政 治 大 學
統 計 學 系 碩 士 班

統 計 計 算 與 模 擬
期 末 報 告
Multivariate Analysis: Midterm Exam

授課教授： 翁 久 幸 教授

研 究 生： 鄭昕東 統碩一 106354023

研 究 生： 曹立諭 統碩一 106354012

研 究 生： 王允立 金融一 106352005

目錄

資料簡介

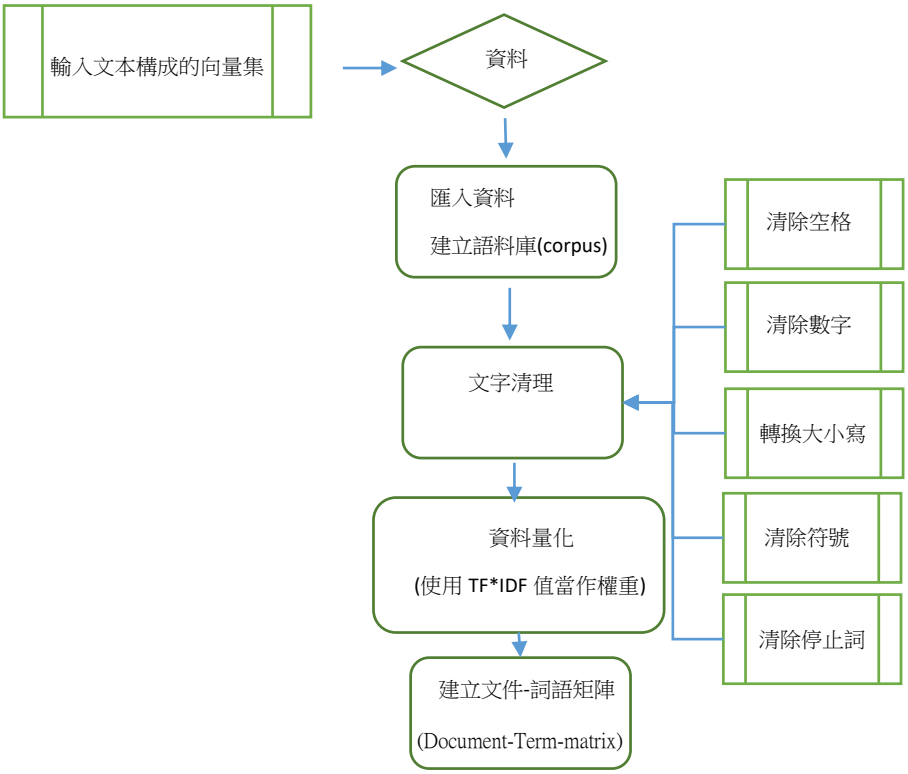
《聯邦黨人文集》或稱《聯邦論》、《聯邦主義議文集》(Federalist Papers)，是18世紀80年代三位美國政治家在制定美國憲法的過程中所寫作的有關美國憲法和聯邦制度的評論文章的合集，共收有85篇文章。這些文章最早連載於紐約地區的報紙，之後在1788年，首次出版了合集，書名為「聯邦黨人」(The Federalist)。此書主要對美國憲法和美國政府的運作原理進行了剖析和闡述，是研究美國憲法的最重要的歷史文獻之一。

《聯邦黨人文集》的作者包括James Madison、Alexander Hamilton 、 John Jay。他們使用了Publius這個筆名。這個名字來源於他們所尊敬的古羅馬執政官Publius Valerius Publicola。Madison常被後人稱為美國憲法之父，並且擔任了第四任美國總統。Hamilton是憲法會議中非常有影響力的一位代表，並成為美國首任財政部長。John Jay則是美國首任聯邦最高法院首席大法官。在文集中，大多數文章是由Hamilton執筆的，Madison也對文集做出了重大的貢獻，而Jay因為患病，只寫了5篇文章。

文集的第10篇和第51篇通常被認為是文集中最有影響力的兩篇作品。其中，第10篇文章提倡建立一個強大的共和國，並包括了各黨派的討論；而第51篇則解釋了分權制度的必要性。文集第84篇也非常重要，因為這篇文章與後來的美國權利法案有著重大的聯繫。

文本分類的前置流程

流程圖:



(1) 文字清理

資料為 77 篇文件，其中 51 篇為 H 所作，14 篇為 M 所作，12 篇未知

(a)清理文檔

“數字”跟文中運用”<p>,<Q>,<c>,…”等來表示標點符號，這些對我們來說沒有幫助所以我們在此去除之。

因為在英文書寫上首字需要大寫，若不進行把字母都轉成小寫的話，會造成同一單字佔據兩個欄位，所以我們把字母都統一轉換成小寫。

“停止詞(stopwords)”泛指不具鑑別力的無意義單詞，舉例像是 i,she,a,your…等字詞，而移除停止詞的概念最早由 Hans Peter Luhan(1957)提出，在此我們使用的是 tm 套件中的”English”停止詞清單去進行停止詞的刪除。

(b)斷詞

因為文檔均為英文，在此我們使用空白進行斷詞處理。

“this is a pen”　== [“this”,”is”,”a”,”pen”]

(2) 資料量化

由於每篇文章的總字數不同，每個字詞在不同文章出現的次數可能深受其影響，無法進行比較，比如說文字 a 在文件 1 中出現 30 次，文字 a 在文件 2 出現 2 次，這樣是否代表文字 a 在文件 1 中比較重要?答案是不一定，若是文件 1 有 10000 個字，文件 2 只有 30 個字，則應該是文字 a 對文件 2 比較重要才對。

另一個問題則是，時常出現在文件中的常用詞，其可能不具備特別的意義，像是"the","a"。

所以若直接使用字的次數來當作權重，可能其結果會較為不良，所以我們權重的方式改採用 TF-IDF，下面將會進行介紹。

(a)TF-IDF 名詞簡介

(i)詞頻（term frequency，TF）:

指的是某一個給定的詞語在該文件中出現的頻率，舉例說，若是在第一篇文章中，共有 100 個字詞，其中"eat"出現在第一篇文章中的次數為 10 次，其 tf 值即為 0.1。

以頻率來看待文字的重要性，讓文章與文章之間比較有比較性，但其仍不適單獨拿來當作相似度評價標準，因為常用詞"the"這些，其頻率在各篇文章中通常出現頻率較高，可能導致分類結果較差。

(ii)逆向文件頻率（inverse document frequency，IDF）:

用來處理常用字的問題，假設總文章數為 20 篇，"eat"在 10 篇文章中出現，其 idf 值即為 log(20/10)。

因為 log 為遞增函數，當"eat"在越多文章中出現，其 log 內的值會越小，其 idf 值反而會低；反之，若越少出現的字詞，其 idf 值就會越大，說明其在全部文檔中具有良好的鑑別能力。

簡單來說，TF 表達了詞語 t 對某篇文件的重要性，IDF 表達了詞語 t 對整個文件集的重要性。

(iii)TF-IDF:

對區別文件最有意義的詞語，應該是那些在文件中出現頻率高，而在整個文件集合的其他文件中出現頻率少的詞語。

(3) 文件-詞語矩陣(Document-Term matrix)

列為文件，行為字詞，中間為其所對應之 TF-IDF 值。

Terms Docs	able	absurd	accident	accordingly	acknowledge
1	0.001	0.004	0.007	0.003	0.005
2	0.000	0.000	0.000	0.000	0.000
3	0.002	0.000	0.000	0.000	0.000

(4) 變數(詞語)選取

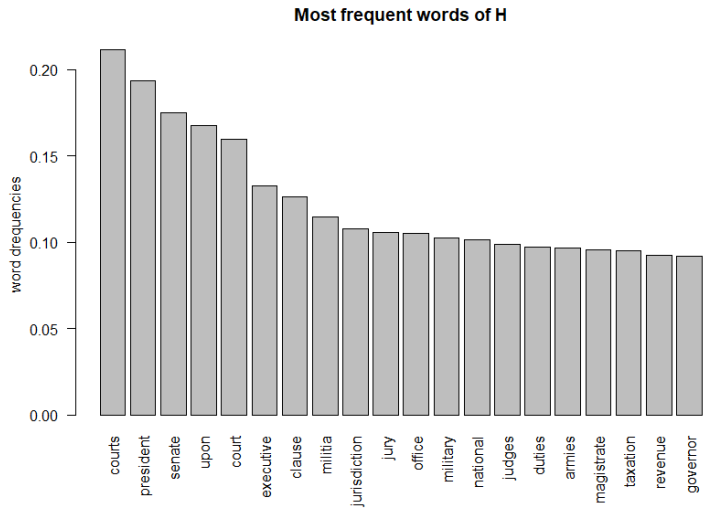
經過上述的文字處理，仍有 8119 個詞語，若直接當作 DTM 去做分類的話，其維度太高，可能效果會較差。

想法(1):

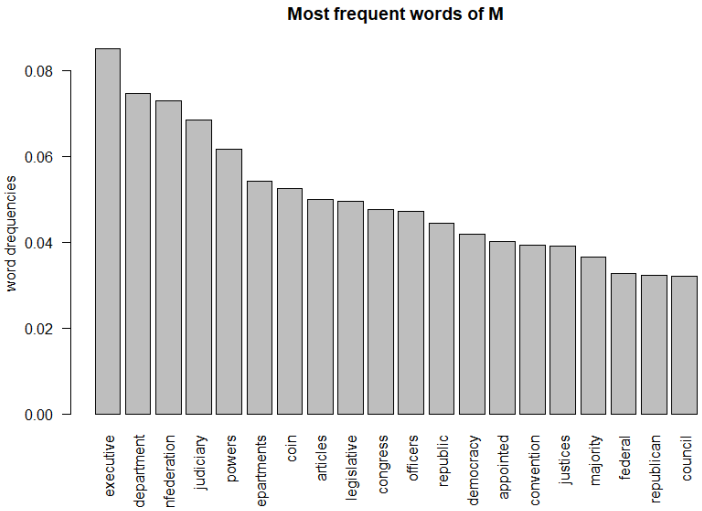
我們在此把 65 篇已知文章，分別對 51 篇 H 作者文件跟 14 篇 M 作者文件做出各自 DTM 矩陣，把 DTM 矩陣，直行加總後進行排序，由大到小找出前 20 個在個別作者文章中較為重要的詞語，因為考量到針對文章主題，可能會有一些共同的重要字詞，這些字詞對於分類作者的慣用字較不具鑑別力，所以我們再把兩位作者的前 20 個重要詞語找互斥項，找出其互斥項(38 個)。

在扣除掉相同的重要詞後，我們認為這些互斥項可以視為兩人寫作習慣上的具有鑑別力的詞語，所以我們把這些互斥項作為我們最終用來分類的 DTM。

(a) HAMILTON　TOP 20



(b) MADISON　TOP 20



(c)兩個作者的文字雲

左圖 HAMILTON TOP 50

右圖 MADISON TOP 50

從文字雲中方便我們看出較高頻的字詞，像是”executive”均為兩者常用字詞。

(d)互斥項

courts	president	senate	upon	court
clause	militia	jurisdiction	jury	office
military	national	judges	duties	armies
magistrate	taxation	revenue	governor	department
nfederation	judiciary	powers	erartments	Coin
articles	legislative	congress	officers	republic
democracy	appointed	convention	justices	majority
federal	republican	council		

想法(2):

取 H-M 差集，差集項 19 項

courts	president	senate	upon	court
clause	militia	jurisdiction	jury	office
military	national	judges	duties	armies
magistrate	taxation	revenue	governor	

分類流程:

步驟一 : 將測試資料集(65 筆)帶入各方法建立模型

步驟二 : 計算其訓練資料之錯誤率

步驟三 : 將測試資料帶入各模型並計算真實錯誤率(True Error Rate)

測試資料集(65 筆共 19 個文字變數)

courts	president	senate	upon	court
clause	militia	jurisdiction	jury	office
military	national	judges	duties	armies
magistrate	taxation	revenue	governor	

分類方法:

(一) 隨機森林

隨機森林是一個集成方法，將數個建立好的模型結果整合在一起，以提升預測的準確性，雖然這方法提供比較好的預測，但他在推論和解適度方面上就會有所限制。隨機森林由好幾個決策樹組合而成，而不同決策樹是由不同抽取的預測變數與觀察值所組成，所以每一棵樹的模型也不盡相同，也正因為如此，是由隨機建立的樹所組成的模型，故稱為隨機森林。

流程:

Step 1 : 將訓練資料集使用拔靴法製造出更多的樣本。

Step 2 : 生成更多的決策樹，每棵決策樹皆由隨機的方式抽取變數及觀察值組成。

Step 3 : 生成的每棵樹都不進行修剪。

Step 4 : 重複 Step 1 – Step 3，獲得 N 棵隨機決策樹。

Step 5 : 將 N 棵樹的預測進行投票，選取最適合的預測。

結果:

Apparent Rate (表面錯誤率: 1.53 %)

	1	2
1	51	0
2	1	13

True Error Rate (真實錯誤率: 0%)

	1	2
1	0	0
2	0	12

(二) 羅吉斯迴歸

羅吉斯迴歸適用於依變數為二元類別的情形，羅吉斯迴歸不需要考慮資料服從常態性假設，帶入 Logistic regression 模型進行計算並算出 Apparent Rate 與 True Error Rate。

結果:

Apparent Rate (表面錯誤率: 1.53 %)

	1	2
1	51	0
2	2	12

True Error Rate (真實錯誤率: 0%)

	1	2
1	0	0
2	2	10

(三) SVM 支持向量機

分類資料是機器學習中的一項常見任務。 假設某些給定的資料點各自屬於兩個類別之一，而目標是確定新資料點將在哪個類中。對於支援向量機來說，資料點被視為 p 維向量，而我們想知道是否可以用 (p-1) 維超平面來分開這些點。這就是所謂的線性分類器。可能有許多超平面可以把資料分類。最佳超平面的一個合理選擇是以最大間隔把兩個類分開的超平面。因此，我們要選擇能夠讓到每邊最近的資料點的距離最大化的超平面。如果存在這樣的超平面，則稱為最大間隔超平面，而其定義的線性分類器被稱為最大間隔分類器，或者叫做最佳穩定性感知器。

流程:

Step 1： 將訓練資料帶入 SVM 模型，並跑出最佳的懲罰項及 gamma 值

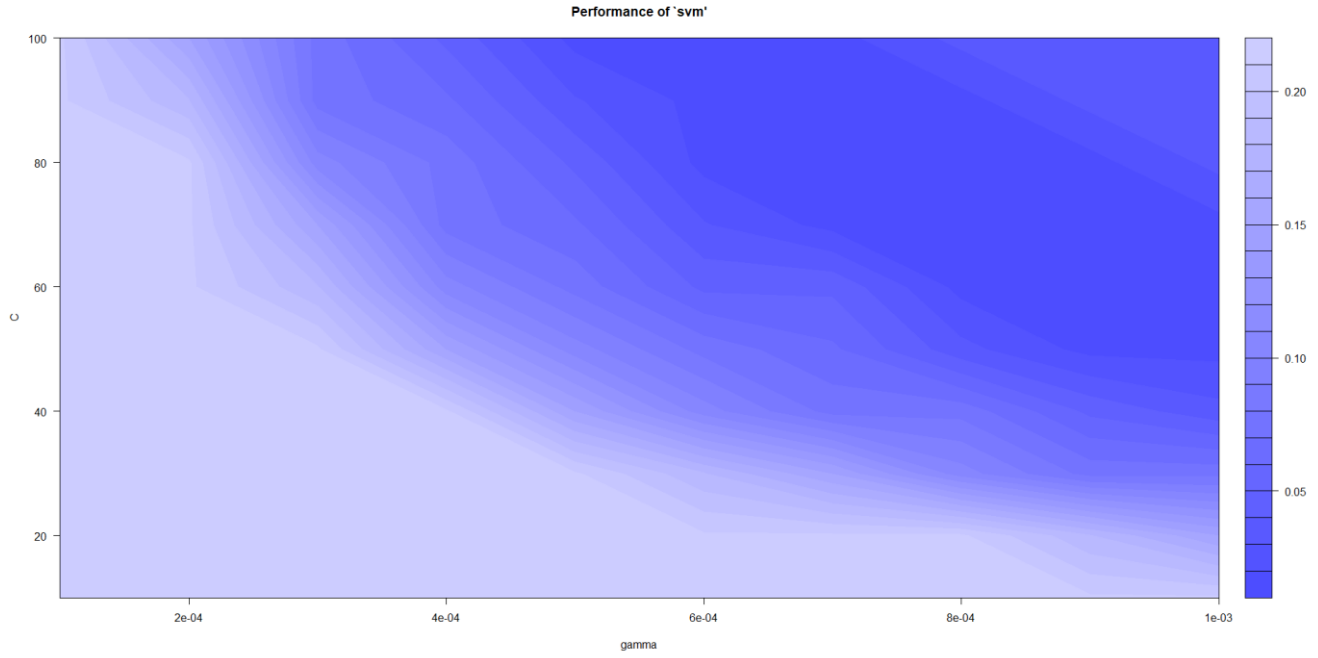
(cost： 在 Lagrange formulation 中的大 C，決定給被分錯資料的懲罰值)

(gamma： 值越大，資料點的影響力範圍越近，對超平面來說，近點的影響力權重較大，也容易造成 overfitting。)

Step 2： 找出最好的懲罰值及 gamma 值帶入模型

Step 3： 將測試資料集帶入模型預測結果

我們從下圖可以觀察出訓練資料集的最適值偏右上方，於是我們選取了 cost = 100 作為被分錯資料的懲罰值，並選取了 gamma 值為 0.001 作為點與點之間的權重。



結果:

Apparent Rate (表面錯誤率: 0 %)

	1	2
1	51	0
2	0	14

True Error Rate (真實錯誤率: 0 %)

	1	2
1	0	0
2	0	12

(四) LDA 線性判別分析

LDA 是一種監督式學習的方法，它的原理是將帶上標籤的數據(點)，通過投影的方法，投影到維度更低的空間中，使得投影後的點，會形成一群一群對應各類別的群體，相同類別的點，將會在投影後的空間更為接近。LDA 的基本概念是將高維的樣本變數空間投影到最佳鑑別向量空間，以達到抽取分類信息和壓縮特徵空間維度的效果，因此它是一種有效的特徵選取方法，使用這種方法能夠使投影後的觀察值在新的空間中有最小的類別內的距離和最大類別間的距離，及該模型在該空間中有最佳的可分離性。

流程:

Step 1 將訓練資料集帶入模型建立出最佳的投影空間

Step 2 找出最佳的線性分類器在投影空間中分割各類別

Step 3 將測試訓練集帶入模型中進行分類

結果:

Apparent Rate (表面錯誤率: 1.53 %)

	1	2
1	50	1
2	0	14

True Error Rate (真實錯誤率: 0 %)

	1	2
1	0	0
2	0	12

(五) 結論:

在建立的模型時候，因為我們已經確定了測試資料集的答案，所以我們會傾向選取適合這些測試資料集的模型方法，所以我們在一開始有試過了很多組變數下去建立模型，發現使用使用了兩個作者的差集的文字(共 19 個變數)進行建模，得到的真實準確率最高，但如果今日我們沒有真實資料的類別時，我們可能就不會傾向選擇此方法，我們認為使用兩個作者各前 20 名的 TF*IDF 文字中互斥的單字，較符合邏輯的去進行分類，因為差集所得到的單字已經沒有另一個作者使用習慣的字詞，會喪失一些重要的資訊，得到的結果可能會不好。

我們進一步的猜測差集項的變數在此題目中會分類好的原因，有可能是因為訓練資料集中的篇數幾乎為 Hamilton 作者所撰寫(65 篇中有 51 篇)，所以在分類時，考慮 Hamilton 作者所習慣的字詞可能會是一個不錯的方法，在使用了四個分類方法所得到的結果中，將 12 篇測試資料集的文章帶入計算，發現其分錯的篇章個數很少，除了羅吉斯迴歸會分錯 2 篇文章外，其餘的方法都全對，可能的原因我們尚未想出一個合理的解釋，但是是一個值得探討的議題。

附錄:Table of Essay

Author	Title	Date
Alexander Hamilton	General Introduction	October 27,1787
	Concerning Dangers from Dissensions Between the States	November 14,1787
	The Same Subject Continued: Concerning Dangers from Dissensions Between the States	November 15,1787
	The Consequences of Hostilities Between the States	November 20, 1787
	The Union as a Safeguard Against Domestic Faction and Insurrection	November 21,1787
	The Utility of the Union in Respect to Commercial Relation and a Navy	November 24,1787
	The Utility of the Union In Respect to Revenue	November 27,1787
	Advantage of the Union in Respect to Economy in Government	November 28,1787
	The Insufficiency of the Present Confederation to Preserve the Union	December 1, 1787
	The Same Subject Continued: The Insufficiency of the Present Confederation to Preserve the Union	December 4, 1787
	The Same Subject Continued: The Insufficiency of the Present Confederation to Preserve the Union	December 5, 1787
	Other Defects of the Present Confederation	December 12, 1787
	The Same Subject Continued: Other Defects of the Present Confederation	December 14, 1787
	The Necessity of a Government as Energetic as the One Proposed to the Preservation of the Union	December 18, 1787
	The Powers Necessary to the Common Defense Further Considered	December 19, 1787
	The Same Subject Continued: The Powers Necessary to the Common Defense Further Considered	December 21, 1787
	The Idea of Restraining the Legislative Authority in Regard to the Common Defense Considered	December 22, 1787
	The Same Subject Continued: The Idea of Restraining the Legislative Authority in Regard to the Common Defense Considered	December 25, 1787
	The Same Subject Continued: The Idea of Restraining the Legislative Authority in Regard to the Common Defense Considered	December 26, 1787
	Concerning the Militia	January 9, 1788
	Concerning the General Power of Taxation	December 28, 1787
	The Same Subject Continued: Concerning the General	January 1, 1788

	Power of Taxation	
	The Same Subject Continued: Concerning the General Power of Taxation	January 2, 1788
	The Same Subject Continued: Concerning the General Power of Taxation	January 2, 1788
	The Same Subject Continued: Concerning the General Power of Taxation	January 5, 1788
	The Same Subject Continued: Concerning the General Power of Taxation	January 5, 1788
	The Same Subject Continued: Concerning the General Power of Taxation	January 8, 1788
	Concerning the Power of Congress to Regulate the Election of Members	February 22, 1788
	The Same Subject Continued: Concerning the Power of Congress to Regulate the Election of Members	February 23, 1788
	The Same Subject Continued: Concerning the Power of Congress to Regulate the Election of Members	February 26, 1788
	The Powers of the Senate Continued	March 7, 1788
	Objections to the Power of the Senate To Set as a Court for Impeachments Further Considered	March 8, 1788
	The Executive Department	March 11, 1788
	The Mode of Electing the President	March 12, 1788
	The Real Character of the Executive	March 14, 1788
	The Executive Department Further Considered	March 15, 1788
	The Duration in Office of the Executive	March 18, 1788
	The Same Subject Continued, and Re-Eligibility of the Executive Considered	March 19, 1788
	The Provision For The Support of the Executive, and the Veto Power	March 21, 1788
	The Command of the Military and Naval Forces, and the Pardoning Power of the Executive	March 25, 1788
	The Treaty Making Power of the Executive	March 26, 1788
	The Appointing Power of the Executive	April 1, 1788
	The Appointing Power Continued and Other Powers of the Executive Considered	April 2, 1788
	The Judiciary Department	May 28, 1788 (book) June 14, 1788 (newspaper)
	The Judiciary Continued	May 28, 1788 (book)

		June 18, 1788 (newspaper)
	The Powers of the Judiciary	June 21, 1788
	The Judiciary Continued, and the Distribution of the Judicial Authority	June 25, 1788 and June 28, 1788
	The Judiciary Continued	July 2, 1788
	The Judiciary Continued in Relation to Trial by Jury	July 5, 1788, July 9, 1788 and July 12, 1788
	Certain General and Miscellaneous Objections to the Constitution Considered and Answered	July 16, 1788, July 26, 1788 and August 9, 1788
	Concluding Remarks	August 13, 1788 and August 16, 1788
John Jay	Concerning Dangers from Foreign Force and Influence	October 31, 1787
	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	November 3, 1787
	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	November 7, 1787
	The Same Subject Continued: Concerning Dangers from Foreign Force and Influence	November 10, 1787
	The Powers of the Senate	March 5, 1788
James Madison	The Same Subject Continued: The Union as a Safeguard Against Domestic Faction and Insurrection	November 22, 1787
	Objections to the Proposed Constitution From Extent of Territory Answered	November 30, 1787
	The Same Subject Continued: The Insufficiency of the Present Confederation to Preserve the Union	December 7, 1787
	The Same Subject Continued: The Insufficiency of the Present Confederation to Preserve the Union	December 8, 1787
	The Same Subject Continued: The Insufficiency of the Present Confederation to Preserve the Union	December 11, 1787
	Concerning the Difficulties of the Convention in Devising a Proper Form of Government	January 11, 1788
	The Same Subject Continued, and the Incoherence of the Objections to the New Plan Exposed	January 12, 1788
	The Conformity of the Plan to Republican Principles	January 18, 1788
	The Powers of the Convention to Form a Mixed Government Examined and Sustained	January 18, 1788
	General View of the Powers Conferred by the Constitution	January 19, 1788

	The Powers Conferred by the Constitution Further Considered	January 22, 1788
	The Same Subject Continued: The Powers Conferred by the Constitution Further Considered	January 23, 1788
	Restrictions on the Authority of the Several States	January 25, 1788
	The Alleged Danger From the Powers of the Union to the State Governments Considered	January 26, 1788
	The Influence of the State and Federal Governments Compared	January 29, 1788
	The Particular Structure of the New Government and the Distribution of Power Among Its Different Parts	January 30, 1788
	These Departments Should Not Be So Far Separated as to Have No Constitutional Control Over Each Other	February 1, 1788
	Method of Guarding Against the Encroachments of Any One Department of Government	February 2, 1788
	Periodic Appeals to the People Considered	February 5, 1788
	The Structure of the Government Must Furnish the Proper Checks and Balances Between the Different Departments	February 6, 1788
	The House of Representatives	February 8, 1788
	The Same Subject Continued: The House of Representatives	February 9, 1788
	The Apportionment of Members Among the States	February 12, 1788
	The Total Number of the House of Representatives	February 13, 1788
	The Same Subject Continued: The Total Number of the House of Representatives	February 16, 1788
	The Alleged Tendency of the New Plan to Elevate the Few at the Expense of the Many	February 19, 1788
	Objection That The Number of Members Will Not Be Augmented as the Progress of Population Demands Considered	February 20, 1788
	The Senate	February 27, 1788
	The Senate Continued	March 1, 1788