説明変数にも誤差を含む回帰とカーネル法

新機軸創発センター　モンテカルロ計算研究グループ
准教授　　伊庭　幸人

1　説明変数にも誤差を含む回帰

回帰分析で説明変数（入力）と応答変数（出力）の両方の誤差を考える問題は，measurement error problem とか「関数関係の推定」と呼ばれることもあるが，一見初等的に見えて，理論的にも実際的にもあなどれない問題として知られている．

従来，この問題については，説明変数の真値に相当する無限個の未知変数が存在することに伴う数理的な難しさが強調されてきた．しかし，それ以外にもこのタイプの問題にはさまざまな面白さがある．特に，数理的な難しさが軽減されるような設定の場合でも，少し複雑な状況を考えると実際の推定計算は容易ではなく，計算統計的な意味で興味ある事例が現われる．

たとえば，中江ら (2009) は説明変数の誤差と応答変数の誤差に相関がある場合を階層ベイズモデルで定式化し，MCMCの一種であるレプリカ交換モンテカルロ法で扱っている．以下では別の例として，カーネル回帰との関係を考える．

2　具体例

以下で念頭におく具体例をベイズモデル（同時確率）の形で与えると以下のようになる．(x_i, y_i) は i 番目のデータの説明変数と応答変数の対，z_i は x_i の真の値を示すパラメータである．データ数を n，説明変数 x_i,z_i はいずれも d 次元ベクトルであるとした．

$$p(x, y, z, f) = p(y|f, z)p(x|z)p(z)p(f) \tag{2.1}$$

$$p(y|z, f) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(y_i - f(z_i))^2}{2\sigma_y^2}\right) \tag{2.2}$$

$$p(x|z) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma_x^2)^{d/2}} \exp\left(-\frac{||x_i - z_i||^2}{2\sigma_x^2}\right) \tag{2.3}$$

上では，あてはめる関数 f の事前分布 $p(f)$ を指定した．通常は f を多項式や三角関数でパラメトリックに表現するが，$p(f)$ として平滑化事前分布を用いた研究もある (Berry et al, (2002))．

z の事前分布 $p(z)$ はある程度自由に与えてよい．次の節の手法が使えるためには $p(x|z)$ がガウス分布であることは本質的ではないが，$p(y|z, f)$ はガウス分布である必要がある．数理的な難しさを避けるために，分散 σ_x^2, σ_y^2 は既知とする．

3　カーネル回帰と説明変数の誤差

いわゆるガウシアンプロセス回帰では f が再生核ヒルベルト空間 $\mathcal{H}$ の元であるとして，次のような形の $p(f)$ を考える．

$$p(f) \propto \exp\left(-\frac{1}{2}\|f\|_{\mathcal{H}}^2\right),\ f \in \mathcal{H}, \tag{3.1}$$

ここで便利なのは罰金項 $\|f\|_{\mathcal{H}}^2$ に対応するカーネル関数 k が存在して，最適解（事後確率最大解）$f^*(\cdot)$ を $\sum_{i=1}^N a_i k(\cdot, x_i)$ の形の線形結合の中で探せばよいことである（リプリゼンター定理）．ここで x_i はデータの説明変数の値で通常の回帰では誤差なしと仮定されている．もちろん，$p(f)$ が通常の平滑化事前分布の場合と同様に，f を折れ線などで離散的に表現して解くこともできるが，カーネルを利用することで高次元の場合や無限階の階差に対応する場合も少ない計算量で扱うことが可能である．

ここでの主題は，これに対応する手法を説明変数に誤差がある場合について考えることである．$\{x_i\}$ のかわりに $\{z_i\}$ を考えることになるが，$\{z_i\}$ は未知変数であり，MCMCで解くとすると計算の各ステップでさまざまな値を取って動き回る量である．従って，リプリゼンター定理を適用することは一見難しそうに思われる．しかし，実は工夫するとカーネル回帰に相当する手法をMCMCに組み込んで推定を行うことができる (赤穂, 伊庭 (2009), Iba and Akaho (2009))．これには複数の方法が考えられるが，われわれが試みた手法では次のことがカギとなる．

> $z^{(t)}$ が（MCMCの過程における）z の現在の値とするとき，提案密度 $q(\cdot|z^{(t)})$ に従って生成された次の候補 z^* の採否は Metropolis-Hastings ratio
>
> $$r = \frac{p(z^*\ , f^{(t+1)}|x, y)\, q(z^{(t)}|z^*)}{p(z^{(t)}, f^{(t+1)}|x, y)\, q(z^*|z^{(t)})}. \tag{3.2}$$
>
> を乱数と比較することで決められる．ここで f の現在の値を $f^{(t+1)}$ とした．ところが，上の r は今のモデルでは $f^{(t+1)}(z^{(t)})$ と $f^{(t+1)}(z^*)$ の関数であり，$f^{(t+1)}$ 全体を知らなくても有限個の点の関数値だけで計算できる．

これを利用すると，無限次元の f 全体をサンプルすることなくMCMC計算が行える．詳細は講演で示す．

参考文献

赤穂昭太郎, 伊庭幸人 (2009), カーネルマルコフ連鎖モンテカルロ法による測定誤差モデル推定, 第12回情報論的学習理論ワークショップ (IBIS2009), 博多.

Berry, S.M. and Carroll,R. J. and Ruppert, D., (2002), Bayesian Smoothing and Regression Splines for Measurement Error Problems, J. Amer. Statistical Assoc., 97, 457, 160-169.

Iba, Y. and Akaho, S. (2009), Gaussian process regression with measurement error, 投稿中.

中江健, 伊庭幸人, 青柳富誌生, 坪泰宏, 深井朋樹 (2009), 位相応答曲線のベイズ推定, 2009年度統計関連学会連合大会, 京田辺.