誤差に相関がある場合のカーネル回帰における バンド幅の推定方法

数理情報学専攻 48166202　荒木 健司
指導教員　清 智也 准教授

1　はじめに

回帰分析において, 観測したデータから非線形性が読み取れるときや, 説明変数やデータ数が多いために 2 次や 3 次の項まで用いた線形回帰が計算量の問題で難しいときに, ノンパラメトリック回帰を行うことがある. 本研究では, 誤差に相関がある場合に, ノンパラメトリック回帰のうちの 1 つである Nadaraya–Watson カーネル推定量のバンド幅を推定する方法を提案する. 誤差に相関がある場合に Nadaraya–Watson カーネル推定量のバンド幅を推定する既存手法として, correlation-corrected cross validation と呼ばれる方法がある. この手法は誤差相関を推定したうえで修正した cross validation を用いる手法であるが, 誤差相関の推定には改善の余地がある. また, 別のノンパラメトリック回帰手法である局所多項式回帰において, 誤差に相関がある場合にバンド幅を推定する手法として correlated direct plug-in と呼ばれる方法がある. これはプラグイン推定量を順次推定していくものであるが, 誤差の相関に様々な仮定を置いており, その仮定を満たさないときに推定が悪くなってしまう可能性がある. そこで, correlated direct plug-in を用いてより正確に誤差相関を推定したうえで, 相関のある部分を取り除いて cross validation を行う手法を提案する.

2　誤差に相関がある場合のカーネル回帰

観測 $(X_i, Y_i)_{i=1}^n; X_i \in \mathbb{R}, Y_i \in \mathbb{R}$ から関数 f を推定する問題を考える. ただし, 誤差を ϵ_i とし,

$$Y_i = f(X_i) + \epsilon_i,$$

であるとし, 説明変数 X_i は等間隔であるとする. 誤差に相関がある場合に, カーネル回帰のうちの 1 つである Nadaraya–Watson カーネルのバンド幅を選択する問題について考える. Nadaraya–Watson カーネルとは, K をカーネル関数, h をバンド幅としたとき, f の推定量 $\hat{f}$ を,

$$\hat{f}(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}$$

とするノンパラメトリックな推定方法である. バンド幅 h の選択方法としては, cross validation (CV) が広く用いられているが, 誤差に相関がある場合にはバンド幅を小さく選択してしまう. そこで, Chu and Marron [2] は modified cross validation (MCV) と呼ばれる手法を提案した. MCV とは,

$$\mathrm{MCV}_l(h) = \frac{1}{n}\sum_{i=1}^n \left(\hat{f}_{l,(-i)}(X_i) - Y_i\right)^2,$$

$$\hat{f}_{l,(-i)}(X_i) = \frac{\sum_{j:|j-i|>l} K\left(\frac{X_j - X_i}{h}\right) Y_j}{\sum_{j:|j-i|>l} K\left(\frac{X_j - X_i}{h}\right)}$$

を最小化するような h を選択する方法である. $l = 0$ のとき MCV は通常の CV と等価である. l を選択する方法として, Brabanter et al. [1] において, correlation-corrected cross validation (CC-CV) と呼ばれるアルゴリズムによって選択する方法が提案されている. CC-CV は bimodal カーネルを用いて f を推定して誤差 ϵ_i を推定し,

$$\left|\frac{\sum_{i=1}^{n-q} \hat{\epsilon}_i \hat{\epsilon}_{i+q}}{\sum_{i=1}^n \hat{\epsilon}_i^2}\right| \leq \frac{\Phi^{-1}(1-\alpha/2)}{\sqrt{n}} \tag{1}$$

を満たす最小の q を l とする手法である. ただし Φ は標準正規分布の累積分布関数である. CC-CV は誤差の相関の推定に改善の余地があると考えられる.

また, 別のカーネル回帰として, 局所線形回帰があげられる. 局所線形回帰は誤差に相関がない場合には Nadaraya–Watson カーネルのバイアスを下げることができることが知られている [3]. 局所線形回帰において誤差に相関がある場合にバンド幅を選択する手法として, Opsomer et al. [4] において, correlated direct plug-in (CDPI) と呼ばれる方法が提案されている. CDPI は誤差の相関に連続性など様々な仮定を置いている. CDPI は説明変数が多次元の場合も加法モデルを仮定することで適用できる.

そこで, CDPI を用いて誤差を推定し, l を選択して MCV を行う手法を提案する. これを automated modified cross validation (AMCV) と呼ぶ. また, 説明変数が d 次元のとき, 説明変数を $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_n$ とし, $\boldsymbol{X}_i =$

$(X_{i1},\cdots,X_{id}), X_{ij}\in\mathbb{R}$ とする．$\boldsymbol{H}=\{h_1,\cdots h_d\}$ をバンド幅としたとき，f の推定量 $\hat{f}$ を，

$$\hat{f}(\boldsymbol{x})=\frac{\sum_{i=1}^{n}K\left(\frac{X_{i1}-x_1}{h_1}\right)\cdots K\left(\frac{X_{id}-x_d}{h_d}\right)Y_i}{\sum_{i=1}^{n}K\left(\frac{X_{i1}-x_1}{h_1}\right)\cdots K\left(\frac{X_{id}-x_d}{h_d}\right)}$$

とし，バンド幅 $\boldsymbol{H}$ を選択する．Y_i の推定値と Y_i の二乗誤差を計算するときに，$\boldsymbol{X}_i$ とのユークリッド距離が L 以下のデータを除外したデータで推定するように MCV を定義し，CDPI を用いて誤差を推定するような手法を提案する．これを説明変数が多次元の場合の AMCV とする．

3　数値実験

3.1　説明変数が 1 次元の場合

真の関数 f は，

$$f(x)=1-6x+36x^2-53x^3+22x^5, \tag{2}$$
$$f(x)=\sin(2\pi x), \tag{3}$$
$$f(x)=x \tag{4}$$

の 3 種類とし，誤差相関は，

$$\forall i\neq j\qquad \mathrm{corr}(\epsilon_i,\epsilon_j)=\frac{1}{3(n|X_i-X_j|)^3}$$

としたとき，各手法の MASE の推定値は表 1 のようになった．誤差相関が不連続な場合は AMCV が最も良い推定をしていると言える．さらに，真の関数の次元が低い場合，誤差の分散が大きい場合にも AMCV を用いると良いことが別の数値実験から確かめられた．

1981 年の東京の日ごとの最高気温のデータ [5] に AMCV を適用したところ，図 1 のようになった．平年値 (1981 年から 2010 年の平均) のデータをよく推定できていると言え，実データにおいても AMCV が最も良い推定をする場合が存在することが確かめられた．

表 1. MASE の推定値

	f: 式 (2)	f: 式 (3)	f: 式 (4)
通常 CV	0.353	0.357	0.187
MCV ($l=1$)	0.0683	0.0601	0.0334
CC-CV	0.0642	0.0616	0.0327
AMCV (提案)	<u>0.0592</u>	<u>0.0569</u>	<u>0.0313</u>
CDPI	0.0627	0.0570	0.0483
ローカルレベルモデル	0.123	0.115	0.0914

3.2　説明変数が 2 次元の場合

真の関数 f は，

$$f(\boldsymbol{x})=f_1(x_1)+f_2(x_2), \tag{5}$$
$$f(\boldsymbol{x})=f_1(x_1)+f_2(x_2)+\sin(2\pi x_1x_2) \tag{6}$$

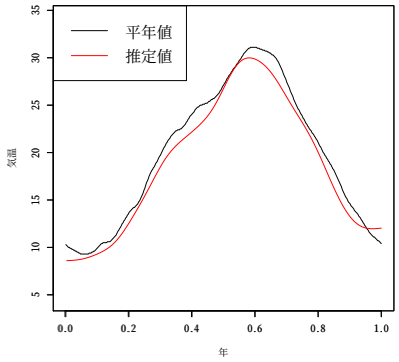

図 1. 日ごとの最高気温の推定

の 2 種類とする．ただし，

$$f_1(x)=1-6x+36x^2-53x^3+22x^5,$$
$$f_2(x)=\sin(2\pi x)$$

である．誤差相関は，

$$\forall i\neq j\qquad \mathrm{corr}(\epsilon_i,\epsilon_j)=\exp(-20|\boldsymbol{X}_i-\boldsymbol{X}_j|)$$

としたとき，各手法の MASE の推定値は表 2 のようになった．加法モデルの仮定が正しくないことが疑われる場合は AMCV を用いればよいことがわかる．

表 2. MASE の推定値

	f: 式 (5)	f: 式 (6)
通常 CV	0.309	0.325
MCV ($L=1/\sqrt{n}$)	0.232	0.256
CC-CV	0.160	0.187
AMCV (提案)	0.148	<u>0.141</u>
CDPI	<u>0.0628</u>	0.246

参考文献

[1] K. Brabanter, J. Brabanter, J. Suykens, and B. Moor: Kernel regression in the presence of correlated errors. *The Journal of Machine Learning Research*, vol. 12 (2011), pp. 1955–1976.

[2] C.-K. Chu and J. Marron: Comparison of two bandwidth selectors with dependent errors. *The Annals of Statistics*, vol. 19 (1991), pp. 1906–1918.

[3] J. Fan: Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, vol. 87 (1992), pp. 998–1004.

[4] J. Opsomer, Y. Wang, and Y. Yang: Nonparametric regression with correlated errors. *Statistical Science*, vol. 16 (2001), pp. 134–153.

[5] 気象庁ホームページ, http://www.jma.go.jp/jma/index.html (参照 2018 年 1 月 19 日)