財團法人台北外匯市場發展基金會

「由媒體報導文字資訊預測 GDP」

結案報告

委託單位：財團法人台北外匯市場發展基金會

承辦單位：國立中央大學

中華民國 108 年 11 月 30 日

目錄

圖目錄

表目錄

1. 前言

隨著科技的進步，現今資料型態已不僅侷限在結構化的數字資料，舉凡文字、圖片、影像、音樂、網頁等非結構化資料，亦廣泛應用於人文社會科學領域。尤其因為人們的情感表達、溝通對話、展望預期與文化法規大多以文字表達，故近年來文字探勘 (text mining) 的分析方法相當受到學界、企業與政策制定者的重視。文字探勘主要是利用統計方法來處理高維的 (high-dimensional) 文字資料，並發掘文件檔案 (document) 中所隱含的資訊。其原屬於自然語言處理 (natural language processing) 領域，但因各種演算法與機器學習方法快速進步，使得愈來愈多的研究開始運用文字探勘技術。

若以文字探勘方法應用於經濟、金融、財務與會計相關議題為例，Armestro et al. (2009)，Jubinski and Tomljanovich (2013)，Sadique et al. (2013)，Balke et al. (2016)，Hansen et al. (2018)，黃裕烈和管中閔 (2019)，Huang and Kuan (2020) 等利用文字探勘技巧分析美國聯邦儲備系統的各類文件，並說明其如何影響或預測總體經濟活動。Tetlock (2007)，Tetlock et al. (2008)，Hoberg and Phillips (2010)，Loughran and McDonald (2011, 2014)，Jegadeesh and Wu (2013)，Manela and Moreira (2015) 則利用此方法分別探討媒體文字與公司年報在金融市場、企業財務與會計等不同議題的應用。Baker et al. (2016)，黃裕烈等 (2019) 透過關鍵字搜尋新聞報導中同時涵蓋經濟、政策與不確定性之三大類別文字，建構衡量經濟政策不確定性 (economic policy uncertainty, EPU) 指標。Hoberg and Phillips (2016) 與 Kelly et al. (2018) 則將文字探勘應用於產業商品分類與專利品質衡量。其他關於文字探勘文獻的綜合論述可參考 Loughran and

McDonald (2016) 與 Gentzkow et al. (2019)。

即時瞭解當前與未來的景氣狀態，對於執行有效的商業決策和經濟政策極為重要，在所有衡量總體經濟活動的變數中，又以國內生產毛額 (gross domestic product, GDP) 的估算與預測最受政策制定者關注。然而，以往經濟學家或政府單位所建構的 GDP 計量模型，通常只蒐集數量資料來預測經濟成長，但隨著文字探勘技術逐漸發展，近期已有文獻特別考慮非結構化的新聞媒體資訊，探討其對總體經濟活動的預測績效。例如，Ardia et al. (2019)，Feuerriegel and Gordon (2018)，Thorsrud (2018) 等分別利用不同的文字探勘與統計降維 (dimension reduction) 方法，針對總體經濟指標的當前情勢與未來動向進行即時預報 (nowcasting) 或預測 (forecasting)，這些研究結果均發現新聞資訊有助於研判當前景氣狀態或預測未來 GDP 變動。

本計畫主要目的為利用文字探勘方法，分析巨量的媒體報導資訊，探討其是否對台灣 GDP 季成長率具有預測能力。然而，若要達成此計畫目的必須解決兩項主要問題：首先，由於媒體文件基本上屬於高維資料，必須將原始文本轉換成結構化資料，因此我們先透過不同的特徵選取 (feature selection) 或降維統計方法對文字資料進行處理。在特徵選取方法上，本計畫利用中文斷詞 (tokenize)、移除停用詞 (stop words)、詞袋 (bag of words) 分析等，刪除無效與無意義詞彙，篩選出新聞文件中的關鍵詞彙。然而，必須強調的是：在處理媒體文字時，本計畫特別融入經濟相關的領域知識 (domain knowledge)，整合經濟預測常用的專業術語，避免發生錯誤估算關鍵詞頻或誤解文章語意的情況，提升文字資料的資訊含量，使文字探勘更具合理性。雖然透過特徵選取找出關鍵詞彙，可以有效降低文字資

料維度，但此資料量仍然非常龐大，無法提供當前或未來經濟情勢明確的訊息。因此本計畫嘗試將關鍵詞彙區分成不同類別，運用統計方法編制新聞擴散指標 (news-based diffusion indexes，NDI)，並以適當的計量模型探討 NDI 對台灣 GDP 成長率的預測表現。

其次，由於每天都有許多篇關於台灣經濟現況或展望的新聞報導，故本計畫所編制的經濟新聞擴散指標之最高頻率為日資料。然而，實質 GDP 成長率卻是季或年資料，且主計總處通常晚 1-2 月才公布最新一季經濟成長率的統計初估值。因此如何善用混合頻率 (mixed frequency) 資料，針對台灣經濟成長率進行即時的預報或預測，亦是本計畫必須解決的問題。

另一方面，為了釐清文字資料是否單獨具有預測能力，我們必須控制數量資料的影響效果，而以往文獻亦發現：同時考慮日、月、季不同頻率的數量資料，可提升對總體經濟變數的預測表現，例如 Aruoba et al. (2009)，Marcellino and Schumacher (2010)，Andreou et al. (2011)，Banbura et al. (2013)，Schorfheide and Song (2015)，黃裕烈等 (2017)。因此本計畫將參考這些文獻，嘗試結合大量文字與數字的混合頻率資料，並利用 Stock and Watson (2002a, 2002b, 2011) 的動態因子模型 (dynamic factor model)，期望能夠即時地預測當前與未來的總體經濟動向。

本計畫的主要研究步驟可歸納如下：從每日媒體文字資料中，利用文字探勘技術，萃取出當前與未來總體經濟活動的資訊，透過適當的計量方法編制新聞擴散指標，並結合大量混合頻率的總體經濟數據，對台灣 GDP 季成長率進行即時預報，同時探討媒體文字資料是否具有預測能力。在預測模型選擇上，本計畫主要參考混合頻率的動

態因子模型。由於本計畫處理的總體變數資料相當龐大，必須透過因子模型予以適度地降維，且因應每日隨時更新的媒體文字與經濟數據，混合頻率模型將可提供更即時的經濟預測。

根據本計畫所編制的新聞擴散指標，我們發現：無論是 NDI 或是構成 NDI 的 6 項文字指標均與領先指標的相關性最高，顯示本計畫所建構的媒體文字指標領先反應總體經濟活動的變化趨勢。雖然 NDI 的波動較同時指標與 GDP 成長率劇烈，但 NDI 的高峰與谷底大致領先國發會所認定的景氣峰谷時點，因此其走勢亦可即時 (real time) 反應台灣 GDP 變化與景氣波動的方向。

在混合頻率模型的樣本外預測表現方面，本計畫實證結果發現：若只考慮媒體文字資訊進行預測，其不一定優於基準模型 (一階自我迴歸模型)。但若加入 GDP 成長率落後項，則可提升其預測能力。然而，最佳預測模型為結合新聞擴散指標的動態因子模型，即同時考慮媒體文字、與 GDP 相關的大量總體經濟數據、以及 GDP 成長率落後項，其表現優於只考慮經濟數據的動態因子模型。此外，由於媒體文字資訊波動較大，若將新聞擴散指標適度平滑化，再加入預測模型中，則可提升媒體文字動態因子模型的預測績效。綜合而言，本計畫所編制的新聞擴散指標有助於即時預測台灣 GDP 成長率，並幫助政府及企業決策者正確研判當前與未來的總體經濟動向。

本計畫內容安排如下：第 2 節說明如何建構新聞擴散指標；第 3 節概述本計畫所採用的混合頻率預測模型；第 4 節為總體資料與新聞擴散指標的基本敘述統計分析；第 5 節為不同計量模型的估計結果，以及樣本外預測表現評估；最後一節為本計畫之結論。

2. 新聞擴散指標建構

由於媒體文字屬於高維資料，假設某一篇新聞報導共有 w (如 100) 個字，而常用中文字有 C (如 5000) 個字，則此新聞報導的資料維度可表示為 C^w (5000^{100})，因此我們可以想像每天與經濟相關的文字資料維度將是非常龐大。若要從巨量的文字資料中萃取出有用的資訊，並利用此資訊進行量化分析，本計畫將採取以下三個步驟：

1. 透過不同的特徵選取或 N 元語法 (N-grams)，將原始文本轉換成結構化資料，篩選出新聞文件中的關鍵符號 (token，如詞彙、片語等)，並進行語義分析 (semantic analysis)。

2. 由於從媒體文字中所萃取出的重要詞彙與語意仍非常龐大，本計畫將利用主成分分析 (principal component analysis, PCA) 或因子分析 (factor analysis) 進行維度縮減 (dimension reduction)，並編制新聞擴散指標。

3. 比較新聞擴散指標與台灣總體經濟變數間的關係，並利用 NDI 建構包含文字與數字資料的混合頻率動態因子模型，對台灣 GDP 成長率進行預測。

2.1 關鍵詞彙與正負語義指標建構

本計畫首先利用字典法 (dictionary-based methods)，針對媒體關鍵文字進行初步分析。我們搜尋台灣主要 4 大報紙的網路新聞，「工商時報」、「蘋果日報」、「中國時報」、「自由時報」等，由於著作權法的關係，本計畫透過撰寫網路爬蟲相關程式，直接從各網頁蒐集歷年每日新聞中符合臺灣經濟成長相關的關鍵詞彙，並將非結構性的文字

資料轉換成統計數據。

在關鍵詞彙選取方面，本計畫結合文字探勘技術與經濟情勢分析的領域知識。我們先蒐集近一年政府相關部門分析未來經濟情勢展望的新聞稿，其中包含主計總處、國家發展委員會、經濟部與財政部等，透過初步的文字雲找出重要的關鍵詞彙 (請參考圖 1-4)。必須特別說明是：由於本計畫主要目的是預測實質 GDP 成長率，雖然金融面總體變數 (如利率、匯率、股價等) 亦會影響實體經濟成長率，但因其傳遞管道較為間接，故我們暫不考慮金融層面的媒體文字訊息，而此議題可以留待後續研究方向。

除利用文字雲找出重要的關鍵詞彙外，本計畫也根據專家意見，利用經濟相關的領域知識，統一新聞文本中的專業術語，以研判 GDP 預測所需的文字資訊。此外，我們參考 Ardia et al. (2019) 與 Thorsrud (2018) 的作法，將與總體經濟活動息息相關的媒體文字劃分成不同的類別或主題 (topics)，其中包括經濟 (GDP) 成長、民間消費、固定投資、出口貿易、外銷訂單、景氣概況、工業生產、批發、零售與餐飲銷售等，並分別針對不同類別訂定關鍵詞彙，以萃取出預測當前或未來經濟情勢的文字資訊。

由於本計畫必須研判未來 GDP 走勢，故對總體經濟活動變化方向的描述，無論在新聞語意分析或擴散指標編制上均至為關鍵。因此我們針對各主題區分出正面詞彙 (如「上升」、「復甦」、「上修」、「樂觀」、「增加」等) 與負面詞彙 (如「下降」、「衰退」、「下修」、「悲觀」、「減少」等)，[1]透過文字探勘技術，我們可以揭露不同新聞主題所持之態度，並建構相關正負語意的指標。

[1] 由於不同主題的關鍵詞彙與正、負面語詞過於繁多，受制於論文的篇幅有限，無法完整列出所有詞彙。

圖 1：主計總處經濟情勢展望新聞稿文字雲

圖 2：國發會景氣概況新聞稿文字雲

圖 3：財政部海關進出口貿易統計新聞稿文字雲

圖 4：經濟部工業生產統計新聞稿文字雲

本計畫蒐集 2003 年 5 月 1 日起至 2018 年 12 月 31 日止各大報紙的網路新聞,[2]並事先排除與經濟不相關的新聞分類 (如娛樂新聞、社會新聞、運動新聞、廣告等)，總共搜尋的新聞篇數為 1,745,703 篇。我們將一篇經濟新聞分成數個段落 (paragraph)，並以段落為主要對象進行語意分析。針對某一主題，利用文字探勘技術 (如 *N*-grams) 與 R 電腦程式的 Jieba 套件，將新聞內容斷詞，並移除停用詞 (如數字、標點符號、啊、嗎等) 與刪除無意義詞彙，找出段落中所包含語意的關鍵文字，將其區分正面、負面或中性段落。若某一主題新聞多數段落都是正 (負) 面內容，則該篇新聞計為對經濟情勢的正 (負) 面報導。[3]因此，我們可以計算第 t 天第 i 個主題的正/負面新聞篇數差距 ($PN_{i,t}$)，當作衡量總體經濟活動的指標：

$$PN_{i,t} = \text{正面篇數}_{i,t} - \text{負面篇數}_{i,t}$$

至此已將非結構的文字資訊轉換成可分析的數字資訊。然而，由於每天與經濟相關的新聞篇數都不一致，不同天的 $PN_{i,t}$ 並不適合比較，因此本計畫將以正反面篇數比例差距 ($Tone_{i,t}$)，作為衡量或預測經濟情勢的指標。即

$$Tone_{i,t} = \frac{\text{正面篇數}_{i,t} - \text{負面篇數}_{i,t}}{\text{正面篇數}_{i,t} + \text{負面篇數}_{i,t}} \times 100 \qquad (1)$$

當 $Tone_{i,t} > 0$ (< 0) 時，表示對第 i 個主題持正面 (負面) 態度。我們

[2] 由於 2019 年部分網路新聞登入方式更改，故本計畫搜尋資料期間至 2018 年底。

[3] 由於中文博大精深，以中文進行語意分析仍可能面臨相當的衡量誤差(特別在少數否定字的處理)，故需對程式進行校正與改進，盡可能降低語意衡量的誤差。然而，少數語句的衡量誤差對該新聞正、負面判別之影響應屬輕微。

也可以將某一主題每天的正/負面新聞篇數加總，建構月資料的 $PN_{i,t}$ 與 $Tone_{i,t}$ 的文字指標。

此外，為了避免發生錯誤估算關鍵詞頻或誤解文章語意的情況，Baker et al. (2016) 建議以人工審查方式來判斷新聞內容，但礙於人力與時間有限，本計畫除抽 20% 樣本，以人工方式進行局部審核內容外，還參考黃裕烈與管中閔 (2019) 所提出的 MAP-PLSA 模型協助篩選，並判斷新聞資訊是否與臺灣經濟情勢相關，更詳細的說明請見黃裕烈、管中閔 (2019) 與黃裕烈等 (2019)。

2.2 主成分分析與新聞擴散指標

為了避免某日特定主題的新聞篇數過少，我們將衡量總體經濟活動的文字指標 ($Tone_{i,t}$) 劃分成六項 ($i = 6$)，即經濟 (GDP) 成長文字指標、景氣動向文字指標、民間消費文字指標、固定投資文字指標、生產銷售文字指標、出口暨訂單文字指標等。然而，如何將這六項文字指標透過合理的統計方法加權成為單一新聞指標，文獻上有許多不同的作法，本計畫採主成分分析 (或因子分析) 方法建構日資料的新聞擴散指標 (news-based diffusion indexes，NDI)，以衡量當前與預測未來的總體經濟情勢。

Stock and Watson (2002a, 2002b) 利用主成分分析方法，從一百多個總體變數中，萃取出主要因子當作指標變數，再利用這些指標變數建立一個簡單線性預測模型，以預測美國未來經濟走勢。徐之強和黃裕烈 (2005) 運用 Bai and Ng (2002) 的因子數目選擇準則，從台灣眾多總體經濟變數中，萃取出重要的共同移動因子，以捕捉景氣循環波

動。本計畫為比較文字資料與數字資料的預測能力，我們都採用主成分分析來降低文字與數字的資料維度，並從中萃取出最重要的因子，分別建構新聞與數據的擴散指標。

我們先簡要說明因子模型的設定與估計方式。若以矩陣方式表示的因子模型如下：

$$\boldsymbol{X} = \boldsymbol{F}\boldsymbol{\Lambda}' + \boldsymbol{\varepsilon}, \qquad (2)$$

其中 $\boldsymbol{X}$ 是 T 期 N 個文字指標所構成的矩陣，$\boldsymbol{X} = (X_1^{'}, \cdots, X_T^{'})$ 為 $T \times N$ 的資料矩陣；$\boldsymbol{F}$ 為 $T \times r$ 的因子矩陣；$\boldsymbol{\Lambda}$ 為 $N \times r$ 維度的因子負載 (factor loading) 矩陣，可視為因子對於 $\boldsymbol{X}$ 的影響程度；$\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_N)'$ 為 $T \times N$ 的干擾項矩陣。因此，$\boldsymbol{X}$ 變動可分成兩個部份，一部份為個別變動 (specific variation) 造成，即 $\boldsymbol{\varepsilon}$；另一則是共同變動 (common variation) 部份，即 $\boldsymbol{F}$，以刻劃共移 (comovement) 現象。

然而，在估計因子模型參數 $\boldsymbol{F}$ 與 $\boldsymbol{\Lambda}$ 的過程中，將會產生認定問題。例如，對任何一個 $r \times r$ 的可逆矩陣 $\boldsymbol{A}$ 而言，

$$\boldsymbol{F}\boldsymbol{\Lambda}' = \boldsymbol{F}\boldsymbol{A}\boldsymbol{A}^{-1}\boldsymbol{\Lambda}' = \ddot{\boldsymbol{F}}\ddot{\boldsymbol{\Lambda}}',$$

其中 $\ddot{\boldsymbol{F}} = \boldsymbol{F}\boldsymbol{A}$ 且 $\ddot{\boldsymbol{\Lambda}}' = \boldsymbol{A}^{-1}\boldsymbol{\Lambda}'$，將此代入 (2) 式，仍然會得到相同的因子模型結構 $\boldsymbol{X} = \ddot{\boldsymbol{F}}\ddot{\boldsymbol{\Lambda}}' + \boldsymbol{\varepsilon}$。換言之，$\boldsymbol{F}$ 與 $\boldsymbol{\Lambda}$ 在估計上產生認定問題。故估計因子模型時，必需加入參數限制條件，才能估算出固定的 $\boldsymbol{F}$ 與 $\boldsymbol{\Lambda}$。以 $r \times r$ 的矩陣 $\boldsymbol{A}$ 而言，內有 r^2 個待估參數，因此我們需要 r^2 條限制式。若我們要求 $\boldsymbol{F}$ 正規化 (normalization)：

$$\frac{\boldsymbol{F}'\boldsymbol{F}}{T} = \boldsymbol{I}_r \text{ ，}$$

則要求了 $r(r+1)/2$ 條限制式。若進一步要求 $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ 必需是一個對角線矩陣 (diagonal matrix)，則又限制了 $r(r-1)/2$ 個條件。若將此二假設所得的限制式加總起來，我們可得到 r^2 條限制式。在這些限制條件下，利用最小平方估計式估算模型參數 $\boldsymbol{F}$ 時，恰好是主成份分析的結果。更詳細的說明請參考附錄 I 或 Stock and Watson (2002a, 2002b), Bai and Ng (2002)。

本計畫利用主成份分析方法，針對不同主題的文字指標，求出影響這些指標的最重要因子，建構出新聞擴散指標。由於日資料的新聞擴散指標波動過於劇烈，較不適合應用於預測 GDP 成長率，故本計畫主要利用月資料的新聞擴散指標，並結合相同頻率的數據資料，以混合頻率動態因子模型來預測 GDP 季成長率。

3. 混合頻率預測模型

由於新聞擴散指標與多數實體經濟指標 (如工業生產指數、海關進出口等) 均為月資料，但 GDP 成長率則為季資料，因此為了探討媒體文字資訊是否對台灣 GDP 季成長率具預測能力，我們必須建立月、季混合頻率資料的預測模型。混合頻率模型優點在於直接納入所有不同頻率的資料，避免因將高頻資料轉換成低頻資料的處理，而忽略了高頻資料中可能蘊藏的有用訊息。此外，混合頻率資料預測模型更可進行即時預報 (nowcasting)，因新的統計調查數據與新聞擴散指標會逐月公布，我們可即時更新模型中的資料與參數估計值，並立刻修正模型預測結果，此優點能協助政府決策者提早掌握經濟動向，進而即時採取適當的因應措施。

3.1 混合頻率模型基本設定

本計畫將以 GDP 季成長率的自我迴歸 (autoregressive, AR) 模型當作基準 (benchmark) 預測模型：

$$y_{t+1} = \alpha + \sum_{i=0}^{q} \beta_i y_{t-i} + \varepsilon_{t+1}, \qquad (3)$$

其中 y_t 為 GDP 季成長率。此模型只考慮季資料訊息，以 GDP 成長率的落後項當作預測變數。

本計畫採用混合頻率資料抽樣（mixed-data sampling, MIDAS）模型為主要的預測模型。MIDAS 模型是由 Ghysels et al. (2004) 所提出，其概念來自分配遞延模型 (distributed lag model)，模型特色主要

透過迴歸參數的多項式權重設定，直接建構高頻與低頻變數之間的關係，無須處理不同頻率間的資料轉換。詳細 MIDAS 模型的綜合論述可參考 Andreou et al. (2011)。

為了探討新聞擴散指標的樣本內配適與樣本外預測表現，我們先建立以下 MIDAS 預測模型：

$$y_{t+1}^{Q} = \alpha + \beta y_{t}^{Q} + \sum_{j=0}^{m-1} \gamma_j NDI_{m-j,t}^{M} + \varepsilon_{t+1}, \qquad (4)$$

其中 NDI 屬於高頻 (月) 資料，每季有 $m=3$ 個月。MIDAS 模型假設高頻資料可以透過多項式函數加總為低頻資料，即

$$W(L^m;\theta^M)NDI_t^M = \sum_{j=0}^{m-1} w_j(\theta^M)NDI_{t-j}^M$$

而 $w_j(\theta^M)$ 多項式函數的設定有許多不同形式，如指數 Almon 多項式 (exponential Almon lag)：

$$w_j(\theta_1,\theta_2) = \frac{\exp\{\theta_1 j + \theta_2 j^2\}}{\sum_{j=1}^{m}\exp\{\theta_1 j + \theta_2 j^2\}},$$

或 Beta 多項式 (Beta lag)：

$$w_j(\theta_1,\theta_2) = \frac{f(j,\theta_1;\theta_2)}{\sum_{j=1}^{m} f(j,\theta_1;\theta_2)},$$

其中 $f(j,\theta_1;\theta_2) = j^{\theta_1-1}(1-j)^{\theta_2-1}\Gamma(\theta_1+\theta_2)/(\Gamma(\theta_1)\Gamma(\theta_2))$，$\Gamma(\theta_1)$ 為 Gamma 函數。

若 GDP 季成長率具有序列相關，我們可考慮動態設定，加入自我迴歸分配遞延項 (autoregressive distributed lag, ADL)，並建立以下

ADL-MIDAS 模型：

$$y_{t+1}^{Q} = \alpha + \sum_{j=0}^{q_y-1} \beta_{j+1} y_{t-j}^{Q} + \gamma \sum_{j=0}^{q_d-1} \sum_{i=0}^{m-1} w_{i+j*m}(\theta^{M}) NDI_{m-i,t-j}^{M} + \varepsilon_{t+1} \qquad (5)$$

ADL-MIDAS 模型的優點是：所有參數皆可透過非線性最小平方法 (nonlinear least squares, NLS) 估計。若樣本內係數 $w_j(\theta^M)$ 估計值顯著異於零，且樣本外預測表現優於 AR 模型，則我們可以推論媒體文字資訊有助於預測未來台灣經濟成長率。

3.2 混合頻率動態因子模型

本計畫除了探討新聞報導的預測能力 (即考慮最基本的 ADL-MIDAS 模型) 外，亦參考 Stock and Watson (2002a, 2002b, 2011) 所提出的動態因子模型。我們將蒐集大量實體經濟的統計數據，利用維度縮減方法，從中萃取出重要的共同變動因子，在控制數字資料的影響下，才能正確分析媒體文字資訊是否有助於預測台灣未來的經濟成長率。

據此，我們提出混合頻率資料的動態因子模型。考慮因子模型有二大好處：一是預測模型可以考量更多的月頻率資訊；二是待估參數數目不會因為加入更多的月頻率變數而膨脹太快。本計畫參考 Marcellino, et al. (2016) 的動態因子模型，建立以下結合文字與數字資料的混合頻率動態因子 (FADL-MIDAS) 模型：

$$\begin{aligned} y_{t+1}^{Q} = \alpha + \sum_{j=0}^{q_y-1} \beta_{j+1} y_{t-j}^{Q} + \gamma \sum_{j=0}^{q_d-1} \sum_{i=0}^{m-1} w_{i+j*m}^{a}(\theta^{M}) NDI_{m-i,t-j}^{M} \\ + \delta \sum_{j=0}^{q_d-1} \sum_{i=0}^{m-1} w_{i+j*m}^{b}(\lambda^{M}) F_{m-i,t-j}^{M} + \varepsilon_{t+1}, \end{aligned} \qquad (6)$$

其中 F 為從大量總體資料中所萃取出的最重要因子，可利用第 2.2 節與附錄 I 所介紹的主成分分析法從月資料中得出，而且我們假設 F 存在序列相關，故在迴歸中加入其落後項，並以 NLS 進行估計。若樣本內係數 $w_j^a(\theta^M)$ 估計值顯著異於零，且其樣本外預測表現優於 (5) 式，則我們可以推論：在控制統計數據因子的影響下，媒體文字資訊仍有助於預測台灣未來的經濟成長率。

由於月資料公布時間早於季資料，因此只要獲得新的月資料訊息，我們可立即估計 MIDAS 模型參數，並對當季 GDP 成長率進行即時的 nowcasting，故我們所利用的資料訊息集合，將不僅限於月頻變數的落後項，亦可考慮將月頻變數的領先項 (lead) 加入 MIDAS 迴歸模型中。因此，本計畫將建立以下加入領先項的混合頻率動態因子模型 (FADLL-MIDAS)：

$$\begin{aligned} y_{t+1}^Q = \alpha &+ \sum_{j=0}^{q_y-1} \beta_{j+1} y_{t-j}^Q \\ &+ \gamma \left[\sum_{i=1}^{l} w_i^a(\theta^M) NDI_{i,t+1}^M + \sum_{j=0}^{q_d-1} \sum_{i=0}^{m-1} w_{i+j*m}^a(\theta^M) NDI_{m-i,t-j}^M \right] \\ &+ \delta \left[\sum_{i=1}^{l} w_i^b(\lambda^M) F_{i,t+1}^M + \sum_{j=0}^{q_d-1} \sum_{i=0}^{m-1} w_{i+j*m}^b(\lambda^M) F_{m-i,t-j}^M \right] + \varepsilon_{t+1} \end{aligned} \quad (7)$$

詳細的說明請參考 Giannone, et al. (2008)。同理，若 (7) 式的樣本外預測表現優異，則新聞報導資訊將有助於預測台灣 GDP 成長率。

4. 資料分析與敘述統計

本計畫蒐集 2003/5/1-2018/12/31 報紙網路新聞，以編制六項文字指標，並利用主成分分析法建構新聞擴散指數。為了避免期初值的影響，以及統一數字資料的起訖期間，本計畫實證分析的月資料期間從 2004 年 1 月到 2018 年 12 月，季資料期間為 2004 年第 1 季到 2018 年第 4 季。在數字資料方面，我們主要蒐集 34 個台灣總體經濟月頻變數的資料，並對季頻的 GDP 成長率進行預測。由於台灣目前尚未建置不同時點的即時 (real-time) 資料庫，故所有數字資料皆屬於修正後的資料。本計畫資料皆取自「主計處中華民國統計資訊網—總體統計資料庫」及「AREMOS 台灣經濟統計資料庫」，詳細的月頻變數選取請參考附錄 II。

在數字資料處理方面，總體經濟變數大多會受到季節性因素的影響，因此本研究以 X-12 ARIMA 方法對總體經濟變數進行季節性調整，以去除季節性變動。由於本計畫主要目的是預測經濟成長率，故我們先對總體變數水準值取對數，再進行單根檢定，驗證該變數是否為定態數列，並按照其結果決定是否需要差分。若總體變數為比率資料，則不對其取對數，直接進行單根檢定，也按照檢定結果決定是否差分。詳細的資料處理方式請參考附錄 II。在主成分分析中，為了避免因單位不同造成估計上的偏誤，我們必須先將 34 個總體變數經過定態處理，再予以標準化，找出最重要的數字資料因子，並比較其與新聞擴散指標的預測能力。

表 1 為新聞擴散指標與構成其六項文字指標的基本統計量，包括 GDP 成長文字指標、景氣動向文字指標、民間消費文字指標、固定投資文字指標、生產銷售文字指標、出口暨訂單文字指標等，每項

指標在實證分析期間共有 180 個觀察值。因所有媒體文字指標的中位數與平均數大於零，故樣本期間正面總體經濟新聞報導的比例多於負面報導的比例。新聞擴散指標與所有類別文字指標的偏態係數均為負，且除景氣動向文字指標外，其峰度值均大於 3，顯示媒體文字指標為負偏態與高峽峰 (leptokurtic) 分配。而 Jarque-Bera 統計量亦顯示：新聞擴散指標與分項文字指標均拒絕常態分配的虛無假設。

表 1 ：新聞擴散指標與其構成項的基本統計量

	生產銷售文字指標	景氣動向文字指標	固定投資文字指標	GDP 成長文字指標	出口暨訂單文字指標	民間消費文字指標	NDI
平均值	0.57	0.33	0.65	0.58	0.66	0.72	0.00
中位數	0.60	0.37	0.71	0.64	0.69	0.76	0.54
標準差	0.25	0.31	0.26	0.18	0.13	0.18	2.05
偏態	-1.00	-0.54	-1.20	-1.63	-1.23	-2.18	-1.62
峰度	4.77	2.33	4.11	5.76	5.18	8.60	6.11
Jarque-Bera 統計量	53.20	12.09	52.34	136.49	81.14	378.45	151.79
p 值	0.00	0.00	0.00	0.00	0.00	0.00	0.00
樣本數	180	180	180	180	180	180	180

註：本研究自行整理

圖 5 為 6 分項文字指標與 6 個重要總體經濟變數的熱力圖 (相關係數矩陣)。從深色區塊可以得知：媒體文字指標彼此之間的相關性較高，而統計數據指標彼此之間的相關性亦高。值得注意是：6 分項文字指標與領先指標的相關性最高，顯示我們所建構的各分類媒體文字指標領先反應總體經濟活動變化。

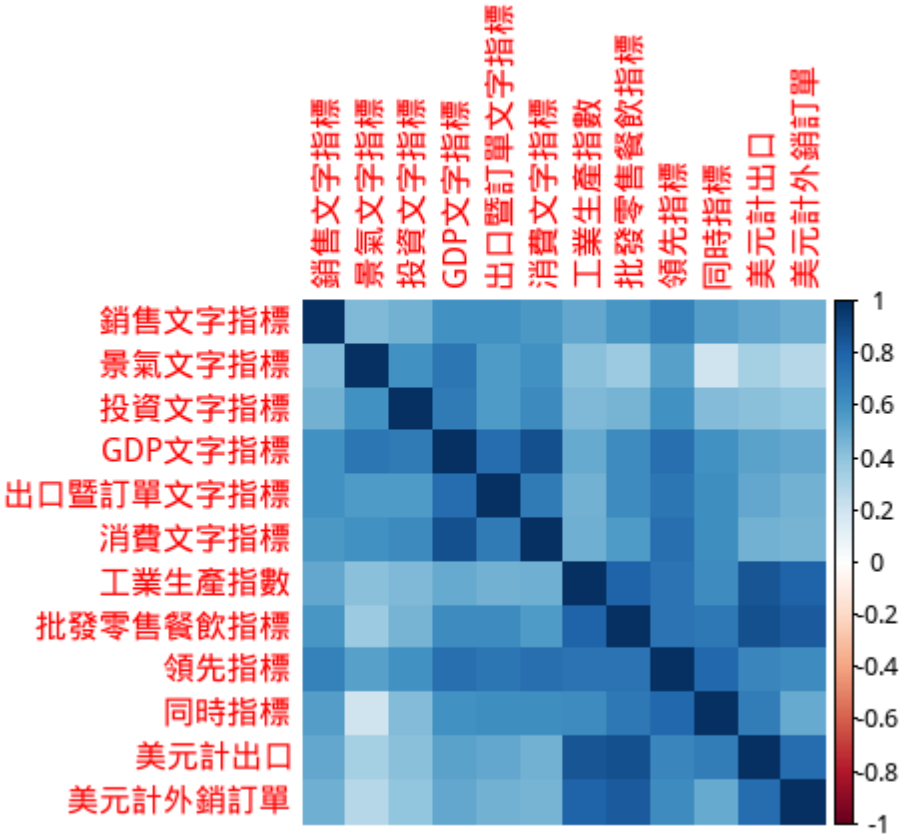


圖 5：各單項文字指標與重要總體變數熱力圖 (相關係數矩陣)

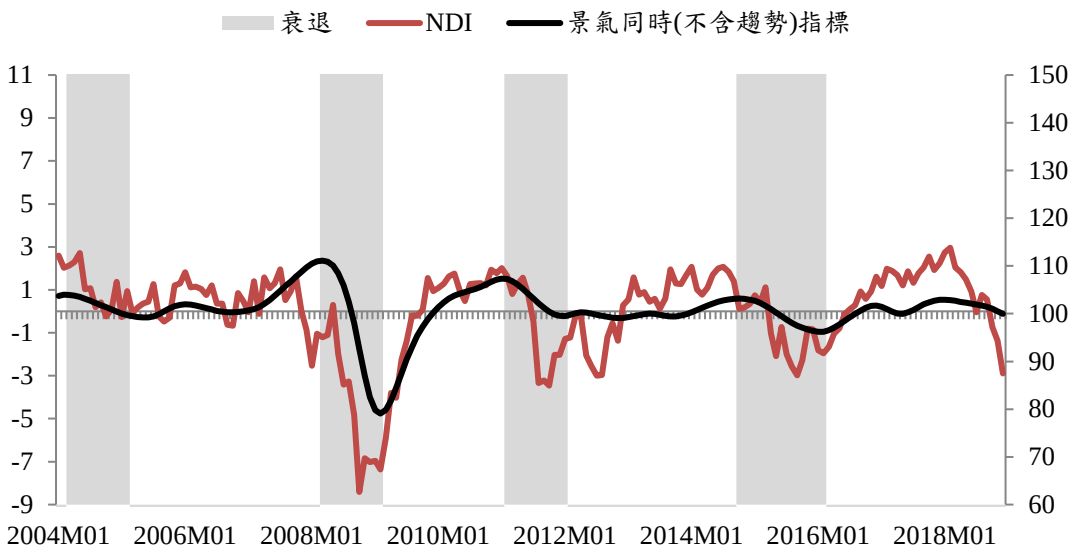
表 2 則為新聞擴散指標與重要總體經濟變數間的相關係數。由表 2 可知：NDI 與工業生產指數、批發零售餐飲營業額、同時指標、領先指標、海關出口、外銷訂單等重要總體變數均呈正相關，且相關係數均超過 0.5，代表 NDI 與總體經濟活動具有相當程度的共移性。其中，NDI 與景氣領先指標相關係數更達 0.81，顯示透過媒體文字所建構的 NDI 指標具有領先特質，有助於我們預測未來台灣經濟成長率。

表 2 ： 新聞擴散指標與重要經濟變數之相關係數矩陣

	NDI	工業生產指數	批發零售餐飲營業額	領先指標	同時指標	海關出口(美元計價)	外銷訂單(美元計價)
NDI	1.00	0.57	0.65	**0.81**	0.61	0.56	0.53
工業生產指數	0.57	1.00	0.81	0.73	0.63	0.86	0.81
批發零售餐飲營業額	0.65	0.81	1.00	0.73	0.71	0.88	0.83
領先指標	**0.81**	0.73	0.73	1.00	0.79	0.66	0.63
同時指標	0.61	0.63	0.71	0.79	1.00	0.69	0.51
出口(美元計)	0.56	0.86	0.88	0.66	0.69	1.00	0.77
外銷訂單(美元計)	0.53	0.81	0.83	0.63	0.51	0.77	1.00

註：本研究自行整理

為了方便比較，我們將新聞擴散指標、景氣領先、同時指標與 GDP 季成長率三者走勢繪於圖 6 和圖 7。其中陰影部份為行政院國家發展委員會所公佈的第 11 次至 14 次景氣循環的衰退期間，陰影區的起始點與終止點分別代表該景氣循環的高峰與谷底。觀察圖 6 和圖 7，我們發現：雖然 NDI 的波動比領先指標、同時指標與 GDP 成長率較為劇烈，但 NDI 的高峰與谷底大致領先國發會所認定的景氣峰谷時點，其走勢亦大致領先同時指標與 GDP 季成長率的變動，且 NDI 數據會比國發會的領先指標先公布，故本計畫所編制的新聞擴散指標可即時預報與預測當前與未來總體經濟動向，並提供政府及企業進行經濟決策的參考依據。



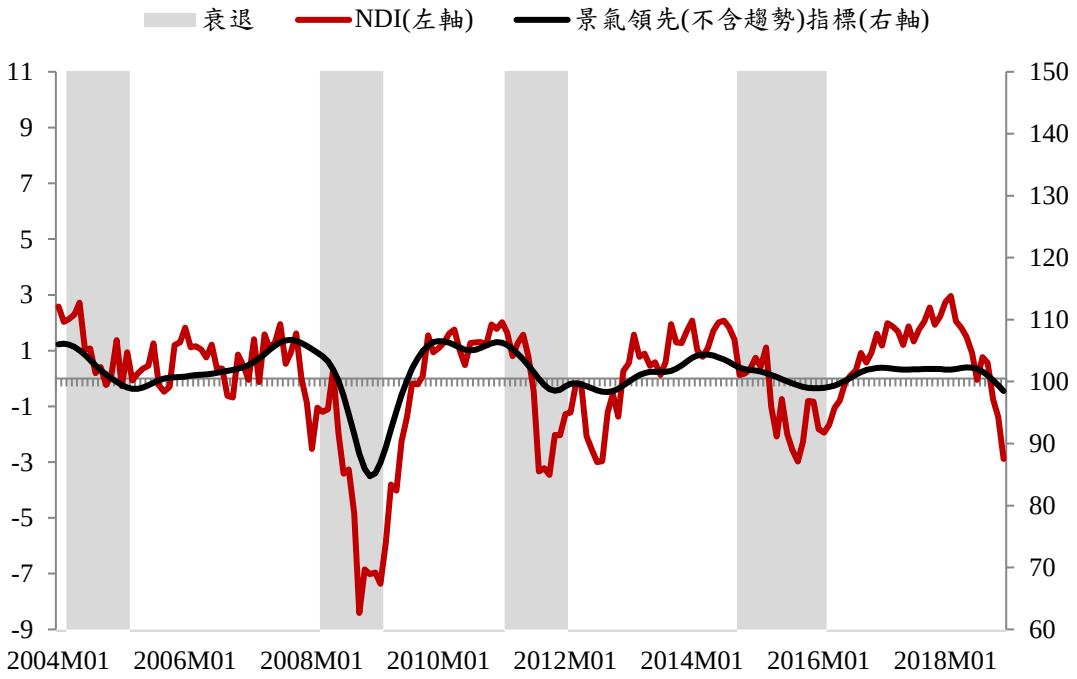


圖 6：新聞擴散指標與景氣指標

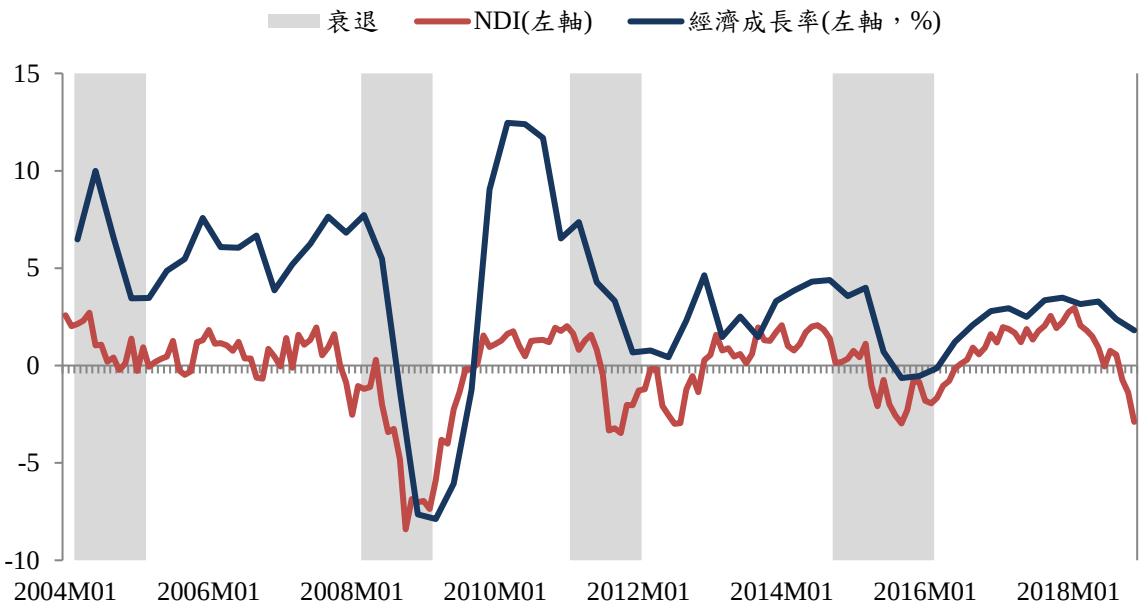


圖 7：新聞擴散指標與 GDP 季成長率

5. 模型估計與預測評估

5.1 參數估計與樣本外預測設計

本節主要評估新聞文字對 GDP 成長率的樣本外預測能力。我們首先針對不同的 MIDAS 預測模型進行估計，然而因為樣本期間包含了金融海嘯時期，造成模型參數估計在金融海嘯前後產生不穩定狀態，進而影響模型的預測表現，故本計畫實證估計期間與樣本外預測評估皆從金融海嘯結束後開始，即 2011-2018 年共 32 個季資料觀察值與 96 個月資料觀察值。

表 3 為後金融海嘯時期，6 種不同 MIDAS 模型的估計結果。被解變數為 GDP 季成長率 (y_t^Q)，解釋變數包括落後 1 期 GDP 季成長率 (y_{t-1}^Q)，月頻率的新聞擴散指標 (NDI_t^M)，3 個月移動平均的新聞擴散指標 ($NDI3m_t^M$)，以及落後 1 期的月頻率數據因子 (DF_{t-1}^M) 所組成。[4] 在不同 MIDAS 模型中，$w_j(.)$ 多項式函數的設定採指數 Almon 形式，並以非線性最小平方法估計模型參數。

表 3 顯示落後 1 期 GDP 季成長率的係數在所有 MIDAS 模型中均為顯著，估計值介於 0.34-0.54。若 MIDAS 模型加入同期新聞擴散指標月資料，我們發現：即使控制落後 1 期 GDP 成長率的影響，新聞擴散指標與 GDP 成長率的關係仍然顯著為正，且 3 個月移動平均 NDI 的估計結果亦相同，從此實證結果可以推論：媒體文字資訊有助於預測台灣 GDP 成長率。然而，若我們加入落後 1 期數據因子的解釋變數，則不論媒體文字或數據因子的係數估計值均不顯著。但這裏必須強調：樣本內配適程度的好壞不一定與樣本外預測

[4] 由於實體經濟統計數據公佈均較媒體文字資料落後，故我們以落後 1 期的數據因子當作解釋變數，詳細說明請參考下節。

表 3 ：不同預測模型重要參數估計結果

模型	β	θ_1	θ_2	θ_3	λ_1	λ_2	λ_3
AR1	0.54						
	(4.33)***						
AR1.DF1	0.51				-0.14	-0.13	0.02
	(4.58)***				(-0.74)	(-0.68)	(0.50)
AR1.NDI	0.37	4.37	-5.03	1.26			
	(3.22)***	(2.74)**	(-2.74)**	(2.79)**			
AR1.NDI.DF1	0.40	-0.05	-0.13	0.02	0.90	-0.70	0.12
	(2.92)***	(-0.27)	(-0.59)	(0.35)	(1.34)	(-1.22)	(1.19)
AR1.NDI3m	0.34	11.25	-13.13	3.26			
	(2.68)**	(1.97)*	(-1.92)*	(1.90)*			
AR1.NDI3m.DF1	0.35	0.72	-0.56	0.09	-0.12	-0.04	-0.001
	(2.39)**	(0.97)	(-0.87)	(0.84)	(-0.66)	(-0.17)	(-0.04)

註 1：括號內數值為 *t* 統計量。

註 2：「*」為顯著水準 10%下顯著，「**」為顯著水準 5%下顯著，「***」為顯著水準 1%下顯著。

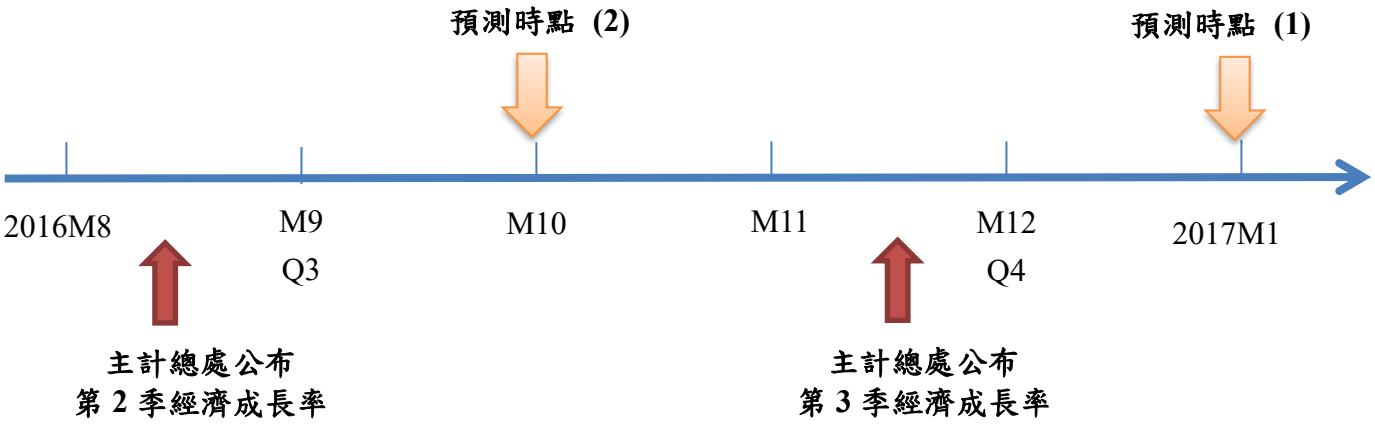


圖 8：混合頻率資料公布與不同預測時點示意圖

表現相符合。

本計畫以一階自我迴歸模型 (AR1) 當作基準預測模型，並將後金融海嘯時期的樣本分成：樣本內估計期間與樣本外預測期間，而兩段期間的觀察值數目與我們處於哪一個預測時點有關，圖 8 為不同預測時點的示意圖。因主計總處於每年 2、5、8、11 月公布上一季 GDP 成長率的初估值，此為央行貨幣政策的重要參考依據。若我們在時點 (1) 對 2016 年第 4 季的 GDP 成長率進行即時預報，則我們已經有前一季 (2016 年第 3 季) 的 GDP 成長率統計值。而在月資料方面，我們在 2016 年 12 月底已蒐集到當月的即時新聞擴散指標，但政府所公布的月頻統計數據則只有上個月 (11 月) 的數字，故數據因子資料將落後 NDI 一個月。因此在預測時點 (1)，我們將採加入領先項月頻變數的 MIDAS 動態因子模型，起始估計期間為 2011Q1 (M1) - 2016Q3 (M12)，並以遞迴 (recursive) 方式不斷加入新的資料，再重新估計與預測。

若在時點 (2) 對 2016 年第 4 季的 GDP 成長率進行預測，當

時我們就只有 2016 年第 2 季的 GDP 成長率統計值，則 AR1 預測模型中的季資料落後項解釋變數為 y_{t-2}^{Q}，[5]我們以 AR1(2) 代表其與預測時點 (1) 的差異。在月資料方面，新聞擴散指標已即時蒐集到當月 (2016M9)，但數據因子資料仍落後 NDI 一個月，此時我們採用傳統的混合頻率動態因子模型預測未來 GDP 成長率。

本計畫只考慮 one-step-ahead 預測，並以 RMSE 與 MAE 作為評估不同模型的樣本外預測表現，其公式分別如下：

$$RMSE = \sqrt{\frac{\sum_{t=T_0+1}^{T_0+h} (\hat{y}_t^Q - y_t^Q)^2}{h}},$$

$$MAE = \frac{\sum_{t=T_0+1}^{T_0+h} \left| \hat{y}_t^Q - y_t^Q \right|}{h},$$

其中 $\hat{y}_t^Q$ 為預測值，h 為預測期數。若 RMSE 或 MAE 越小，代表模型預測越準確。

5.2 混合領先頻率動態因子模型預測表現

我們首先探討加入領先資訊 MIDAS 模型的預測能力，樣本外預測期間為 2016Q4-2018Q4，共九期 ($h=9$)。圖 9 與圖 10 為不同模型在預測時點 (1) 的樣本外預測走勢，其中 AR1.tgdp3m.DF1 模型是將 3 個月移動平均的經濟成長文字指標取代新聞擴散指標，目的是探討若採用與經濟 (GDP) 成長直接相關的媒體文字指標，其預測表現是否會與 3 個月移動平均的新聞擴散指標 (AR1.NDI3m.DF1

[5] 另一種解釋為：此時我們的基準模型為 AR2，但限制 y_{t-1}^{Q} 的係數為零。

模型) 不相上下？從圖 10 我們可知：兩模型對未來 GDP 成長率的預測值相當接近。

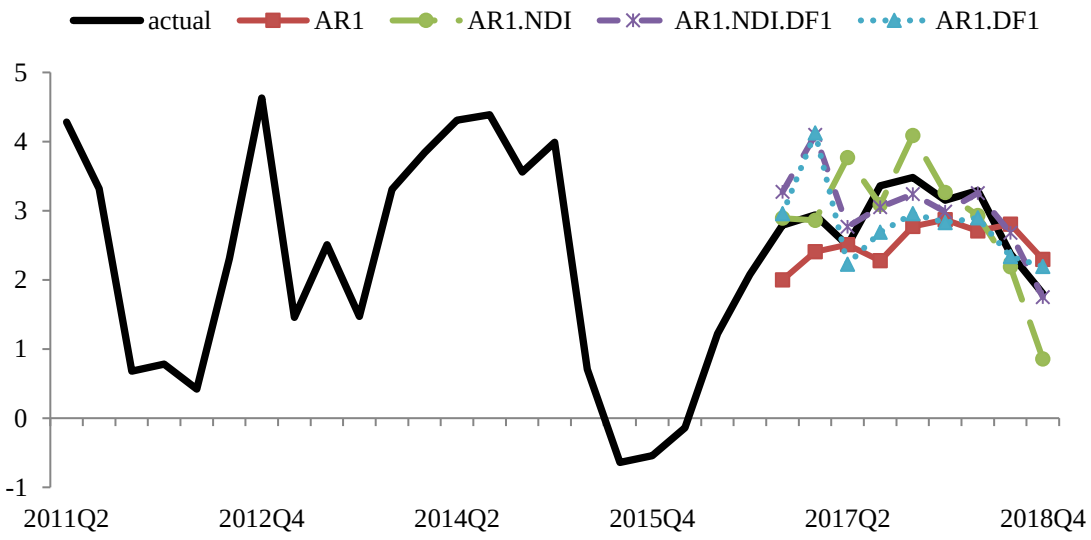


圖 9：不同模型在預測時點 (1) 的樣本外預測走勢 (I)

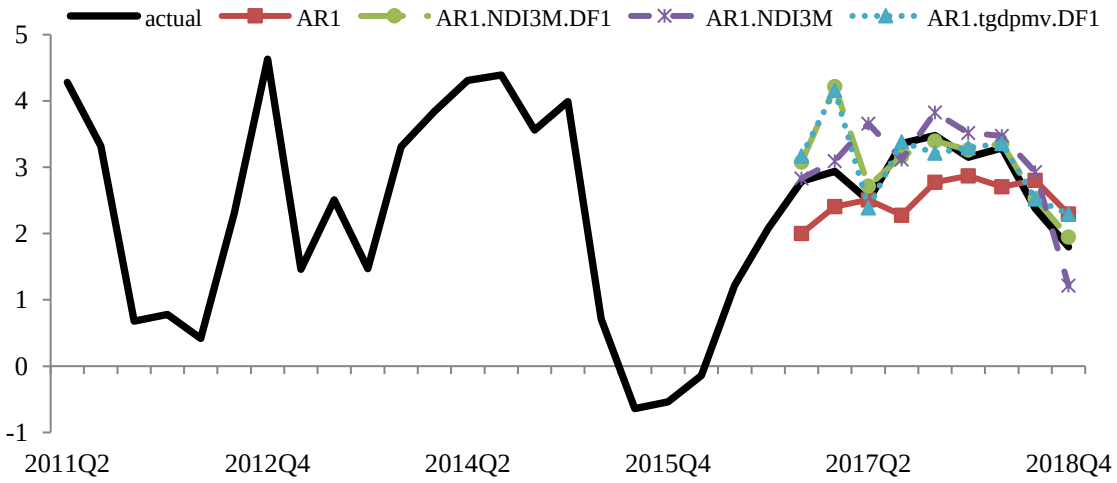


圖 10：不同模型在預測時點 (1) 的樣本外預測走勢 (II)

表 4 為評估不同 MIDAS 模型在時點 (1) 的樣本外預測表現。實證結果發現：若只考慮媒體文字資訊對 GDP 成長率進行即時預報，如 NDI 或 NDI3m 模型，其預測表現不一定優於基準模型 (AR1)，兩模型的 RMSE 相對於 AR1 模型的 RMSE 均大於 1。雖然其 MAE 較 AR1 模型為小，但亦是除 AR1 模型外，樣本外預測表現較差的兩個模型。

若預測模型中加入 GDP 成長率的落後項，則可提升媒體文字的預測績效。如 AR1.NDI 與 AR1.NDI3m 模型之 RMSE 均較 NDI 與 NDI3m 模型為小。但值得注意是：加入數據因子的 AR1 模型，如 AR1.DF1，其樣本外預測表現不一定優於考慮媒體文字的 AR1 模型 (如 AR1.NDI 與 AR1.NDI3m)。此隱含本計畫所建構的新聞擴散指標，其預測能力不輸於大量統計數據所建構的因子指標。

表 4 ：在預測時點 (1) 的樣本外預測表現

模型	RMSE	相對 AR1 的 RMSE 比率	MAE	相對 AR1 的 MAE 比率
AR1	0.62	1.00	0.55	1.00
NDI	0.65	1.05	0.50	0.92
AR1.NDI	0.59	0.96	0.44	0.80
AR1.NDI.DF1	0.46	0.75	0.33	0.61
AR1.DF1	0.54	0.88	0.44	0.81
AR1.NDI3m.DF1	**0.46**	**0.74**	**0.28**	**0.52**
AR1.NDI3m	0.51	0.83	0.40	0.74
NDI3m	0.62	1.00	0.51	0.94
AR1.tgdp3m.DF1	0.47	0.77	0.32	0.58

註：在預測時點 (1) 中，NDI 為當季所有三個月的媒體文字資料，DF 為當季前兩個月的數據因子資料，GDP 成長率為前一季的資料。

若我們同時考慮前一期 GDP 成長率、新聞擴散指標與統計數據因子等三類變數，其預測表現優於其他只考慮某單項或兩項解釋變數的 MIDAS 模型，此隱含結合不同頻率資料、媒體文字與眾多經濟數據的動態因子模型，其預測表現優於只考慮單一頻率或只蒐集經濟數據的預測模型。此外，由於媒體文字資訊的波動較大，若先將新聞擴散指標予以平滑化處理，如考慮 3 個月移動平均的新聞擴散指標變數，則可提升混合頻率動態因子模型的預測表現。綜合而言，加入媒體文字資訊能增進 MIDAS 模型對經濟成長的即時預報與預測的準確度。

5.3 混合頻率動態因子模型預測表現

若在時點 (2) 預測 2016Q4 的 GDP 成長率，此時我們擁有的資訊集合為 2016Q3 (M9) 之前所有季、月頻率的文字與數字資料，則所有 MIDAS 預測模型均不包含變數的領先資訊，即 2016Q4 的月頻資料訊息。圖 11 與圖 12 為不同 MIDAS 模型在時點 (2) 的 GDP 成長率預測走勢，其中以 AR1(2).DF1 預測走勢波動較大，預測表現受極端值影響大。

表 5 為評估不同 MIDAS 模型在時點 (2) 的樣本外預測表現。我們發現：只要 MIDAS 模型中包含媒體文字變數，其預測表現均優於 AR1(2).DF1 模型。顯示在時點 (2) 的情況下，新聞擴散指標較統計數據因子的預測能力為佳，且更能即時準確預測未來 GDP 成長率。另一方面，與在時點 (1) 的實證結果一致，若我們先將新聞

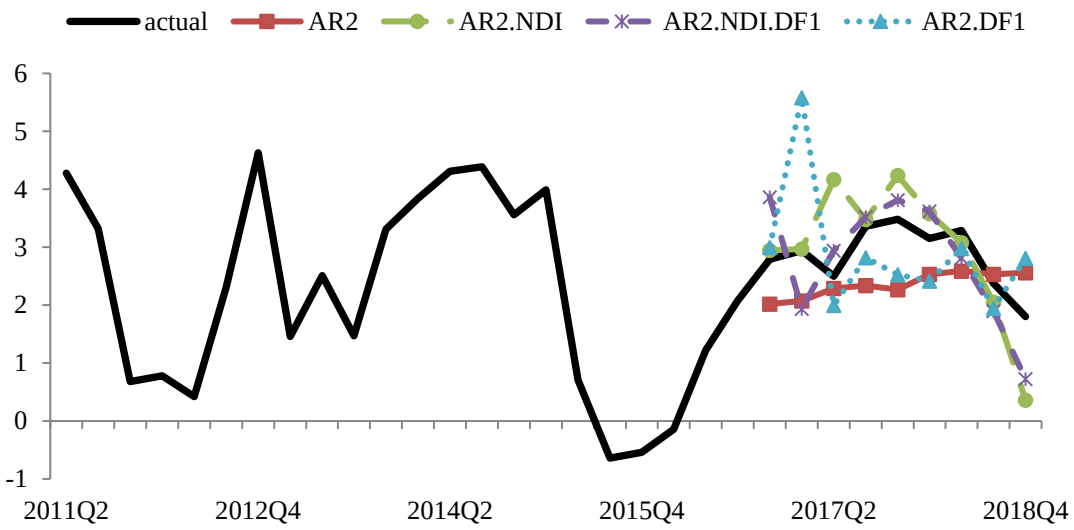


圖 11：不同模型在預測時點 (2) 的樣本外預測走勢 (I)

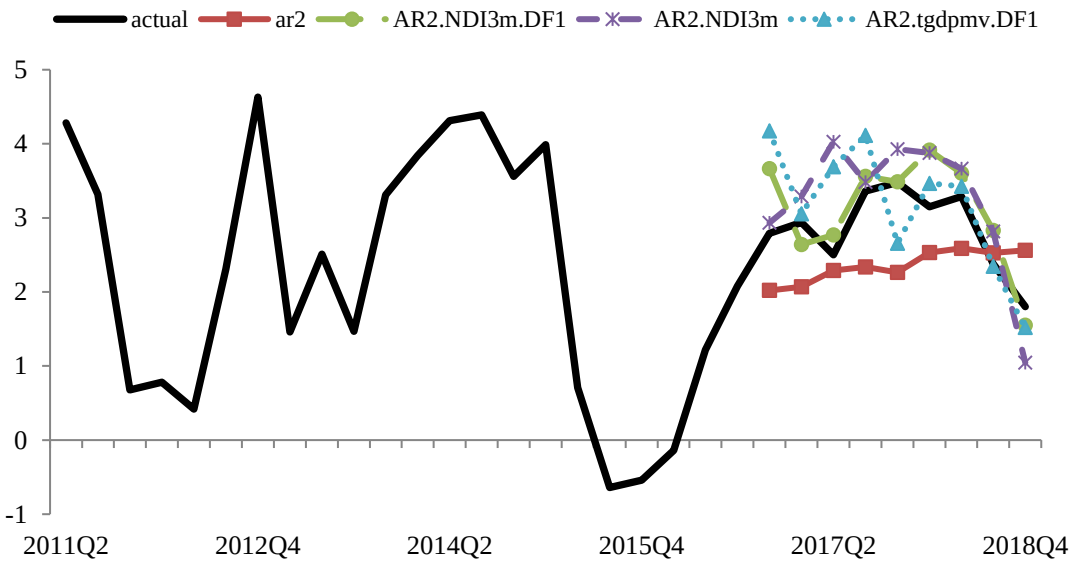


圖 12：不同模型在預測時點 (2) 的樣本外預測走勢 (II)

擴散指標予以平滑化處理，則其 (AR1(2).NDI3m.DF1) 預測表現優於其他混合頻率動態因子模型。但若只以經濟成長文字指標作為預測變數，其預測績效不如包含多項媒體文字主題的新聞擴散指標。

綜合上述實證結果，我們發現：本計畫所編制的新聞擴散指標可即時反應當前與未來總體經濟的動向，若結合大量文字與數字資料，並考慮混合頻率的動態因子模型，將有助於提升 GDP 成長率的預測表現。

表 5 ：在預測時點 (2) 的樣本外預測表現

模型	RMSE	相對 AR1(2)的 RMSE 比率	MAE	相對 AR1(2)的 MAE 比率
AR1(2)	0.77	1.00	0.70	1.00
AR1(2).NDI	0.80	1.04	0.57	0.81
AR1(2).NDI.DF1	0.69	0.90	0.61	0.87
AR1(2).DF1	1.07	1.38	0.81	1.16
AR1(2).NDI3m.DF1	**0.46**	**0.60**	**0.38**	**0.54**
AR1(2).NDI3m	0.68	0.87	0.54	0.77
AR1(2).tgdp3m.DF1	0.73	0.94	0.56	0.80

註：在預測時點 (2) 中，NDI 為前一季所有三個月的媒體文字資料，DF 為前一季前兩個月的數據因子資料，GDP 成長率為前兩季的資料。

此外，在評估不同 MIDAS 模型的樣本外預測表現，本計畫對多項式函數 $w_j(.)$ 的設定均採指數 Almon 形式，我們也嘗試採用其他不同函數的設定形式，如 Beta 多項式，並分別在時點 (1) 和時點 (2) 探討包含新聞擴散指標模型的樣本外預測表現，其實證結論與 5.2 和 5.3 節一致，即本計畫所編制的新聞擴散指標有助於增進 GDP 預測的準確性。

6. 結論

目前台灣政府單位或民間智庫主要根據經濟數據，設定適當的總體計量模型，對 GDP 成長率進行預測，但無充分利用一些非結構化資料，以提供即時的經濟預測，並幫助決策者研判當前的景氣動向。本計畫主要目的為利用文字探勘方法，分析大量的媒體報導資訊，並結合不同頻率的經濟統計數據，利用混合頻率動態因子模型，探討媒體文字是否對台灣 GDP 成長率具有預測能力。

本計畫主要貢獻如下：首先，我們透過文字探勘技術，針對大量的新聞報導與不同的經濟主題進行正反面語意分析，並建構新聞擴散指標。實證結果發現：無論新聞擴散指標或是其構成的 6 項文字指標均能領先反應台灣總體經濟活動的變化趨勢，且新聞擴散指標的高峰與谷底亦大致領先國發會所認定的景氣峰谷時點，故媒體文字資訊可確實捕捉台灣 GDP 成長率變化與景氣波動方向。

其次，本計畫利用混合頻率動態因子模型，探討媒體文字是否可提升傳統以數字資料為基礎之計量模型的預測能力。實證結果發現：若只考慮媒體文字資料，其預測表現不一定優於基準模型。但若同時考慮 GDP 成長率落後項，與 GDP 相關的大量經濟數據，以及新聞擴散指標等混合頻率資料，則可提升媒體文字的預測表現。此外，由於文字資訊的波動較大，若將新聞擴散指標適度平滑化，將可增進混合頻率動態因子模型的預測績效。

對經濟政策執行效益而言，由於政府公布的統計數據較為延遲，無法即時反應當前經濟動向。本計畫利用每天媒體文字資訊編制新聞擴散指標，可立即掌握當前與未來總體經濟情勢，幫助政府及企業適時採行有效的決策，並控制景氣衰退可能產生的風險。據我們所

知，目前尚未有任何機構定時公布與景氣動向有關的媒體文字指標，若相關單位想要建構類似的網站，只要有適當的資訊工程人員配合，本研究團隊願意貢獻相關的研究成果，相信對即時研判台灣景氣動向應有所助益。

本計畫未來可能的延伸方向：(1) 因本計畫主要專注於台灣實質部門的經濟數據與文字探勘分析，針對實質經濟成長率進行預測，並未考慮大量日資料的金融變數，故未來可以加入金融變數的動態因子，如金融情勢指標 (FCI)，探討其是否能提升本計畫模型的預測表現[6]。(2) 本計畫提出結合文字探勘技術與混合頻率資料模型的想法可應用於預測其他總體經濟變數，如物價、就業、經濟信心等。此外，運用類似的語意分析方法，進一步將文字訊息依不同程度劃分，如偏向正面、偏向負面等，並建立股、匯市交易者的情緒指標，以提供相關金融機構參考。這些議題均可留待後續研究。

[6] 金融情勢指數 (Financial Conditions Index, FCI) 是根據眾多金融變數所編制而成的綜合指標，藉以描述整體金融市場寬鬆與緊縮的狀況。理論上，貨幣政策通常可以透過債券、股票與外匯市場等傳遞管道，進而影響實質面總體經濟活動，詳細說明與預測能力分析請參考管中閔等 (2014)。

參考文獻

1. 徐之強、黃裕烈 (2005)，「景氣基準循環指數之檢討與修訂」，經濟建設委員會委託研究計畫。

2. 黃裕烈、徐士勛、陳志成 (2017)，「臺灣景氣動向指數之建構與分析」，台北外匯市場發展基金會委託研究計畫。

3. 黃裕烈、葉錦徽、陳重吉 (2019)，「臺灣不確定指標之建構」，經濟論文叢刊，即將出版。

4. 黃裕烈、管中閔 (2019)，「美國聯準會會議紀要的文字探勘與臺灣經濟變數預測」，經濟論文叢刊，47，363-391。

5. 管中閔、徐之強、黃裕烈、徐士勛 (2014)，「臺灣金融情勢指數與總體經濟關係」，臺灣經濟預測與政策，44，103-132。

6. Andreou E., E. Ghysels, and A. Kourtellos (2011), "Forecasting with Mixed-Frequency Data," Oxford Handbook of Economic Forecasting, edited by M. P. Clements and D. F. Hendry, Oxford University Press.

7. Ardia, D., K. Bluteau, and K. Boudt (2019), "Questioning the News about Economic Growth: Sparse Forecasting Using Thousands of News-Based Sentiment Values," *International Journal of Forecasting*, forthcoming.

8. Aruoba, S. B., F. X. Diebold, and C. Scotti (2009), "Real-Time Measurement of Business Conditions," *Journal of Business and Economic Statistics*, 27, 417-27.

9. Armesto, M. T, H. Ruben, M. Owyang, and J. Piger (2009), "Measuring the Information Content of the Beige Book: A Mixed Data

Sampling Approach," *Journal of Money, Credit and Banking*, 41, 35-55.

10. Bai, J. (2003), "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-171.

11. Bai, J. and S. Ng (2002), "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.

12. Baker, Scott R, Bloom, Nicholas, & Davis, Steven J. (2016), "Measuring Economic Policy Uncertainty," *Quarterly Journal of Economics*, 131, 1593-1636.

13. Balke, N. S., M. Fulmer, and R. Zhang (2016), "Incorporating the Beige Book into a Quantitative Index of Economic Activity," *Journal of Forecasting*, 36, 497-514.

14. Banbura, M., D. Giannone, M. Modugno, and L. Reichlin (2013), "Now-Casting and the Real-Time Data Flow," Handbook of Economic Forecasting, 2A, 195-237.

15. Feuerriegel, S. and J. Gordon (2018), "News-Based Forecasts of Macroeconomic Indicators: A Semantic Path Model for Interpretable Predictions," *European Journal of Operational Research*, forthcoming.

16. Gentzkow, M., B. T. Kelly, and M. Taddy (2019), "Text as Data", *Journal of Economic Literature*, forthcoming.

17. Giannone,D., L. Reichlin, and D. Small (2008), "Nowcasting: The Real-Time Informational Content of Macroeconomic Data," *Journal of Monetary Economics*, 55, 665-676.

18. Ghysels, E., P. Santa-Clara and R. Valkanov (2004), The MIDAS

Touch: Mixed Data Sampling Regressions, Discussion Paper UNC/ UCLA.

19. Hansen, S., M. McMahon, and A. Prat (2018), "Transparency and Deliberation within the FOMC: A Computational Linguistics Approach," *Quarterly Journal of Economics*, 133, 801-870.

20. Hoberg G. and G. Phillips (2010), "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis," *Review of Financial Studies,* 23, 3773-3811.

21. Hoberg, G. and G. Phillips (2016), "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, 124, 1423-1465.

22. Huang, Y. L. and C. M. Kuan (2020), "Economic Prediction with the FOMC Minutes: An Application of Text Mining," *International Review of Economics and Finance*, forthcoming.

23. Jegadeesh, N. and D. Wu (2013), "Word Power: A New Approach for Content Analysis," *Journal of Financial Economics*, 110, 712-729.

24. Jubinski, D. and M. Tomljanovich (2013), "Do FOMC Minutes Matter to Markets? An Intraday Analysis of FOMC Minutes Releases on Individual Equity Volatility and Returns," *Review of Financial Economics*, 22, 86-97.

25. Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2018), "Measuring Technological Innovation over the Long Run," Working Paper.

26. Loughran, T. and B. McDonald (2011), "When Is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of*

Finance, 66, 35-65.

27. Loughran, T. and B. McDonald (2014), "Measuring Readability in Financial Disclosures," *Journal of Finance*, 69, 1643-1671.

28. Loughran, T. and B. McDonald (2016), "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, 54, 1187-1230.

29. Manela, A. and A. Moreira (2015), "News Implied Volatility and Disaster Concerns," *Journal of Financial Economics*, 123, 137-162.

30. Marcellino, M., M. Porqueddu, and F. Venditti (2016), "Short-Term GDP Forecasting with a Mixed-Frequency Dynamic Factor Model with Stochastic Volatility," *Journal of Business & Economic Statistics*, 34, 118-127.

31. Marcellino, M. and C. Schumacher (2010), "Factor MIDAS for Nowcasting and Forecasting with Ragged-Edge Data: A Model Comparison for German GDP," *Oxford Bulletin of Economics and Statistics*, 72, 518-550.

32. Sadique, S., F. In, M. Veeraraghavan, and P. Wachtel (2013), "Soft Information and Economic Activity: Evidence from the Beige Book," *Journal of Macroeconomics*, 37, 81-92.

33. Schorfheide, F. and D. Song (2015), "Real-Time Forecasting With a Mixed-Frequency VAR," *Journal of Business & Economic Statistics*, 33, 366-380.

34. Stock, J. H. and M. W. Watson (2002a), "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the*

American Statistical Association, 97, 1167-1179.

35. Stock, J. H. and M.W. Watson (2002b), "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147-162.

36. Stock, J. H. and M. W. Watson (2011), "Dynamic Factor Models," Oxford Handbook of Forecasting, edited by M. P. Clements and D. F. Hendry, Oxford University Press.

37. Tetlock, P. (2007), "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *Journal of Finance*, 62, 1139-1168.

38. Tetlock, P., M. Saar-Tsechansky, and S. Macskassy (2008), "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *Journal of Finance*, 62, 1139-1168.

39. Thorsrud, L. A. (2018), "Words are the New Numbers: A Newsy Coincident Index of the Business Cycle," *Journal of Business & Economic Statistics*, forthcoming.

附錄 Ⅰ

學術文獻常利用主成分分析或因子分析來處理巨量資料，希望可以藉由維度縮減方法，萃取出最重要的因子，並進一步預測未來總體經濟變數的走勢。令 x_{it} 為第 i 個數據或文字變數在 t 時點的觀察值，其中 $i=1,\cdots,N$ 而 $t=1,\cdots,T$，則因子分析模型可表示為：

$$\underbrace{\begin{bmatrix} x_{11} & \cdots & x_{N1} \\ \vdots & \ddots & \vdots \\ x_{1T} & \cdots & x_{NT} \end{bmatrix}}_{\boldsymbol{X}\ T\times N} = \underbrace{\begin{bmatrix} f_{11} & \cdots & f_{r1} \\ \vdots & \ddots & \vdots \\ f_{1T} & \cdots & f_{rT} \end{bmatrix}}_{\boldsymbol{F}\ T\times r} \times \underbrace{\begin{bmatrix} \ell_{11} & \cdots & \ell_{N1} \\ \vdots & \ddots & \vdots \\ \ell_{1r} & \cdots & \ell_{Nr} \end{bmatrix}}_{\boldsymbol{\Lambda}'\ r\times N} + \underbrace{\begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{N1} \\ \vdots & \ddots & \vdots \\ \varepsilon_{1T} & \cdots & \varepsilon_{NT} \end{bmatrix}}_{\boldsymbol{\varepsilon}\ T\times N},$$

其中 f_{jt} 代表第 j 個主要因子在 t 時點的數值，$j=1,2,\cdots,r$；r 為真正因子個數，且 $r<N$；ℓ_{ij} 稱為因子負載，表示第 j 個主要因子對第 i 個數量或文字變數的影響力；而 ε_{it} 則為干擾項。在此模型中，我們只能觀察到數據或文字變數 $\boldsymbol{X}$ 的資料，但因子 $\boldsymbol{F}$、因子負載 $\boldsymbol{\Lambda}$ 以及干擾項 $\boldsymbol{\varepsilon}$ 均無法被觀察到。

為了估算 $\boldsymbol{F}$ 以及 $\boldsymbol{\Lambda}$，傳統的因子分析大多假設 $\boldsymbol{F}$ 與 $\boldsymbol{\varepsilon}$ 沒有序列相關，直到近期，學術文獻才將此假設放寬。若我們允許 ε_{it} 有序列相關，則上述模型便稱為近似因子模型 (approximate factor model)。由於 $\boldsymbol{F}$ 與 $\boldsymbol{\Lambda}$ 估計上存在認定問題，為解決此問題有兩種不同方式：一是要求 $\boldsymbol{F}$ 符合正規化條件，$\boldsymbol{F}'\boldsymbol{F}/T=\boldsymbol{I}_r$，且 $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}$ 必須是一個對角線矩陣；二是要求 $\boldsymbol{\Lambda}$ 符合正規化條件，$\boldsymbol{\Lambda}'\boldsymbol{\Lambda}/N=\boldsymbol{I}_r$，且 $\boldsymbol{F}'\boldsymbol{F}$ 必須一個對角線矩陣。

考慮以下最小平方估計式的目標函數：

$$\min_{F,\Lambda} Q(\boldsymbol{F},\boldsymbol{\Lambda}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \boldsymbol{\lambda}_i' \underline{\boldsymbol{f}}_t)^2 = \frac{1}{NT}\sum_{t=1}^{T}\left(\underline{\boldsymbol{x}}_t - \boldsymbol{\Lambda}\,\underline{\boldsymbol{f}}_t\right)'\left(\underline{\boldsymbol{x}}_t - \boldsymbol{\Lambda}\,\underline{\boldsymbol{f}}_t\right) \quad \text{(A.1)}$$

$$\text{s.t.} \quad \frac{\boldsymbol{\Lambda}'\boldsymbol{\Lambda}}{N} = \boldsymbol{I}_r$$

給定 $\boldsymbol{\Lambda}$ 之下，$\underline{\boldsymbol{f}}_t$ 必須滿足以下的一階條件：

$$\frac{\partial Q}{\partial \underline{\boldsymbol{f}}_t} = \frac{2}{NT}\left[\boldsymbol{\Lambda}'\underline{\boldsymbol{x}}_t - (\boldsymbol{\Lambda}'\boldsymbol{\Lambda})\underline{\boldsymbol{f}}_t\right] = \boldsymbol{0}$$

故 $\underline{\boldsymbol{f}}_t$ 為 $\boldsymbol{\Lambda}$ 的函數，即 $\underline{\boldsymbol{f}}_t^* = (\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\underline{\boldsymbol{x}}_t$。將 $\underline{\boldsymbol{f}}_t^*$ 代入目標函數可得：

$$Q(\boldsymbol{F}(\boldsymbol{\Lambda}),\boldsymbol{\Lambda}) = \frac{1}{NT}\sum_{t=1}^{T}\left(\underline{\boldsymbol{x}}_t'\underline{\boldsymbol{x}}_t - \underline{\boldsymbol{x}}_t'\boldsymbol{\Lambda}\left(\boldsymbol{\Lambda}'\boldsymbol{\Lambda}\right)^{-1}\boldsymbol{\Lambda}'\underline{\boldsymbol{x}}_t\right)$$

由於矩陣 $\boldsymbol{\Lambda}^*$ 必須滿足 $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}/N = \boldsymbol{I}_r$ 限制式，並且使上式極小化，此等同於 $\boldsymbol{\Lambda}^*$ 會使下式極大化：

$$\text{tr}\left[\boldsymbol{\Lambda}'\left(\sum_{t=1}^{T}\underline{\boldsymbol{x}}_t\underline{\boldsymbol{x}}_t'\right)\boldsymbol{\Lambda}\right] = \text{tr}\left[\boldsymbol{\Lambda}'\left(\boldsymbol{X}'\boldsymbol{X}\right)\boldsymbol{\Lambda}\right]$$

因此我們可證明 $\boldsymbol{\Lambda}^*$ 的最小平方估計式 $(\bar{\boldsymbol{\Lambda}}^*)$ 為 $\sqrt{N}$ 乘上 $N\times N$ 矩陣 $\boldsymbol{X}'\boldsymbol{X}$ 中最大 r 個特徵值所對應的特徵向量形成之集合 (Bai and Ng, 2002)：

$$\bar{\boldsymbol{\Lambda}}^* = \sqrt{N}\left[\bar{\boldsymbol{\xi}}_1,\cdots,\bar{\boldsymbol{\xi}}_r\right], \quad \text{(A.2)}$$

其中 $\bar{\boldsymbol{\xi}}_i$ 為第 i 個特徵值所對應的特徵向量。

在 $\boldsymbol{\Lambda}'\boldsymbol{\Lambda}/N = \boldsymbol{I}_r$ 的限制條件下，將 $\boldsymbol{\Lambda}^*$ 代入 $\underline{\boldsymbol{f}}_t^* = (\boldsymbol{\Lambda}'\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\underline{\boldsymbol{x}}_t$，則 $\boldsymbol{F}$ 的最小平方估計式為：

$$\bar{\boldsymbol{F}}^* = \frac{\boldsymbol{X}\bar{\boldsymbol{\Lambda}}^*}{N}, \quad \text{i.e.,} \quad \bar{\underline{\boldsymbol{f}}}_t^* = \bar{\boldsymbol{\Lambda}}^{*\prime}\underline{\boldsymbol{x}}_t,\ \forall t \quad \text{(A.3)}$$

令 $\tilde{\boldsymbol{V}}^r$ 為對角矩陣，其對角線元素為 $\boldsymbol{XX}'/(TN)$ 所對應的最大 r 個特徵值，則我們可以證明：

$$\tilde{\boldsymbol{V}}^r = \frac{\tilde{\boldsymbol{\Lambda}}^{r'}\tilde{\boldsymbol{\Lambda}}}{T} = \frac{\bar{\boldsymbol{F}}^{r'}\bar{\boldsymbol{F}}^r}{N}, \qquad \bar{\boldsymbol{F}}^r = \tilde{\boldsymbol{F}}^r\left(\tilde{\boldsymbol{V}}^r\right)^{1/2}, \qquad \tilde{\boldsymbol{\Lambda}}^r = \bar{\boldsymbol{\Lambda}}^r\left(\tilde{\boldsymbol{V}}^r\right)^{1/2}$$

而估計式 $\bar{\boldsymbol{F}}^r$ 與 $\tilde{\boldsymbol{F}}^r$ 彼此有一對應關係。換言之，利用目前的估計方法可以一致性地估算因子所形成的空間，但不能估算出真正的因子。而 Bai (2003) 也證明 $\tilde{\boldsymbol{F}}$ 與 $\tilde{\boldsymbol{\Lambda}}$ 最小平方估計式 (A.2)-(A.3) 的大樣本性質。

附錄 II

本計畫蒐集以月頻率為主的總體變數，但排除商品與金融面的價格變數 (如消費者物價、利率、匯率、股價等)，總共包括 34 個總體經濟數據變數，詳細資料名稱請參考下表。

本計畫先對資料進行單根檢定，驗證該變數是否為定態數列，並按照其結果決定是否需要差分。歸納資料的處理方法共分三類，分別為 (1) 一階差分；(2) 取自然對數後再一階差分；(3) 取自然對數再二階差分。個別變數轉換資料的方式請參考下表。

變數名稱	轉換方式
生產量指數-最終需要財(2016=100)-月(指數)	2
生產量指數-投資財(2016=100)-月(指數)	2
生產量指數-消費財(2016=100)-月(指數)	2
生產量指數-生產財(2016=100)-月(指數)	3
生產量指數-電力及燃氣供應業(2016=100)-月(指數)	3
銷售量指數-製造業(2016=100)-月(指數)	2
存貨量指數-製造業(2016=100)-月(指數)	2
生產量指數-製造業(2016=100)-月(指數)	2
十五歲以上民間人口-月(千人)	2
就業人口數-月(千人)	2
勞動人口數-月(千人)	2
失業人口數-月(千人)	2
勞動參與率-月(%)	1

變數名稱	轉換方式
失業率-月(%)	1
出口總值(含復出口)-美元-月(百萬美元)	2
進口總值(含復進口)-美元-月(百萬美元)	2
進口-貿易結構-資本設備-美元-月(百萬美元)	2
外銷訂單金額總計-美元-月(百萬美元)	3
貨幣總計數-M1A-期底-月-台幣(百萬)	3
貨幣總計數-M1B-期底-月-台幣(百萬)	3
貨幣總計數-M2-期底-台幣-月(百萬)	2
貨幣總計數-貨幣機構以外各部門持有通貨-月底數-月-台幣(百萬)	2
中央銀行-負債與權益-通貨發行額-月-台幣(百萬)	2
準備貨幣-全體貨幣機構以外各部門持有通貨-庫存現金-月底數-月-台幣(百萬)	2
貨幣總計數-存款貨幣-支票存款-月底數-月-台幣(百萬)	2
貨幣總計數-存款貨幣-活期存款-月底數-月-台幣(百萬)	2
貨幣總計數-存款貨幣-活期儲蓄存款-月底數-月-台幣(百萬)	2
貨幣總計數-準貨幣合計-月底數-月-台幣(百萬)	2
貨幣總計數-準貨幣-定期及定期儲蓄存款-月底數-月-台幣(百萬)	2
貨幣機構-放款與投資合計-期底餘額-月-台幣(百萬)	2
貨幣機構-放款-期底餘額-月-台幣(百萬)	3
外匯存底-美元-月(億美元)	2
同時指標不含趨勢指數-月(指數)	2
領先指標不含趨勢指數-月(指數)	2