レポート「信用調査データを用いた雇用傾向の把握」で用いられているモデリング

中谷 太洋 *

2023 年 9 月 20 日

本資料は『信用調査データを用いた雇用傾向の把握（2022 年 12 月データ）』(`https://www.tdb-di.com/posts/2023/08/e2023083001.php`) で公開されたレポートのガウス過程回帰による欠測値の補完方法について数理的な背景を詳細にまとめたものである。

一般に統計や機械学習モデルでは、観測値の組 $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ $(i = 1, \cdots, n)$ について、$\boldsymbol{x}_i$ を入力として $\boldsymbol{y}_i$ を出力する関数 $f(\boldsymbol{x})$ を推定する。現実的には観測値は誤差を伴うため、$\boldsymbol{y} = f(\boldsymbol{x}) + \varepsilon$ $(\varepsilon \sim N(0, \sigma^2))$ として推定を行う。ガウス過程回帰はこの $f(\boldsymbol{x})$ を非線形かつベイズ的に推定する。

まずガウス過程の 1 つの仮定として、$f(\boldsymbol{X}) \sim N(m(\boldsymbol{X}), V(\boldsymbol{X}, \boldsymbol{X}))$ とする。これは、モデルが一次元の入力に対して一次元の出力を行い、誤差がない場合の出力が平均 $m(\boldsymbol{X}_i)$、分散 $V(\boldsymbol{X}_i, \boldsymbol{X}_i)$ の正規分布にしたがうことを表している。これに伴って観測値全体の同時分布が多変量正規分布で表されることになる。この仮定は便利で、モデルの入力の次元が増えても平均行列と 2 点間の関係の記述だけで観測値の同時分布が求まる。よって、連続的な無限次元の入力でも機能する。ガウス過程回帰の非線形かつベイズ的に関数を推定するという特徴は、以上の仮定とその推定方法によるものである。先の記述から、$f(\boldsymbol{X}) \sim N(m(\boldsymbol{X}_i), V(\boldsymbol{X}_i, \boldsymbol{X}_i))$ として事前分布からランダムに乱数を生成する。ここでは乱数を関数値とみなすことで、関数 f そのものを抽出していることになる。乱数の組をいくつも生成すれば、異なる関数が得られる。ここで $V(X_i, X_i)$ の要素について、ガウス過程回帰ではカーネル関数（共分散関数）$k(\boldsymbol{x}_i, \boldsymbol{x}_i)$ を用いる。カーネル関数には、動径基底関数（RBF カーネル、ガウスカーネル）、線形カーネル、Matern カーネルなど様々な種類があり、これらの選択やその組み合わせによって直線や曲線の表現が可能になる。ここからはカーネル行列を $K(\boldsymbol{x}_i, \boldsymbol{x}_i)$ と表す。

次に事後分布について考える。ガウス過程回帰では、関数を乱数列としているため、明示的なパラメータがない。つまり、通常の統計モデルや機械学習モデルのようにパラメータと入力値の和積でその入力に対する出力が得られない。しかし、ガウス過程の仮定から新しい入力点に対する出力を求めることができる。新しい入力ベクトルを $\boldsymbol{x}^*$、対応する出力を $f(\boldsymbol{x}^*)$ とすると、出力の同時分布は、

$$\begin{pmatrix} y \\ f(x^*) \end{pmatrix} \sim N[\begin{pmatrix} m(\boldsymbol{X}_i) \\ m(\boldsymbol{X}^*) \end{pmatrix}, \begin{pmatrix} K(\boldsymbol{X}_i, \boldsymbol{X}_i) + \sigma^2 I & K(\boldsymbol{X}_i, \boldsymbol{X}^*) \\ K(\boldsymbol{X}_i, \boldsymbol{X}^*)^T & K(\boldsymbol{X}^*, \boldsymbol{X}^*) \end{pmatrix}]$$

として表すことができる。分散共分散行列のカーネル関数では、既存の入力と新しい入力と関心がある $f(\boldsymbol{x}^*)$ の分布は、y が与えられたもとでの条件付き分布となるため、

$$\begin{aligned} f(\boldsymbol{X}^*)|\boldsymbol{y} \sim N(&m(\boldsymbol{X}^*) + K(\boldsymbol{X}_i, \boldsymbol{X}^*)^T (K(\boldsymbol{X}_i, \boldsymbol{X}_i) + \sigma^2 \boldsymbol{I})^{-1} (\boldsymbol{y} - m(\boldsymbol{X}_i)), \\ &K(\boldsymbol{X}^*, \boldsymbol{X}^*) - K(\boldsymbol{X}_i, \boldsymbol{X}^*)^T (K(\boldsymbol{X}_i, \boldsymbol{X}_i) + \sigma^2 \boldsymbol{I})^{-1} K(\boldsymbol{X}_i, \boldsymbol{X}^*)) \end{aligned}$$

* 博士前期課程 2 年

と表せる。計算を簡便化させるために予め観測から平均を引いて中心化することが多い。実際のモデルは関数の出力に誤差を加えたものであるため、予測分布は、

$$\boldsymbol{y}^*|\boldsymbol{y} \sim N(m(\boldsymbol{X}^*) + K(\boldsymbol{X}_i, \boldsymbol{X}^*)^T(K(\boldsymbol{X}_i, \boldsymbol{X}_i) + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - m(\boldsymbol{X}_i)),$$
$$K(\boldsymbol{X}^*, \boldsymbol{X}^*) - K(\boldsymbol{X}_i, \boldsymbol{X}^*)^T(K(\boldsymbol{X}_i, \boldsymbol{X}_i) + \sigma^2\boldsymbol{I})^{-1}K(\boldsymbol{X}_i, \boldsymbol{X}^*) + \sigma^2)$$

となる。本レポートの分析では、各企業に対してガウス過程回帰を適用し、各四半期の中心となる月の値を取得している。つまり、1 年前までの直近 10 年で毎年調査が入っている企業を対象にしていることから最低でも 10 点の入出力の組をもってガウス過程回帰のモデルを構築し、毎年 2、5、8、11 月の値を取得している。値の取得には不確実性を考慮して、平均曲線上の値とその ±1 標準偏差の値を用いている。最終的に、平均、+1 標準偏差、-1 標準偏差のそれぞれで指標を算出する。全体的に LOCF よりもなだらかな結果となった。

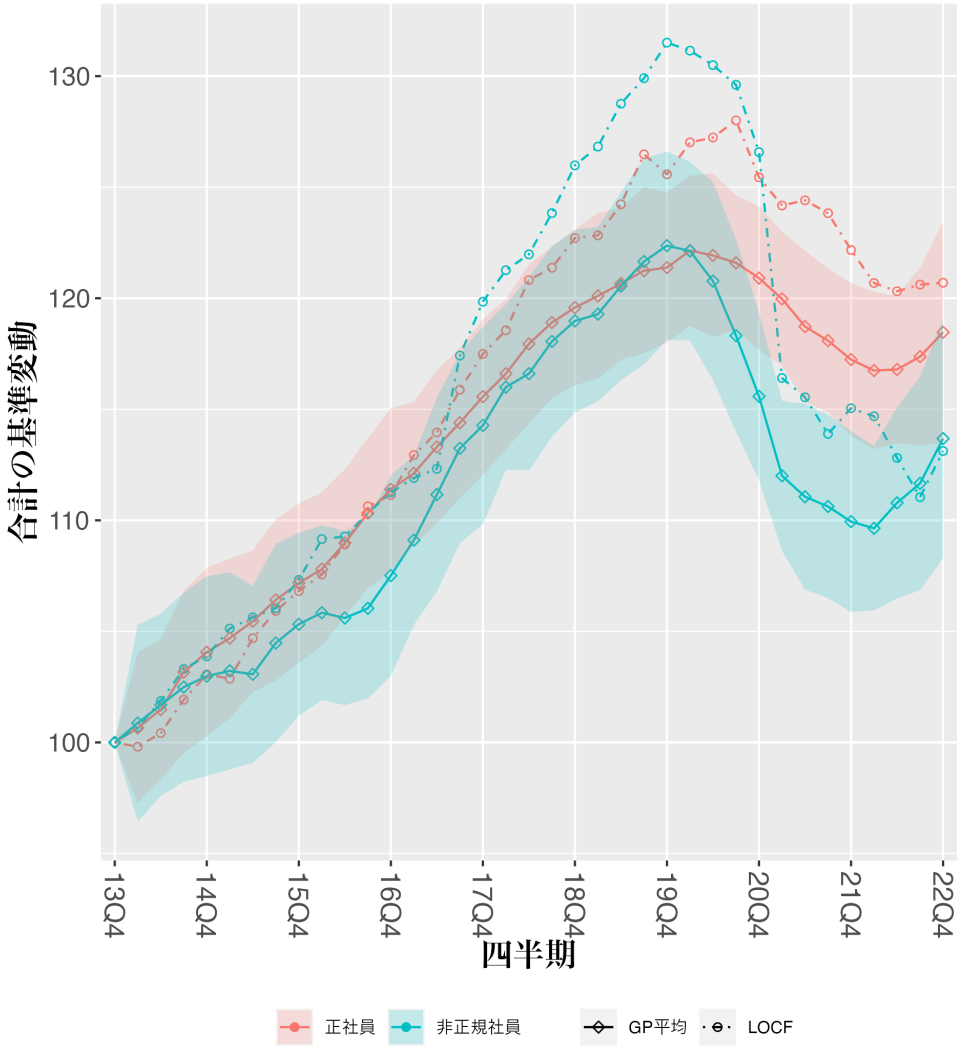

図 1　2022 年第 4 四半期の合計の基準変動