外れ値検出
Outlier Detection

外れサンプル検出
Outlier Sample Detection

明治大学 理工学部 応用化学科
データ化学工学研究室　金子 弘昌

外れ値検出とは？

✓データセットの分布から外れたデータを検出する

✓もちろんデータセットにはばらつきがあるが、ばらつき過ぎていると
考えられるデータを外れ値とする

✓ロバストな方法、時系列データに適した方法、複数の変数を考慮する
方法もある

- ３σ法
- Hampel identifier
- 平滑化(スムージング)による外れ値検出
- データ密度の推定による外れ値 (外れサンプル) 検出

３σ法

✓ある一つの変数のデータがベクトルで与えられているとき、
その平均値から標準偏差 (σ) の３倍以上離れている値を
外れ値とする

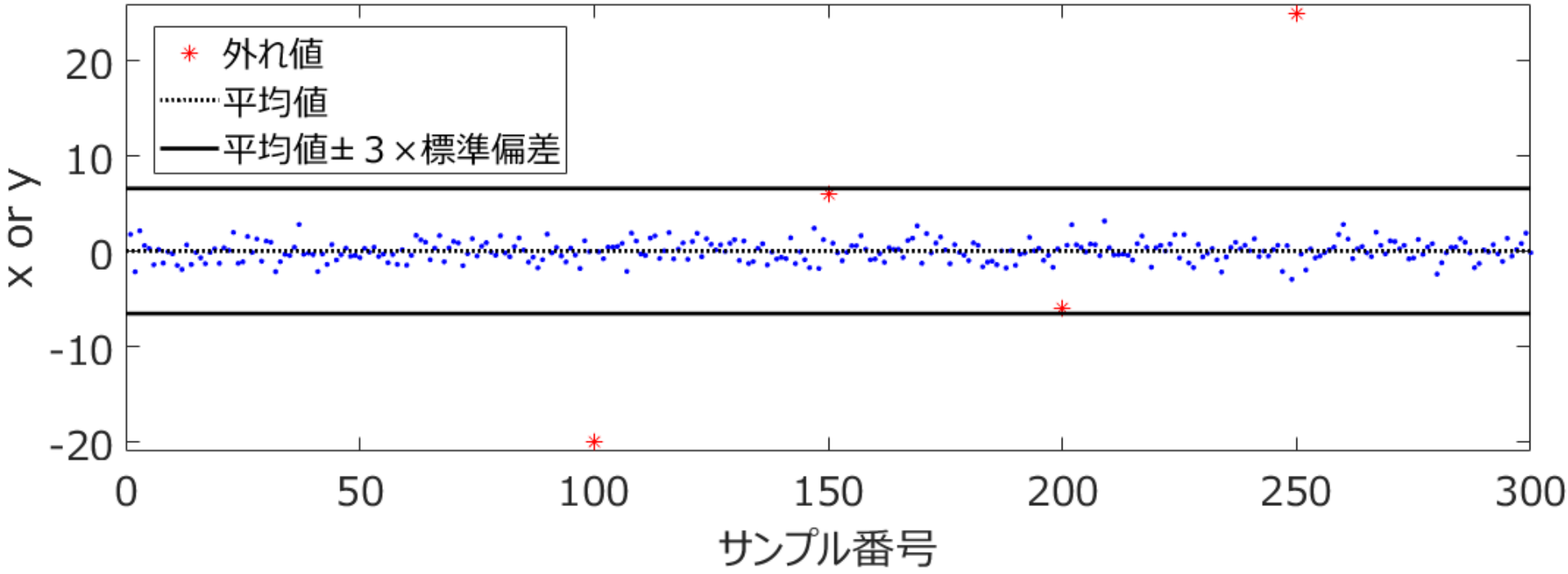
- 閾値は、平均値±3σ
- 変数は説明変数 x でも 目的変数 y でも OK

✓データが正規分布に従うことを仮定している

３σ法の例

✓正規分布に従うデータを乱数で発生させ、意図的に４つの外れ値を混ぜて、それらを検出できたか確認

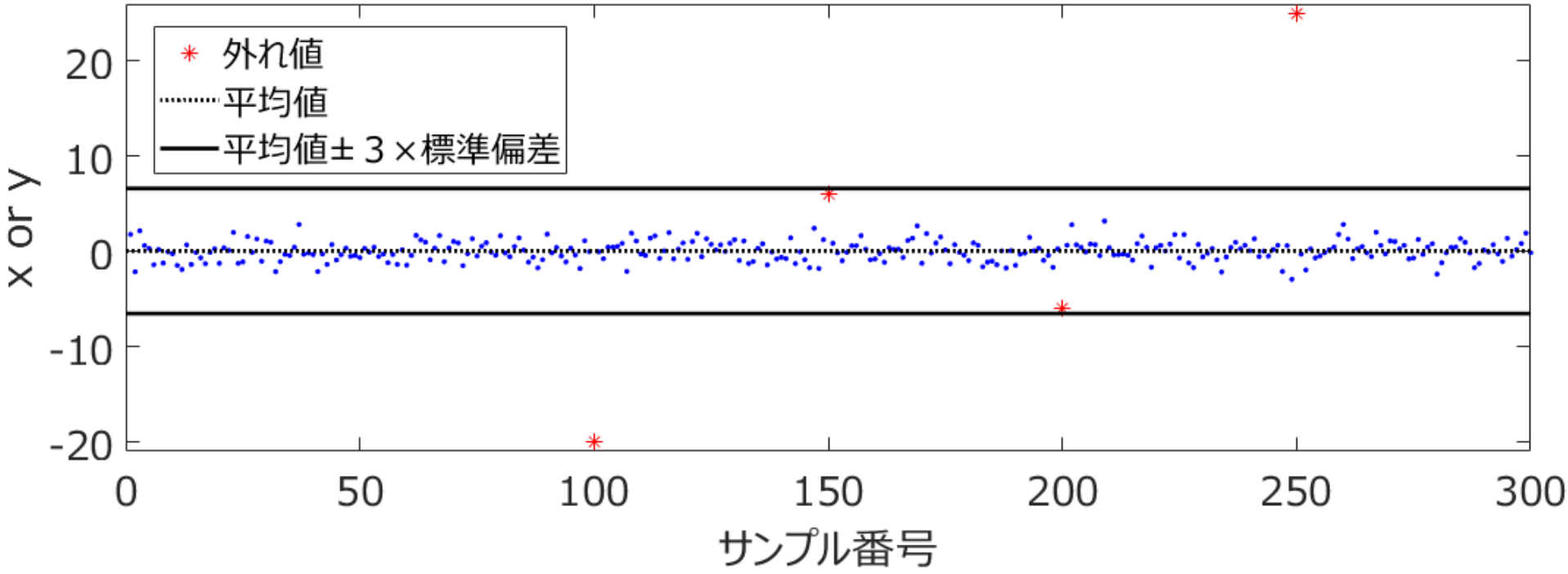
✓４つの外れ値のうち、２つを検出できた



３σ法の問題点

✓外れ値を含んだベクトルで、平均や標準偏差が計算され、
平均値や標準偏差が外れ値の影響を受けてしまう

- 下の例では、外れ値の影響を受けて標準偏差が大きくなってしまい、
２つしか外れ値を検出できていない



Hampel Identifier

✓平均値や標準偏差が外れ値の影響を受ける、という問題を解決するために開発された手法

✓以下のように、平均値と標準偏差をそれぞれロバストな統計量に置き換える

- 平均値 → 中央値
- 標準偏差 → 中央絶対偏差の1.4826倍
  - 1.4826 は、正規分布に従うデータのときに、標準偏差に等しくなるよう補正するための係数

✓ロバストについては、こちら http://datachemeng.com/robustmodel/

✓閾値は、中央値±3×1.4826×中央絶対偏差

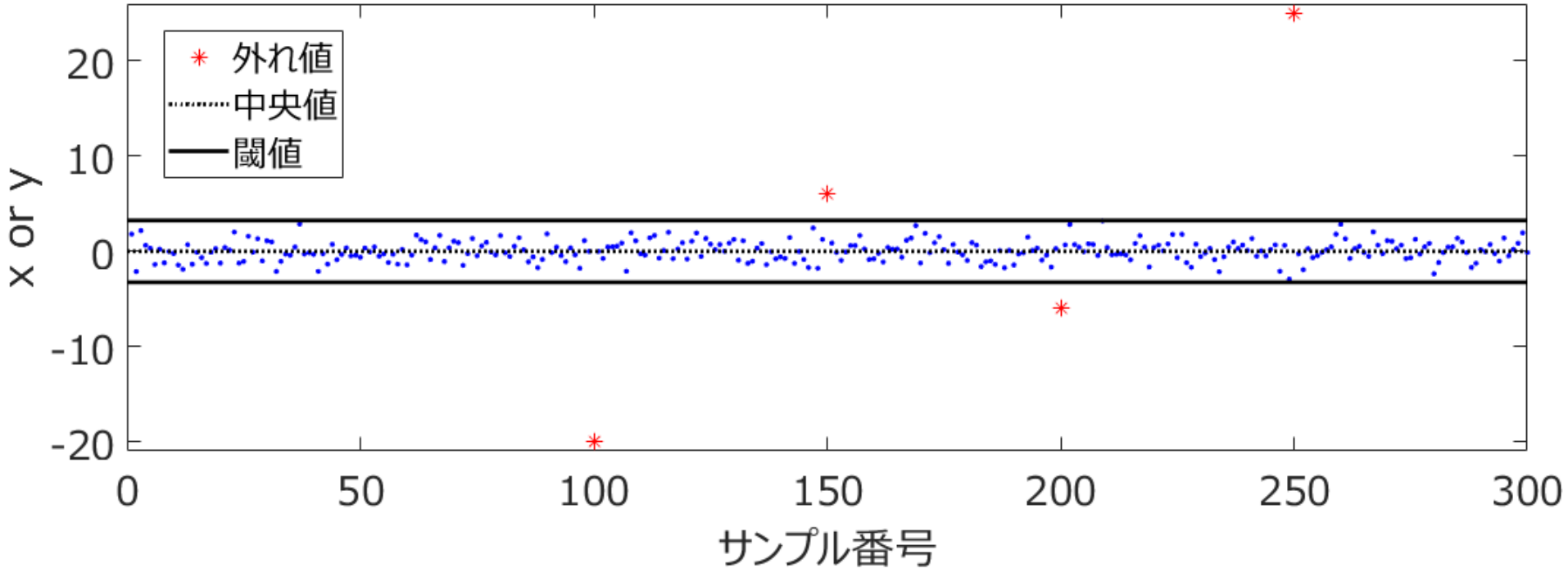
✓変数は説明変数 x でも 目的変数 y でも OK

✓データが正規分布に従うことを仮定している

Hampel Identifierの例

✓正規分布の上限・下限付近に閾値がある

✓４つとも外れ値を検出できた♪



平滑化(スムージング)による外れ値検出

✓時系列データの外れ値検出で有効な方法

✓ある一つの変数のデータがベクトルで与えられているとき、
平滑化 (スムージング) を行う

✓平滑化についてはこちら http://datachemeng.com/preprocessspectratimeseriesdata/

✓平滑化する前とした後とで差をとる

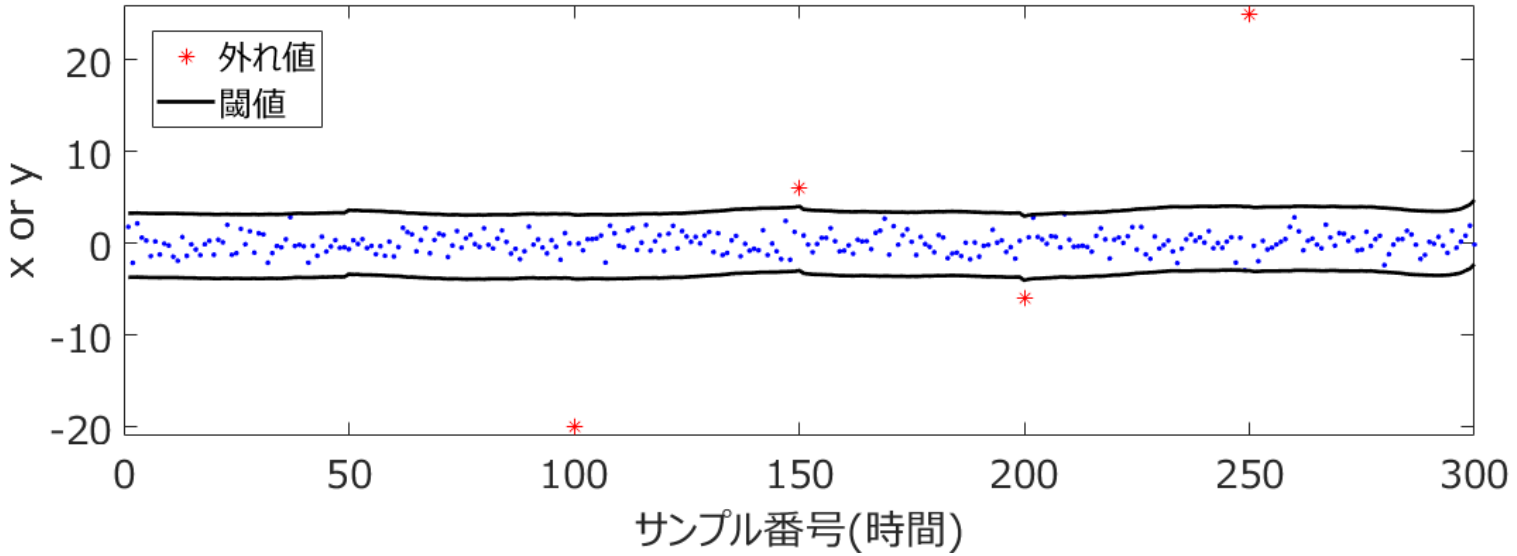
✓その差に対して、３σ法や Hampel identifier で外れ値を検出する

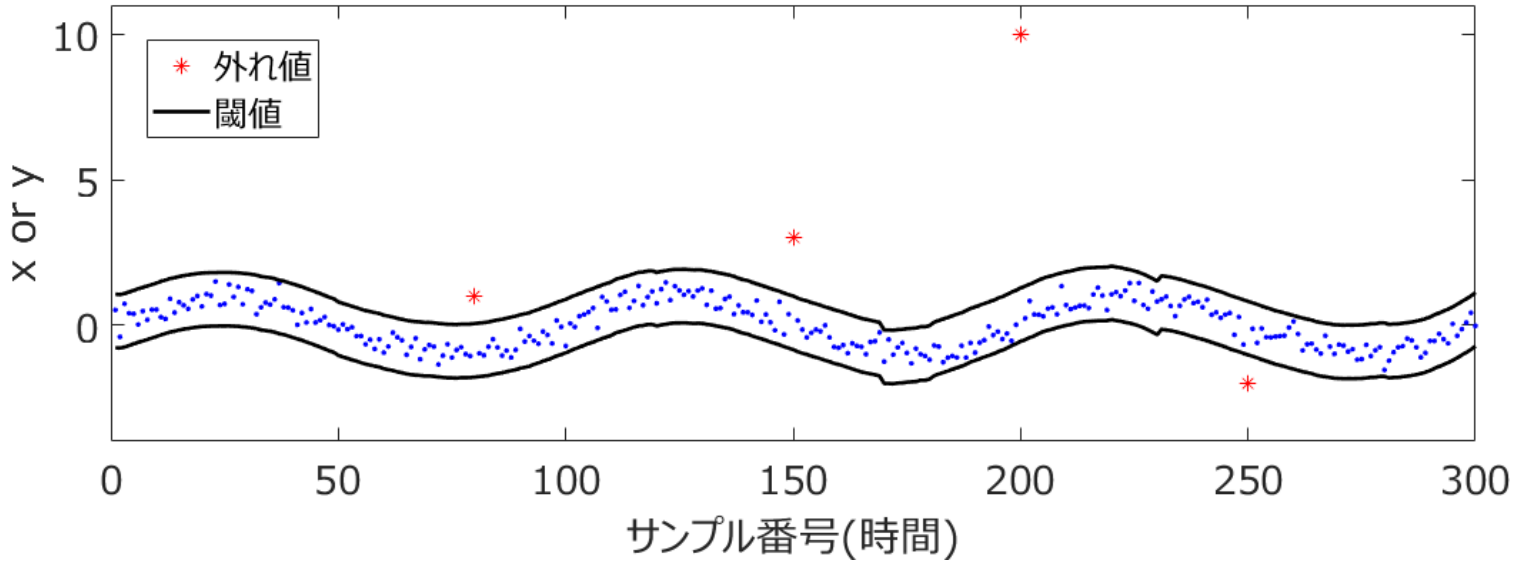
- 平滑化によって、変数の時間変化を考慮した外れ値検出が可能
- ３σ法より Hampel identifier の方がロバストに外れ値検出できる

✓変数は説明変数 x でも 目的変数 y でも OK

平滑化(スムージング)による外れ値検出の例

✓Hampel identifier を用いた例





データ密度による外れ値(外れサンプル)検出

- ✓３σ法、Hampel identifier、平滑化(スムージング)による外れ値検出 は一つの変数に対して外れ値検出をする方法
- ✓複数の変数があるときは、一つの変数ずつ外れ値を検出する必要がある
- ✓これでは、複数の変数を同時に考慮した外れ値検出ができない

- ✓複数の変数を同時に用いる方法の一つに、データ密度による外れ値検出がある
- ✓各サンプルのデータ密度を計算して、データ密度の低いサンプルを検出する

- ✓外れ値というか、外れサンプルを検出できる

データ密度の推定方法

✓k最近傍法(k-Nearest Neighbor, k-NN)

- k-NNについてはこちら http://datachemeng.com/knn/
- k 個の距離の平均が大きいほど、データ密度が低い
  → 外れサンプル

✓One-Class Support Vector Machine (OCSVM)

- OCSVM についてはこちら https://datachemeng.com/ocsvm/