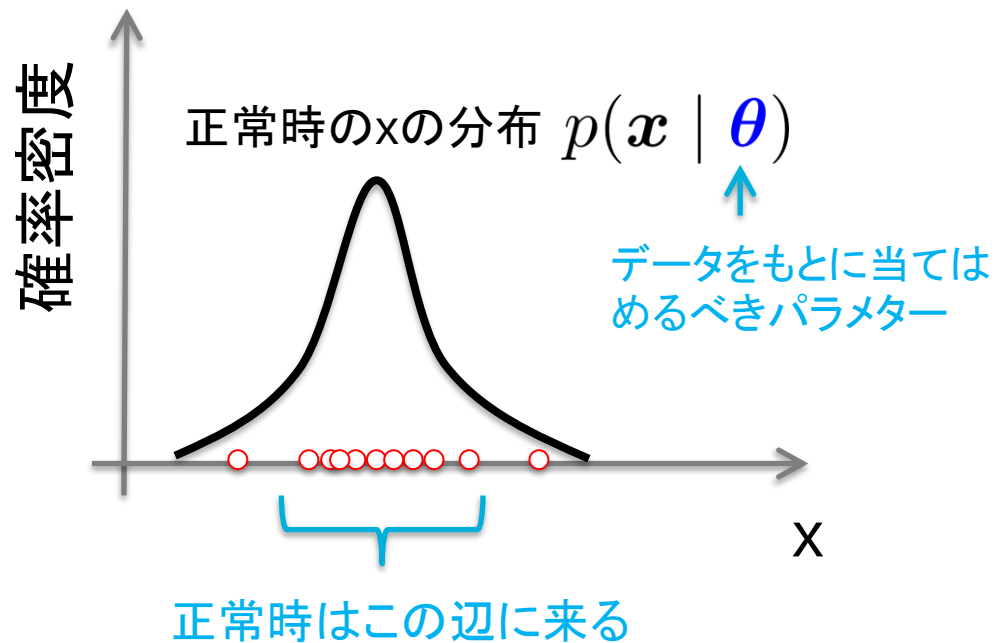
■ しきい値設定問題

- 実数値の異常度から、異常・正常の判断をどのように行うか
- 異常標本を使ってどういう指標を使って性能評価を行うか

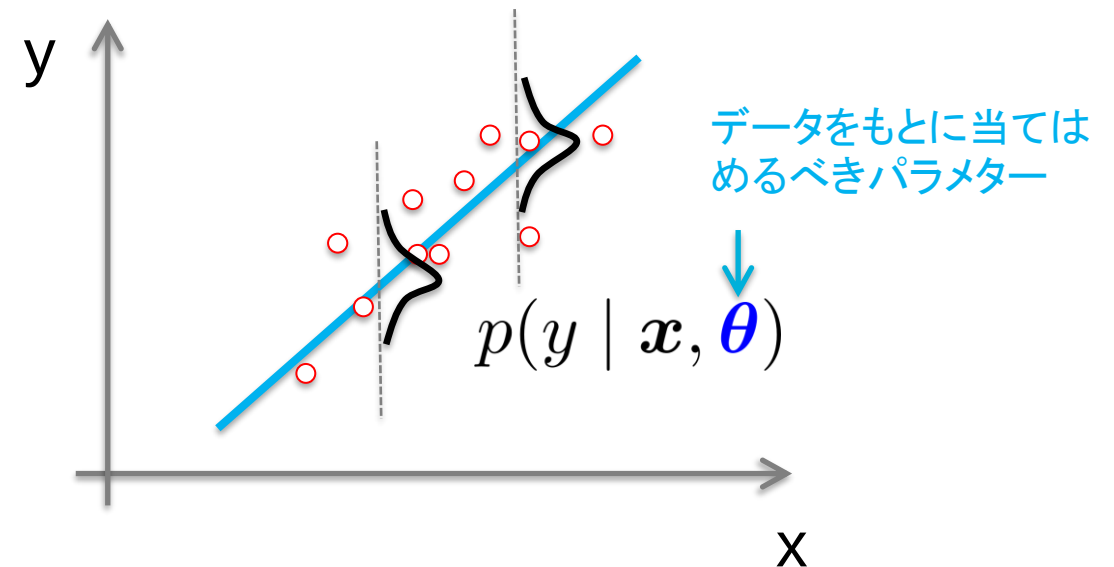
分布推定問題は、観測値の確率分布をデータから求める問題

- 最も重要だが、データが高次元だと一般に簡単でない
- 適切なモデルの選択のためには幅広い知識が必要

観測値が x のみの場合 (密度推定)



x と y が観測される場合 (回帰)



異常度設計問題において、異常度の基本は $-\ln p$

■ x のみが観測される場合 (密度推定)

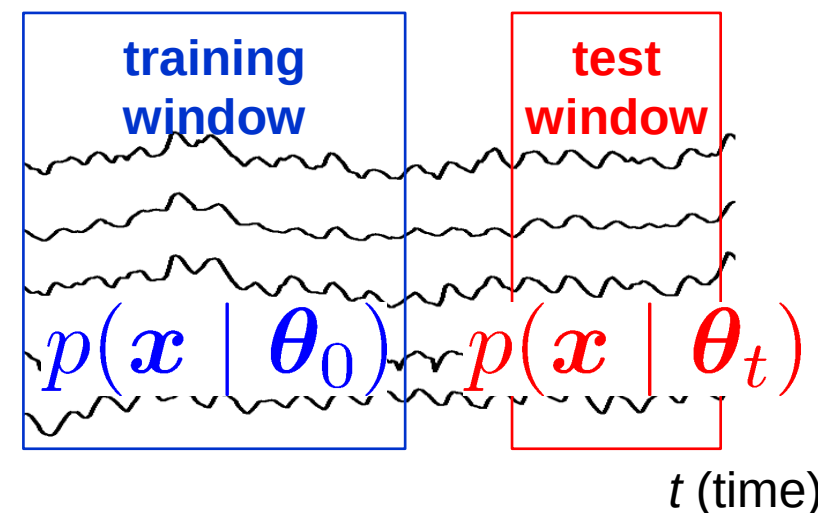
- $\text{score}(\mathbf{x}) \propto -\ln p(\mathbf{x} | \mathcal{D})$
 - ✓ $\mathcal{D}$ は観測データ。「データから決めた分布」という意味。以下同じ
 - ✓ 正規分布の場合、マハラノビス距離と等価 (後述)

■ x と y の組が観測される場合 (回帰、分類)

- $\text{score}(\mathbf{x}, y) \propto -\ln p(y | \mathbf{x}, \mathcal{D})$
 - ✓ 正規分布を使った線形回帰の場合、2乗誤差と等価 (後述)

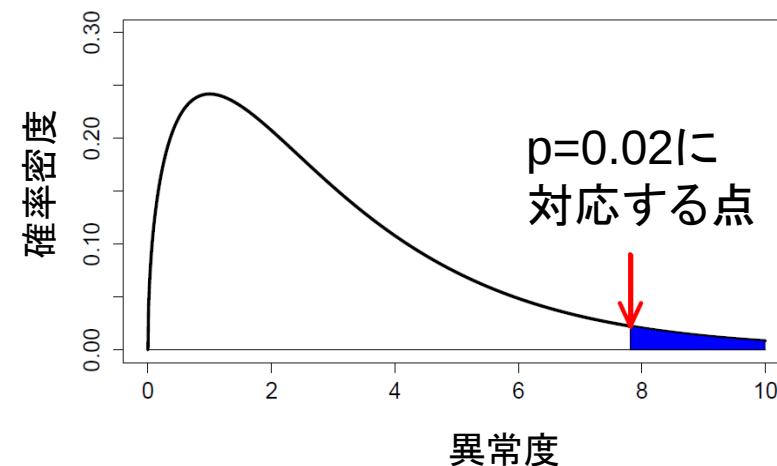
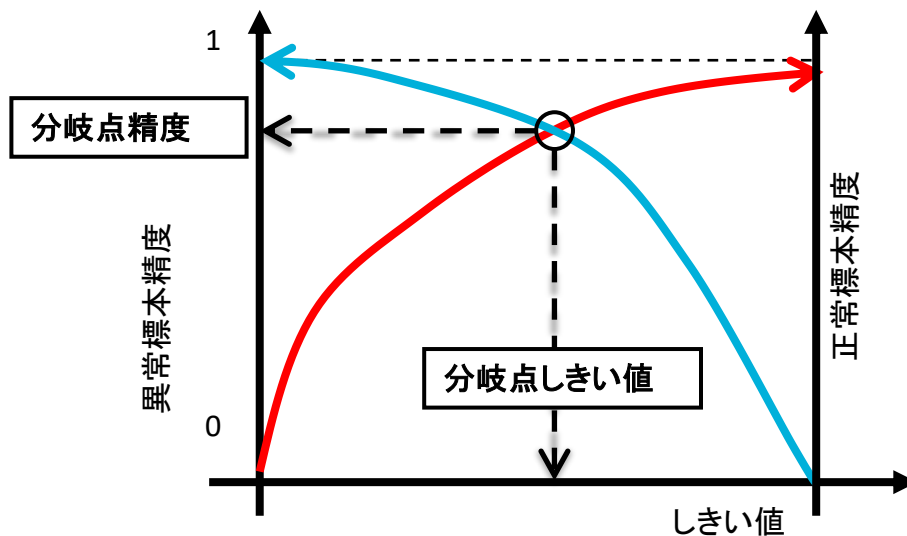
■ 変化度を知りたい場合

- $\text{score} \propto -\ln \frac{p(\mathbf{x} | \theta_0)}{p(\mathbf{x} | \theta_t)}$
 - ✓ いわゆる対数尤度比。最適性が保証されている
(→ 井手、杉山、異常検知と変化検知)



しきい値の設定にはだいたい2つのやり方がある

- 異常標本の数が多い場合
 - テストデータに対し異常度(スカラー)を計算してみて、競合精度プロット(contrastive accuracy plot)を描く
 - 分岐点精度やF値を最大にするようにしきい値を選ぶ
- 異常標本がほとんどない場合
 - テストデータに対し異常度を計算してみる。うまく異常度が設計されていれば、左側に偏った分布になっているはずなので、それにガンマ分布(カイ2乗分布)を当てはめる
 - 当てはめた分布の $p=0.02$ などに対応する値をしきい値とする
- 統計学の本にある理論的な(漸近)分布の大半は実測に合わない。そのままでは使うのではなく、計算した異常度に当てはめ直す、などの工夫が必要
 - 例: ホテリングT2統計量におけるF分布の自由度



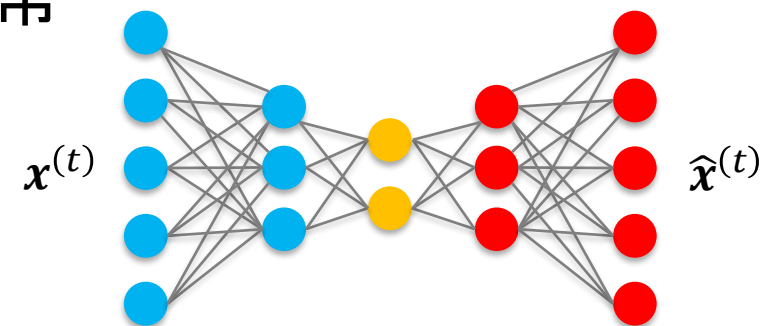
異常検知でありがちな誤解

- 「異常検知は教師なし学習なので、異常標本は不要」
 - 正常標本で密度推定するのはその通りだが、異常標本がないと閾値の評価ができない
- 「異常検知は分類問題なので、その辺の分類器を使えば楽勝」
 - クラス間バランスが圧倒的に違うので、二値分類としての扱いは失敗することが多い
- 「予測値を計算できれば異常検知は完了」
 - 予測値だけを計算できても、統計的有意性が評価できない。値(期待値)に加えて分布(予測分布)を求める必要がある
- 「深層学習は古典モデルに圧勝している」
 - 画像や言語と違い、ノイジーなセンサーデータでは、何も考えずにLSTMとかを使ってもまともな結果が出ないことが多い。人手をかけて慎重に特徴量を設計する必要がある
 - 多くの深層学習器は確率分布を与えないので異常検知に使いにくいことが多い
 - ブラックボックスになりがちなため、既存のエンジニアの知見を活かせないことが多い

「予測値を計算できれば異常検知は完了」ではない。 再構成誤差 (reconstruction error) による異常検知について

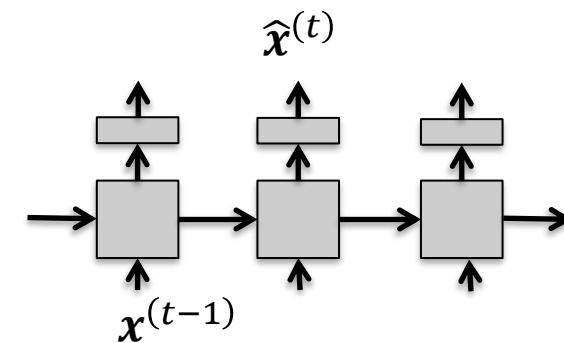
■ 最近の深層学習系の異常検知モデルの典型的な異常度モデル $a(x^{(t)}) = ||\hat{x}^{(t)} - x^{(t)}||^2$

- $x^{(t)}$: t 番目の実測値。 $\hat{x}^{(t)}$ はその推定値。
- 自己符号化器 (autoencoder) による再構成誤差
- 再帰型ニューラルネット (RNN) による予測誤差



■ 計算された数値の統計的有意度を調べるためには、予測値についての確率分布が必要

- 再構成誤差だけで押し通す場合、ずれの大きさについての事前知識が必要
- 画像や言語といった応用の場合、データが「直接読める」ので、確率的予測の必要性はあまり認識されないが、一般の産業応用では確率的予測が絶対必須



Agenda

■ 基礎編

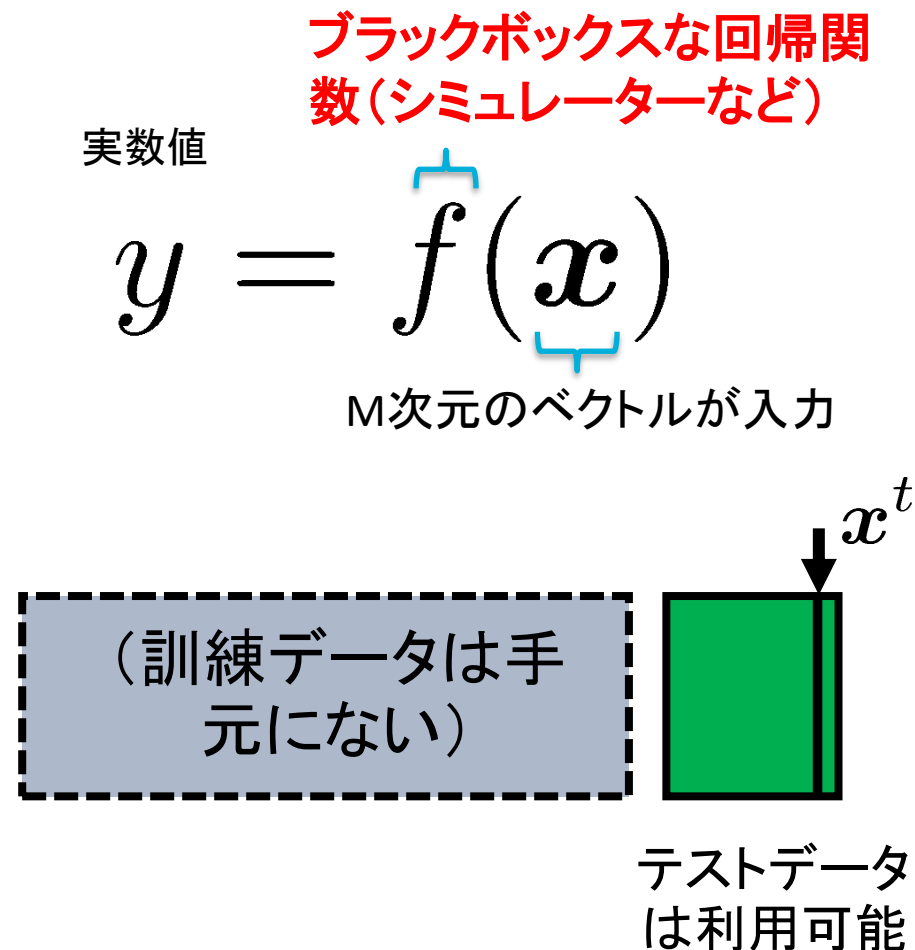
- 機械学習の問題設定
- 異常検知の基礎

■ 最近の話題

- 異常検知における解釈可能性 [Idé et al., AAAI 21]
- 確率的イベントに対する因果分析 [Idé et al., NeurIPS 21]

ブラックボックスの回帰モデルが与えられているときに、異常寄与度の計算をやりたい

- **タスク:** 予測値 $f(x)$ が実測値 y と相当の乖離があるときに、各入力変数のそれぞれに対して、その乖離への寄与度を計算する
- **背景:** これまで知られているほとんどの Explainable AI (XAI) の研究は、ブラックボックスモデル $f(x)$ 自体を説明することを目標としていた。乖離(異常)を説明するものはほぼない。
- **解決策:** 尤度補償法という手法を新たに開発、実データで効果を確認
 - IBM で製品化されている
 - ✓ TRIRIGA Building Insights, IBM OpenScale



ブラックボックス予測モデルからの異常寄与度計算問題 全体の流れ

入力

異常または大きな乖離を生じているテスト標本(の集まり)

$$\{(x^t, y^t)\}$$

ブラックボックスな予測(回帰)モデル

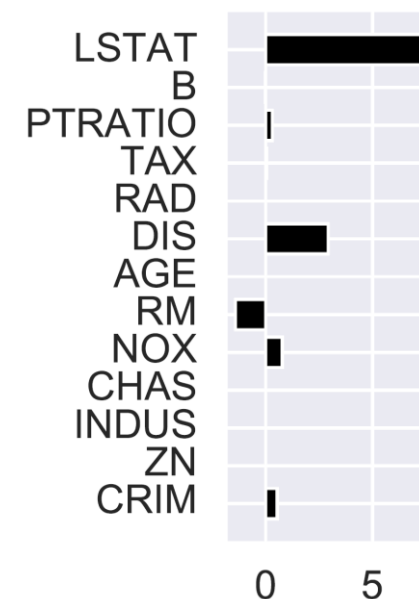
$$y = f(x)$$

尤度補償法

出力

あるテスト標本 (x^t, y^t) の近傍での、乖離に対する各入力変数の責任度:

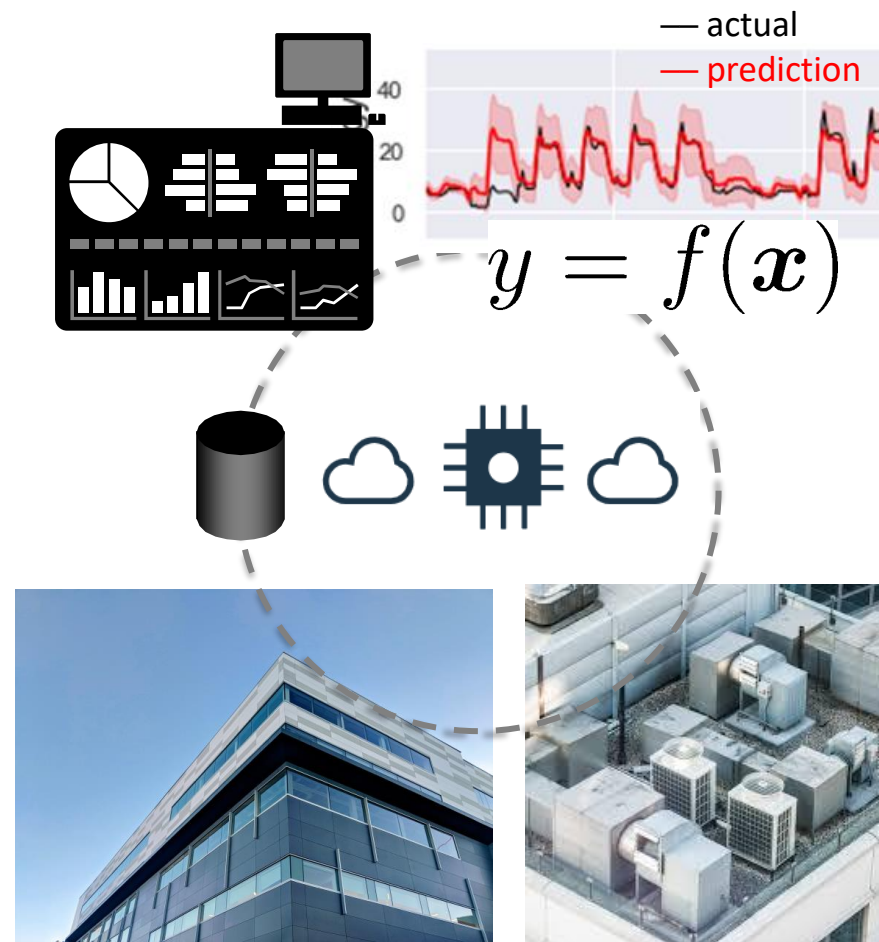
$$\delta_1, \dots, \delta_M$$



(ボストン住宅データでの例)

具体的な応用シナリオの例 ビルのエネルギー管理

- 想定するデータ
 - y : あるビルのエネルギーの総消費量の実測値
 - x : エネルギーに関係しそうな測定値
 - ✓ 外気温、湿度、曜日、月、部屋の埋まり具合、など
- エネルギー消費量を計算するブラックボックスなモデル $f(x)$ が実装されているとする
 - 多くの企業が、ビルのエネルギー管理システムを提供しており、 $f()$ はその一部として提供されるのが典型的
 - ビル管理者は $f()$ の内部に対する詳細な知識はないが、実測値 (x, y) を自由に得ることはできる
- ビル管理者は、直近のデータを定期的に眺めて、モデルと実測の乖離を確認し、もし乖離がある場合、ビル設備の問題を疑う
 - この場合、予測に使っているのではないことに注意



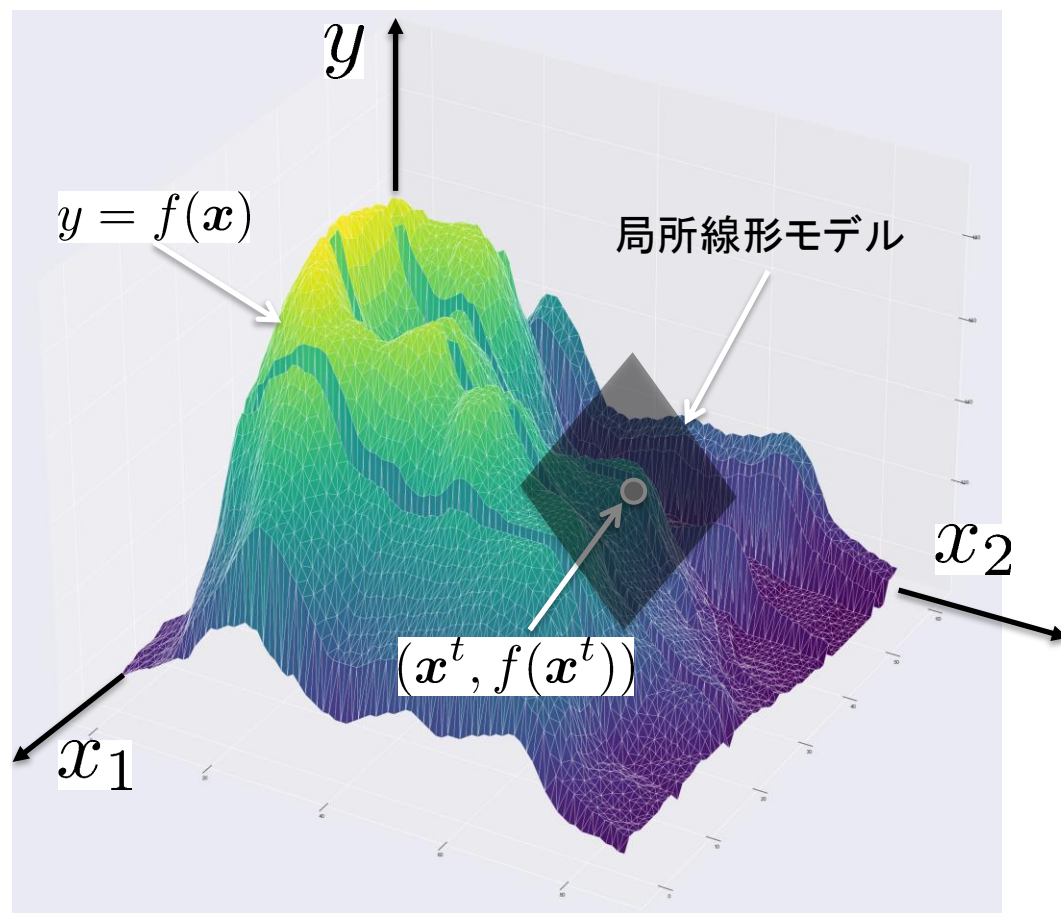
ブラックボックスモデルに対する説明可能性の既存手法

- ほとんどすべてが深層学習による分類モデル(しかも画像)を対象にしている。回帰モデルに使えるものは非常に少ない
- 回帰モデルに対しては、ほぼすべてが、予測値 $f(x)$ を与えた理由を説明しようとしている。
 - 「 $x = x'$ という点に対し、値 $f(x')$ を返した理由」を、各変数の寄与度として出力
- 回帰については主な手法は2つ
- LIME* [Ribeiro 18]
 - ある入力 $x = x^t$ の近傍で局所的に無作為標本を生成、そこで $f(\cdot)$ を近似する回帰平面を推定
 - 回帰係数($\approx$ 傾き)が寄与度
- シャープリー値 [Strumbelj+ 14]
 - あらゆる可能な周辺化分布に対して、各変数が、予測値 $f(x^t)$ をどれだけ「動かしたか」を計算
 - 計算のためには x に対する真の分布(または経験分布)が必要

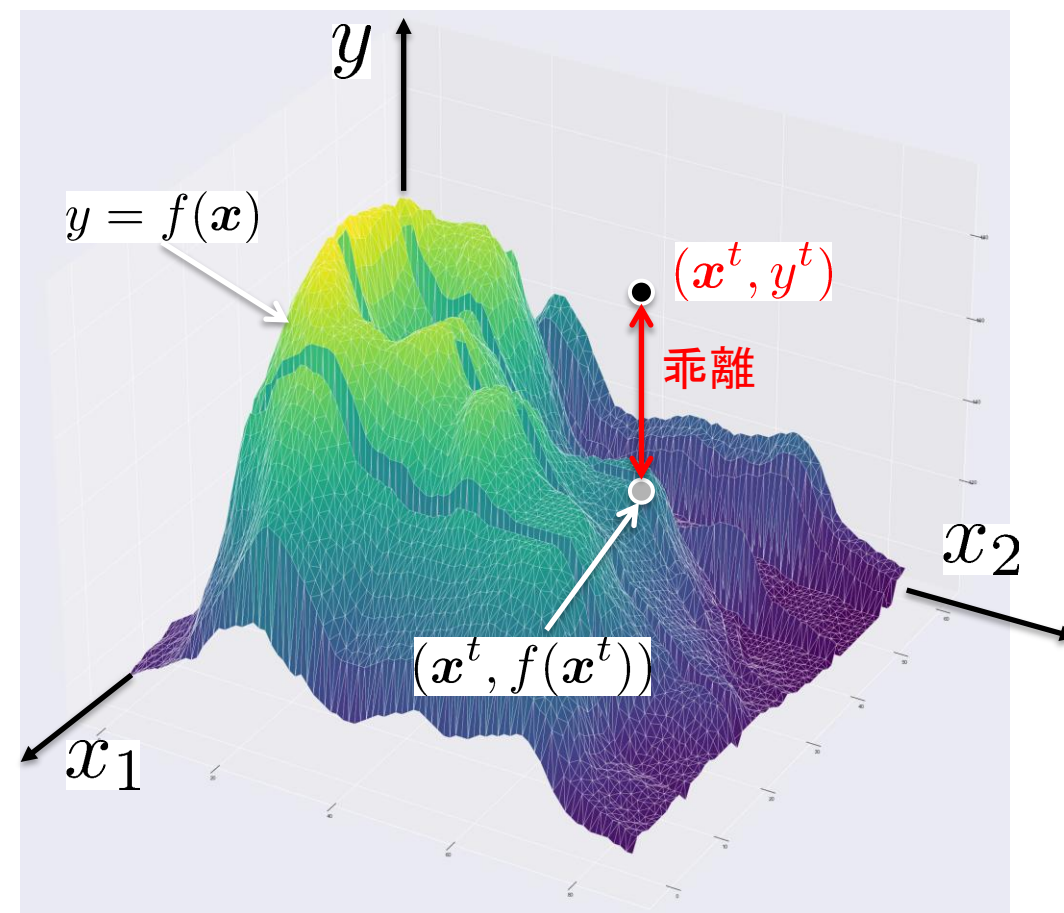
* Local Interpretable Model-agnostic Explanations

LIMEは実測値 y とは無関係に、回帰曲面を局所的に近似する平面を求めているだけ。明らかに異常診断には使えない

$f(x)$ を解釈するための局所的な回帰平面

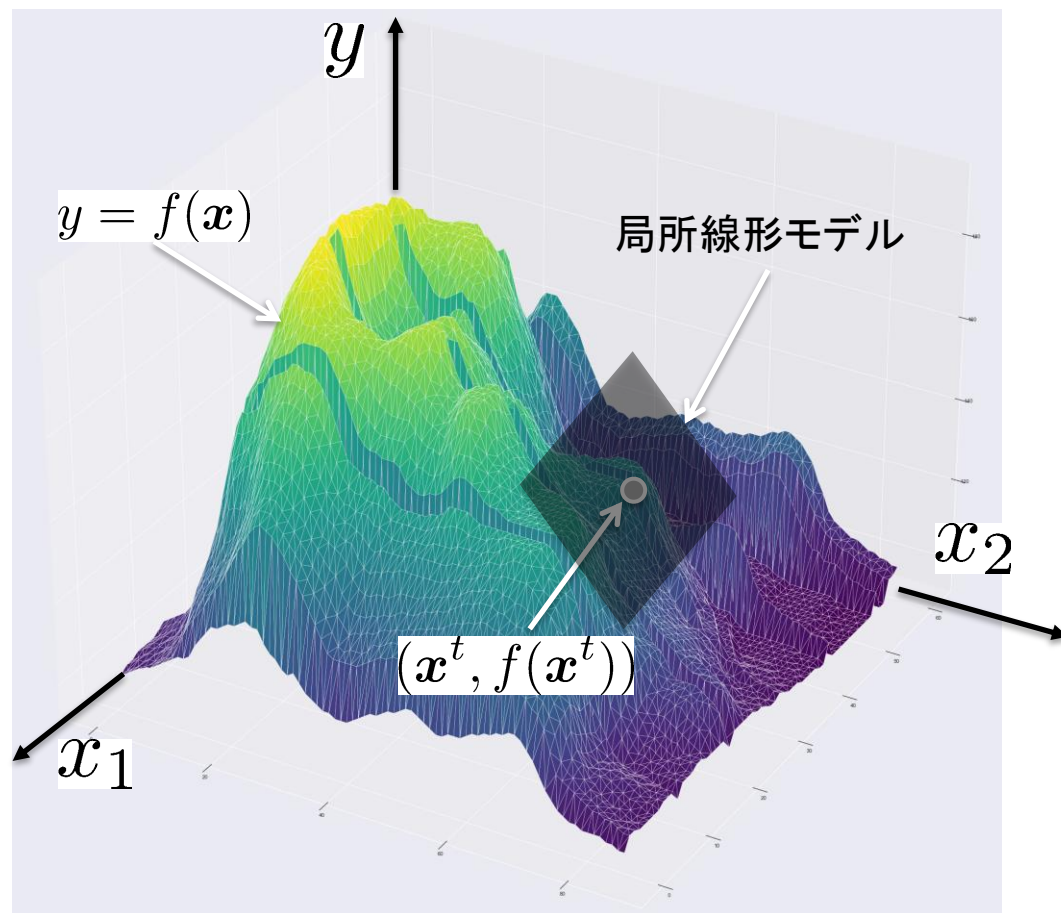


異常寄与度の場合、乖離 $f(x) - y$ を説明したい

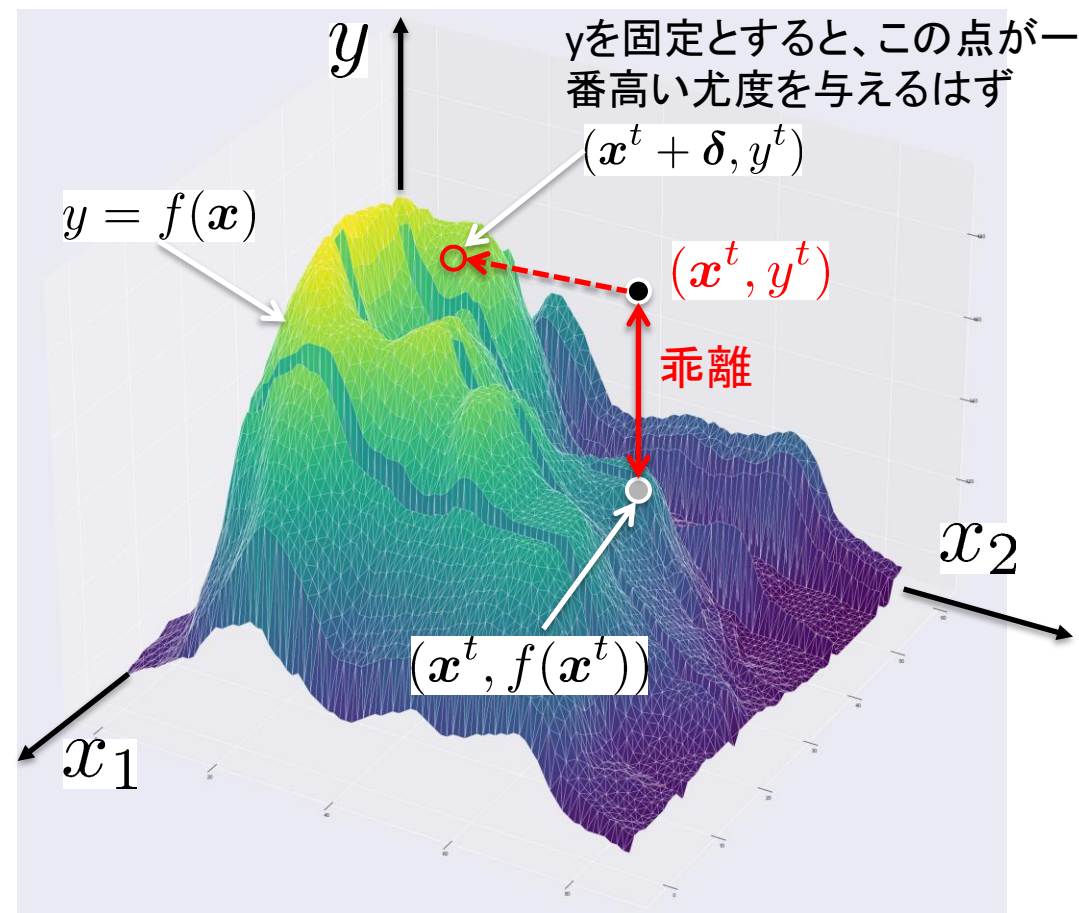


尤度補償法の基本的な考え方: 縦方向(y方向)の乖離を、水平方向の乖離に読み替える

$f(x)$ を解釈するための局所的な回帰平面



δ : 責任度 (尤度補償変位)



尤度補償法における寄与度のスコアは、「尤度を最も高められるような変位」として定義される

- **尤度補償変位 δ** : ある観測点 x^t において、 $x^t + \delta$ が可能な最大の尤度を獲得するような変位 δ

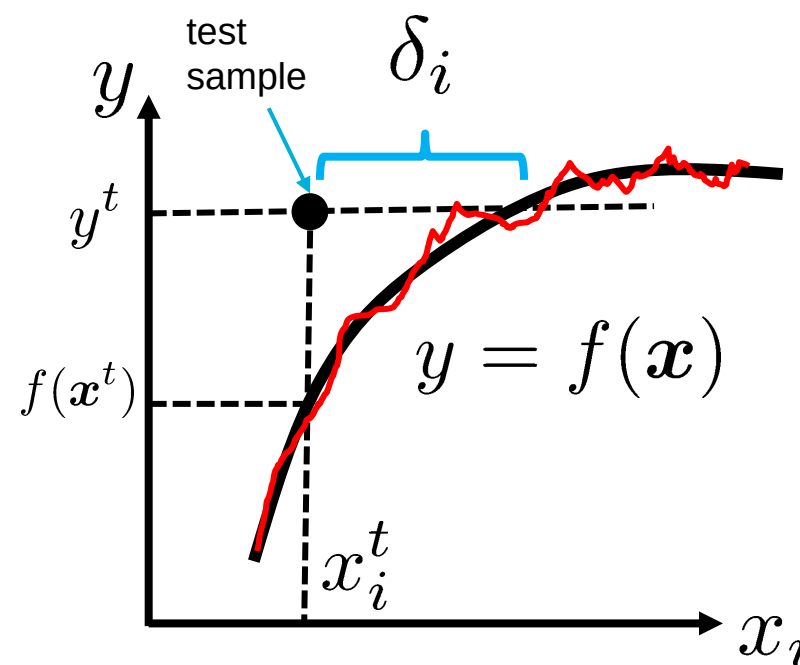
- 対数尤度は $\ln p(y^t | f(x^t))$ はテスト標本 (x^t, y^t) の当てはまりの良さを表現している
- ゆえ、 $x^t + \delta$ は「 y^t を実現するための本来あるべき入力」
- δ が大きいほど本来の姿からかけ離れている

- The main optimization problem

$$\delta = \arg \max_{\delta} \left[\underbrace{\ln \{p(y^t | f(x^t + \delta))\}}_{\text{Gaussian}} \underbrace{p(\delta)}_{\text{elastic net}} \right],$$

$$\min_{\delta} \left\{ \frac{1}{N_{\text{test}}} \sum_{t=1}^{N_{\text{test}}} \frac{[y^t - f(x^t + \delta)]^2}{2\sigma_t^2} + \frac{1}{2} \lambda \|\delta\|_2^2 + \nu \|\delta\|_1 \right\}, \quad (\lambda \text{ and } \nu \text{ are constant})$$

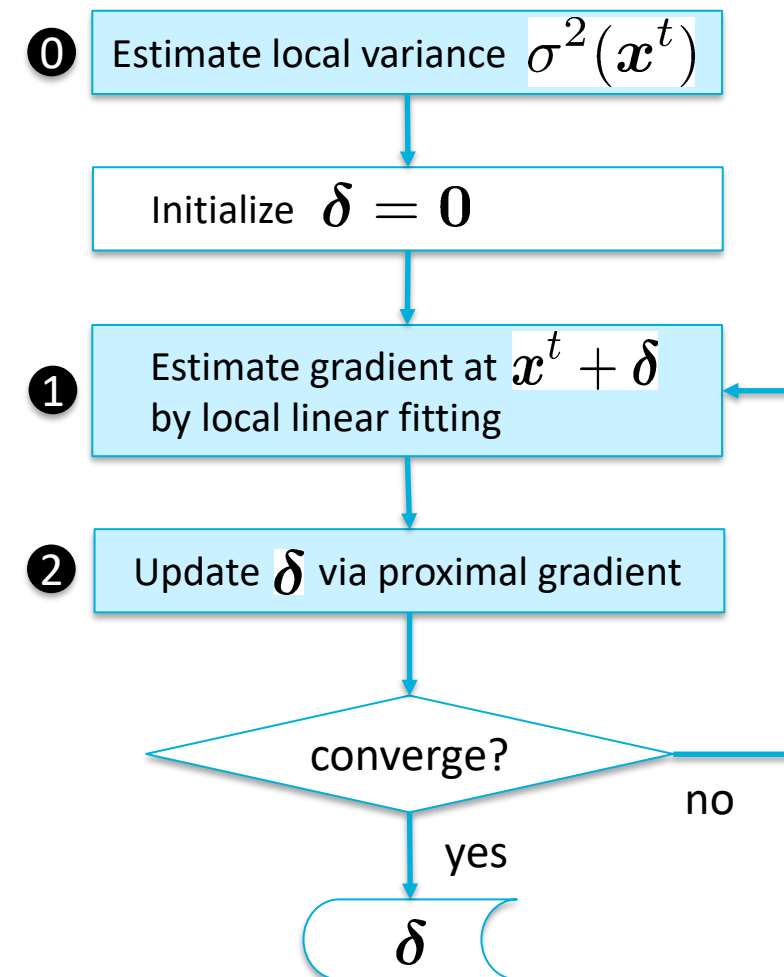
ざっくり言って、尤度補償変位というのは、「縦じゃなくて横に測った偏差」のこと



尤度補償スコアの計算手順(詳細略)

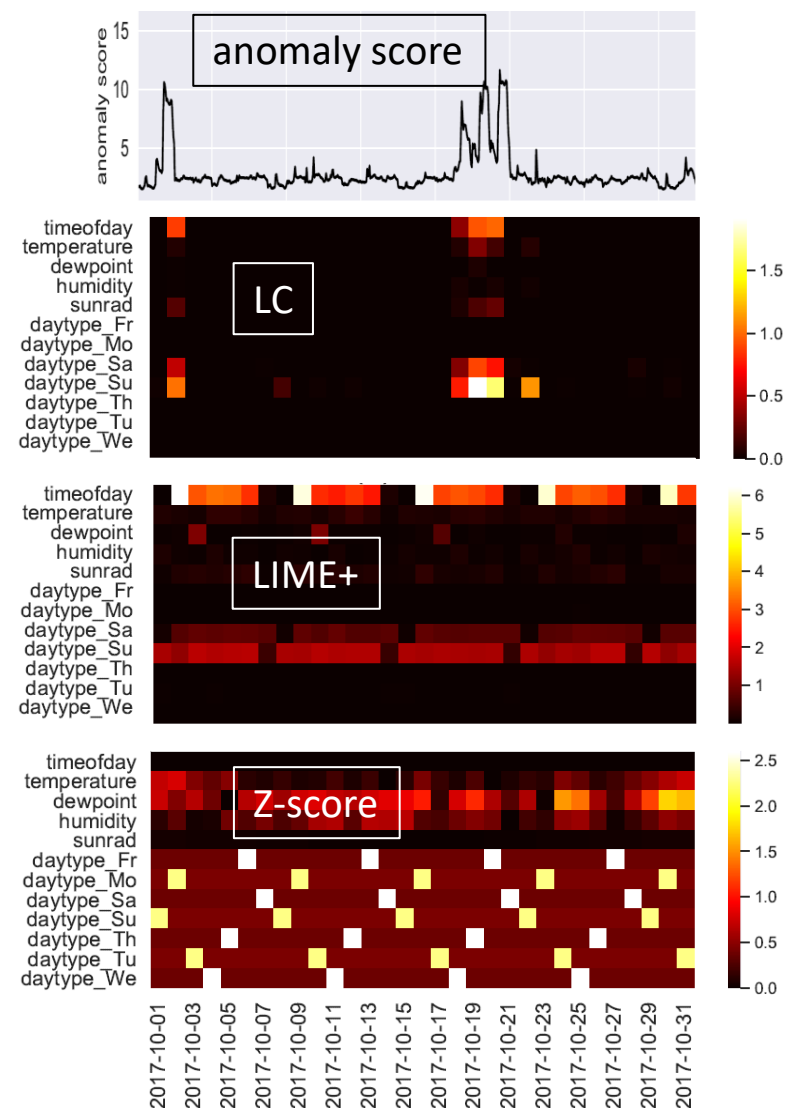
- 注意: $f(\mathbf{x})$ はブラックボックスなシミュレータ。関数として連続でも滑らかでもないかもしれない
- 0. 局所的に関数の分散を推定する(1回のみ)
 - テストデータを使って、 $f(\mathbf{x})$ の予測値のばらつきを評価する
- 1. $f(\mathbf{x})$ の局所的な勾配を推定
- 2. 尤度を最大化するように δ を更新
 - 近接勾配法を使う

反復



実際のビルのデータでの評価結果

- 1か月程度のビルのデータ(インドの某オフィスビル)
 - y : エネルギー消費量
 - x : 曜日 (one-hot encoding)、外気温、湿度、日照量、など
- 一番上が異常度 $\equiv$ 実測と予測との乖離の度合い。提案手法(LC)のみがまとめた結果
 - 「異常度が上がったのは、その日の“週末っぽさ”が妙に高かった」 $\rightarrow$ 実はインドの特別な連休だった
- 寄与度スコアは24時間の平均を1日ごとに計算
 - 図はその絶対値をプロット
 - 比較した手法
 - ✓ Zスコア $a_i(x^t) = (x_i^t - \text{mean}_i) / \text{stddev}_i$
 - ✓ LIMEの改良版。f(x) じゃなくて、f(x)-y に対して計算したLIMEスコア
- LIME+は値の乖離を無視。Zスコアは定義から y に依存せず、したがって乖離を無視
 - しましま模様は曜日のOne-hotエンコーディングの副作用



Agenda

■ 基礎編

- 機械学習の問題設定
- 異常検知の基礎

■ 最近の話題

- 異常検知における解釈可能性 [Idé et al., AAAI 21]
- 確率的イベントに対する因果分析 [Idé et al., NeurIPS 21]

イベントの時系列が与えられた時、イベント同士の因果診断(event causal diagnosis)をやりたい

■ データ: イベント時系列

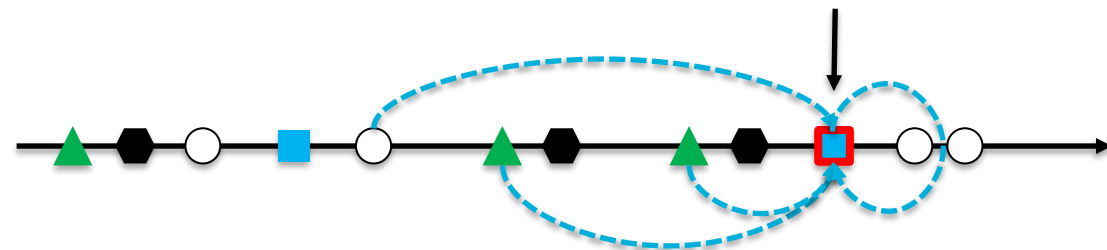
- 一般に、(イベント種、時間スタンプ) の組

$$\mathcal{D} = \{(t_0, d_0), (t_1, d_1), \dots, (t_N, d_N)\}$$

✓ t_n : 時間スタンプ

✓ d_n : イベント種 $\{1, \dots, D\}$

Event log



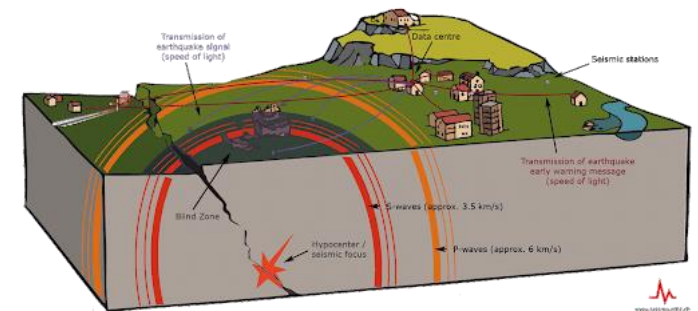
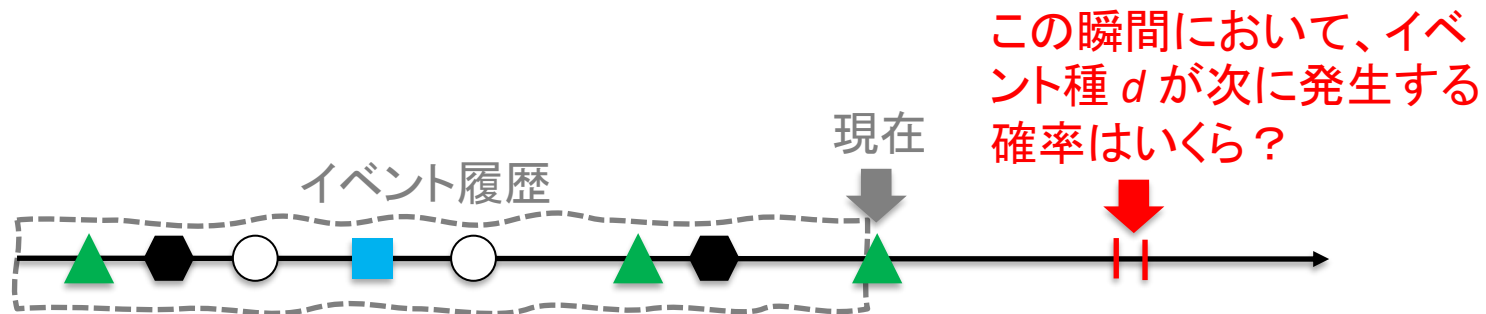
■ 典型的応用: コンピュータシステムの監視業務

- 個別のモジュールから、大量の警告・エラーイベントが出される
- 何が何を引き起こしたのか理解するのは難しい
- イベント間の因果関係が分かれば、直ちに対処策の検討に移ることができる



地球物理学の分野では、本震と余震の関連づけのために自己励起過程 (aka ホークス過程) と呼ばれる確率モデルが使われてきた

- 因果分析という観点で問題領域は非常に近い
- 自己励起過程では「イベント強度関数」が中心的な役割を果たす
 - 過去履歴とイベント種が与えられた時、未来のある時点でイベントが起こる確率密度



* Picture source: Swiss Seismological Service,
<http://www.seismo.ethz.ch/en/home/>

- なお、日本の統計数理研究所のグループが指導的役割を果たしている
 - ✓ Y. Ogata, "Seismicity analysis through point-process modeling: A review." Seismicity patterns, their statistical significance and physical meaning (1999): 471-507.

自己励起過程(ホークス過程)ではイベント強度を、自発的発生確率と励起確率の線形結合で表す

■ ホークス過程での強度関数の表現

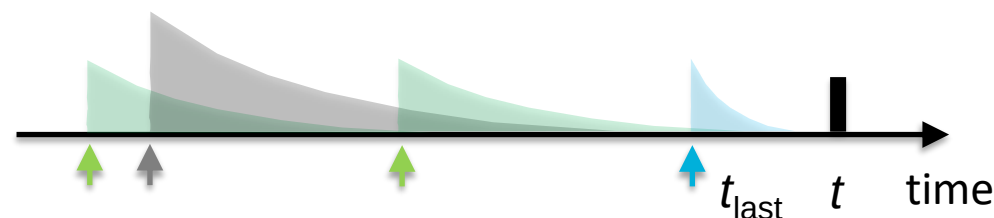
- (発生強度) = (自発発生効果) + (因果効果)
 - ✓ (因果効果) = (イベント種間の因果関係) × (時間減衰)
- イベント種 d のイベントの強度の式

$$\lambda_d(t \mid \mathcal{H}_{t-}) = \underbrace{\mu_d}_{\text{基本強度} \rightarrow \text{自発発生効果}} + \sum_{i:t_i < t} \underbrace{A_{d,d_i}}_{\text{インパクト行列: タイプ間の因果関係を表現}} \underbrace{\phi_d(t - t_i)}_{\text{減衰関数 (d に依存) } \rightarrow \text{トリガー効果}}$$

基本強度
→ 自発発生効果

減衰関数 (d に依存)
→ トリガー効果

インパクト行列: タイプ間の因果関係を表現



よく使われる減衰モデル

$$\phi_d(\tau) = \beta_d \exp(-\beta_d \tau)$$

$$\phi_d(\tau) = \frac{\eta \beta_d}{(1 + \beta_d \tau)^{\eta+1}}$$

■ モデルパラメーターは、通常、最尤推定で求められる

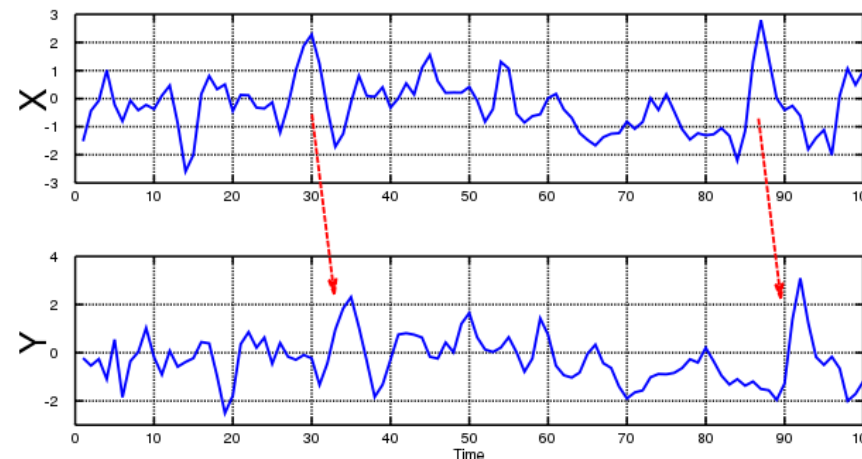
- 基本強度 $(\mu_1, \dots, \mu_D)$, 減衰定数 $(\beta_1, \dots, \beta_D)$
- インパクト行列 A

機械学習の観点でのホークス過程研究の2つの重要なマイルストーン

- 下限最大化法 (MM法) による反復的最尤推定法の発明 [Veen and Schoenberg 08]
 - Newton法は形式的に書き下せるが、数値計算上の深刻な問題があった
 - 対数尤度のJensen下限を最大化する手法 (EM算法と類似)
- グレンジャー因果学習との関連の認識 ([Eichler 17] ほか)
 - ホークス過程のインパクト行列の最尤推定は、イベント種同士のグレンジャー因果行列を学習していることと等価

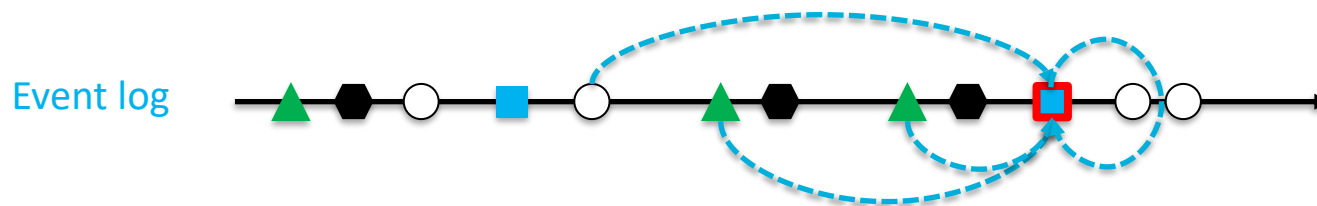
グレンジャー因果

- 時系列同士の因果関係を定義する方法のひとつ
- ある遅れを伴って時間相関を持てば、因果ありと判断
- 右の例では、XはYの原因 (cause) の可能性が高い



イベント因果診断のための要件: ホークス過程の学習において、スパースにインパクト行列を学習すること

- あるイベントインスタンスに対し、潜在的には、過去のあらゆるインスタンスが因果的に関係しうる
 - 桶屋が儲かったのは風が吹いたから？
 - ムルソーがアラブ人を撃ったのは、太陽が眩しかったから？（カミュ『異邦人』）



- したがって、「適切な割り切り」が実用上極めて重要。
- ホークス過程のスパース最尤推定、そしてそれによるイベント系列に対するスパース・グレンジャー因果モデリングの技術確立したい

ホークス過程の最尤推定: 下限最大化法 (MM法) の概要

Log likelihood

$$L = \sum_{n=1}^N \left\{ \ln \lambda_{d_n}(t_n \mid \mathcal{H}_{n-1}) - \int_{t_{n-1}}^{t_n} du \lambda_{d_n}(u \mid \mathcal{H}_{n-1}) \right\}$$

ジェンセンの不等式で下界を求める

$$\begin{aligned} \ln \lambda_{d_n}(t_n \mid \mathcal{H}_{n-1}) &= \ln \left\{ \mu_{d_n} + \sum_{i=0}^{n-1} A_{d_n, d_i} \phi_{d_n}(t_n - t_i) \right\} \\ &\geq q_{n,n} \ln \frac{\mu_d}{q_{n,n}} + \sum_{i=0}^{n-1} q_{n,i} \ln \frac{A_{d_n, d_i} \phi_d(t_n - t_i)}{q_{n,i}} \end{aligned}$$

4つのパラメーターそれぞれに最適化問題がある

個々のイベントのトリガー確率

$\{q_{n,i}\}$

インパクト行列

A

基底強度

$\mu_1, \dots, \mu_D$

減衰度

$\beta_1, \dots, \beta_D$

A, μ, β が既知とした時の最適解

$$q_{n,i} = \begin{cases} \frac{\mu_{d_n}}{\mu_{d_n} + \sum_{i=0}^{n-1} A_{d_n, d_i} \phi_{d_n}(t_n - t_i)} & i = n \\ \frac{A_{d_n, d_i} \phi_{d_n}(t_n - t_i)}{\mu_{d_n} + \sum_{i=0}^{n-1} A_{d_n, d_i} \phi_{d_n}(t_n - t_i)} & i \neq n \end{cases}$$

これが話の中心
→ next section



$\{q_{n,i}\}$ が既知とした時、 L_2 正則化の下で解析解が知られている

収束まで反復

(おさらい)トリガー確率のMM解の導出: ジェンセンの不等式の最もタイトな下界を求める

■ 対数尤度において関係する部分を抜き出す ($n = 1, \dots, N$)

$$\circ \max_{\mathbf{q}_n} \left\{ q_{n,n} \ln \frac{\mu_d}{q_{n,n}} + \sum_{i=0}^{n-1} q_{n,i} \ln \frac{A_{d_n,d_i} \phi_d(t_n - t_i)}{q_{n,i}} \right\} \quad \text{subject to} \quad \sum_{i=0}^n q_{n,i} = 1; \forall i, q_{n,i} > 0$$

$$\checkmark \mathbf{q}_n = [q_{n,0}, \dots, q_{n,n}]^\top$$

○ ラグランジュ乗数で制約を取り込んだ目的関数

$$\mathcal{L} = q_{n,n} \ln \frac{\mu_d}{q_{n,n}} + \sum_{i=0}^{n-1} q_{n,i} \ln \frac{A_{d_n,d_i} \phi_d(t_n - t_i)}{q_{n,i}} - \lambda \left(\sum_{i=0}^n q_{n,i} - 1 \right)$$

■ 目的関数は上に凸なので唯一の最大値を持つ

○ (証明) $q_{n,i}$ について2回微分すると $-1/q_{n,i}$ が出る。これは常に負。

■ $\mathcal{L}$ を1回微分して0とおいたものが最適性の条件

$$\circ \text{解は} \quad \ln \frac{\mu_d}{q_{n,n}} = \text{constant}, \quad \ln \frac{A_{d_n,d_i} \phi_d(t_n - t_i)}{q_{n,i}} = \text{constant}$$

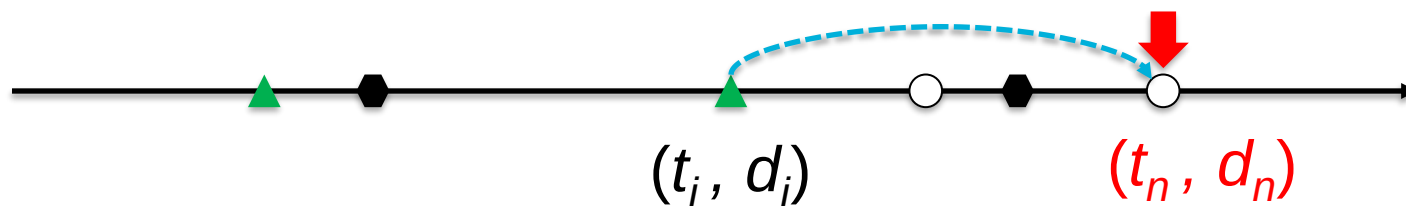
✓ 総和が1の条件を考慮すると前ページで示した解に

$\{q_{n,i}\}$ はイベント同士の因果トリガー確率と解釈できる

- 前ページで導いた解は第 i 番目のイベントが第 n 番目のイベントを引き起こした確率と解釈できる。

$$q_{n,i} = \begin{cases} \frac{\mu_d}{\mu_d + \sum_{i=1}^{n-1} A_{d,d_i} \phi_{d_n}(t - t_i)} \\ \frac{A_{d,d_i} \phi_{d_n}(t - t_i)}{\mu_d + \sum_{i=1}^{n-1} A_{d,d_i} \phi_{d_n}(t - t_i)} \end{cases}$$

- この「因果トリガー確率」をインスタンスレベルの因果分析に使う



“I just got an event (t_n, d_n) .
Tell me which event caused
that particular event?”

(おさらい) 基底強度 $\mu = [\mu_1, \dots, \mu_D]^\top$ のMM解の導出

- 対数尤度において μ に関する項を抜き出し L_2 正則化項を付す

$$\circ \quad L = \sum_{n=1}^N \left\{ q_{n,n} \ln \frac{\mu_{d_n}}{q_{n,n}} - \mu_{d_n} \Delta_{n,n-1} \right\} - \frac{1}{2} \nu_\mu \|\mu\|_2^2, \quad \text{where } \Delta_{n,n-1} \triangleq t_n - t_{n-1}$$

- 2次導関数は常に負なので、上に凸、唯一の最大値が存在

- 1次微分を0とおいて最適性の条件を得る

$$\circ \quad 0 = \frac{\partial L}{\partial \mu_k} = \sum_{n=1}^N \delta_{d_n, k} \left\{ \frac{q_{n,n}}{\mu_k} - \Delta_{n,n-1} \right\} - \nu_\mu \mu_k, \quad \text{where } \delta_{d_n, k} \text{ is Kronecker's delta}$$

- これは2次方程式なので簡単に解が求められる

$$D_k^\mu = \sum_{n=1}^N \delta_{d_n, k} \Delta_{n,n-1}, \quad N_k^\mu = \sum_{n=1}^N \delta_{d_n, k} q_{n,n}, \quad \mu_k = \frac{1}{2\nu_\mu} \left(-D_k^\mu + \sqrt{(D_k^\mu)^2 + 4\nu_\mu N_k^\mu} \right).$$

(おさらい) 減衰率 $\beta = [\beta_1, \dots, \beta_D]^\top$ のMM解: 一般的な解の表現

- 対数尤度において β に関する項を抜き出し L2 正則化項を付す

$$L = \sum_{n=1}^N \sum_{i=0}^{n-1} \left\{ q_{n,i} \ln \frac{\phi_{d_n}(\Delta_{n,i})}{q_{n,i}} - A_{d_n, d_i} h_{n,i} \right\} - \frac{1}{2} \nu_\beta \|\beta\|_2^2$$

$$h_{n,i} \triangleq \int_{t_{n-1}}^{t_n} du \phi_{d_n}(t - t_i).$$

$$\Delta_{n,i} \triangleq t_n - t_i$$

- 1回微分して0と置いて最適解を求める

$$0 = \frac{\partial L}{\partial \beta_k} = \sum_{(n,i)} \left\{ q_{n,i} \frac{\partial \ln \phi_{d_n}(\Delta_{n,i})}{\partial \beta_k} - A_{d_n, d_i} \frac{\partial h_{n,i}}{\partial \beta_k} \right\} - \nu_\beta \beta_k,$$

- 無次元の減衰関数 $\varphi(\cdot)$ を $\phi_d(u) = \beta_d \varphi(\beta_d u)$ のように作っておくと一般解が次のように書けることが分かる

$$\beta_k = \frac{1}{2\nu_\beta} \left(-D_k^\beta + \sqrt{(D_k^\beta)^2 + 4\nu_\beta N_k^\beta} \right),$$

$$N_k^\beta = \sum_{(n,i)} \delta_{d_n, k} q_{n,i} = \sum_{n=1}^N \delta_{d_n, k} (1 - q_{n,n})$$

$$D_k^\beta = \sum_{(n,i)} \delta_{d_n, k} \left\{ A_{k, d_i} \frac{\partial h_{n,i}}{\partial \beta_k} - q_{n,i} \frac{\varphi'(\beta_k \Delta_{n,i})}{\varphi(\beta_k \Delta_{n,i})} \right\}.$$

(おさらい) 減衰率 $\beta = [\beta_1, \dots, \beta_D]^\top$ の MMM 解: 指数分布およびべき分布の場合の解

■ Exponential distribution

○ $\varphi(u) = \exp(-u)$

$$D_k^\beta = \sum_{n=1}^N \delta_{d_n, k} \sum_{i=0}^{n-1} \left[q_{n,i} \Delta_{n,i} + A_{k,d_i} \frac{\partial h_{n,i}}{\partial \beta_k} \right],$$

$$\frac{\partial h_{n,i}}{\partial \beta_k} = \delta_{d_n, k} \left[\Delta_{n,i} e^{-\beta_k \Delta_{n,i}} - \Delta_{n-1,i} e^{-\beta_k \Delta_{n-1,i}} \right].$$

■ Power distribution

○ $\varphi(u) = \eta(1+u)^{-\eta-1}$

$$D_k^\beta = \sum_{n=1}^N \delta_{d_n, k} \sum_{i=0}^{n-1} \left[\frac{(\eta+1)q_{n,i}\Delta_{n,i}}{1+\beta_k\Delta_{n,i}} + A_{k,d_i} \frac{\partial h_{n,i}}{\partial \beta_k} \right],$$

$$\frac{\partial h_{n,i}}{\partial \beta_k} = \delta_{k,d_n} \left\{ \frac{\eta\Delta_{n,i}}{(1+\beta_k\Delta_{n,i})^{\eta+1}} - \frac{\eta\Delta_{n-1,i}}{(1+\beta_k\Delta_{n-1,i})^{\eta+1}} \right\}.$$

[Idé et al., NeurIPS 21] で何を新たに論じたか

- L0 正則化に基づく新しいインパクト行列のスパース学習法を考えた
 - 既存の方法は実はスパースな解を与えないことを「証明」
- インスタンスレベルとイベント種レベルの疎な(=解釈性の高い)因果分析が同時に行えることを示した

インパクト行列 A についての既存の「スパース」推定算法は、実はスパースな解を与えない(！)

- 対数尤度においてインパクト行列 A に関する部分

$$\begin{aligned}\Psi(A) &\triangleq \sum_{n=1}^N \sum_{i=0}^{n-1} q_{n,i} \ln A_{d_n, d_i} - \sum_{n=1}^N \sum_{i=0}^{n-1} A_{d_n, d_i} \int_{t_{n-1}}^{t_n} dt \phi(t - t_i) - \frac{1}{2} \nu_A \|A\|_2^2 \\ &= \sum_{k=1}^D \sum_{l=1}^D (Q_{k,l} \ln A_{k,l} - H_{k,l} A_{k,l}) - \frac{1}{2} \nu_A \|A\|_2^2 \quad \text{added L2 regularizer}\end{aligned}$$

$$Q_{k,l} \triangleq \sum_{(n,i)} \delta_{d_n, k} \delta_{d_i, l} q_{n,i}$$

$$H_{k,l} \triangleq \sum_{(n,i)} \delta_{d_n, k} \delta_{d_i, l} h_{n,i}$$

$$h_{n,i} \triangleq \int_{t_{n-1}}^{t_n} du \phi_{d_n}(t - t_i).$$

- 既存の「スパース」推定法はこれに L_1 or $L_{2,1}$ 正則化項を付して解くが...

$$\sum_{k=1}^D \sum_{l=1}^D (Q_{k,l} \ln A_{k,l} - H_{k,l} A_{k,l}) - \frac{1}{2} \nu_A \|A\|_2^2 - \tau \|A\|_p$$

Proof: Simply because $\ln 0 = -\infty$ and thus 0 is not allowed (easy!)

- Theorem 1: For $p \geq 1$, this problem is convex and has a unique solution. The solution **cannot be sparse**, i.e., $A_{k,l} \neq 0$, if $Q_{k,l} \neq 0$ and $\nu_A \neq 0$.

対数発散を回避してスパースな解を求める「まともな」方法はあるか？

Introducing L_0 -regularized problem with “ ϵ -sparsity”

- Proposed problem of our Hawkes-Granger framework:

$$\sum_{k=1}^D \sum_{l=1}^D \left(Q_{k,l} \ln A_{k,l} - H_{k,l} A_{k,l} - \frac{1}{2} \nu_A A_{k,l}^2 \right) - \tau \| \mathbf{A} \|_0$$

vectorized version $\rightarrow$

$$\max_{\mathbf{x}} \left\{ \sum_m \Psi_m(x_m) - \tau \| \mathbf{x} \|_0 \right\}, \quad \Psi_m(x_m) \triangleq \left(g_m \ln x_m - h_m x_m - \frac{\nu_A}{2} x_m^2 \right)$$

- Singularity remains at zero
- We introduce a “zero-ness” parameter ϵ and solve:

$$\max_{\mathbf{x}} \sum_m \{ \Psi_m(x_m) - \tau I(x_m > \epsilon) \}$$

Semi-analytic solution exists:
Rare example of “solvable”
 L_0 -regularized problem.

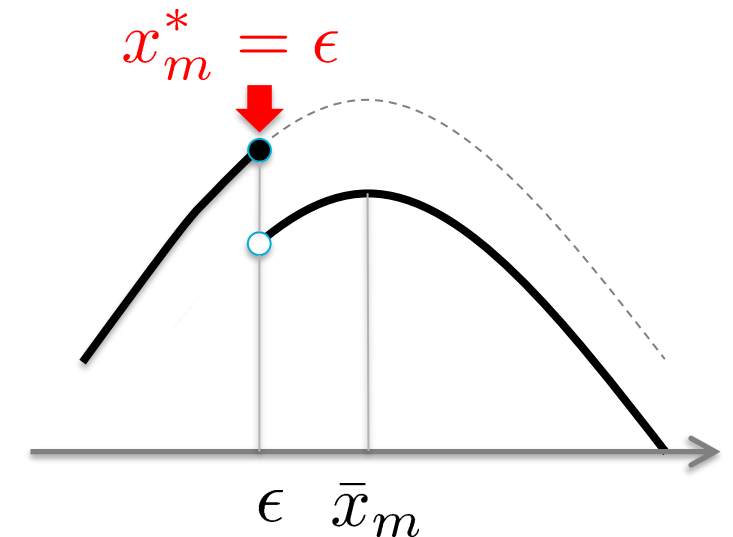
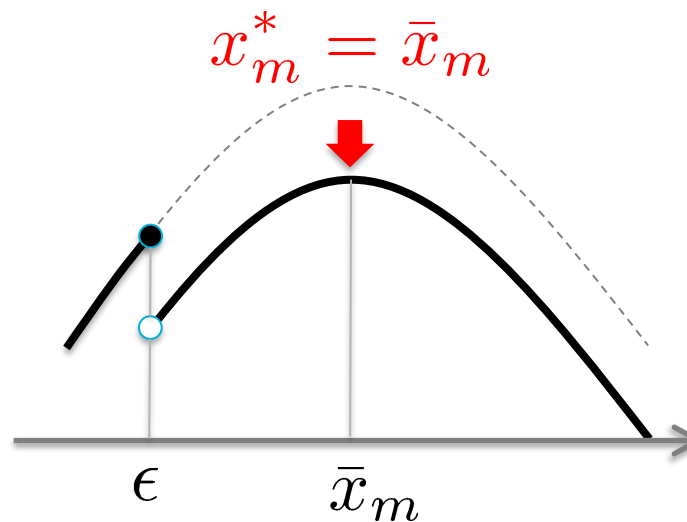
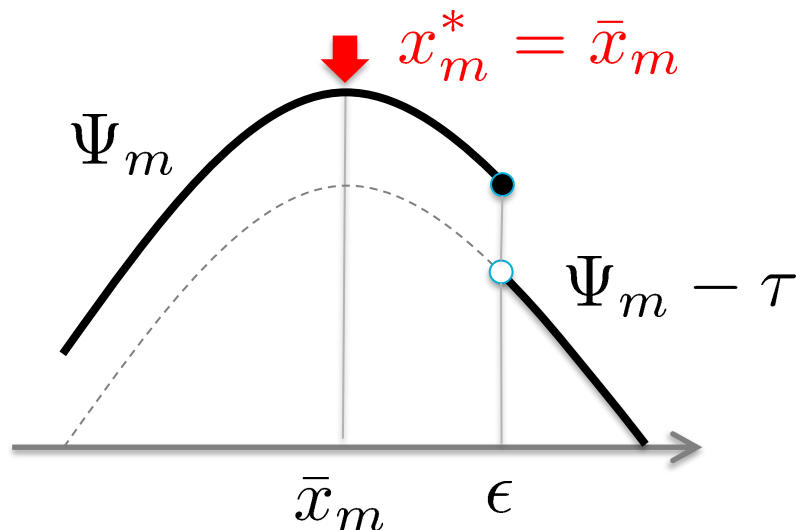
- c.f. [Phan&Ide SDM19] for the first proposal of ϵ -sparsity.

この場合の L_0 -正則化問題は(組み合わせ論的問題に悩むことなく)うまく解くことができる

- Objective function has a jump at $x_m = \epsilon$

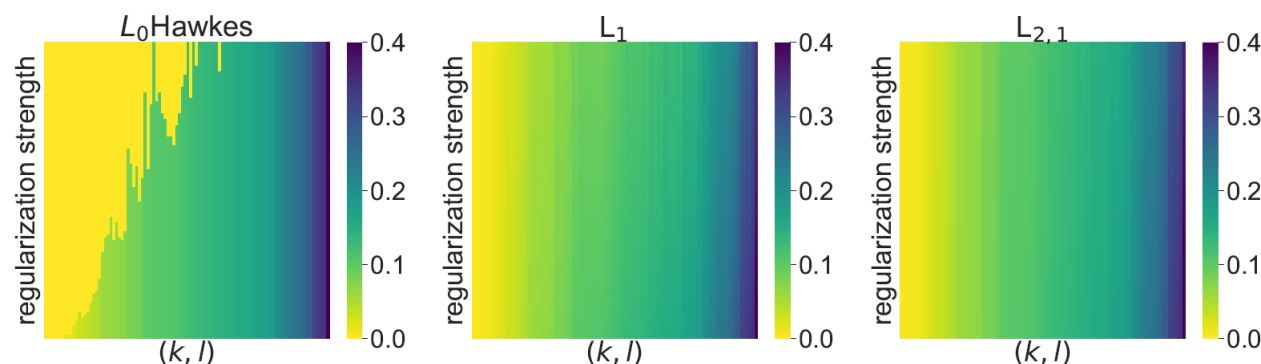
$$\max_x \sum_m \{ \Psi_m(x_m) - \tau I(x_m > \epsilon) \}$$

- The solution covers the three cases below
 - Analytic solution using KKT conditions $\rightarrow$ paper (easy)



インパクト行列のスパースさの数値比較

- インパクト行列を横長ベクトルにつぶして、各行にプロット。
 - 各行は異なる正則化強度に対応する



まっ黄色が0にあたる部分。

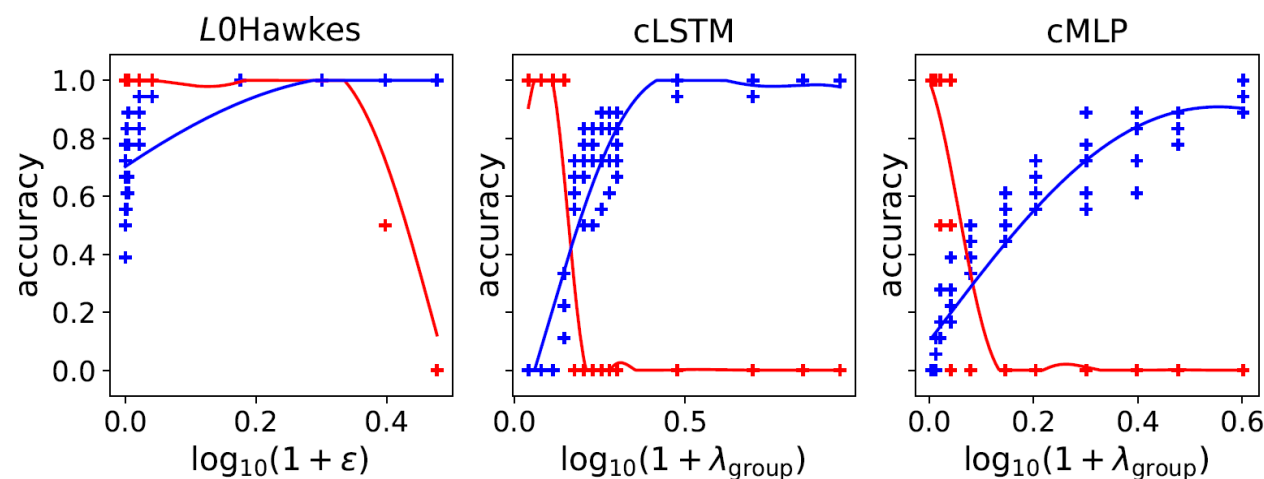
既存手法だと「黄色っぽい」だけ(非ゼロの小さな値)

- 提案手法 L_0 Hawkes は「まっ黄色」の部分がちゃんと存在＝スパース解
 - 既存の「疎な」ホークス過程では0は数学的にありえない。
 - これまでの「疎な」結果は数値誤差の程度の意味しかない

時系列データを対象にしたニューラル因果モデルとの比較: イベントデータを時系列データに直すのは危険

- イベント時系列を、窓内イベント数の時系列に変換、ニューラル因果モデルと比較
 - 注: ニューラル因果モデル [Tank et al. 21]はイベント種同士の因果関係を学習するが、インスタンスの因果関係との同時学習はできない。あくまで部分的比較
 - cMLP: component-wise multi-layer perceptron
 - ✓ VAR (vector autoregressive) モデルによるGranger因果モデルのニューラル拡張
 - cLSTM: component-wise long short-term memory
 - ✓ 状態空間モデルのニューラル拡張としてのLSTMを、さらにGranger因果分析用に変更したもの

■ 因果関係が既知のシミュレーションデータでの実験結果 (競合精度プロット*)



* contrastive accuracy plot

Table 1: Break-even accuracies in Fig. 3.

L_0 Hawkes	cLSTM	cMLP
1.00 ± 0.11	0.43 ± 0.09	0.31 ± 0.10

まとめ 今日話したこと

- 機械学習の問題設定を物理学と対比しつつ説明
- 異常検知のモデリング手順の基本を説明
- 異常検知の文脈での「説明可能なAI」の最新事例
- 最後に、実応用上で非常に重要な(しかし古典的な生存分析以上にはなかなか行けなかった)イベント因果分析に対する最新の成果を紹介