國立中興大學應用數學系

學士班專題

利用函數型變係數模型校正 PM 2.5 預報

Calibrating PM 2.5 forecasts by functional varying coefficient models

指導教授：陳律閎 Lu-Hung Chen

學生：周軒正 Hsuan-Cheng Chou

中華民國一零九年九月

摘要

以函數型主成分分析及回歸分析方法來改良修正現行 PM2.5 空氣品質預報系統準確度。

函數型主成分分析是近代函數型資料與長期追蹤資料分析最重要的基礎工具之一，其估計過程仰賴局部多項式平滑估計、特徵分解等技巧；而回歸分析則是現今在提到分析數據資料時最常使用的方法之一。本專題將在 Python 上以函數型主成分分析結合回歸分析，並使用國立中興大學環境工程學系之空氣品質預報系統（PM2.5 濃度預報部分）作為基礎來進行修改，並與該舊預報系統進行比較。

關鍵詞：多維度函數型主成分分析、多維度回歸分析、參數估計、平滑器、PM2.5、預報系統、Python

目錄

第一章 緒論

1.1 研究目的

修正本校之環境工程學系使用自己的資料庫建置的空氣品質預報系統（PM2.5 部分）的物理模型數值輸出結果。使用函數型主成分分析作為估計方法，回歸分析作為模型建立技巧，並將結果與原系統做預測準確度比較。

而使用回歸分析找出的β_0、β_1則為模型在物理上系統性的偏誤。找出來的β_0、β_1對未來在發展物理模型也會有幫助，可以讓實驗者知道模型在哪個部分有出現錯誤，也提供給了他們精進改善的方向。

1.2 文獻回顧

在 Berrocal、Gelfand、Holland[1]有關空氣品質的文章中，作者們使用了平滑降維的方法把預測模型設為：

$$Y(\boldsymbol{s},t)=\tilde{\beta}_0(\boldsymbol{s},t)+\beta_{1,t}\tilde{x}(\boldsymbol{s},t)+\epsilon(\boldsymbol{s},t)\ , \tag{1}$$
$$\epsilon(\boldsymbol{s},t)\sim N(0,\tau^2)\ ,\ \tilde{\beta}_0(\boldsymbol{s},t)=\beta_{0,t}+\beta_0(\boldsymbol{s},t)$$

其中$\tilde{x}(\boldsymbol{s},t)$為整理後的數據、$\boldsymbol{s}$為位置、$t$為時間；$\epsilon$為誤差；$\beta_0$、$\beta_1$為係數；$Y$為預測值。

而數據的整理方法，因為該預測想以美國各地區切成一塊塊的"網格"來做分析，因此把收集到的數據合併到一個個網格內，並把$\tilde{x}(\boldsymbol{s},t)$定為：

$$\tilde{x}(\boldsymbol{s},t)=\sum_{k=1}^{g}w_k(\boldsymbol{s},t)x(B_k,t)$$

其中w為權重(受網格內資料點數目、時間影響)；B為網格；g為切割的網格數量；k為網格編號。

而這種方法容易在進行資料合併為$\tilde{x}$時出現轉換問題，以及當某個時間、地點的資料出現遺失的情況時我們會需要對遺失的資料另外補上。這些問題都會導致整體模型的準確度下降、複雜度提升。因此我們選用函數型主成分分析作為建立模型中的估計分法，因此方法不僅不用特別對數據進行合併成網格的型式，在有資料遺失的情形下也對整體結果影響不大，非常適合做為改良此模型的方法。

第二章 實現方法

2.1 模型建立

在 Senturk、Muller[2]的文章中，使用了函數型主成分分析來對一維數據進行模型建立，主要模型為：

$$E\{Y(t)|X(t)\} = \beta_0(t) + \beta_1(t)X(t) \quad (2)$$

其中$X(t)$為收集到的數據；$E\{Y(t)|X(t)\}$為在收集到的數據為$X(t)$之情況下所得到的結果$Y(t)$的期望值(也就是我們的預測值)；$\beta_0(t)$、$\beta_1(t)$為變係數函數(相對於一維回歸分析中的係數β_0、β_1)。

將(2)式中的符號以(1)式中的符號做代換的話，我們可以得到以下的模型：

$$E\{Y(\boldsymbol{s},t)|X(\boldsymbol{s},t)\} = \beta_0(\boldsymbol{s},t) + \beta_1(\boldsymbol{s},t)X(\boldsymbol{s},t) \quad (3)$$

而為了找出$\beta_0(t)$、$\beta_1(t)$，我們先分別將X與Y的平均函數(mean finction)以及變異函數(covariance function)定為：

$$u_X(\boldsymbol{s},t) = EX(\boldsymbol{s},t)$$

$$u_Y(\boldsymbol{s},t) = EY(\boldsymbol{s},t)$$

$$G_{XX}\big((\boldsymbol{s_1},t_1),(\boldsymbol{s_2},t_2)\big)$$

$$= cov\{X(\boldsymbol{s_1},t_1),X(\boldsymbol{s_2},t_2)\} = \sum_m \rho_m\, \phi_m(\boldsymbol{s_1},t_1)\phi_m(\boldsymbol{s_2},t_2)$$

$$G_{YY}\big((\boldsymbol{s_1},t_1),(\boldsymbol{s_2},t_2)\big)$$

$$= cov\{Y(\boldsymbol{s_1},t_1),Y(\boldsymbol{s_2},t_2)\} = \sum_k \lambda_k\, \psi_k(\boldsymbol{s_1},t_1)\psi_k(\boldsymbol{s_2},t_2)$$

$$G_{XY}\big((\boldsymbol{s_1},t_1),(\boldsymbol{s_2},t_2)\big)$$

$$= cov\{X(\boldsymbol{s_1},t_1),Y(\boldsymbol{s_2},t_2)\} = \sum_k\sum_m E(\xi_m\zeta_k)\, \phi_m(\boldsymbol{s_1},t_1)\psi_k(\boldsymbol{s_2},t_2)$$

其中ρ_m、λ_k為特徵值(eigenvalues)；ϕ_m、ψ_k為特徵函數(eigenfunctions)；ξ_m、ζ_k為主成分。

有了這些式子，我們就可以重新將$\beta_0(\boldsymbol{s},t)$、$\beta_1(\boldsymbol{s},t)$的定義寫成：

$$\beta_0(\boldsymbol{s},t) = \frac{G_{XY}\big((\boldsymbol{s},t),(\boldsymbol{s},t)\big)}{G_{XX}((\boldsymbol{s},t),(\boldsymbol{s},t))}$$

$$\beta_1(\boldsymbol{s},t) = u_Y(\boldsymbol{s},t) - \beta_0(\boldsymbol{s},t)u_X(\boldsymbol{s},t)$$

再分別將每一項估計出來，就可以將$\beta_0(\boldsymbol{s},t)$、$\beta_1(\boldsymbol{s},t)$代回模型，完成$X(\boldsymbol{s},t)$對$Y(\boldsymbol{s},t)$的關係式，並用於預測。

而為了讓我們的空氣品質數據(三維)可以使用此模型，我們必須對此模型作延伸。

由 Hsu[3]及 Chen and Jiang[4]的文章中，我們可以得知當$X(\boldsymbol{t})$為多維度資料時，只要分別在估計$u(\boldsymbol{s},t)$、$G_{XX}(\boldsymbol{s},t)$、ϕ_m、ψ_k、ξ_m、ζ_k時皆把為$\boldsymbol{s}$、t以向量的型式來表達使用，就可以在合理的情況下把模型擴展到多維度上。

第三章 資料分析

3.1 延伸模型模擬

為了測試(2)式是否可以在多維度的情況之下運行，我們使用 Python 來進行模擬，並使用 Hsu[3]的套件來做為估計工具。套件中的 Lpr 函式為局部線性迴歸的實現。pca 類別為主成份分析的實現。其平均函數、共變異函數和變異數估計的部分皆套用 Lpr 函式，最後重建$X(\boldsymbol{t})$則由 Resturct_Fun 函式實現。

在模擬中，模型的相關設定如下：

$$U_{ij} = X_i(\boldsymbol{t}) + \epsilon = u(\boldsymbol{t}) + \sum_{k=1}^{2} \xi_{ik}\phi_k(\boldsymbol{t}) + \epsilon$$

$$V_{ij} = Y_i(\boldsymbol{t}) + \epsilon$$

$$Y_i(\boldsymbol{t}) = \beta_0(\boldsymbol{t}) + \beta_1(\boldsymbol{t})X_i(\boldsymbol{t}) = 2t^2 + 0.5tX_i(\boldsymbol{t})$$

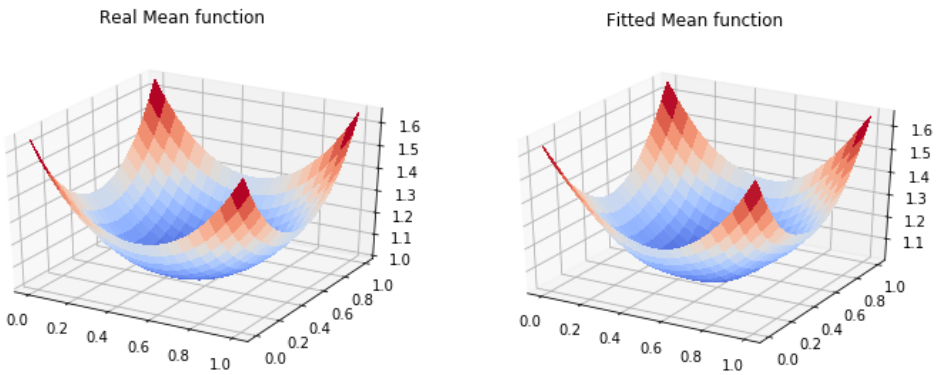
其中

$$u(\boldsymbol{t}) = \exp\{(\boldsymbol{t} - 0.5)'(\boldsymbol{t} - 0.5)\}$$

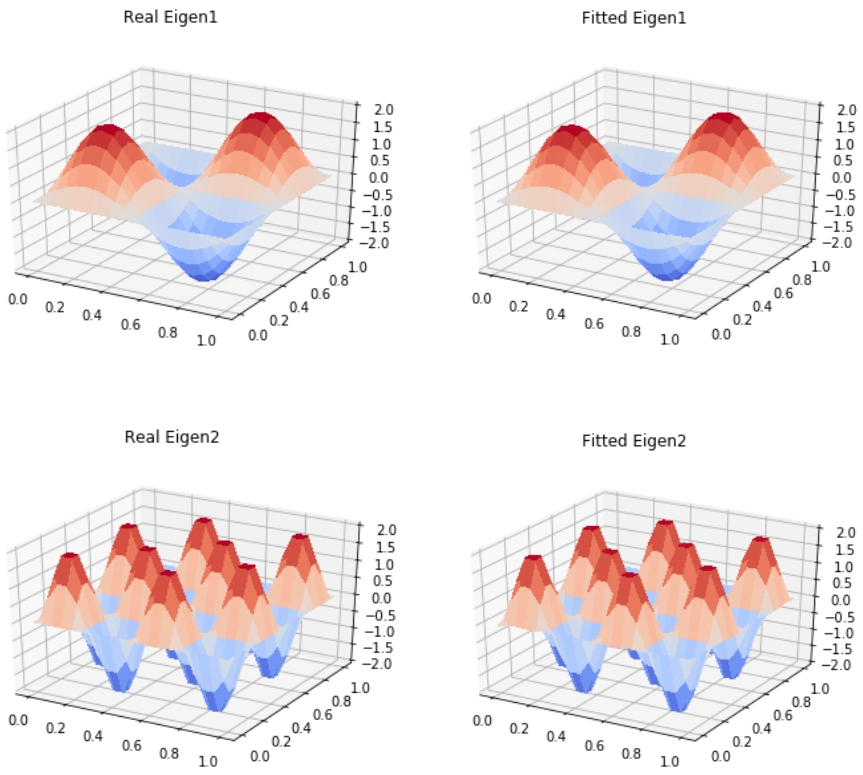
$$\phi_k(\boldsymbol{t}) = \prod_{i=1}^{2}\{\frac{\sin(2k\pi t_i)}{2}\}$$

$$\beta_0(\boldsymbol{t}) = 2t^2 \quad ,\beta_1(\boldsymbol{t}) = 0.5t$$

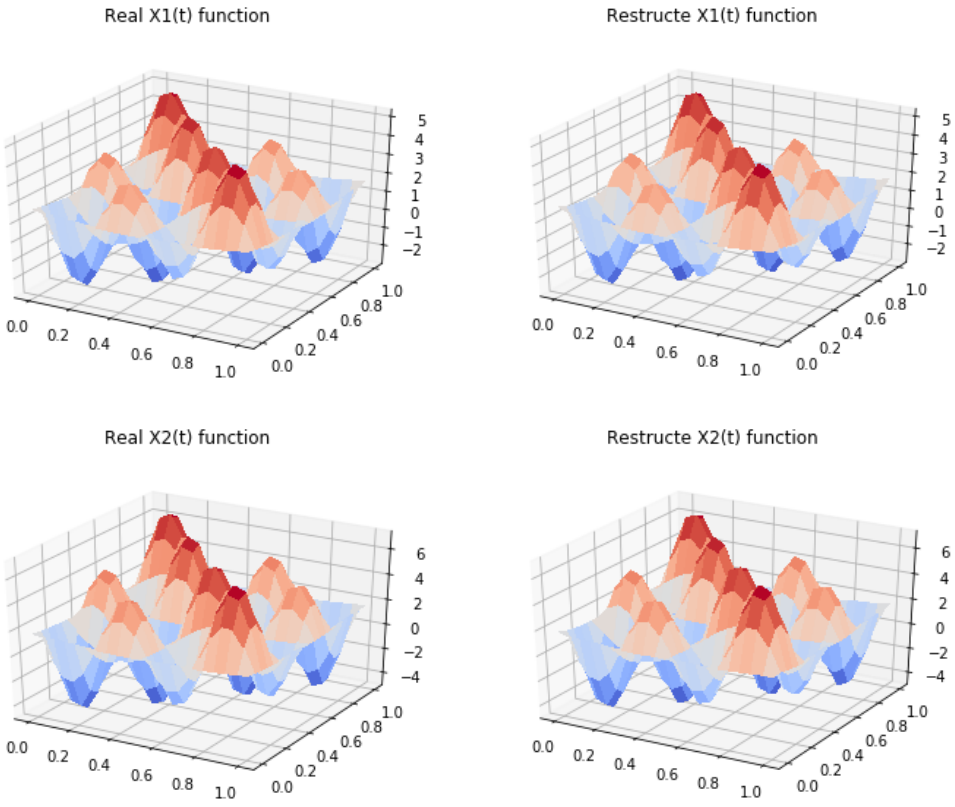
而做出來的結果，真實的與估計的$u_X(\boldsymbol{t})$如下：



真實的與估計的$\phi_k(t)$如下：

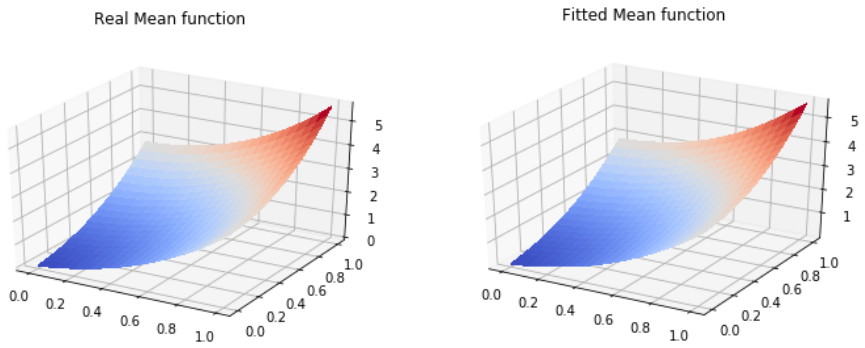


真實的與重建的$X(t)$如下：

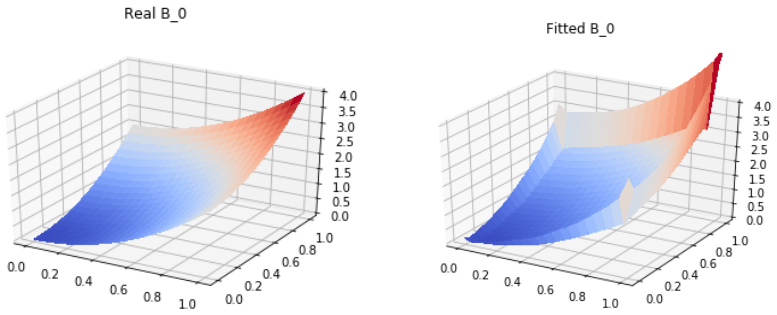


$X(t)$的 MISE 誤差為：0.0016

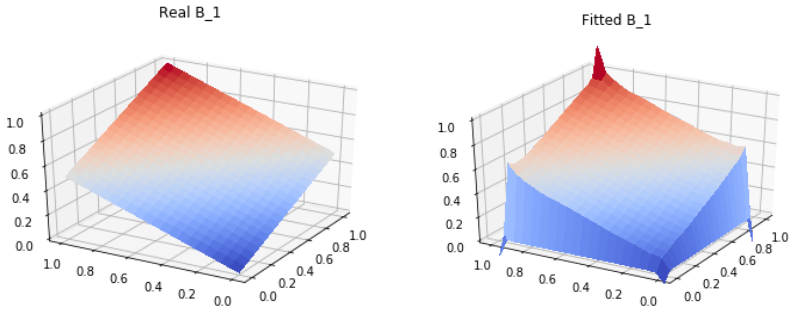
真實的與估計的$u_Y(t)$如下：



真實的與估計的β_0如下：



真實的與估計的β_1如下：



可以發現估計的β_0及β_1在靠近邊界的地方較容易出現誤差，而在非界的地方則是估計的與原函數幾乎相同，而這也符合我們使用的局部多項式回歸估計法的特性。

而β及$\hat{Y}$的各項誤差表現如下：

	β_0	β_1	β	$\hat{Y}$
MADE	0.010062	0.058932	0.068994	0.184923
WASE	0.002010	0.063109	0.065119	:
UASE	0.032162	0.015777	0.047939	

而若以Y_{mean}來取代$\hat{Y}$進而估算誤差的話，MADE 為 0.198868，大於我們用$\hat{Y}$估出來的 0.188521。由此可知，本方法具有處理多維度資料的能力，以及確實有可能做出更高的準確度。

3.2 實際分析資料

(進行中，待用環工系 PM2.5 濃度資料模擬完成後會放在我的 github 上；https://github.com/ChouHsuan-Cheng)

第四章 結論

(進行中，待用環工系 PM2.5 濃度資料模擬完成後會放在我的 github 上；https://github.com/ChouHsuan-Cheng)

參考文獻

[1] Veronica J. Berrocal , Alan E. Gelfand , and David M. Holland. Space-Time Data fusion Under Error in Computer Model Output: An Application to Modeling Air Quality.

[2] Damla¸SENTÜRK and Hans-Georg MÜLLER. Functional Varying Coefficient Models for Longitudinal Data.

[3] Chai-Yung Hsu. A Python Package for Fast Algorithm of Multi-dimensional FunctionalPrinciple Component Analysis.

[4] Lu-Hung Chen and Ci-Ren Jieng. Muti-dimensional functional principal component analysis. Statistic and Computing , 27(1181-1192) , 2017.