2010-03-16

Extensions of Nearest Shrunken Centroid Method for Classification

Tomohiko Funai
Brigham Young University - Provo

Extensions of Nearest Shrunken Centroid Method for Classification

Tomohiko Funai

A project submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Bruce G. Schaalje, Chair
David Whiting
John Lawson

Department of Statistics

Brigham Young University

April 2010

ABSTRACT

Extensions of Nearest Shrunken Centroid Method for Classification

Tomohiko Funai

Department of Statistics

Master of Science

Stylometry assumes that the essence of the individual style of an author can be captured using a number of quantitative criteria, such as the relative frequencies of noncontextual words (e.g., or, the, and, etc.). Several statistical methodologies have been developed for authorship analysis. Jockers et al. (2009) utilize Nearest Shrunken Centroid (NSC) classification, a promising classification methodology in DNA microarray analysis for authorship analysis of the Book of Mormon. Schaalje et al. (2010) develop an extended NSC classification to remedy the problem of a missing author. Dabney (2005) and Koppel et al. (2009) suggest other modifications of NSC. This paper develops a full Bayesian classifier and compares its performance to five versions of the NSC classifier using the Federalist Papers, the Book of Mormon text blocks, and the texts of seven other authors. The full Bayesian classifier was superior to all other methods.

Keywords: machine learning, discriminant analysis, authorship, attribution, fudge factor, shrinkage

ACKNOWLEDGMENTS

First, I would like to thank my advisor, Professor Bruce G. Schaalje, for his continual support. He was always there to listen to me and to give me advice. He is most responsible for helping me complete the writing of this project. I would also like to thank all of the members of graduate committee, Professor David Whiting and Professor John Lawson, who gave me valuable insight and support.

I also would like to thank my wife, Masami. Without her, I could not have finished this project. She was always there to meet with me and talk about my ideas and to help me by asking good questions. This project is a testament of our support for each other.

CONTENTS

CHAPTER 1

INTRODUCTION

The task of determining the true author of a text of disputed authorship has been a topic of interest for a quite some time (Koppel et al. 2009). Several methods for authorship attribution have been developed to improve the accuracy of the authorship analysis. With the development of computers, the capacities for analyzing large quantities of data have enhanced this effort. One notable advancement in authorship attribution methodology is machine learning. Since that advancement, numerous types of machine learning approaches have been introduced, including the naive Bayes classifier. These methods have worked quite well in many complex real-world classification problems (Zhang 2004).

Nearest shrunken centroid (NSC) classification (Tibshirani et al. 2002, 2003) is a promising classification methodology used in DNA microarray analysis. In genomic studies, a huge number of inputs (genes) are measured on relatively few individuals. The goal is to classify new individuals based on their gene intensities. Similar issues are present in authorship attribution since there can be hundreds of literary features measured for relatively few blocks of text.

In a recent study, Jockers et al. (2009) attempt to determine the chapter-by-chapter probabilities that seven potential authors wrote or contributed to writing the Book of Mormon. Using NSC, they conclude that three main authors contributed to the Book of Mormon: Sidney Rigdon, an early leader in the Mormon church, Solomon Spalding, a writer of historical fantasies, and Oliver Cowdery, who acted as a scribe during the "translation" of the Book of Mormon.

Schaalje et al. (2010) argue that the approach proposed by Jockers et al. (2009) is problematic. A key problem is the issue of actual authors not being represented in the training data. Jockers et al. (2009) assume "a closed set of candidate authors." Making modifications to NSC to determine missing authors, Schaalje et al. (2010) are able to more correctly determine each potential author's authorship probability for each chapter.

However, the methods employed by Schaalje et al. (2010) may not be optimal. For example, Dabney (2005) conjectures that the shrinkage methods employed by Tibshirani et al. (2002) are unhelpful in classifying new individuals. Koppel et al. (2009) claim that a full Bayesian approach is more effective in authorship identification than the approach of Tibshirani et al. (2002). While there are many methods that could be used for authorship analysis, we are interested in whether the claims of Dabney (2005) and Koppel et al. (2009) can be verified. The purpose of this paper is to investigate the methods of Dabney (2005) and Koppel et al. (2009) in order to improve the performance of NSC classification in authorship analysis.

CHAPTER 2

LITERATURE REVIEW

2.1 Authorship Analysis Methodologies

The scientific approach to authorship analysis is first proposed by Mendenhall in 1887, "in which single numeric function of a text is sought to discriminate between the authors" (Koppel et al. 2009). With the advancement of technologies and the development of statistical and computational methods, authorship analysis has gained greater attention in recent decades. Most recently, the machine learning approach has emerged. There have been numerious types of machine learning approaches, including neural networks (Li et al. 2006), boosted trees (Koppel et al. 2009), k-nearest neighbors (Kjell et al. 2006), rule learners (Forsyth and Holmes 1996), naive Bayes (Peng et al. 2006), random forests (Dabney 2005), support vector machines (Li et al. 2006), Winnow (Genkin et al. 2007), and Bayesian regression (Genkin et al. 2007).

The naive Bayes classifier uses Bayes theorem together with the assumption of independent features. Naive Bayes classifier has worked quite well in many complex real-world situations. Caruana and Niculescu-Mizil (2006) show that the naive Bayes classifier outperformed other approaches, such as boosted trees and random forests.

2.2 DNA Microarray Analysis

DNA microarray analysis is used to classify samples as well as identify genes associated with biological processes of interest. In microarray analysis, an arrayed series of thousands of microscopic spots of DNA features is used to measure gene expression levels. Once target genes are identified, further investigations are carried out to validate the relationship of their biological functions to the process of interest (Tan et al. 2003).

Microarray analysis, however, poses a large number of problems, including multiple comparisons. Even if the P-value assigned to a gene indicates that it is statistically significant, the high number of genes on an array inflates the Type I error rate. However, if the Bonferroni correction is applied, very few genes are found to be statistically significant (Tan et al. 2003). Thus, Nadon and Shoemaker (2002) and Nakagawa (2004) argue that the Bonferroni correction causes the power of a microarray to decrease to "unacceptable levels."

Many classification methods have been designed for microarray data. These include random forests, support vector machines, CART, neural networks, generalized models, and the Linear Discriminant Analysis (Dabney 2005). The Linear Discriminant Analysis (LDA), a classical method that has been shown to perform well compared to the aforementioned methods (Dudoit et al. 2002; Lee et al. 2005), will be discussed.

2.3 Linear Discriminant Analysis

LDA, first developed by Fisher (Rencher 2002), has proven useful in DNA microarray analysis (Dudoit et al. 2002; Lee et al. 2005). Each class is characterized by its mean vector or "centroid." Then unknown samples are assigned to the class with the closest centroid. By doing this, the misclassification error is minimized (Mardia et al. 1979).

Suppose that the goal is to classify an unknown sample into one of the K classes. Suppose also that there are n_k training samples from class k where $k = 1, \cdots, K$. In each sample, there is an $m \times 1$ expression profile. LDA assumes that expression profiles from class k are distributed as multivariate normal, $MVN(\mu_k, \Sigma)$, with the additional assumption that all variance-covariance matrices Σ_k are equal. Given prior probability π_k that the unknown sample comes from class k, Bayes theorem states that the posterior distribution that profile x comes from class k is

$$\begin{aligned} P(Y=k|X=x) &\propto f(x|\theta) \times \pi_k && (2.1) \\ &\propto \exp\left\{(x-\mu_k)'\Sigma^{-1}(x-\mu_k) - 2\log(\pi_k)\right\}, && (2.2) \end{aligned}$$

where $(x-\mu_k)'\Sigma^{-1}(x-\mu_k)$ is the Mahalanobis distance between x and μ_k (Rencher 2002). LDA assigns the sample to class $\hat{y}$ where

$$\hat{y} = \arg\min_k \left[(x-\mu_k)'\Sigma^{-1}(x-\mu_k) - 2\log(\pi_k) \right]. \tag{2.3}$$

If it is assumed that all genes are independent, the classification criterion can be simplified to

$$\hat{y} = \arg\min_k \left[\sum_{i=1}^{m} \left(\frac{x_i - \mu_{ik}}{\sigma_i} \right) - 2\log(\pi_k) \right]. \tag{2.4}$$

LDA estimates μ_{ik} as $\bar{x}_{ik} = \frac{1}{n_k}\sum_{j=1}^{n_k} x_{ijk}$ and Σ as the pooled within-group covariance matrix. Test statistics such as the F statistic (Dudoit et al. 2002) can be computed to identify informative genes. LDA is commonly thought of as a nearest centroid classifier (Dabney 2005). If the priors are equal, the nearest centroid method simply takes the gene expression profile of each sample and computes its squared distance from each of the class centroids. The predicted class is the one whose centroid is closest to the profile of the test sample. Although the LDA method has good classification properties, it can still yield high misclassification rates (Tibshirani et al. 2002, 2003).

2.4 Nearest Shrunken Centroid Classification

Tibshirani et al. (2002) propose the nearest shrunken centroid (NSC) classifier, a modification of the LDA classification method with independent gene expression values (equation 2.4). Although much statistical development has occurred in the last 6–7 years, the NSC classification method continues to be used in numerous microarray analyses (Volinia et al. 2006).

To reduce the misclassification rate, Tibshirani et al. (2002) propose the estimation of μ_k with "shrunken centroids" rather than sample centroids. If prior probabilities are equal, the class with the closest shrunken centroid is the predicted class for the unknown sample.

Shrunken Centroids

NSC shrinks the class centroids toward the overall centroid after standardizing by the within-class deviation for each gene. Let x_{ijk} be the expression for gene $i = 1, \cdots, p$ in sample $j = 1, \cdots, n_k$ of class k. Instead of using F statistics to identify the most informative genes, NSC uses soft thresholding. Let

$$\bar{x}_{ik} = \sum_{j=1}^{n_k} \frac{x_{ijk}}{n_k}$$

be the mean for gene i in class k and

$$\bar{x}_i = \sum_{k=1}^{K} \sum_{j=1}^{n_k} \frac{x_{ijk}}{n_k}$$

be the overall mean for gene i. Then a t-statistic, d_{ik}, for gene i, comparing class k to the overall mean, is defined as

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \cdot (s_i + s_0)}, \tag{2.5}$$

where

$$s_i^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{j=1}^{n_k} (x_{ij} - \bar{x}_{ik})^2, \tag{2.6}$$

and

$$m_k = \sqrt{1/n_k + 1/n}, \text{ where } n = \sum_{k=1}^{K} n_k. \tag{2.7}$$

The $m_k \times s_i$ in equation 2.5 is the estimated standard error of $\bar{x}_{ik} - \bar{x}_i$. However, s_0 in the denominator of equation 2.5 is a *fudge factor*. If $m_k \times s_i$ is very small, d_{ik} would be inflated to a large value. The fudge factor is intended to guard against very large d_{ik} statistics for very small standard errors (Tibshirani et al. 2002, 2003). The fudge factor is usually obtained as the median value of s_i over the set of genes. The fudge-factor adjusted centroid is then

$$\bar{x}_{ik} = \bar{x}_i + m_k(s_i + s_0)d_{ik}. \tag{2.8}$$

Tibshirani et al. (2002, 2003) further propose to shrink d_{ik} toward zero using *soft thresholding*. In this process, the absolute value of each d_{ik} is reduced by the quantity Δ (called the shrinkage parameter),

$$d'_{ik} = \text{sign}(d_{ik})(|d_{ik}| - \Delta)_+. \tag{2.9}$$

The subscript + means the positive part. That is,

$$(|d_{ik}| - \Delta)_+ = \begin{cases} |d_{ik}| - \Delta & |d_{ik}| > \Delta \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the shrunken centroids are

$$\tilde{x}_{ik} = \bar{x}_i + m_k(s_i + s_0)d'_{ik}. \tag{2.10}$$

If Δ causes d_{ik} to shrink to zero for all classes, then the centroid component for gene i is $\bar{x}_i$ for all classes. Thus, gene i does not contribute to the nearest centroid classification. A small Δ allows more genes to contribute to classification.

Under certain conditions, shrinking centroids using soft thresholding produces better estimates of the true centroid (Donoho and Johnstone 1994). Many genes are eliminated if Δ is chosen appropriately.

Cross Validation

To find the optimal Δ, cross-validation is performed. Cross-validation involves repeatedly leaving out a small percentage of the training data, then using the trained model to predict the membership for the observations that were left out (Tibshirani et al. 2002). The misclassification rate for the observations that were left out is utilized as the criterion for the shrinkage parameter Δ. Tibshirani et al. (2003) recommend choosing the largest value of Δ that achieves the minimal classification error because this leads to the simplest model.

Usefulness of NSC in Microarray Analysis

The NSC classification method has been successfully applied in many DNA microarray classification problems and has become a standard method. NSC is favorable because it is easy to implement and interpret, and seems to produce good results (Tibshirani et al. 2002, 2003).

Application of NSC to Authorship Analysis

LDA and NSC are frequently used in authorship analysis (Moon et al. 2006). Since authorship analysis involves hundreds of literary features measured in only a few blocks of text to identify the correct author, the objective is similar to the microarray analysis.

In authorship analysis, the literary features are usually relative frequencies of noncontextual words (words with primarily grammatical functions such as "the," "of," and "about"). The goal is to assign a text of unknown authorship to one of the potential authors. To conform with the notation of Schaalje et al. (2010), we let x_{ijl} be the value of authorship feature $j(j = 1, \cdots, r)$ for text block $l(l = 1, \cdots, n_i)$ of author $i(i = 1, \cdots, m)$.

Jockers et al. (2009) consider 130 literary features (for complete list of literary features used, see Jockers et al. 2009) and seven potential authors. Let x^* denote the vector of features for a new text and π_i denote the prior probability that the new text is written by ith author. Let $f_k(x)$ be the multivariate density function for author k. It is assumed that the true author of the new text is unknown even though it is known that one of the m candidates is the true author. It is also assumed that the densities follow the multivariate normal distribution and that the features are mutually independent. Then the posterior probability that author k is the author of the new text is

$$p\left(k|x^*\right) = \frac{\pi_k f_k(x^*)}{\pi_i \sum_{i=1}^{m} f_i(x^*)} \tag{2.11}$$

$$= \frac{\pi_k \exp\left[-\frac{1}{2}\sum_{j=1}^{r}\left(\frac{x_j^* - \mu_{kj}}{\sigma_j}\right)^2\right]}{\sum_{i=1}^{m}\pi_i \exp\left[-\frac{1}{2}\sum_{j=1}^{r}\left(\frac{x_j^* - \mu_{ij}}{\sigma_j}\right)^2\right]}, \tag{2.12}$$

where σ_j and μ_{ij} in equation 2.12 are the common variance for feature j across the authors and the mean of author i for feature j, respectively. The quantity σ_j is estimated by s_j, the pooled within-author variance for the feature j. The quantity μ_{ij} is estimated with $\tilde{x}_{ij}$, the threshold and shrunken sample mean.

Jockers et al. (2009) classify each of the 239 chapters in the Book of Mormon based on the author having the highest posterior probability, $\hat{p}(i|x^*)$. That is, they choose author $\hat{h}$, where

$$\hat{h} = \arg\max\left[\hat{p}(i|x^*)\right]. \tag{2.13}$$

The seven potential authors are Oliver Cowdery, Parley Pratt, Sidney Rigdon, Solomon Spalding, Isaiah and/or Malachi (from the Bible), Henry Wadsworth Longfellow, and Joel Barlow.

Schaalje et al. (2010) claim that there are three problematic issues with the NSC classification model proposed by Jockers et al. (2009). The key issue is the assumption of a closed set of potential authors. Schaalje et al. (2010) suggest an extension of the NSC model to improve the results. They propose that the posterior probability that the kth potential author is the true author of the new text should be calculated as

$$p\left(k|x^*\right) = \frac{\pi_k \exp\left[-\frac{1}{2}\sum_{j=1}^{r}\left(\frac{x_j^* - \tilde{x}_{kj}}{s_j}\right)^2\right]}{\left(\sum_{i=1}^{m}\pi_i \exp\left[-\frac{1}{2}\sum_{j=1}^{r}\left(\frac{x_j^* - \tilde{x}_{ij}}{s_j}\right)^2\right]\right) + \pi_{m+1}\exp\left[-\frac{1}{2}\sum_{j=1}^{r} a_j^2\right]}, \tag{2.14}$$

where

$$a_j = \min\left(\max_i \left|\frac{x_j^* - \tilde{x}_{ij}}{s_j}\right|, \lambda\right), \tag{2.15}$$

and λ is a tuning constant "representing the maximum allowed distance for components of x^* from corresponding components of the centroid of the latent author." The additional term in the denominator in equation 2.14, $\pi_{m+1} \exp\left[-\frac{1}{2}\sum_{j=1}^{r} a_j^2\right]$, takes into account the possibility of a missing author. This extra term adjusts the ratio to become a proper probability density function and avoids inflation (Schaalje et al. 2010). The posterior probability that the latent author is the true author of the new text, x^*, could be calculated as

$$p\,(m+1|x^*) = \frac{\pi_{m+1} \exp\left[-\frac{1}{2}\sum_{j=1}^{r} a_j^2\right]}{\left(\sum_{i=1}^{m} \pi_i \exp\left[-\frac{1}{2}\sum_{j=1}^{r}\left(\frac{x_j^* - \tilde{x}_{ij}}{s_j}\right)^2\right]\right) + \pi_{m+1} \exp\left[-\frac{1}{2}\sum_{j=1}^{r} a_j^2\right]}. \tag{2.16}$$

Table 2.4 shows the classification results for the original NSC (Jockers et al. 2009) and the extended NSC (Schaalje et al. 2010). The original NSC classified most of the chapters to Isaiah, Rigdon, and Spalding, while the extended NSC classified almost three-fourths of all of the chapters to the latent author. It is interesting that Isaiah is a strong contender for the Book of Mormon authorship since some of the Book of Mormon chapters are also verbatim to the Isaiah chapters in the Bible. So we conclude that the author of the Book of Mormon is still unknown. The extended NSC classification does not agree with the Spalding-Rigdon theory (Schaalje et al. 2010).

2.5 Classification methods closely related to NSC

Classification to the Nearest Centroid

Dabney (2005) propose the classification to the nearest centroid method (ClaNC). ClaNA does not use fudge factors, does not shrink centroids and carries out the class-specific gene selection.

Table 2.1: Classification results for original NSC (Jockers et al. 2009) and extended NSC (Schaalje et al. 2010).

Author	Original NSC	Extended NSC
Isaiah	68 (28.45%)	35 (14.64%)
Early/Late Rigdon	96 (40.17%)	17 (7.11%)
Spalding	37 (15.48%)	8 (3.35%)
Smith	29 (12.13%)	2 (0.84%)
Cowdery	4 (1.67%)	2 (0.84%)
Pratt	5 (2.09%)	0 (0.00%)
Latent	NA	175 (73.22%)

Dabney (2005) claims that while the fudge factor can denoise the data and stabilize the statistics, it also might increase the misclassification error. Tibshirani et al. (2003) include s_0 in the denominator for equation 2.5 to "guard against the possibility of large d_{ik} arising by chance from genes with low expression values." Dabney (2005) worries that fudge factors may remove many informative genes with moderate differences.

Ignoring the fudge factor, the centroid is obtained as

$$\bar{x}_{ik} = \bar{x}_i + m_k s_i d'_{ik}, \tag{2.17}$$

where d'_{ik} is from equation 2.9.

Dabney (2005) also proposes removing the shrinkage parameter, Δ. He notes that the "shrinkage procedure makes the class centroids look more similar to each other. It is unclear why this should be expected to make it easier to distinguish the classes from each other." In other words, shrinkage makes groups less identifiable since the centroids are similar to each other.

In the NSC classification, the gene selection process is accomplished by calculating

$$\tilde{d}_{ik} = d_{ik} I\left(|d_{ik}| > \Delta\right). \tag{2.18}$$

Note that in a regular NSC analysis, the gene selections are *class-nonspecific*. Dabney (2005) proposes the *class-specific* threshold based on

$$\tilde{d}_{ik} = d_{ik} I(|d_{ik}| > \Delta_k). \tag{2.19}$$

Therefore, the shrunken centroid using ClaNC becomes

$$\tilde{x}_{ik} \quad = \quad \bar{x}_i + m_k \left(s_i + s_0\right) \tilde{d}_{ik}. \tag{2.20}$$

The class-specific threshold is performed because a particular class, say k^*, might be heterogeneous.

Dabney (2005) also suggests that if shrinkage is to be used at all, centroids should be shrunken toward a different overall centroid. He suggests shrinking the centroids towards the overall centroid for classes, instead of towards the overall centroid for features. However, Dabney does not make this argument mathematically explicit.

Full independent Bayesian classification

The NSC classification method is often referred to as a "naive Bayes" method, because it assumes that the features are independent and it pretends thats the sample centroids and variances are the true values of the corresponding parameters (Koppel et al. 2009; Langseth and Nielsen 2006). Independence of each feature is assumed since there are not enough degrees of freedom to estimate all of the variance components.

The full independent Bayesian prediction can be developed in the classification analysis to take into account the uncertainties in the unknown parameters of the multivariate normal likelihood, μ_{ij} and σ_j^2. The procedure would involve specifying priors for the parameters, using MCMC (Gelman et al. 2004) to sample from the posterior distributions based on the training data and then sampling from the posterior distributions of class probabilities for each test sample. Comparison of these last posterior distributions would lead to a classification decision. One advantage of this method is that the shrinkage would be automatic; it would be inherent in the Bayesian paradigm. The only problem is that of selecting features. The general idea of the Bayesian classification is discussed by Shi et al. (2003).

CHAPTER 3

METHODS AND RESULTS

3.1 Shrinkage toward class means

In this section, we investigate Dabney's (2005) idea to shrink the centroids towards class means rather than feature means. We formulate equations for the shrinkage factors.

Shrinkage towards class means of $\bar{x}_{ij}$ can be accomplished using

$$\tilde{x}_{ij} = \bar{x}_i + q_{ij} \times \tilde{d}_{ij}, \tag{3.1}$$

where $\tilde{d}_{ij}$ is analogous to equation 2.5 and q_{ij} is the standard error for $\bar{x}_{ij} - \bar{x}_i$. To find the standard error for $\bar{x}_{ij} - \bar{x}_i$, recall that the data vector for individual l in class i is distributed as a multivariate normal, $X_{il} \sim MVN(\theta_i, \mathbf{D})$, where θ_i is the mean vector for class i and

$$\mathbf{D} = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_r^2 \end{bmatrix}.$$

The standard error q_{ij} can be simplified to

$$\begin{aligned} q_{ij} &= \sqrt{Var\left(\bar{x}_{ij} - \bar{x}_i\right)} \\ &= \sqrt{Var\left(\bar{x}_{ij}\right) + Var\left(\bar{x}_i\right) - 2Cov\left(\bar{x}_{ij}, \bar{x}_i\right)}. \end{aligned}$$

Next, we derive each component of $Var\left(\bar{x}_{ij} - \bar{x}_i\right)$:

$$
\begin{aligned}
Var\left(\bar{x}_{ij}\right) &= Var\left(\sum_{l=1}^{n_i} x_{ijl}\right) \\
&= \frac{1}{n_i^2} Var\left(x_{ijl}\right) = \frac{\sigma_j^2}{n_i}, \\
Var\left(\bar{x}_i\right) &= Var\left(\frac{1}{rn_i}\sum_{j=1}^{r}\sum_{l=1}^{n_i} x_{ijl}\right) \\
&= \frac{1}{r^2 n_i^2} Var\left(\sum_{j=1}^{r}\sum_{l=1}^{n_i} x_{ijl}\right) \\
&= \frac{1}{r^2 n_i} Var\left(\sum_{j=1}^{r} x_{ijl}\right) = \frac{1}{r^2 n_i}\sum_{j=1}^{r}\sigma^2, \quad \text{and} \\
Cov\left(\bar{x}_{ij}, \bar{x}_i\right) &= Cov\left(\bar{x}_{ij}, \frac{1}{r}\sum_{a\neq j}\bar{x}_{ia} + \frac{\bar{x}_{ij}}{r}\right) \\
&= Cov\left(\bar{x}_{ij}, \frac{1}{r}\sum_{a\neq j}\bar{x}_{ia}\right) + Cov\left(\bar{x}_{ij}, \frac{\bar{x}_{ij}}{r}\right) \\
&= 0 + \frac{1}{r} Var\left(\bar{x}_{ij}\right) = \frac{1}{rn_i}\sigma_j^2.
\end{aligned}
$$

$Cov\left(\bar{x}_{ij}, \frac{1}{r}\sum_{a\neq j}\bar{x}_{ia}\right) = 0$ since D is a diagonal covariance matrix. Therefore,

$$
\begin{aligned}
q_{ij} &= \sqrt{\frac{1}{n_i}\sigma_j^2 + \frac{1}{r^2 n_i}\sum_{j=1}^{r}\sigma_j^2 - 2\left(\frac{1}{rn_i}\sigma_j^2\right)} \qquad (3.2) \\
&= \sqrt{\frac{\sum_{j=1}^{r}\sigma_j^2}{r^2 n_i} + \left(1 - \frac{2}{r}\right)\frac{\sigma_j^2}{n_i}}. \qquad (3.3)
\end{aligned}
$$

Thus, $\tilde{d}_{ij}$ is

$$
\tilde{d_{ij}} = \left(\frac{\bar{x}_{ij} - \bar{x}_i}{q_{ij}}\right) I\left(\left|\left(\frac{\bar{x}_{ij} - \bar{x}_i}{q_{ij}}\right)\right| > \Delta\right). \qquad (3.4)
$$

Combining these results with equation 3.1, we obtain

$$
\tilde{x}_{ij} = \bar{x}_i + \sqrt{\frac{\sum_{j=1}^{r}\sigma_j^2}{r^2 n_i} + \left(1 - \frac{2}{r}\right)\frac{\sigma_j^2}{n_i}} \times \left(\frac{\bar{x}_{ij} - \bar{x}_i}{q_{ij}}\right) I\left(\left|\left(\frac{\bar{x}_{ij} - \bar{x}_i}{q_{ij}}\right)\right| > \Delta\right). \qquad (3.5)
$$

3.2 Full independent Bayesian Approach to Classification

In this section, we develop the full independent Bayesian approach to classification by choosing appropriate priors and a classification strategy based on posterior distributions of class probabilities.

The *logit* link function was used so that the data span all real numbers. That is, $X \in (-\infty, \infty)$. We assume that the *logit* transformed vector of features X_{il} for individual l in class i is distributed as multivariate normal:

$$X_{il} \sim MVN\left(\theta_i, diag\left(\sigma_j^2\right) = \mathbf{D}\right).$$

Hence,

$$f(X_{il};\theta) = \frac{1}{(2\pi)^{r/2}\,|\mathbf{D}|^{1/2}} \exp\left[-\frac{1}{2}\left(X_{il}-\theta_i\right)'\mathbf{D}^{-1}\left(X_{il}-\theta_i\right)\right],$$

where $\theta_i \in (-\infty, \infty)$ and $\sigma_j^2 > 0$. Notice that the off-diagonal elements in the variance covariance matrix $\mathbf{D}$ are all zero. Thus, we assume that each feature has a different variance and that the features are independent.

The prior for θ_i is $\theta_i \sim MVN\left(\mu\mathbf{J}, \tau^2\mathbf{I}\right)$, where $\mathbf{I}$ is the identity matrix and $\mathbf{J}$ is a vector of one. That is,

$$\pi(\theta_i) \quad = \quad \left((2\pi)^{r/2}|\tau^2\mathbf{I}|^{1/2}\right)^{-1} \exp\left[-\frac{1}{2\tau^2}\left(\theta_i-\mu\mathbf{J}\right)'\left(\theta_i-\mu\mathbf{J}\right)\right].$$

$\mu\mathbf{J}$ and $\tau^2\mathbf{I}$ specify that all feature means have the same prior mean, μ, with prior variance τ^2.

The prior for σ_j^2 is an inverse Gamma distribution: $\sigma_j^2 \sim IG\left(\alpha, \beta\right)$. That is,

$$\pi(\sigma_j^2) \quad = \quad \left(\Gamma(\alpha)\beta^{\alpha}\right)^{-1}\left(\sigma_j^2\right)^{-(\alpha+1)} \exp\left[-\frac{1}{\beta\sigma_j^2}\right].$$

The inverse gamma distribution can take many different shapes over the positive real numbers. Therefore, the inverse gamma prior seems appropriate for σ_j^2.

The complete conditionals for θ_i and σ_j^2 can be found in closed form. The data are distributed as a multivariate normal distribution. That is,

$$f(X_{il};\theta) = \frac{1}{(2\pi)^{r/2}\,|D|^{1/2}} \exp\left[-\frac{1}{2}\left(X_{il}-\theta_i\right)' \mathbf{D}^{-1}\left(X_{il}-\theta_i\right)\right],$$

where $\theta_i \in (-\infty,\infty)$ and $\sigma_j^2 > 0$. The priors are

$$\begin{aligned}
\pi(\theta_i) &= \left((2\pi)^{r/2}|\tau^2\mathbf{I}|^{1/2}\right)^{-1} \exp\left[-\frac{1}{2}\left(\theta_i-\mu\mathbf{J}\right)'\left(\tau^2\mathbf{I}\right)^{-1}\left(\theta_i-\mu\mathbf{J}\right)\right] \quad \text{and} \\
\pi(\sigma_j^2) &= \left(\Gamma(\alpha)\beta^{\alpha}\right)^{-1}\left(\sigma_j^2\right)^{-(\alpha+1)} \exp\left[-\frac{1}{\beta\sigma_j^2}\right].
\end{aligned}$$

Thus, the joint distribution is

$$\begin{aligned}
\pi\left(X|\theta\right) &= \prod_{i=1}^{m}\prod_{l=1}^{n_i} \frac{1}{(2\pi)^{r/2}\,|\mathbf{D}|^{1/2}} \exp\left[-\frac{1}{2}\left(X_{il}-\theta_i\right)'\mathbf{D}^{-1}\left(X_{il}-\theta_i\right)\right] \\
&\times \prod_{i=1}^{m}\left((2\pi)^{r/2}|\tau^2\mathbf{I}|^{1/2}\right)^{-1} \exp\left[-\frac{1}{2}\left(\theta_i-\mu\mathbf{J}\right)'\left(\tau^2\mathbf{I}\right)^{-1}\left(\theta_i-\mu\mathbf{J}\right)\right] \\
&\times \prod_{j=1}^{r}\left(\Gamma(\alpha)\beta^{\alpha}\right)^{-1}\left(\sigma_j^2\right)^{-(\alpha+1)} \exp\left[-\frac{1}{\beta\sigma_j^2}\right] \\
&= \left(\frac{1}{(2\pi)^{r/2}\,|\mathbf{D}|^{1/2}}\right)^{-\sum_{i=1}^{m} n_i} \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}\left(X_{il}-\theta_i\right)'\mathbf{D}^{-1}\left(X_{il}-\theta_i\right)\right] \\
&\times \left((2\pi)^{r/2}|\tau^2\mathbf{I}|^{1/2}\right)^{-m} \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\left(\theta_i-\mu\mathbf{J}\right)'\left(\tau^2\mathbf{I}\right)^{-1}\left(\theta_i-\mu\mathbf{J}\right)\right] \\
&\times \left(\Gamma(\alpha)\beta^{\alpha}\right)^{-r}\left(\prod_{j=1}^{r}\sigma_j^2\right)^{-(\alpha+1)} \exp\left[-\frac{1}{\beta\sum_{j=1}^{r}\sigma_j^2}\right].
\end{aligned}$$

The complete conditional for θ_i is then as follows:

$$\pi\left[\theta_i\right] \propto \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}\left(X_{il}-\theta_i\right)'\mathbf{D}^{-1}\left(X_{il}-\theta_i\right) - \frac{1}{2}\sum_{i=1}^{m}\left(\theta_i-\mu\mathbf{J}\right)'\left(\tau^2\mathbf{I}\right)^{-1}\left(\theta_i-\mu\mathbf{J}\right)\right].$$

Note that $(\tau^2\mathbf{I})^{-1} = \frac{1}{\tau^2}\mathbf{I}$. In addition, all the summations with respect to index i can be dropped. Thus, the above equation becomes

$$\begin{aligned}
\pi[\theta_i] &\propto \exp\left[-\frac{1}{2}\sum_{l=1}^{n_i}(X_{il}-\theta_i)'\,\mathbf{D}^{-1}(X_{il}-\theta_i) - \frac{1}{2\tau^2}(\theta_i-\mu\mathbf{J})'(\theta_i-\mu\mathbf{J})\right] \\
&= \exp\left[-\frac{1}{2}\sum_{l=1}^{n_i}\left(X_{il}'\mathbf{D}^{-1}X_{il} - 2\theta_i'\mathbf{D}^{-1}X_{il} + \theta_i'\mathbf{D}^{-1}\theta_i\right) - \frac{1}{2\tau^2}\left(\theta_i'\theta_i - 2\mu'\theta_i + \mu\mathbf{J}'\mu\mathbf{J}\right)\right] \\
&\propto \exp\left[-\frac{(-2\tau^2\theta_i'\mathbf{D}^{-1}\sum_{i=1}^{n_i}X_{il} + \tau^2 n_i\theta_i'\mathbf{D}^{-1}\theta_i + \theta_i'\mathbf{I}\theta_i - 2\theta_i'\mu\mathbf{J})}{2\tau^2}\right] \\
&= \exp\left[-\frac{(\theta_i'(\tau^2 n_i\mathbf{D}^{-1}+\mathbf{I})\,\theta_i - 2\theta_i'(\tau^2\mathbf{D}^{-1}\sum_{l=1}^{n_i}X_{il} + \mu\mathbf{J}))}{2\tau^2}\right].
\end{aligned}$$

Thus,

$$\pi[\theta_i] \sim MVN\left(\left(\tau^2 n_i\mathbf{D}^{-1}+\mathbf{I}\right)^{-1}\left(\tau^2\mathbf{D}^{-1}\sum_{l=1}^{n_i}X_{il}+\mu\mathbf{J}\right), \left(\tau^2 n_i\mathbf{D}^{-1}+\mathbf{I}\right)^{-1}\tau^2\right).$$

Let $N = \sum_{i=1}^{m} n_i$. The complete conditional for σ_j^2 is as follows:

$$\begin{aligned}
\pi\left[\sigma_j^2\right] &\propto \left(\prod_{j=1}^{r}\sigma_j^2\right)^{-N/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}(X_{il}-\theta_i)'\,\mathbf{D}^{-1}(X_{il}-\theta_i)\right] \\
&\times \left(\prod_{j=1}^{r}\sigma_j^2\right)^{-(\alpha+1)}\exp\left[-\frac{1}{\beta\sum_{j=1}^{r}\sigma_j^2}\right].
\end{aligned}$$

Once again, any summation and product operators with index j can be removed. Thus, the above equation resolves to

$$\begin{aligned}
\pi\left[\sigma_j^2\right] &\propto \left(\sigma_j^2\right)^{-N/2}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}(X_{il}-\theta_i)'\,\mathbf{D}^{-1}(X_{il}-\theta_i)\right] \\
&\times \left(\sigma_j^2\right)^{-(\alpha+1)}\exp\left[-\frac{1}{\beta\sigma_j^2}\right] \\
&= \left(\sigma_j^2\right)^{-((\alpha+N/2)+1)}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}(X_{il}-\theta_i)'\,\mathbf{D}^{-1}(X_{il}-\theta_i) - \frac{1}{\beta\sigma_j^2}\right] \\
&= \left(\sigma_j^2\right)^{-((\alpha+N/2)+1)}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}\left(X_{il}'\mathbf{D}^{-1}X_{il} - 2\theta_i'\mathbf{D}^{-1}X_{il} + \theta_i'\mathbf{D}^{-1}\theta_i\right) - \frac{1}{\beta\sigma_j^2}\right].
\end{aligned}$$

Note that

$$\begin{aligned}
X_{il}'\mathbf{D}^{-1}X_{il} &= y_{1il}\frac{1}{\sigma_1^2}X_{1il}+\cdots+X_{ril}\frac{1}{\sigma_r^2}X_{ril}=\sum_{j=1}^{r}X_{jil}\frac{1}{\sigma_j^2}X_{jil},\\
\theta_i'\mathbf{D}^{-1}X_{il} &= \theta_{1i}\frac{1}{\sigma_1^2}X_{1il}+\cdots+\theta_{ri}\frac{1}{\sigma_r^2}X_{ril}=\sum_{j=1}^{r}\theta_{ji}\frac{1}{\sigma_j^2}X_{jil}, \text{ and}\\
\theta_i'\mathbf{D}^{-1}\theta_i &= \theta_{1i}\frac{1}{\sigma_1^2}\theta_{1i}+\cdots+\theta_{ri}\frac{1}{\sigma_r^2}\theta_{ri}=\sum_{j=1}^{r}\theta_{ji}\frac{1}{\sigma_j^2}\theta_{ji}.
\end{aligned}$$

Moreover, the summation on index j can be removed. Thus, the complete conditional for σ_j^2 finally becomes

$$\begin{aligned}
\pi\left[\sigma_j^2\right] &\propto \left(\sigma_j^2\right)^{-((\alpha+N/2)+1)}\exp\left[-\frac{1}{2}\sum_{i=1}^{m}\sum_{l=1}^{n_i}\left(X_{jil}\frac{1}{\sigma_j^2}X_{jil}-2\theta_{ji}\frac{1}{\sigma_j^2}X_{jil}+\theta_{ji}\frac{1}{\sigma_j^2}\theta_{ji}\right)-\frac{1}{\beta\sigma_j^2}\right]\\
&= \left(\sigma_j^2\right)^{-((\alpha+N/2)+1)}\exp\left[-\frac{1}{\sigma_j^2}\left(\frac{\sum_{i=1}^{m}\sum_{l=1}^{n_i}\left(X_{jil}X_{jil}-2\theta_{ji}X_{jil}+\theta_{ji}\theta_{ji}\right)}{2}+\frac{1}{\beta}\right)\right].
\end{aligned}$$

Hence,

$$\pi\left[\sigma_j^2\right] \sim IG\left(\alpha+N/2,\left(\frac{\sum_{i=1}^{m}\sum_{l=1}^{n_i}\left(X_{jil}X_{jil}-2\theta_{ji}X_{jil}+\theta_{ji}\theta_{ji}\right)}{2}+\frac{1}{\beta}\right)^{-1}\right).$$

Thus, the closed-form nature of all complete conditionals is verified.

The Gibbs sampling algorithm (also called *alternating conditional sampling algorithm*) was used to generate a sequence of samples from these complete conditional distributions (Gelman et al. 2004).

For each posterior Gibbs sample for μ_{ij} and σ_j^2 (denoted by $\hat{\mu}_{ij}$ and $\hat{\sigma}_j^2$), the posterior probability that author k is the true class of a new sample of an unknown class with observed vector x^* is calculated. This is obtained by modifying equation 2.14:

$$p\left(k|x^*\right)=\frac{\pi_k\exp\left[-\frac{1}{2}\sum_{j=1}^{r}\left(\frac{x_j^*-\hat{\mu}_{ij}}{\hat{\sigma}_j^2}\right)^2\right]}{\left(\sum_{i=1}^{m}\pi_i\exp\left[-\frac{1}{2}\sum_{j=1}^{r}\left(\frac{x_j^*-\hat{\mu}_{ij}}{\hat{\sigma}_j^2}\right)^2\right]\right)+\pi_{m+1}\exp\left[-\frac{1}{2}\sum_{j=1}^{r}a_j^2\right]}, \tag{3.6}$$

where

$$a_j=\min\left(\max_i\left|\frac{x_j^*-\hat{\mu}_{ij}}{\hat{\sigma}_j^2}\right|,\lambda\right). \tag{3.7}$$

Note that equation 3.6 still takes into account a missing class. The approximate posterior probability for the latent (missing) class can be obtained by modifying equation 2.16:

$$p\left(m+1|x^{*}\right)=\frac{\pi_{m+1}\exp\left[-\frac{1}{2}\sum_{j=1}^{r}a_j^2\right]}{\left(\sum_{i=1}^{m}\pi_i\exp\left[-\frac{1}{2}\sum_{j=1}^{r}\left(\frac{x_j^{*}-\hat{\mu}_{ij}}{\hat{\sigma}_j^2}\right)^2\right]\right)+\pi_{m+1}\exp\left[-\frac{1}{2}\sum_{j=1}^{r}a_j^2\right]}. \tag{3.8}$$

Fifty thousand Gibbs samples were generated from the complete conditionals for θ_i and σ_j^2. In each iteration of the Gibbs sampling, posterior probabilities for each candidate and latent class were calculated. Thus, at the end of Gibb's sampling algorithm, each class's distribution of posterior probabilities was obtained. The classification was performed in three ways: 10%, 25%, and 50% quantiles of the distributions of posterior probabilities were obtained for each candidate and latent class. For each quantile, the class whose quantile is the highest was chosen as the predicted class of the sample.

The specification of prior distributions for the authorship classification problem was as follows:

$$\begin{aligned}\theta_i &\sim MVN\left(-5.703732\mathbf{J}, 1.366177\mathbf{I}\right) \quad \text{and} \\ \sigma_j^2 &\sim IG\left(3, 0.1\right).\end{aligned}$$

For this prior configuration, θ_i has a mean vector of length 108 with -5.703732 in each element. The variance of θ_i is 108×108 diagonal matrix with diagonal element 1.366177. This implies that, by the empirical rule, about 99.7% of all θ_i fall in the range [-9.210239,-2.197225]. The inverse of *logit* transformation of this range is [0.0001,0.1]. The mean and variance of the prior distributions of σ_j^2 are 5 and 25 respectively on the *logit* scale. This is a noninformative prior that gives more weight to the data than the prior.

3.3 Application to Hamilton Federalist Papers

Comparisons among (1) the extended NSC proposed by Schaalje et al., (2) the extended NSC without fudge factor, (3) the extended NSC without shrunken centroid, (4) the extended

NSC with class-specific threshold, (5) the extended NSC shrunken in the different direction, and (6) the full independent Bayesian classification methods were performed. These six methods were used to determine authorship for the Federalist Papers written by Alexander Hamilton.

Two studies were performed to compare the methods' performance. One study compared all six methods using training texts from Joseph Smith, early and late Sidney Rigdon, Solomon Spalding, Oliver Cowdery, and Parley Pratt (for information on these potential authors, see Appendix B in Schaalje et al. 2010). Note that Hamilton was not included in the training candidate set. Thus, the methods should classify Hamilton texts to the latent nonincluded author. Training texts by these authors were used to classify all 51 Hamilton Federalist Papers.

The same things were done in the second study, but it also included Hamilton texts in the training set. The 51 Hamilton Federalist Papers were divided into two sets. Twenty-six texts were used as a training data along with texts by other authors in the candidate set. The others were used as test texts. Thus, the classifiers should identify Hamilton as the true author for these 25 test texts.

For both studies, equal prior probabilities were assigned to all potential authors. Relative frequencies of 108 noncontextual words were extracted from each of the training texts using PERL script. SAS programming software was used for data management and R software was used for data processing. These codes can be found in the Appendix C. The criterion for classification was the candidate (or latent author) with the highest posterior probability (equation 2.13), except for the full independent Bayesian classification method that used the highest 10th, 25th, and 50th percentile of the distribution of posterior probabilities.

Hamilton not in the training set

Figure 3.1 shows the classification results from the first study where Hamilton was not included as one of the potential authors. The horizontal axis indicates the 51 Federalist Papers in order of publication. Each vertical bar represents the classification result for that paper. Table 3.1 shows the misclassification rate for each method. In the extended NSC classification, the classifier correctly classified 43 Federalist Papers to one or more latent authors as the true author (a misclassification rate of 15.69%). Early Rigdon and Pratt were incorrectly selected as true authors for three and five Federalist Papers, respectively.

Table 3.1: Misclassification for all methods without Hamilton in the training text.

Method	Misclassification Rate
Extended NSC (Schaalje et al. 2010)	15.69 %
Extended NSC without Fudge Factor	39.22 %
Extended NSC without Shrunken Centroid	74.51 %
Extended NSC with Class-Specific Threshold	17.65 %
Extended NSC with Shrinkage towards Class means	21.57 %
Full independent Bayesian Classification (10th percentile)	0.00 %
Full independent Bayesian Classification (25th percentile)	0.00 %
Full independent Bayesian Classification (50th percentile)	0.00 %

In the extended NSC without the fudge factor, only 31 Federalist Papers were correctly classified to one or more latent authors. In addition to early Rigdon and Pratt, who were classified as the authors of five and seven Federalist Papers respectively, Cowdery was classified as the author of eight Federalist Papers. The misclassification rate was 39.22%, which is more than twice as much as the extended NSC method. The fudge factor seems to play a vital role in the author classification.

Cowdery was the most dominant author for the extended NSC without shrunken centroid method. Cowdery was classified to be the true author for 29 Federalist Papers– almost a half of Hamilton Federalist Papers—and only 13 Federalist Papers were classified to the latent author(s). The misclassification rate for this particular method was 74.51%, which is almost five times more than the extended NSC method. Likewise, the misclassification

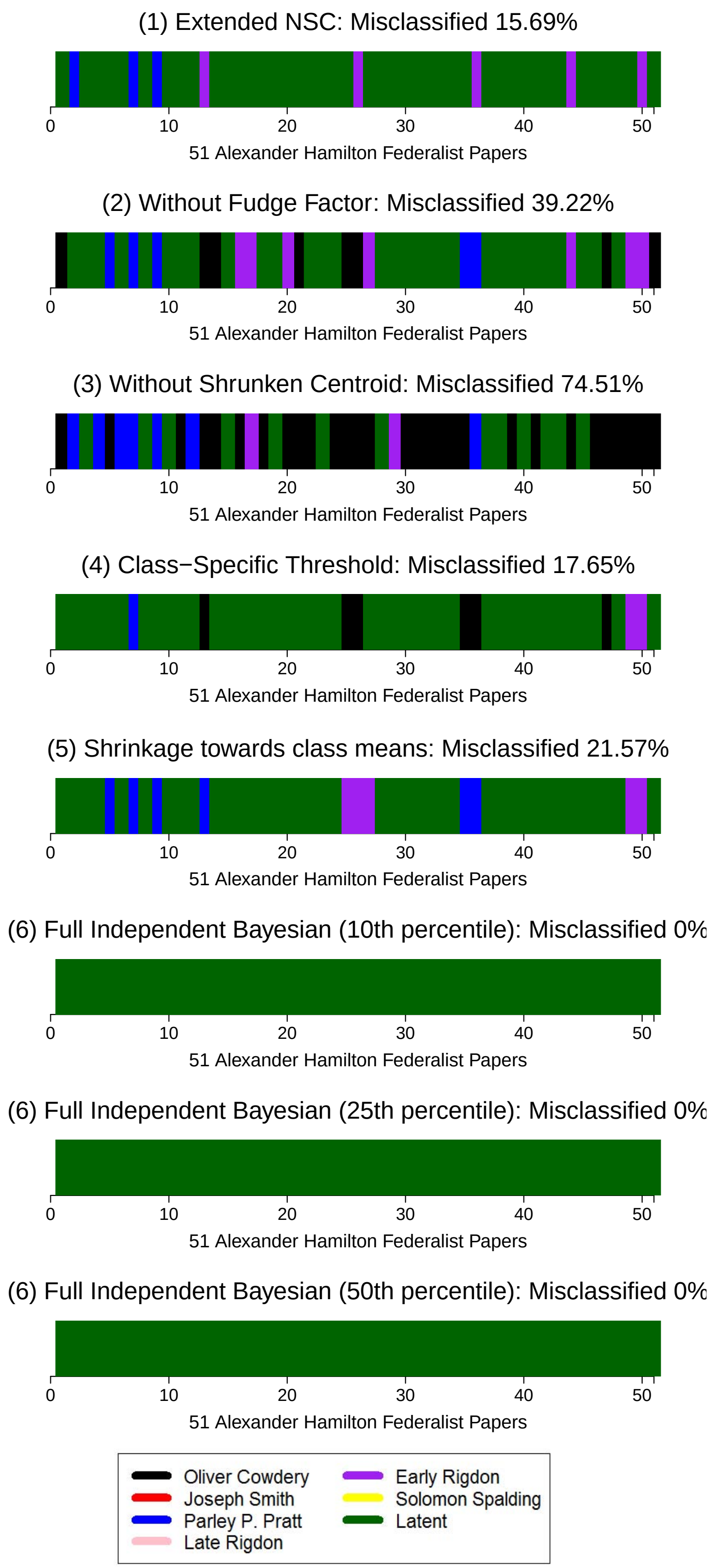


Figure 3.1: Classification results for 51 Hamilton Federalist Papers using (1) extended NSC proposed by Schaalje et al. (2010), (2) extended NSC without fudge factor, (3) extended NSC without shrunken centroid, (4) extended NSC with class-specific threshold, (5) extended NSC shrunken in the different direction, and (6) full independent Bayesian classification using 10%, 25%, and 50% quantiles.

rate for this method was the highest of the six methods. These results clearly show that applying shrinkage to the centroids should be recommended in stylometry even more than the usage of a fudge factor.

Using extended NSC with class-specific thresholding, 42 texts were correctly classified to latent author(s). Cowdery, Pratt, and early Rigdon were selected as the authors of five, one, and two of the Federalist Papers, respectively. Compared to the extended NSC, the misclassification rate increased from 15.69% to 17.65%; an increase of 1.96%. Therefore, class-specific thresholding does not seem to provide any improvement. This result is counterintuitive; we would expect the misclassification for the extended NSC with class-specific thresholding to be at least equal to the misclassification for the extended NSC developed by Schaalje et al. (2010) since the class-specific threshold considers all permutation of Δs, including the permutation that the extended NSC uses. This is caused by the difficulty in cross-validation. Suppose we considered the class-specific threshold for the six authors with Δs ranging from 0 to 2.00. Suppose further that Δ values are incremented by 0.2. This implies that there would be $\frac{n!}{(n-r)!} = \frac{11!}{5!} = 332{,}640$ possible permutation of Δs. This number is already large on its own. Moreover, if Δs were incremented by 0.1, the vector of Δ specified is too large for the R software. Therefore, to perform the cross-validation in a timely manner and to avoid problems with R, we must reduce the number of permutations. However, we may then miss the optimal set of values. For this reason, it is possible for the extended NSC with class-specific thresholding to have higher misclassification rate than the extended NSC (Schaalje et al. 2010).

The inflation of the misclassification rate for the extended NSC with the class-specific threshold method was not as high as the inflation of the extended NSC without a fudge factor or the extended NSC without shrunken centroids. It is apparent that class-specific thresholding does not provide much improvement to the classifier. Moreover, considering the fact that cross-validation for the Δ_is takes a considerable amount of time, this method is not appealing.

For the extended NSC with centroids shrunken in a different direction, 40 Federalist Papers were classified to latent author(s), whereas six and five Federalist Papers were classified to Pratt and early Rigdon, respectively. This implies that the misclassification rate was 21.57%. This method was superior to the extended NSC without a fudge factor and the extended NSC without shrunken centroids in terms of the misclassification. Nevertheless, it was not as accurate as the extended NSC with class-specific thresholds; the misclassification rate for the extended NSC shrunken towards class means is 1.96% higher than the NSC with class-specific thresholds. The difference in the misclassification between the original extended NSC and this method is 5.88%. Changing the direction of shrinkage does not seem to improve the accuracy of the classification.

The full independent Bayesian approach was the only method that showed improvement from the original extended NSC. Figure 3.2 shows a convergence plot for the MCMC procedure. It shows the first 300 Gibbs samples for θ_1, the feature representing the mean of the word "*a*" for Oliver Cowdery. θ_1 converges approximately after the 30th Gibbs sample. The cutoff point for burn-in was set at 100th Gibbs sample.

Regardless of which percentile was used for classification, all 51 Federalist Papers were classified to one or more latent authors, resulting in 0% misclassification. It is clear that the full independent Bayesian method is superior to other methods, including the extended NSC proposed by Schaalje et al. (2010).

Nine Federalist Papers that were misclassified by the extended NSC are papers nos. 6, 12, 13, 15, 22, 35, 70, 78, and 84. Posterior distributions of posterior probabilities for all of these Federalist Papers can be found in Appendix B.1. Figure 3.3 shows each author's (including the latent author's) posterior distributions of posterior probabilities for Federalist Paper No. 35. Although there was no misclassification for the full independent Bayesian method (regardless of which quantile was used), some posterior probabilities for other authors were high. Therefore, even though the full independent Bayesian method seemed too good to be true, it still occasionally allocated high probabilities to other authors. Similar results can be seen for other Federalist Papers that were misclassified by the extended NSC (see Appendix B.1).

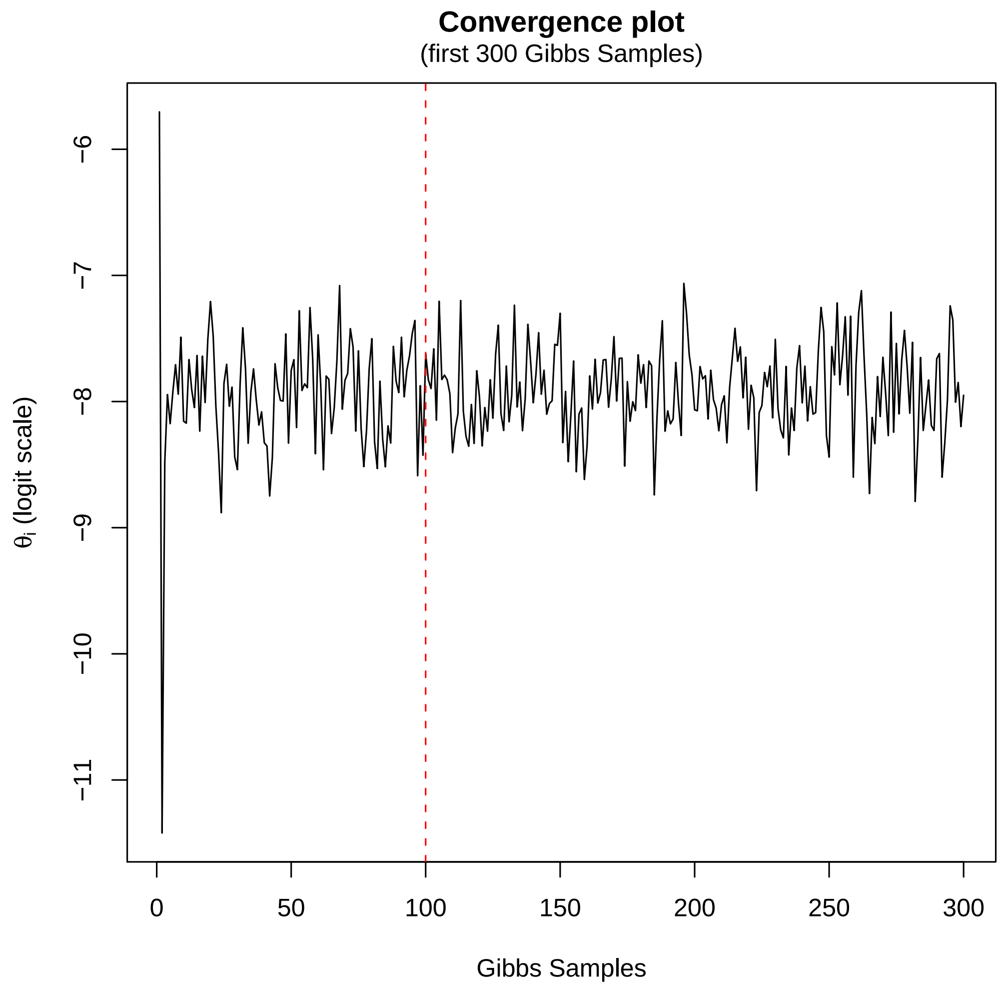


Figure 3.2: First 300 Gibbs samples from θ_1, posterior θ for feature representing the word "*a*" for Oliver Cowdery. Red vertical line indicate the cutoff point for burn.

Hamilton included in the training set

Figure 3.4 shows the classification results from the second study where Hamilton was included as one of the potential authors. The horizontal axis indicates the 25 randomly chosen Hamilton Federalist Papers sorted by their publication date. Each vertical bar represents the classification result for a paper. Table 3.2 shows the misclassification rate for all of

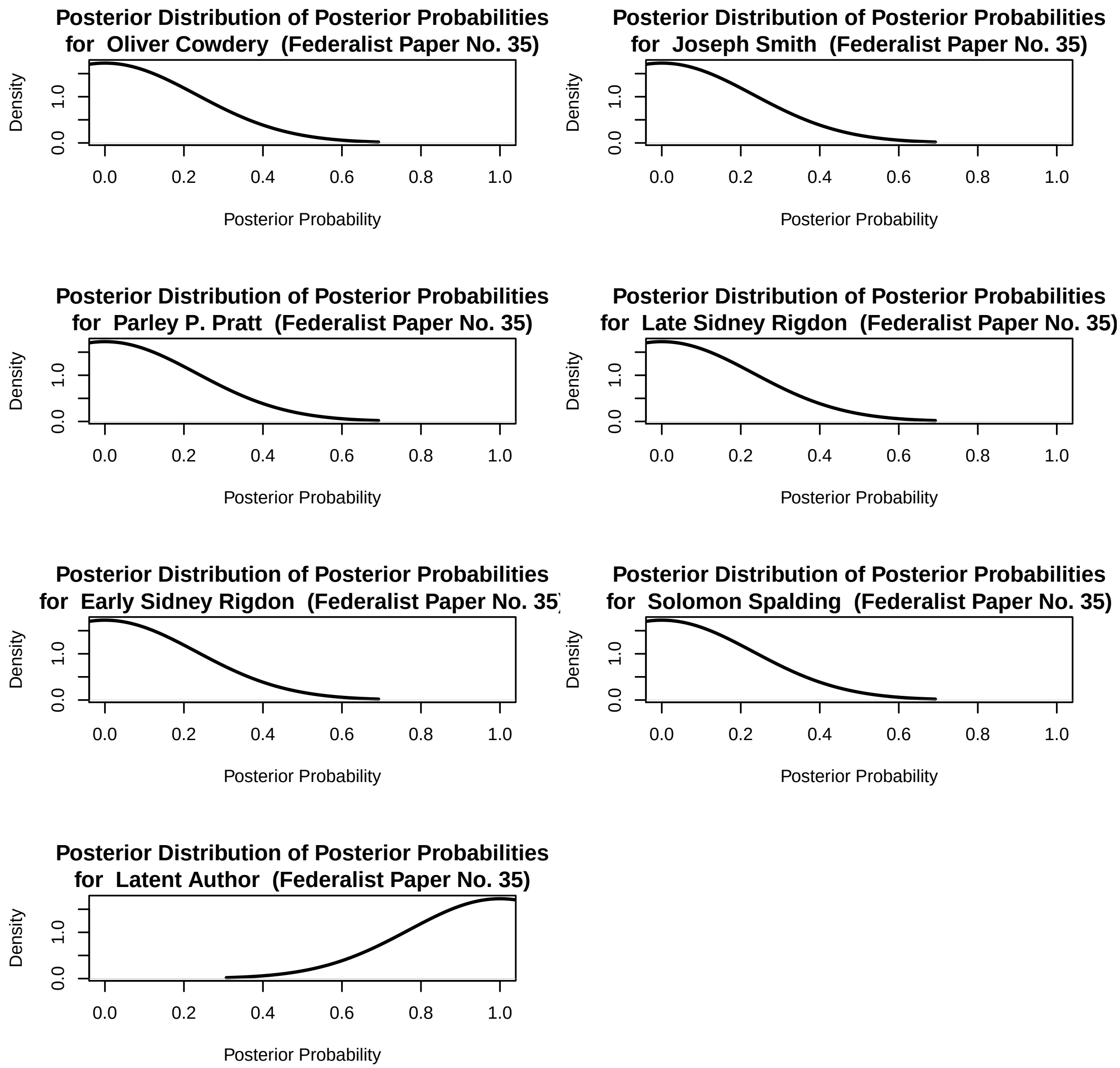


Figure 3.3: Posterior distribution of posterior probabilities for each potential author and latent author for Federalist Paper No. 35.

the methods. In the extended NSC classification, four Federalist Papers were misclassified to one or more latent authors. The misclassification rate was 16%. Using the extended NSC without a fudge factor, the misclassification rate was 24% (six were misclassified). All misclassified Federalist Papers were classified to the latent author(s). The extended NSC without a fudge factor is inferior to the extended NSC.

Table 3.2: Misclassification for all methods with Hamilton in the training text.

Method	Misclassification Rate
Extended NSC (Schaalje et al. 2010)	16.00 %
Extended NSC without Fudge Factor	24.00 %
Extended NSC without Shrunken Centroid	44.00 %
Extended NSC with Class-Specific Threshold	12.00 %
Extended NSC with Shrinkage towards Class Means	16.00 %
Full independent Bayesian Classification (10th percentile)	0.00 %
Full independent Bayesian Classification (25th percentile)	0.00 %
Full independent Bayesian Classification (50th percentile)	0.00 %

Using the extended NSC without shrinkage, the misclassification rate was 44% (11 were misclassified), even more inferior than the extended NSC without a fudge factor. All of the misclassified Federalist Papers were classified to the latent author. This result is consistent with the results from testing without Hamilton in the training texts.

The misclassification rates for the extended NSC with class-specific threshold was 12% (three were misclassified). Latent authors were chosen as the true author of the misclassified Federalist Papers. Unlike previous study where we had Hamilton in the training set, the misclassification performs better than the extended NSC. We can see that, although computing difficulties associated with cross-validation are present, depending on the range of Δ values that were used for the cross-validation, it is possible to obtain better results using a different method than the extended NSC.

The extended NSC shrunken towards the class means misclassified four Federalist Papers (16% misclassification rate). This result has the same misclassification rate as the original extended NSC. Out of the four Federalist Papers that were misclassified using the

Figure 3.4: Classification results for 25 Hamilton Federalist Papers with Hamilton in the candidate set using (1) extended NSC proposed by Schaalje et al. (2010), (2) extended NSC without fudge factor, (3) extended NSC without shrunken centroid, (4) extended NSC with class-specific threshold, (5) extended NSC shrunken in the different direction, and (6) full independent Bayesian classification using 10%, 25%, and 50% quantiles.

extended NSC with centroids shrunken towards class means, one was classified to early Rigdon and three were classified to the latent author.

Figure 3.5 shows the first 300 Gibbs samples for θ_1, the feature representing the mean of the word "*a*" for Oliver Cowdery. θ_1 converges approximately after the 50th Gibbs sample. The cutoff point for burn-in was set at the 100th sample.

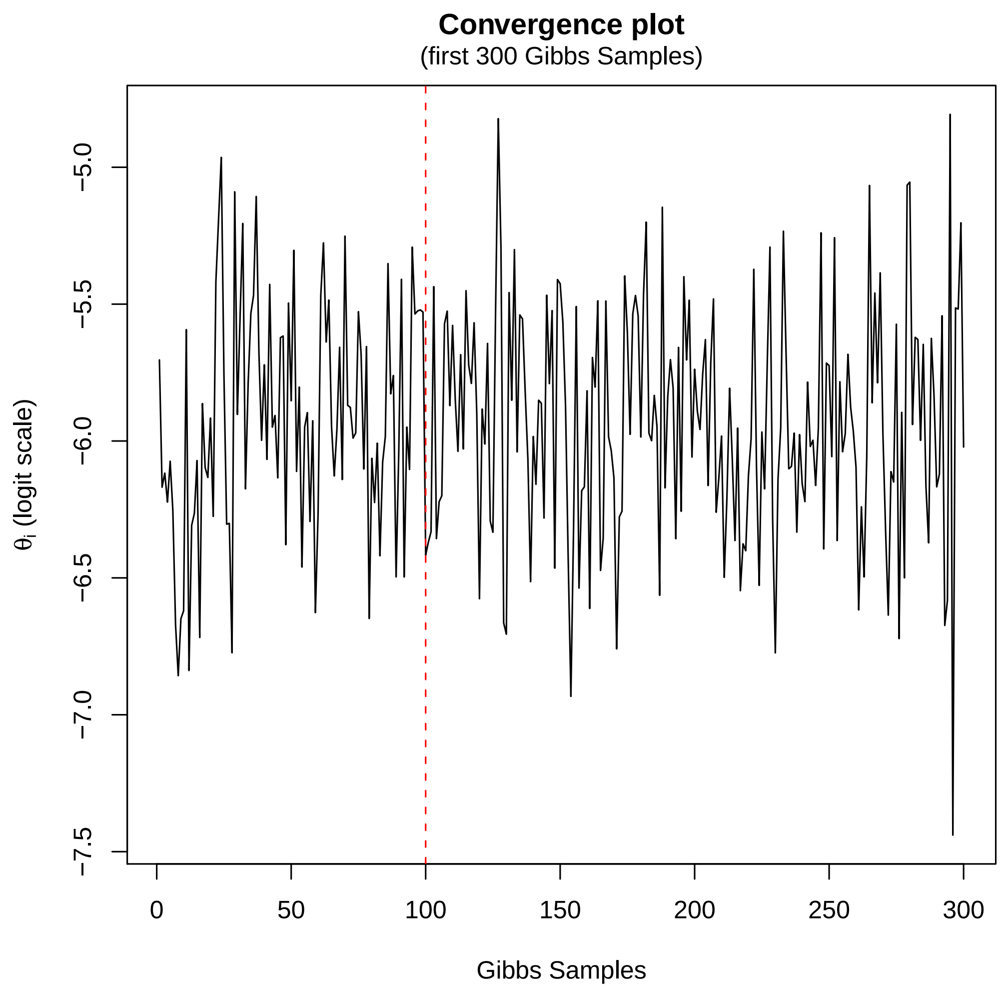


Figure 3.5: First 300 Gibbs samples from θ_1, posterior θ for feature representing the word "*a*" for Oliver Cowdery. Red vertical line indicate the cutoff point for burn.

Figure 3.4 clearly shows that the full independent Bayesian approach to classification is superior to all methods discussed in this paper. This result was consistent with the study where Hamilton is not in the training texts. Regardless of whether classification was determined by 10th, 25th, or 50th percentile, the full independent Bayesian approach had 0% misclassification rate.

Four Federalist Papers that were misclassified by the extended NSC are papers nos. 8, 17, 26, and 34. All of these Federalist Papers were classified to one or more latent authors. Posterior distributions of posterior probabilities for all of these Federalist Papers can be found in Appendix B.2. Figure 3.6 shows each of the author's (including the latent author's) posterior distributions of posterior probabilities for Federalist Paper No. 8. The latent author is the major contender for authorship after Hamilton, which is interesting because the extended NSC classified paper no. 8 to the latent author. The full independent Bayesian method still occasionally allocated high probability to the latent author as the extended NSC did.

3.4 Application to Book of Mormon

The full independent Bayesian classification method was used to attribute authors for successive 5,000-word sections of the Book of Mormon. Blocks of texts were used as the test texts instead of chapters in the Book of Mormon since the chapters vary so greatly in length. Schaalje et al. (2010) discuss this in detail:

> Moreover, high correct classification rates for training texts do not necessarily guarantee that test texts will be correctly classified. Most importantly, even though classification rates might be minimally affected by sample size, posterior probabilities of authorship will certainly be affected by variation in sample size.

Word blocks that are essentially verbatim quotations of Isaiah or Malachi were removed from our test texts. Once again, the *logit* transformation was applied to the relative frequencies of 108 noncontextual words. The potential authors were Joseph Smith, early and late Sidney Rigdon, Solomon Spalding, Oliver Cowdery, and Parley Pratt.

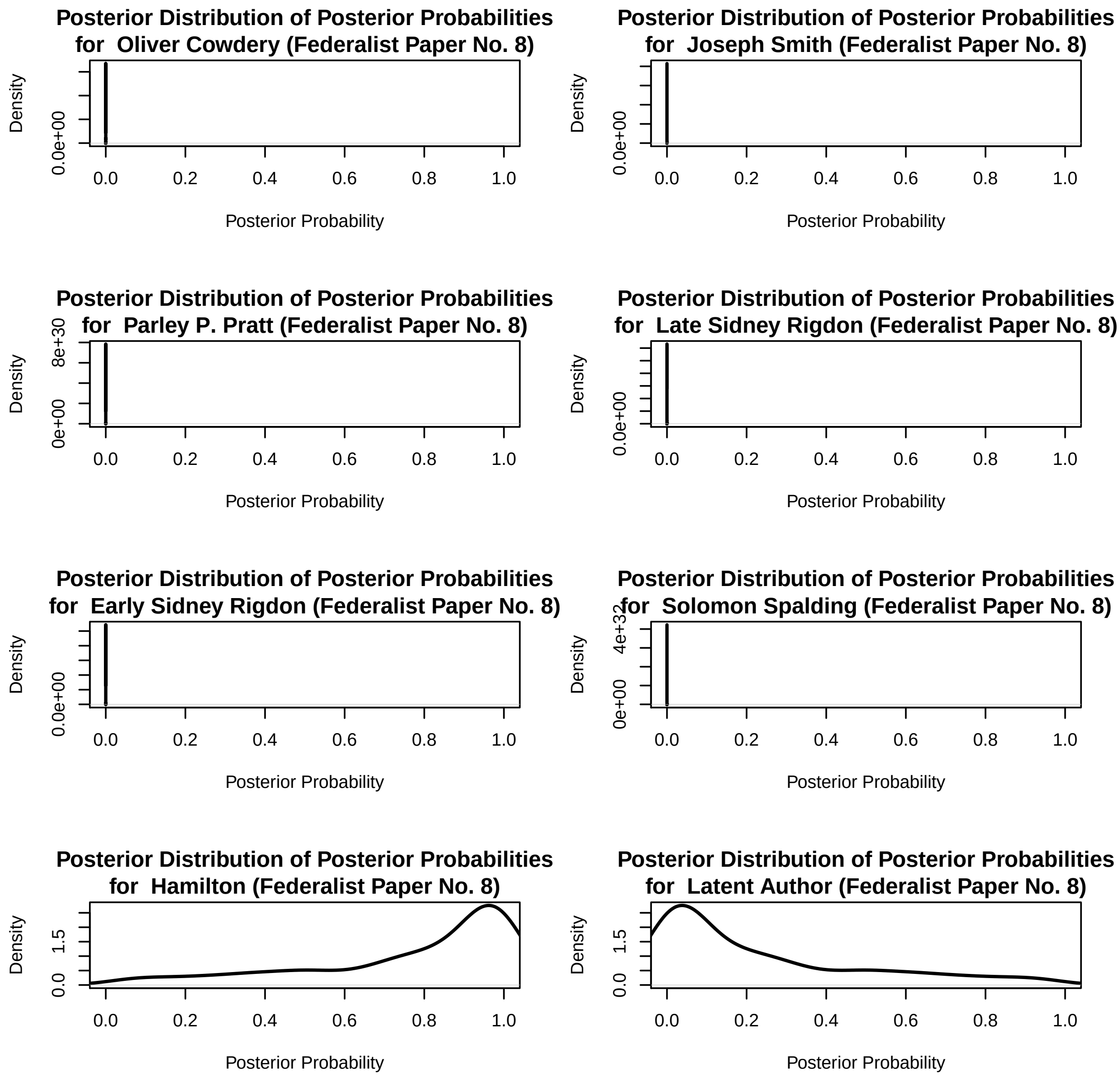


Figure 3.6: Posterior distribution of posterior probabilities for each potential author and latent author for Federalist Paper No. 8.

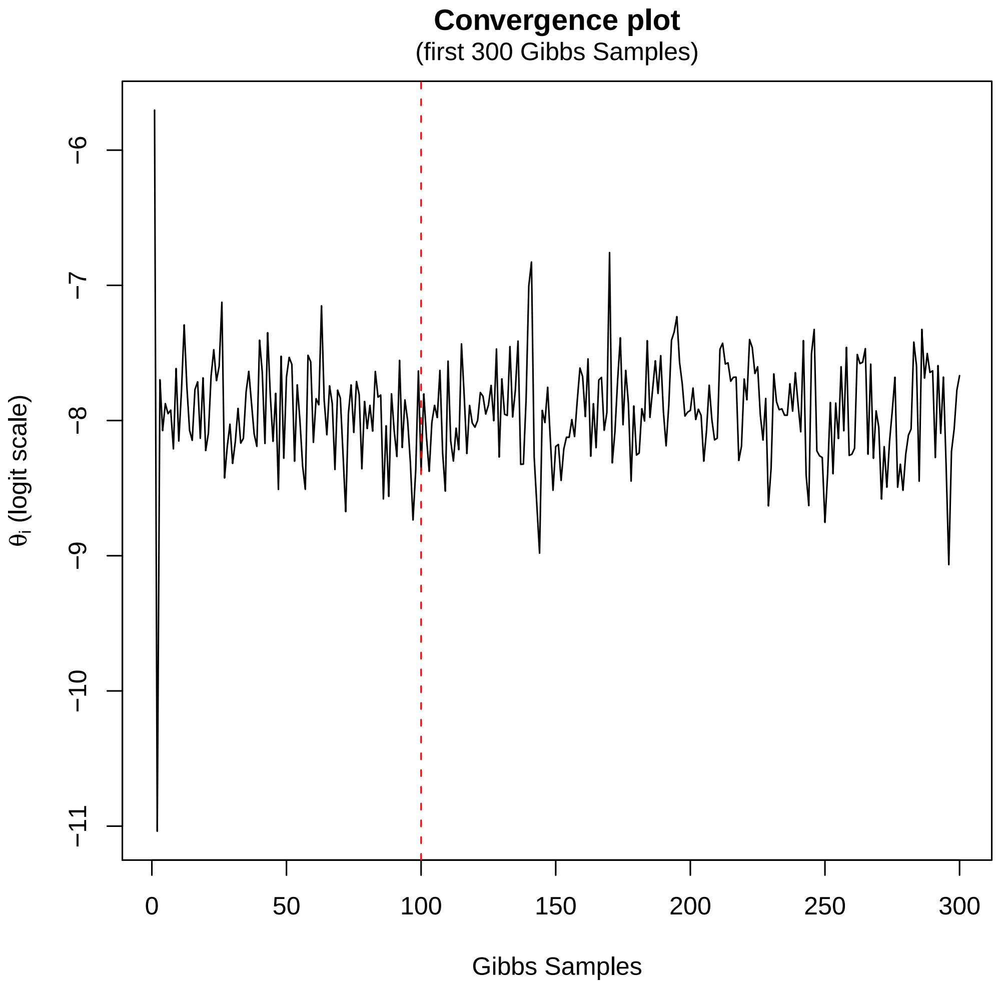


Figure 3.7: First 300 Gibbs samples from θ_1, posterior θ for feature representing the word "*a*" for Oliver Cowdery. Red vertical line indicate the cutoff point for burn.

Figure 3.7 shows the first 300 Gibbs samples for θ_1, the feature representing mean of the word "*a*" for Oliver Cowdery. θ_1 converges approximately after the 30th Gibbs sample. The cutoff point for burn-in was set at the 100th sample.

All of the 27 blocks of texts were classified to one or more latent authors regardless of the quantile used.

Appendix B.3 gives the posterior distributions of posterior probability for all of the authors for all of the 5,000-word blocks from the Book of Mormon. Figure 3.8 shows the posterior distribution of posterior probability for all of the authors for the first 5,000-word blocks from the Book of Mormon. The posterior distribution for a latent author has two modes at around 0 and 1, the second of which has the higher peak. Except for Cowdery and early Rigdon, the posterior distributions of the potential authors have distributions concentrated at zero. Cowdery and early Rigdon have right skewed posterior distributions with long tails. Even though these results indicate that Cowdery and Rigdon were the strongest contenders out of the six potential authors, the latent author was a much more likely contender.

Consistent with previous stylometric analyses of the Book of Mormon, this analysis shows that there is little stylometric support for the Spalding-Rigdon theory of the Book of Mormon authorship (Schaalje et al. 2010). The styles do not match Rigdon, Spalding, or any of the other candidates as claimed by Jockers et al. (2009).

3.5 Reapplication to Federalist Papers

Using only the full independent Bayesian approach, we revisited the Federalist Papers to further investigate the validity of this approach. Three studies were performed: (1) the classification for 25 Hamilton Federalist Papers using Hamilton, James Madison, and John Jay in the training set, (2) the classification for five Federalist Papers that are known to be written by Jay and three Federalist Papers that are indicated as jointly written by Hamilton and Madison, and (3) the classification for 12 Federalist Papers, which are historically of disputed authorship. We once again performed the classification on the Federalist Papers based on the 10%, 25%, and 50% quantiles of the posterior distribution of the author probabilities.

Classificatio==n for 25 Hamilton Federalist Papers using Hamilton, Madison, and Jay in the training set

Figure 3.9 shows the first 300 Gibbs samples for θ_1, the feature representing the mean of the word "*a*" for Hamilton. θ_1 converged approximately after the 30th Gibbs sample. The cutoff for burn-in was set at the 100th sample.

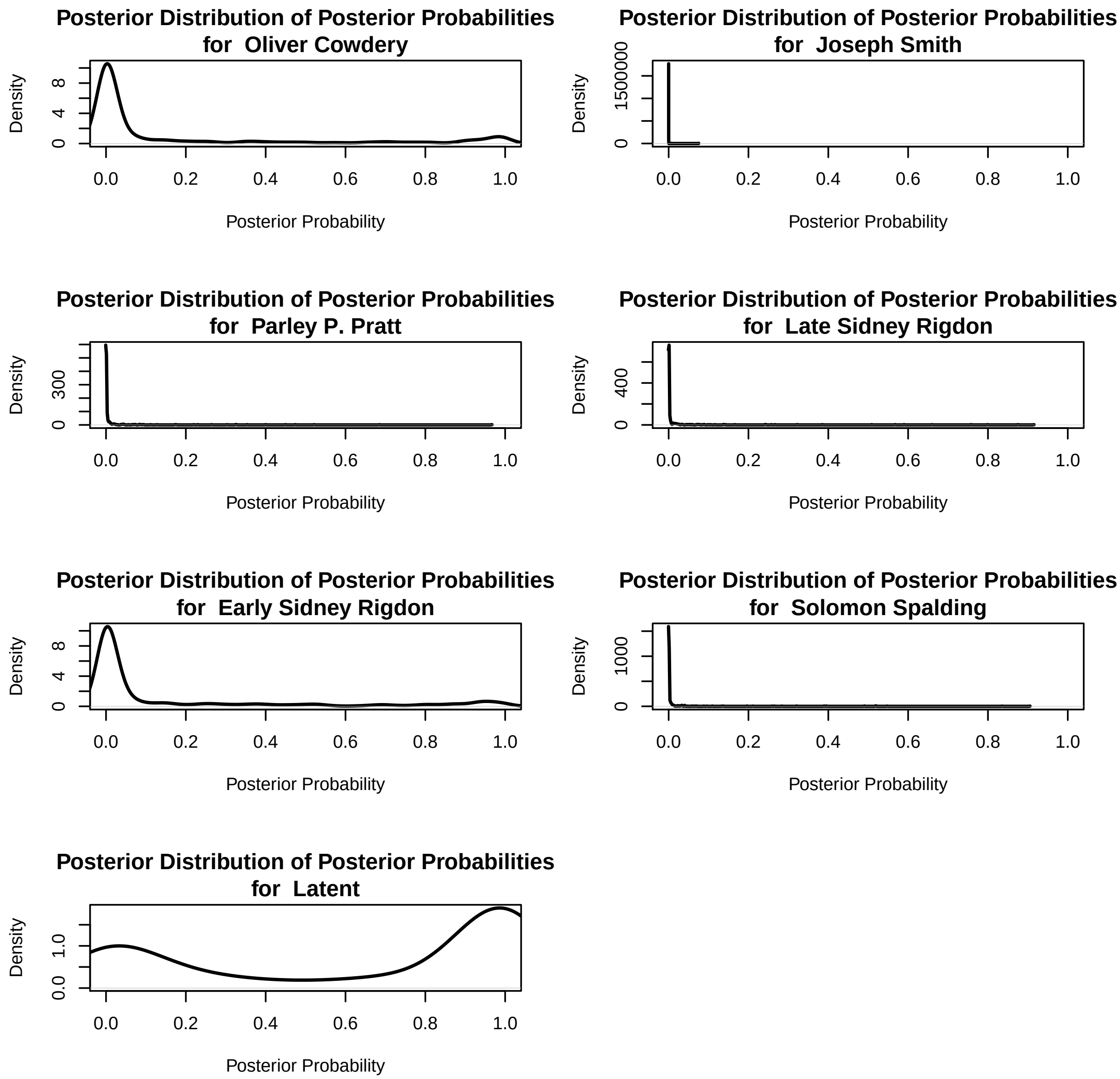


Figure 3.8: Posterior distribution of posterior probabilities for each authors for the first 5,000-words block of texts from Book of Mormon.

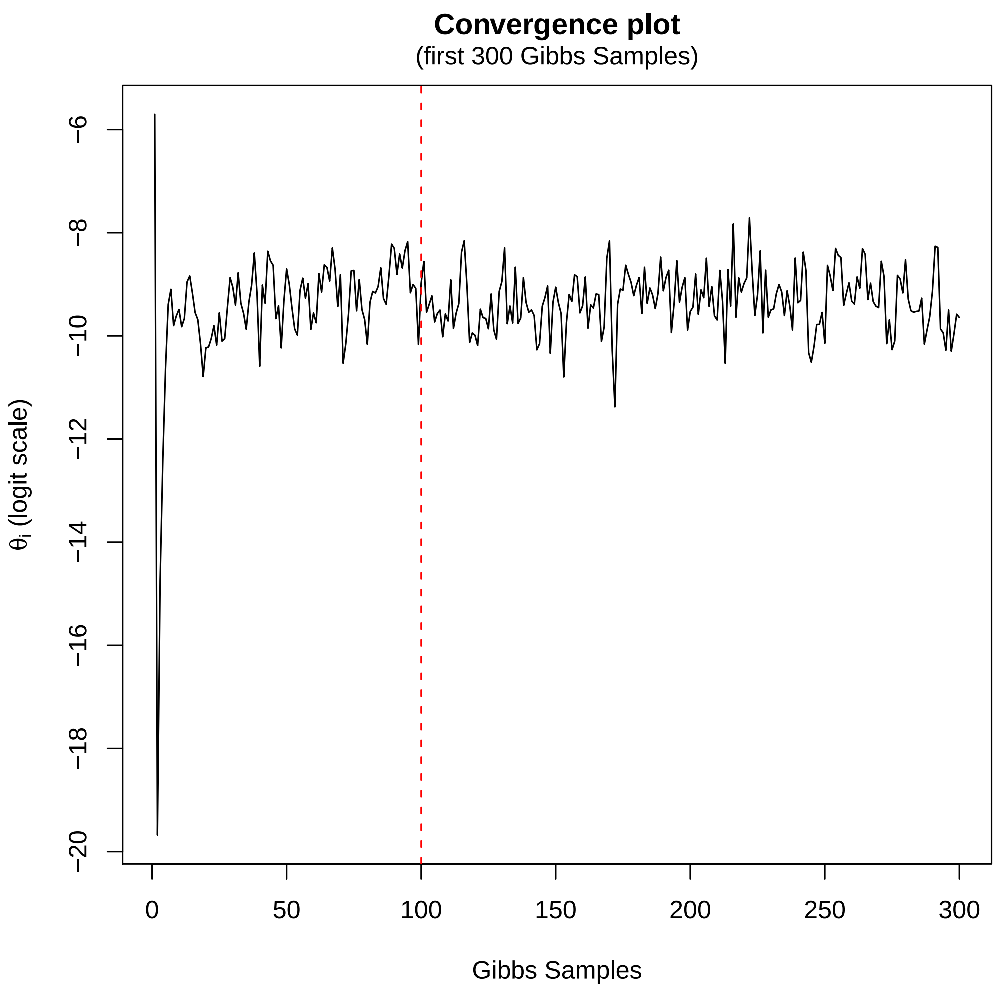


Figure 3.9: First 300 Gibbs samples from θ_1, posterior θ for feature representing the word "*a*" for Hamilton. Red vertical line indicate the cutoff point for burn.

Figure 3.10 shows the classification results for 25 Hamilton Federalist Papers using texts from Madison, Jay, Hamilton. The horizontal axis shows 25 randomly selected Hamilton Federalist Papers sorted by the publication date. Each vertical bar represents the classification result for that paper. The results were the same regardless of which quantiles were used as the classification strategy. Out of the 25 Federalist Papers, 23 papers were correctly assigned to Hamilton (92%) and two were incorrectly assigned to one or more latent authors (8%).

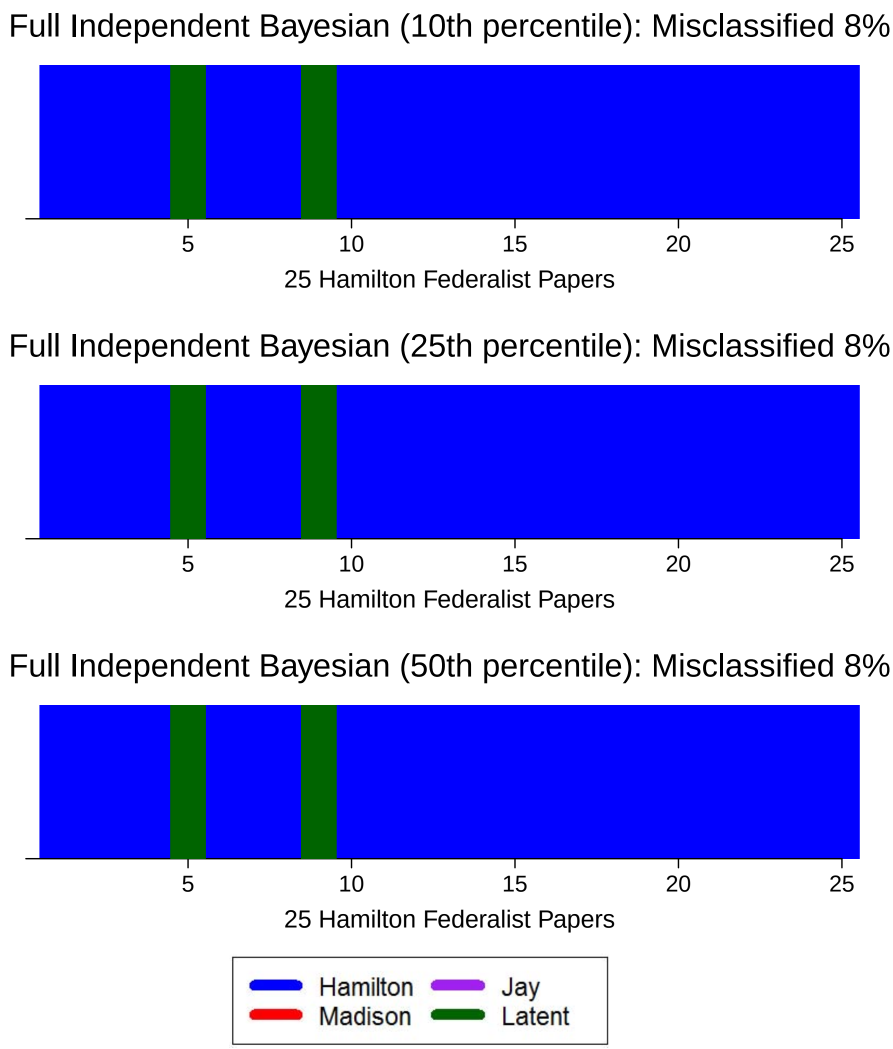


Figure 3.10: Classification results for 25 Hamilton Federalist Papers using texts from Madison, Jay, and Hamilton.

The misclassified Federalist Papers were papers nos. 11 and 37. The misclassification rate was 8%. The performance was inferior to the previous study. However, the overall performance is still excellent.

Figure 3.11 shows the posterior distribution of posterior probabilities for all authors for Federalist Paper No. 51. Madison and the latent author have a posterior distribution concentrated at zero. On the other hand, Hamilton and Madison have a right skewed posterior distribution with a long tail and a left skewed posterior distribution with a long right tail, respectively. From these results, it can be said that Madison was a contender, though not a strong one, for authorship of Federalist Paper No. 51 after Hamilton.

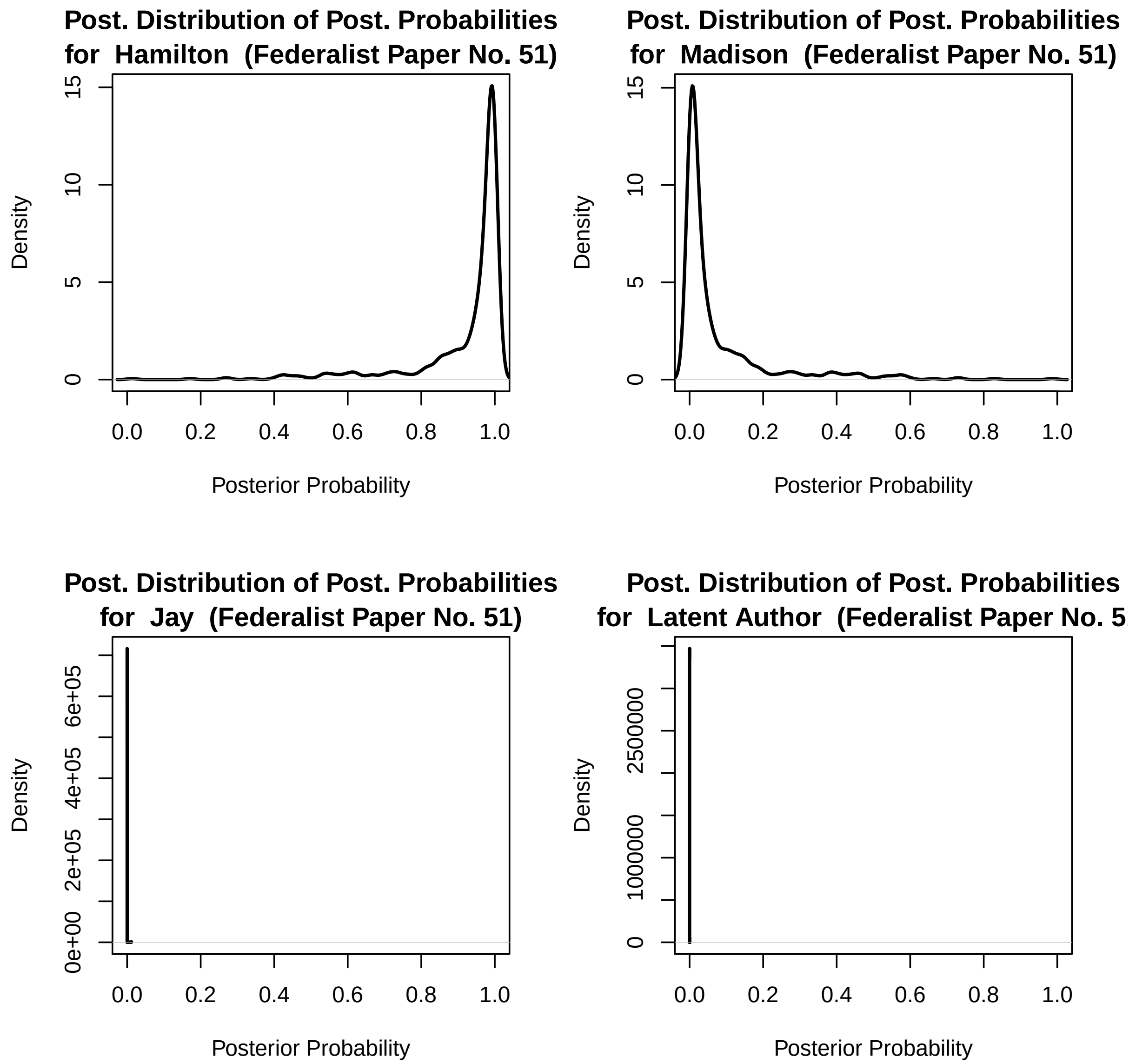


Figure 3.11: Posterior distribution of posterior probabilities for all authors for Federalist Paper No. 51.

Classification for five Jay Federalist Papers and three joint Federalist Papers

Figure 3.12 shows the first 300 Gibbs samples for θ_1, the feature representing the mean of the word "*a*" for Hamilton. θ_1 converges approximately after the 30th Gibbs sample. The cutoff for burn-in was set at the 100th sample.

Figure 3.13 shows the classification results for five Federalist Papers that are known to be written by Jay and three Federalist Papers that are indicated as jointly written by Hamilton and Madison. The horizontal axis shows corresponding Jay Federalist Papers and joint Federalist Papers. Each vertical bar represents the classification result for that paper. Therefore, we expected the first five papers (corresponding to papers nos. 2-5 and 64) to be classified to a latent author since we did not include Jay in the training set, and the last three papers nos. 18–20 to be classified to Madison since Douglass G. Adair, an American historian who researched authorship of disputed Federalist Papers, concluded that these are Madison's writings alone: "Madison had certainly written all of the essays himself, including in revised form only a small amount of pertinent information submitted by Hamilton from his rather sketchy research on the same subject" (Colbourne 1998).

For the Jay papers, the results were different depending on which quantile was used as the criterion for classification. For 10% and 25% quantiles, the full independent Bayesian classifier assigned four out of the five Jay papers to the latent author but one was misclassified to Madison (paper no. 5). Using the 50% quantile, none of the five Jay papers were misclassified; each paper was assigned to one or more latent authors.

Appendix B.4 shows the posterior distributions of posterior probability for Hamilton, Madison, and the latent author for all of the Jay papers. Figure 3.14 shows the posterior distribution of posterior probabilities for Hamilton, Madison, and the latent author for Federalist Paper No. 5. This is the paper for which Madison was chosen as the true author using 10% and 25% quantiles, while the latent author was chosen using the 50% quantile. The posterior distributions for Madison and the latent author are very similar.

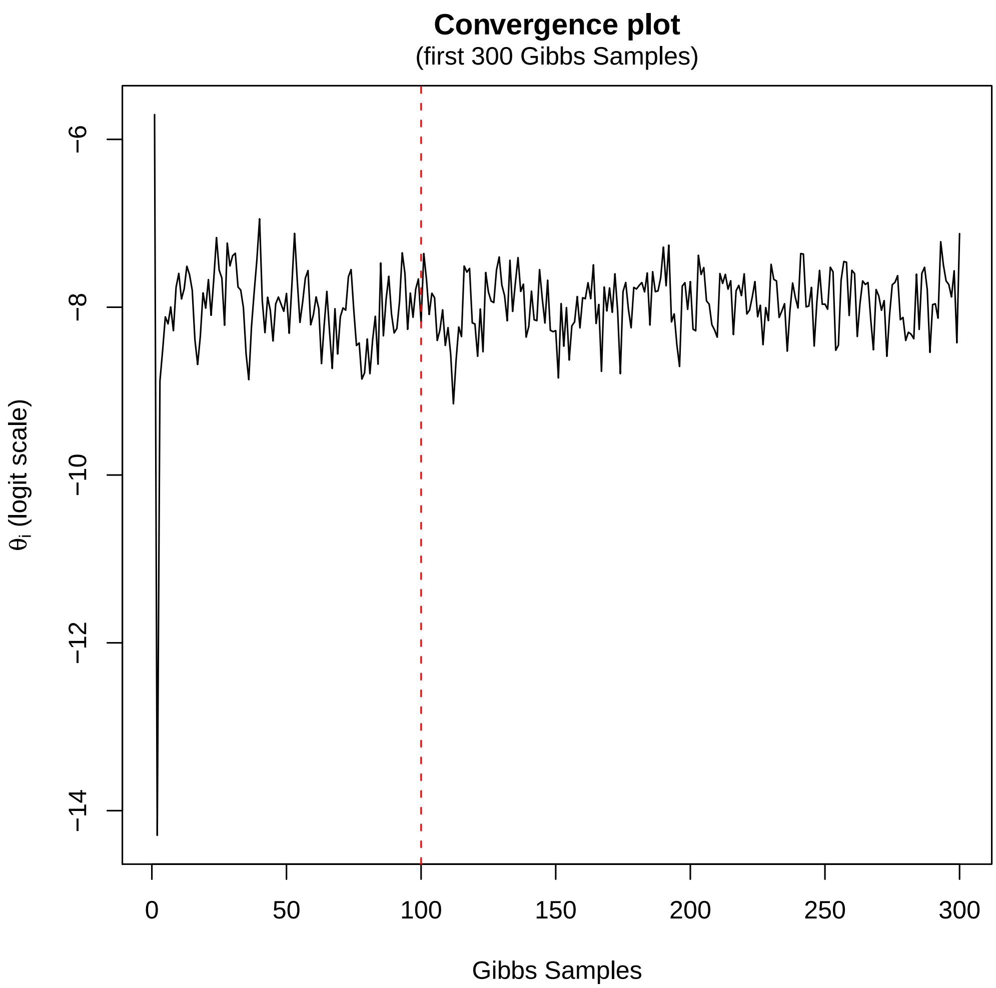


Figure 3.12: First 300 Gibbs samples from θ_1, posterior θ for feature representing the word "*a*" for Hamilton. Red vertical line indicate the cutoff point for burn.

Out of the three Federalist Papers that were said to be jointly written by Hamilton and Madison (Colbourne 1998), two were classified to Madison and one was classified to one or more latent authors. This result was consistent for all quantiles that were used. According to this result, the conclusion that Madison wrote Federalist Papers No. 18–20 is supported only with papers nos. 18 and 19 (Colbourne 1998).

Figure 3.15 shows the posterior distributions of posterior probability for Hamilton, Madison, and the latent author for paper no. 19. It is apparent from Figure 3.15 that

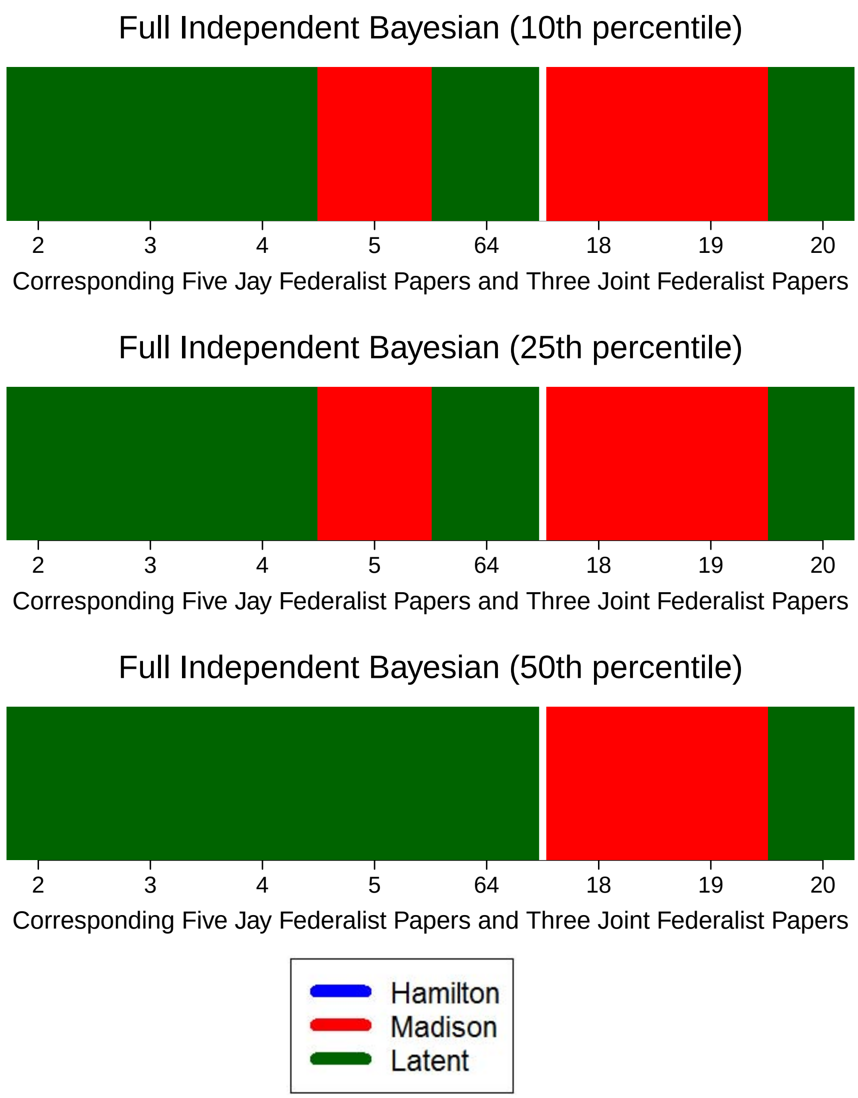


Figure 3.13: Classification results for five Federalist Papers that are known to be written by Jay and three Federalist Papers that are indicated as jointly written by Hamilton and Madison.

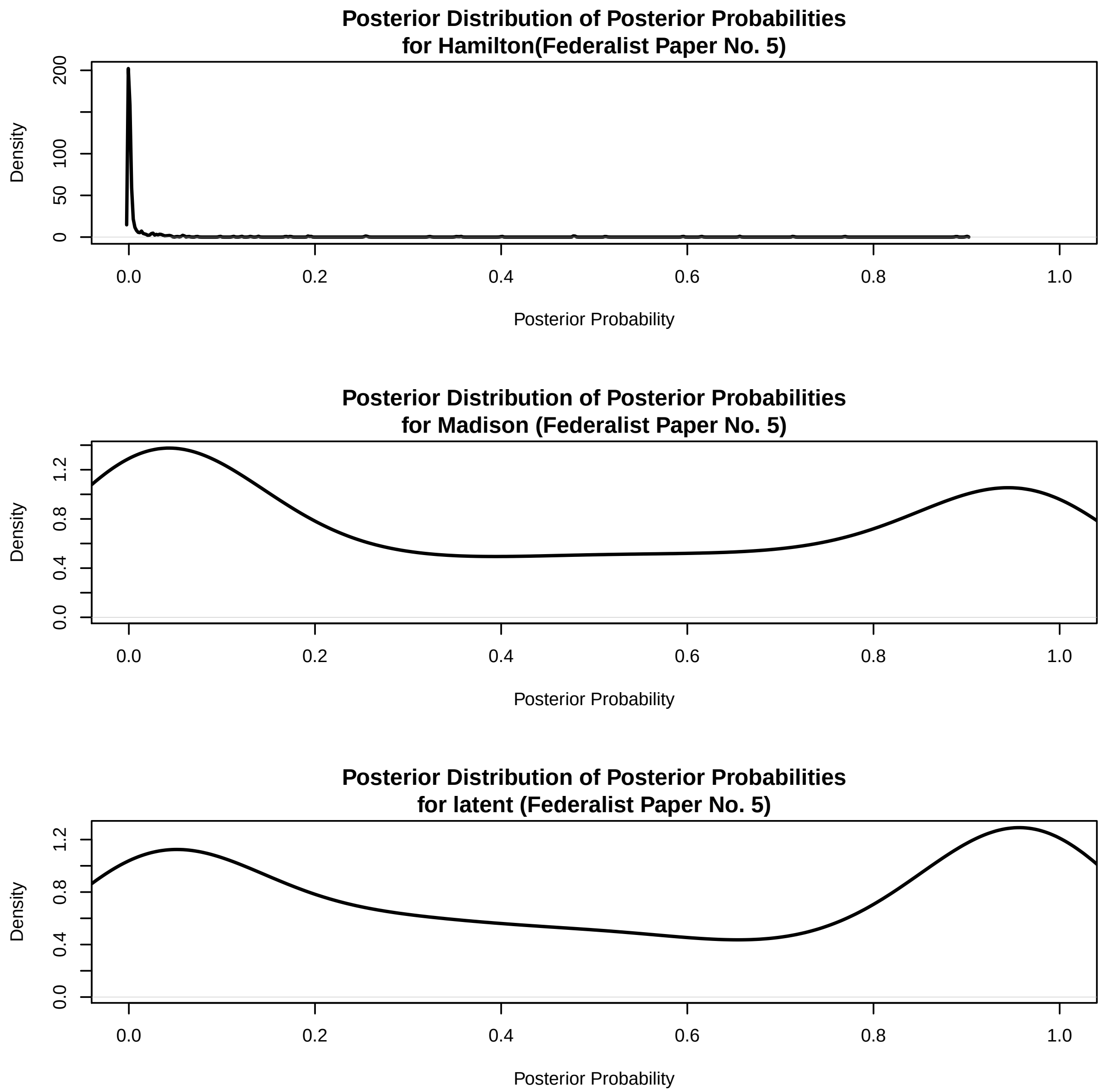


Figure 3.14: Posterior distribution of posterior probabilities for Hamilton, Madison, and latent author for Federalist Paper No. 5.

Madison is the true author of this paper. Appendix B.5 shows the posterior distributions of posterior probability for the other joint Federalist Papers.

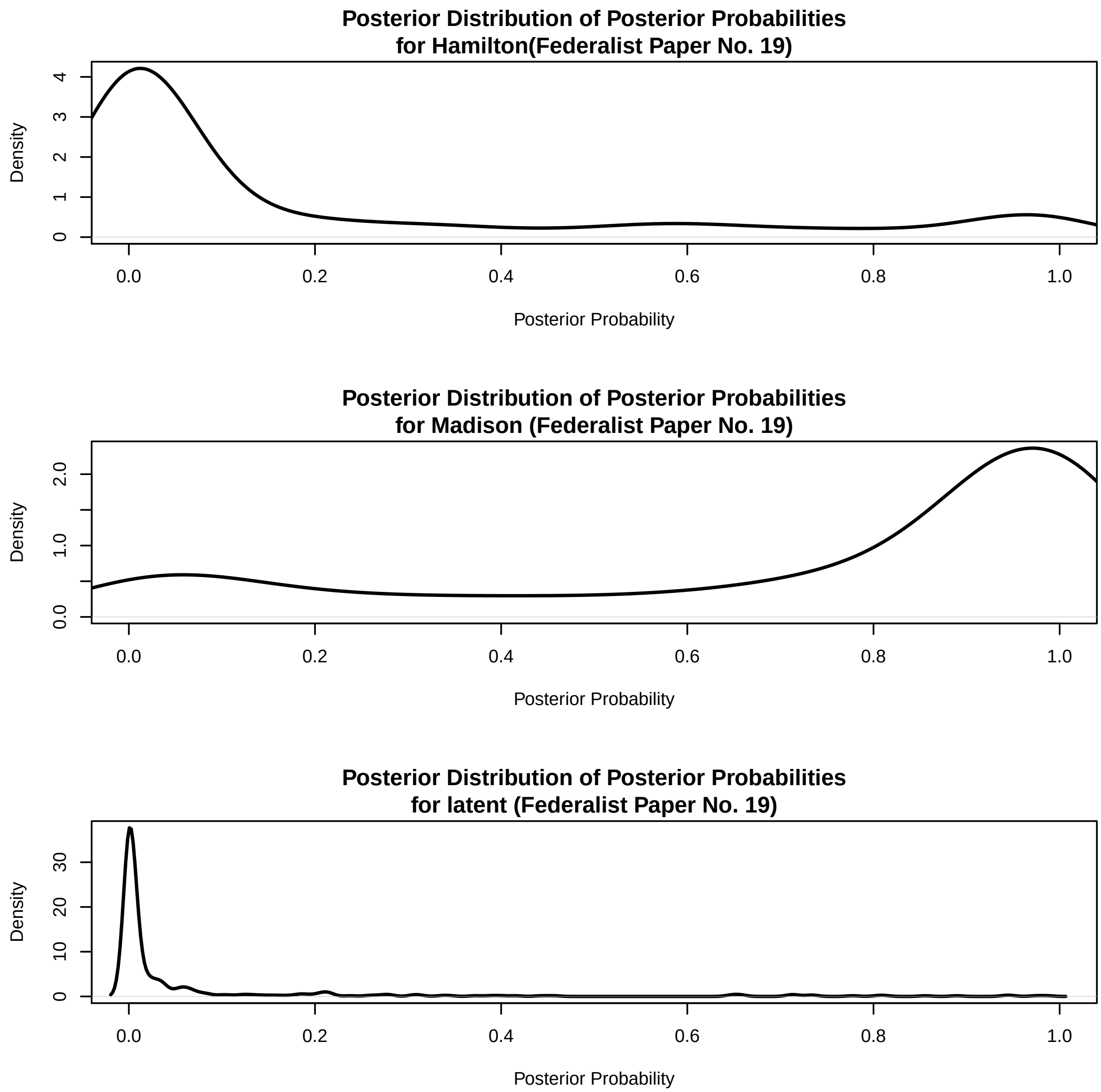


Figure 3.15: Posterior distribution of posterior probabilities for all authors for Federalist Paper No. 19.

Figure 3.16 shows the first 300 Gibbs samples for θ_1, the feature representing the mean of the word "*a*" for Hamilton. θ_1 converges approximately after the 30th Gibbs sample. The cutoff for burn-in was set at the 100th sample.

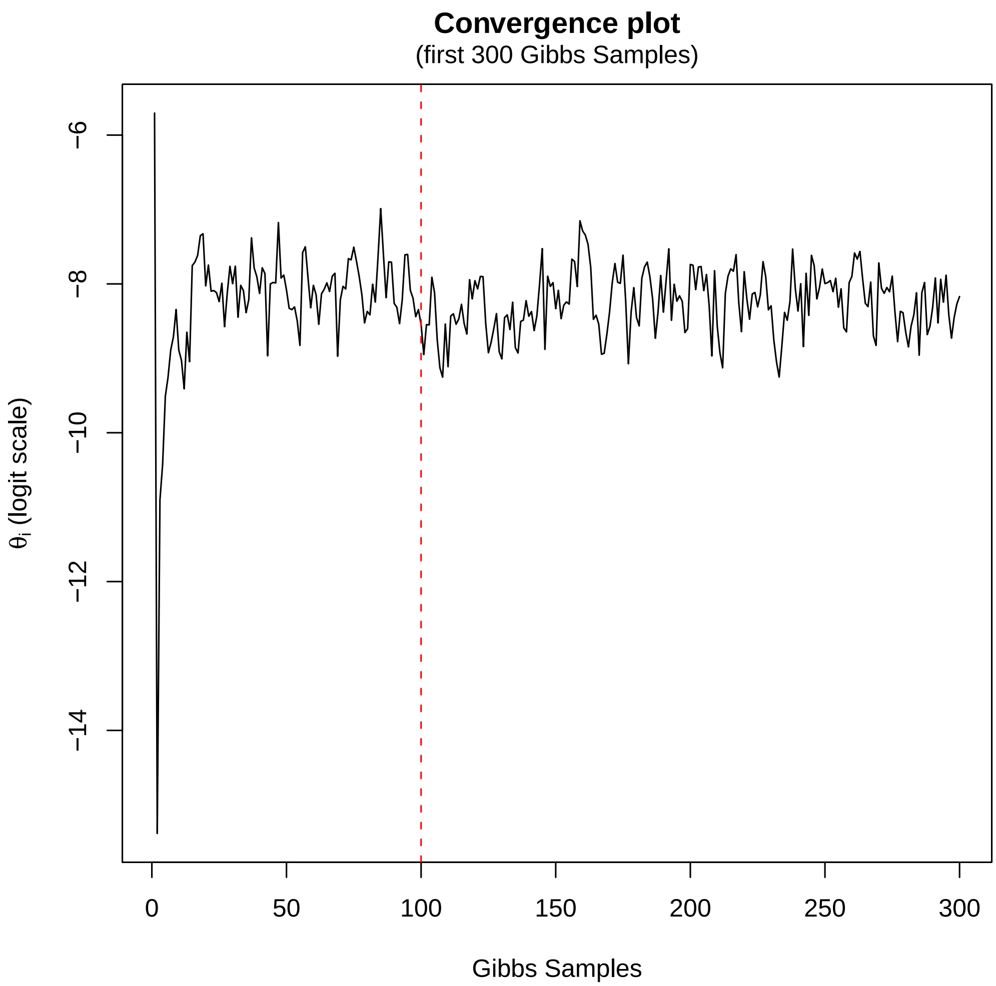


Figure 3.16: First 300 Gibbs samples from θ_1, posterior θ for feature representing the word "*a*" for Hamilton. Red vertical line indicate the cutoff point for burn.

Twelve disputed Federalist Papers are papers nos. 49–58, 62, and 63. Modern scholarly consensus leans towards Madison as the author of all twelve (Colbourne 1998). The results were consistent for all quantiles that were used for classification. All of the disputed

Federalist Papers were classified to Madison. The full independent Bayesian approach to classification agrees with the scholarly consensus.

Appendix B.6 shows the posterior distribution of posterior probability for the 12 disputed Federalist Papers using Hamilton, Madison, and Jay as training texts. Figure 3.17 shows the posterior distribution of posterior probability for paper no. 52. Madison's posterior distribution is heavily left skewed with a mode around 1. On the other hand, Hamilton's and the latent author's posterior distributions are heavily right skewed with a mode around 0.

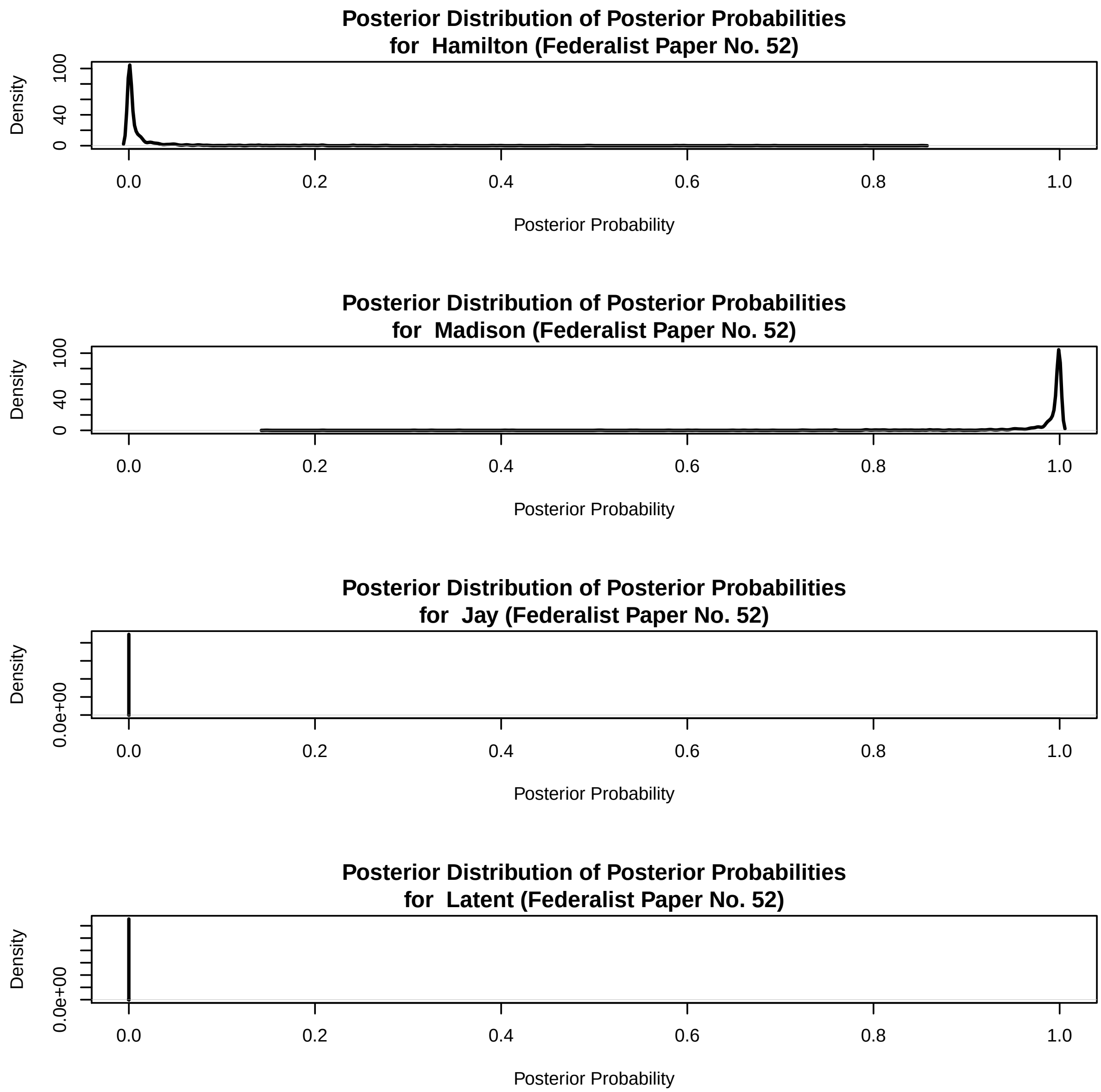


Figure 3.17: Posterior distribution of posterior probabilities for all authors for Federalist Paper No. 52.

CHAPTER 4

CONCLUSION

We have shown that it is possible to shrink the centroids towards class means instead of feature means and to formulate a full independent Bayesian approach to classification. We have also shown that the NSC classification does not account for latent authors and could lead to implausible conclusions. While the extended NSC procedure is a useful tool for the classification analysis of high dimensional data, modifications of its components (i.e., fudge factor, centroid shrinkage, and shrinkage towards class means) do not generally improve the accuracy of classification. It is possible to improve the accuracy using the class-specific threshold as long as the range of Δ values are specified appropriately. The importance of the fudge factor and the shrinkage towards feature means is illustrated emphatically in our studies since the misclassification dramatically increased relative to the original extended NSC. Dabney's (2005) idea of shrinking centroids towards the class means does not change the effectiveness of the stylometric analysis. The most important conclusion of our studies is that the full independent Bayesian approach to the classification, based on Gibbs sampling, is the most effective out of the six methods that we have studied.

Application of the full independent Bayesian approach to the Book of Mormon showed that authorship of the Book of Mormon is still elusive. The results certainly do not support the Spalding-Rigdon theory of the Book of Mormon authorship (Schaalje et al. 2010).

Further application of the full independent Bayesian approach to the Federalist Papers showed that the method can classify well even if the true author is not included in the training set. It seems reasonable to conclude that the full independent Bayesian approach to classification can be used in any authorship analysis text, and its application is not limited to the Book of Mormon and the Federalist Papers.

Areas of further study include performing a simulation study to generalize the classification to more than just the Federalist Papers and the Book of Mormon, analyzing the sensitivity of prior specifications in the full independent Bayesian approach, developing mechanisms to determine goodness-of-fit for each classifier, and formulating a robust classifier relative to the block size (Schaalje et al. 2010). Also, we could formulate the Bayesian classifier using Dirichlet prior. Incorporating this prior will no longer necessitate the *logit* transformation nor independence assumption of the features. It would also be interesting to see if the bivariate nature of the posterior distributions of posterior probabilities is a general phenomenon in authorship analysis.

BIBLIOGRAPHY

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society Series B-Methodology*, 57, 289–300.

Caruana, R., and Niculescu-Mizil, A. (2006), "An Empirical Comparison of Supervised Learning Algorithms," in *Proceedings of the 23rd International Conference on Machine learning*, pp. 161–168.

Colbourne, T. (1998), *Fame and the Founding Fathers: Essays of Douglass Adair*, Indiannapolis: Liberty Fund.

Dabney, A. R. (2005), "Classification of Microarrays to Nearest Centroids," *Bioinformatics*, 21, 4148–4154.

Donoho, D. L., and Johnstone, J. M. (1994), "Ideal Spatial Adaptation by Wavelet Shrinkage," *Biometrika*, 81, 425–455.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002), "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data," *Journal of the American Statistical Association*, 97, 77–87.

Forsyth, R., and Holmes, D. (1996), "Feature-Finding for Text Classification," *Literary and Linguistic Computing*, 11, 4.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis (2nd Edition)*, New York: Chapman & Hall/CRC.

Genkin, A., Lewis, D., and Madigan, D. (2007), "Large-scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, 49, 291–304.

Hastie, T., T., R., R., and Friedman, J. (2001), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer, New York.

Holmes, D. I. (1992), "A Stylometric Analysis of Mormon Scripture and Related Texts," *Journal of the Royal Statistical Society Series A-Statistics in Society*, 155, 91–120.

Jockers, M. L., Witten, D. M., and Criddle, C. S. (2009), "Reassessing Authorship of the Book of Mormon using Delta and Nearest Shrunken Centroid Classification," *Literary and Linguistic Computing*, 23, 465–491.

Johnson, V. E. (2004), "A Bayesian χ^2 Test for Goodness-of-Fit," *The Annals of Statistics*, 32, 2361–2384.

Kjell, B., Woods, W. A., and Frieder, O. (2006), "Information Retrieval using Letter Tuples with Neural Network and Nearest Neighbor Classifiers," *IEEE International Conference on Systems, Man and Cybernetics*, 2, 1222–1225.

Koppel, Argamon, and Shimoni (2002), "Automatically Categorizing Written Texts by Author Gender," *Literary and Linguistic Computing*, 17, 401–412.

Koppel, M., Schler, J., and Argamon, S. (2009), "Computational Methods in Authorship Attribution," *Journal of The American Society for Information Science and Technology*, 60, 9–26.

Langseth, H., and Nielsen, T. D. (2006), "Classification using Hierarchical Naive Bayes Models," *Springer Netherlands*, 63, 135–159.

Lee, J. W., Lee, J. B., Park, M., and Song, S. H. (2005), "An Extensive Comparison of Recent Classification tools applied to microarray data," *Computational Statistics & Data Analysis*, 48, 869–885.

Li, J., Zheng, R., and Chen, H. (2006), "From Fingerprint to Writeprint," *Communities of the ACM*, 49, 76–82.

Mardia, K. V., Kent, J. T., and Bibb, J. M. (1979), *Multivariate Analysis*, Academic Press.

Moon, T. K., Howland, P., and Gunther, J. H. (2006), "Document Author Classification using Generalized Discriminant Analysis," *Proceedings of the Workshop on Text Mining, SIAM International Conference on Data Mining.*

Nadon, R., and Shoemaker, J. (2002), "Statistical Issues with Microarrays: Processing and Analysis," *TRENDS in Genetics*, 18, 265–271.

Nakagawa, S. (2004), "A Farewell to Bonferroni: the Problems of Low Statistical Power and Publication Bias," *Behavioral Ecology*, 15, 1044–1045.

Peng, F., Schuurmans, D., and Wang, S. (2006), "Augumenting Naive Bayes Text Classifier with Statistical Language Models," *Information Retreival*, 7, 317–345.

Rencher, A. C. (2002), *Methods of Multivariate Analysis*, John Wiley & Sons, Inc.

Schaalje, G. B., Fields, P. J., Roper, M., and Snow, G. L. (2010), "Extensions to the Nearest Shrunken Centroid Classification Method, with Special Reference to Book of Mormon Authorship," *Literary and Linguistic Computing*, in press.

Shi, J., Murray-Smith, R., and Titterington, D. (2003), "Bayesian Regression and Classification Using Mixtures of Gaussian Processes," *International Journal of Adaptive Control Signal Process*, 17, 149–161.

Sims, G. E., Jun, S.-R., Wu, G. A., and Kim, S.-H. (2009), "Alignment-Free Genome Comparison with Feature Frequency Profiles (FFP) and Optimal Resolutions," *Proceedings of the National Academy of Science*, 106(8), 2677–2682.

Tan, P. K., Downey, T. J., Jr., E. L. S., Pin Xu, D. F., Dimitrov, D. S., Lempicki, R. A., Raaka, B. M., and Cam, M. C. (2003), "Evaluation of Gene Expression Measurements from Commerical Microasaray Platforms," *Nucleic Acids Research*, 31, 5676–5684.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002), "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression," *Proceedings of the National Academy of Sciences*, 99, 6567–6572.

—— (2003), "Class Prediction by Nearest Shrunken Centroids, with Application to DNA Microarrays," *Statistical Science*, 18, 104–117.

Volinia, S., Calin, G. A., Liu, C.-G., Ambs, S., Cimmino, A., Petrocca, F., Visone, R., Iorio, M., Roldo, C., Ferracin, M., Prueitt, R. L., Yanaihara, N., Lanza, G., Scarpa, A., Vecchione, A., Negrini, M., Harris, C. C., , and Croce, C. M. (2006), "A microRNA Expression Signature of Human Solid Tumors Defines Cancer Gene Targets," *Proceedings of the National Academy of Sciences*, 103, 2257–2261.

Weiss, R. (1995), "An Approach to Bayesian Sensitivity Analysis," *Journal of the Royal Statistical Society Series B-Methodology*, 58(4), 739–750.

Zhang, H. (2004), "The Optimality of Naive Bayes," in *Proceedings of the Seventeenth Florida Artificial Intelligence Research Society Conference*, pp. 562–567.

APPENDICES

APPENDIX A

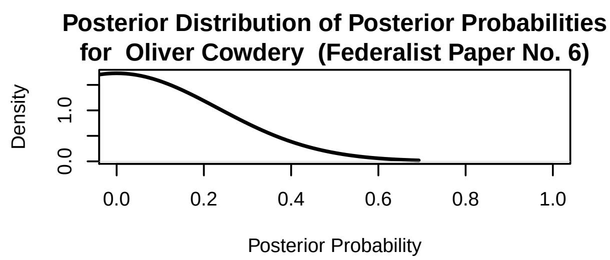
LIST OF NON-CONTEXTUAL LITERARY FEATURES

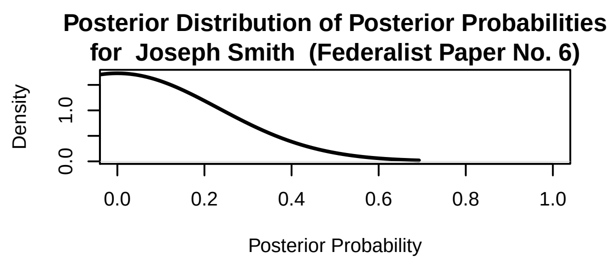
Non-Contextual Words
a, after, again, against, all, among, an, and, are, as,
at, away, be, because, been, before, but, by, came, children,
come, day, did, do, down, earth, even, every, father, for,
forth, from, go, great, had, hand, have, he, her, him,
his, I, if, in, into, is, it, king, know, land,
made, man, many, may, me, men, might, more, my, no,
not, now, O, of, on, one, or, our, out, over,
pass, people, power, said, say, shall, should, so, son, that,
the, their, them, then, there, therefore, these, they, things, this,
those, thus, time, to, up, upon, us, was, we, were,
which, who, will, with, words, would, you, your

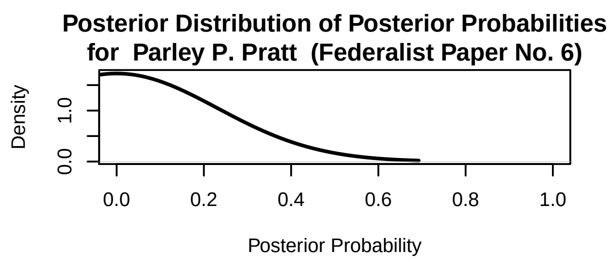
APPENDIX B

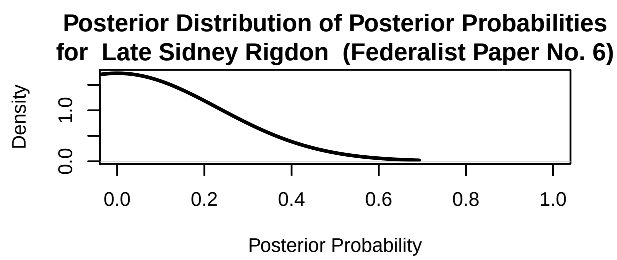
B.1 Hamilton Federalist Papers that were misclassified by extended NSC (without Hamilton in the training set)

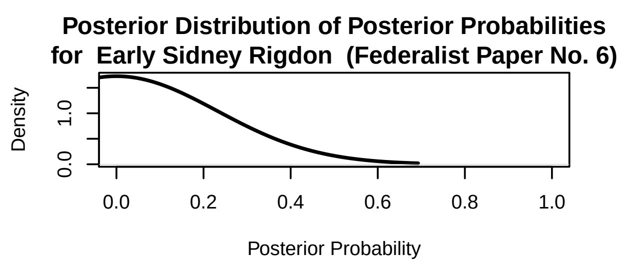
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 6

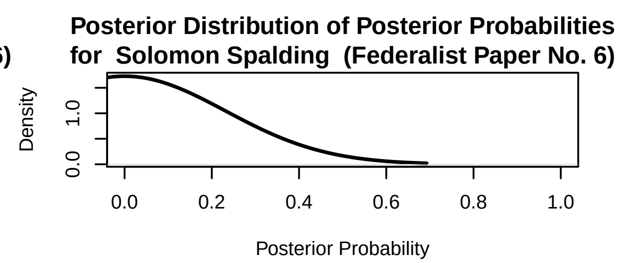


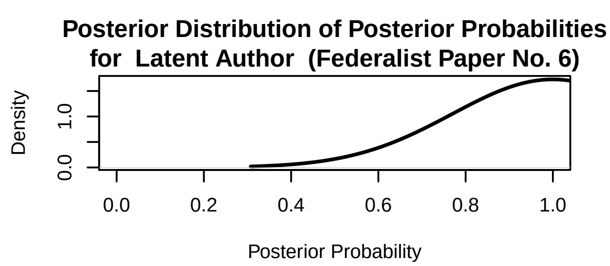




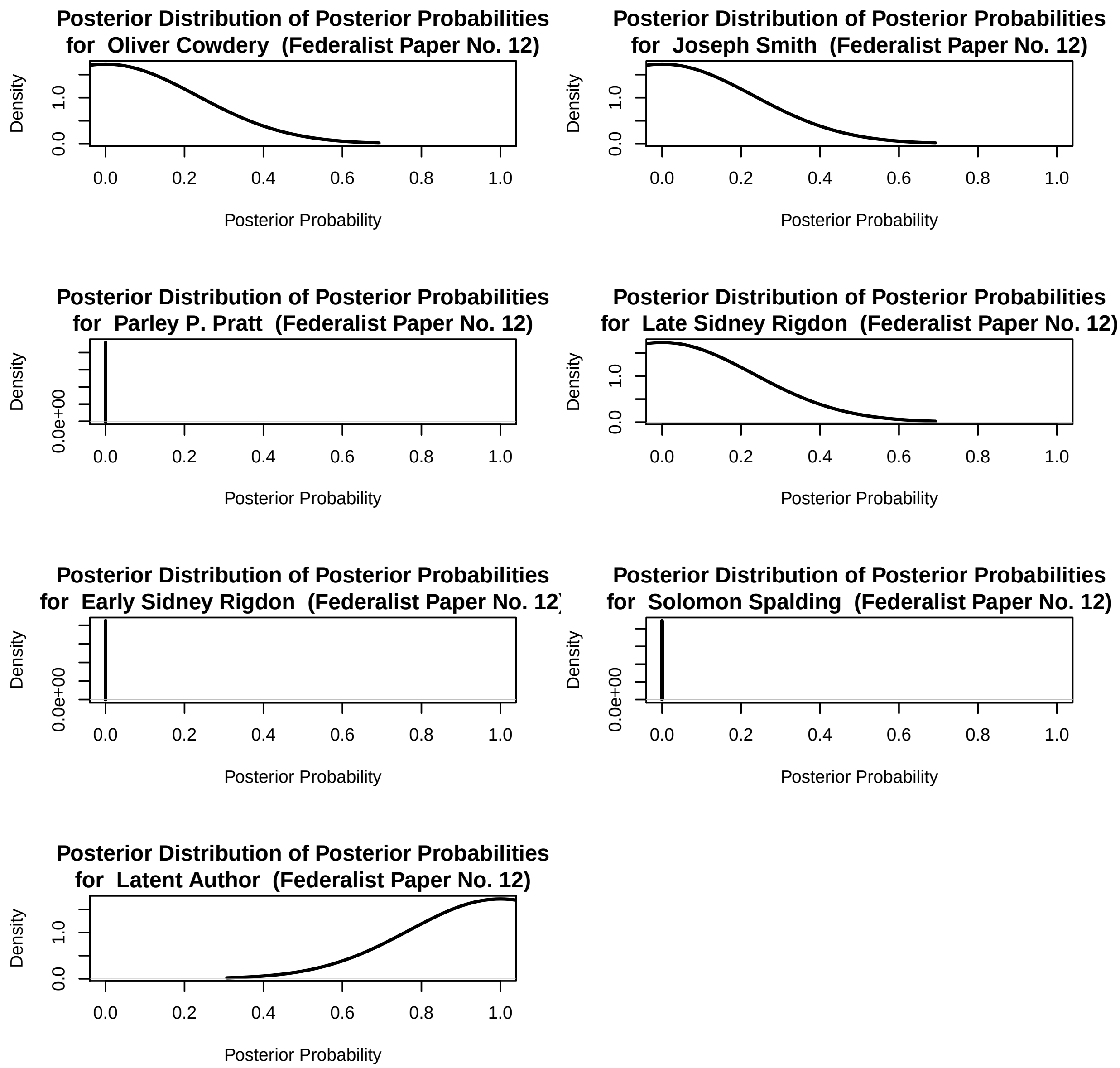




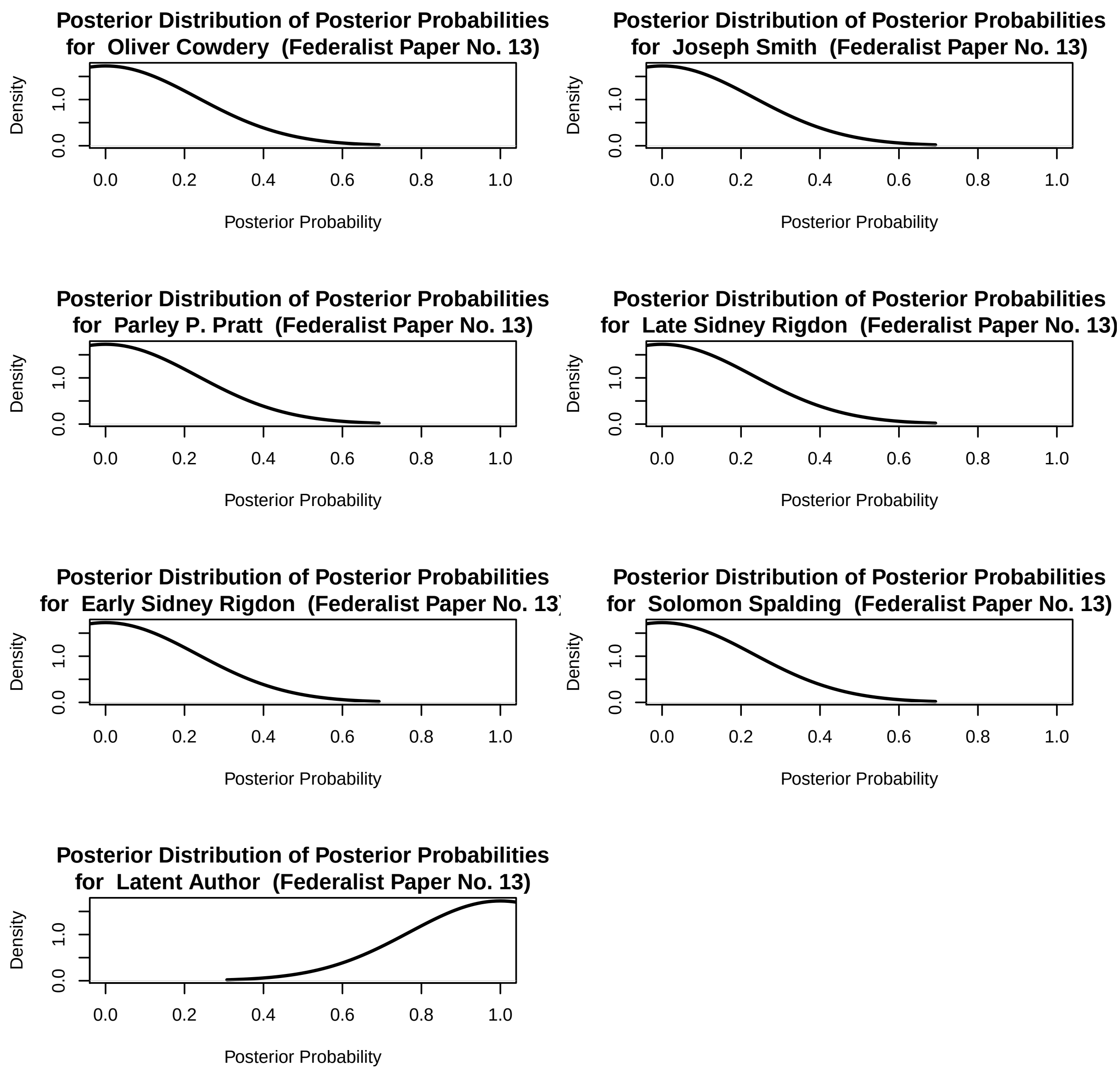




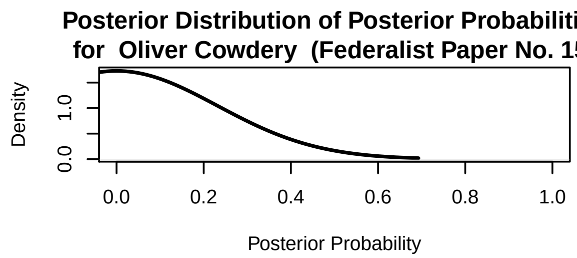
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 12

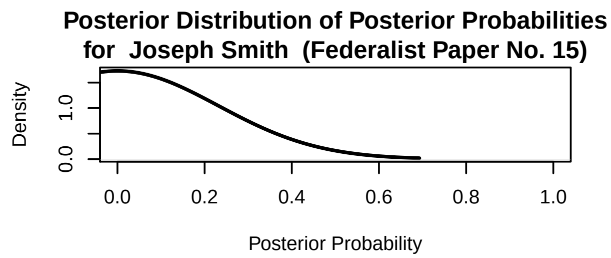


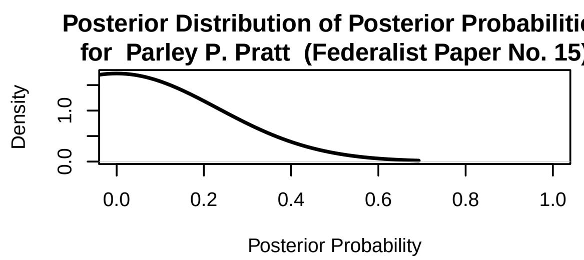
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 13

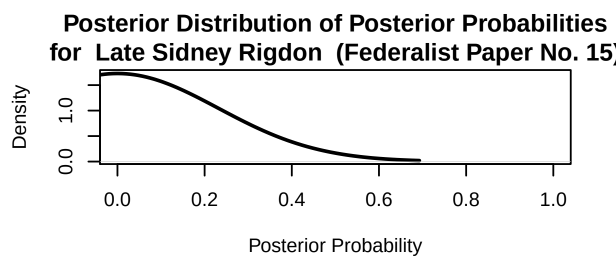


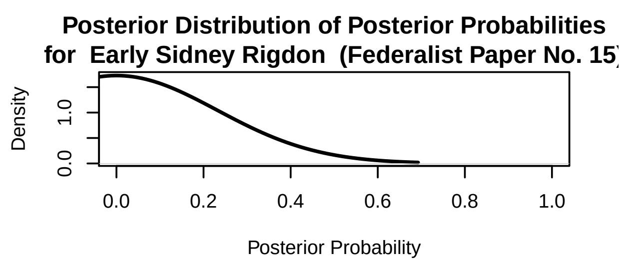
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 15

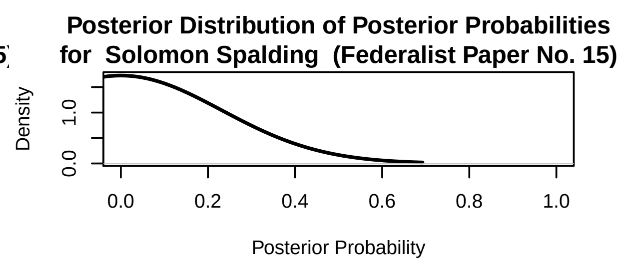


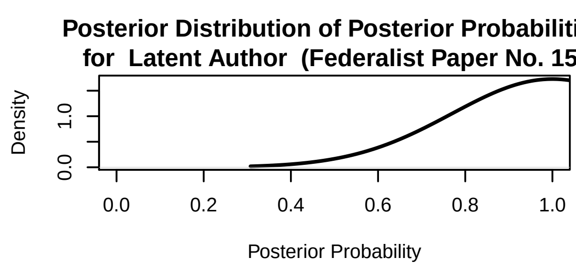




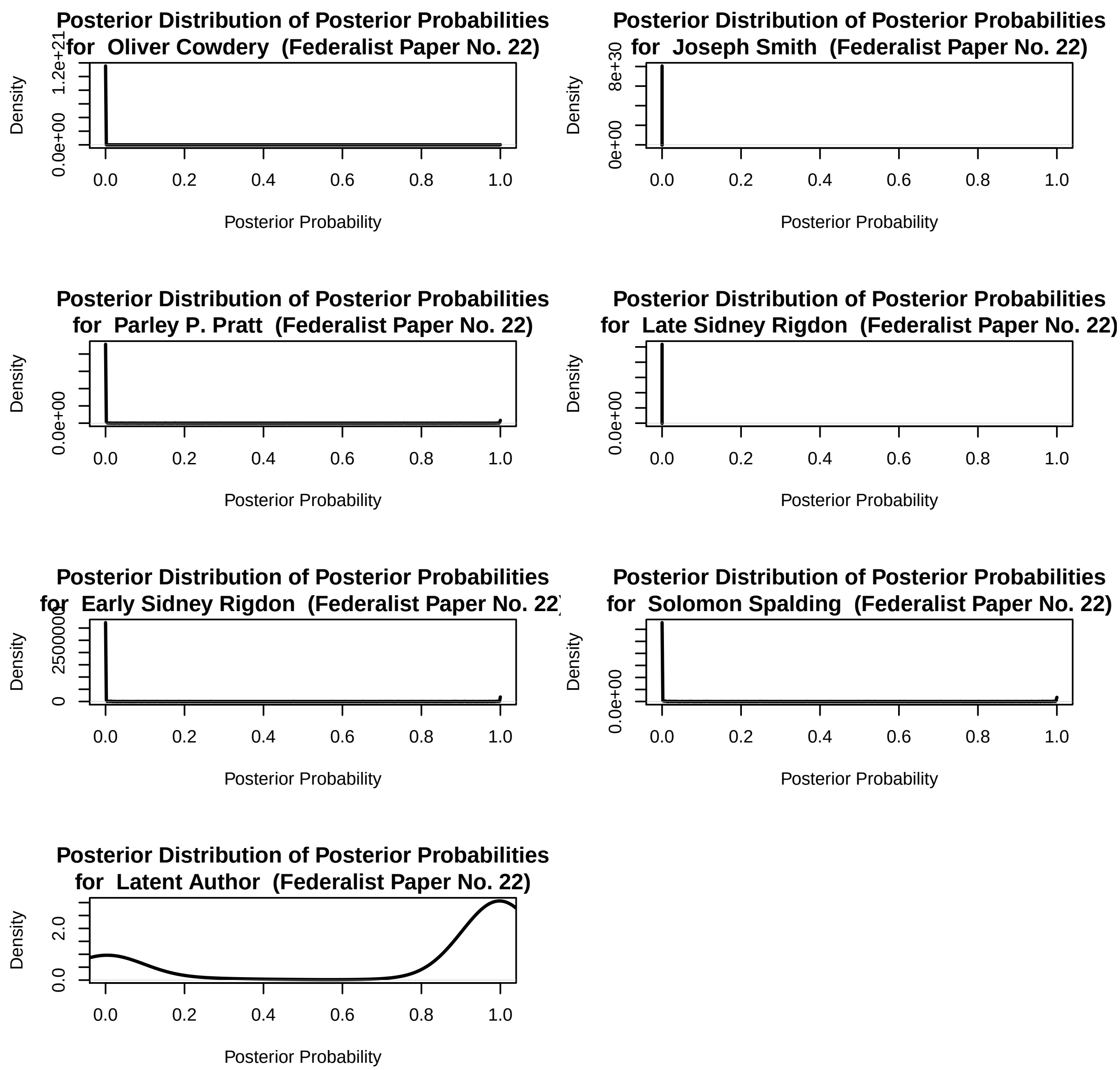




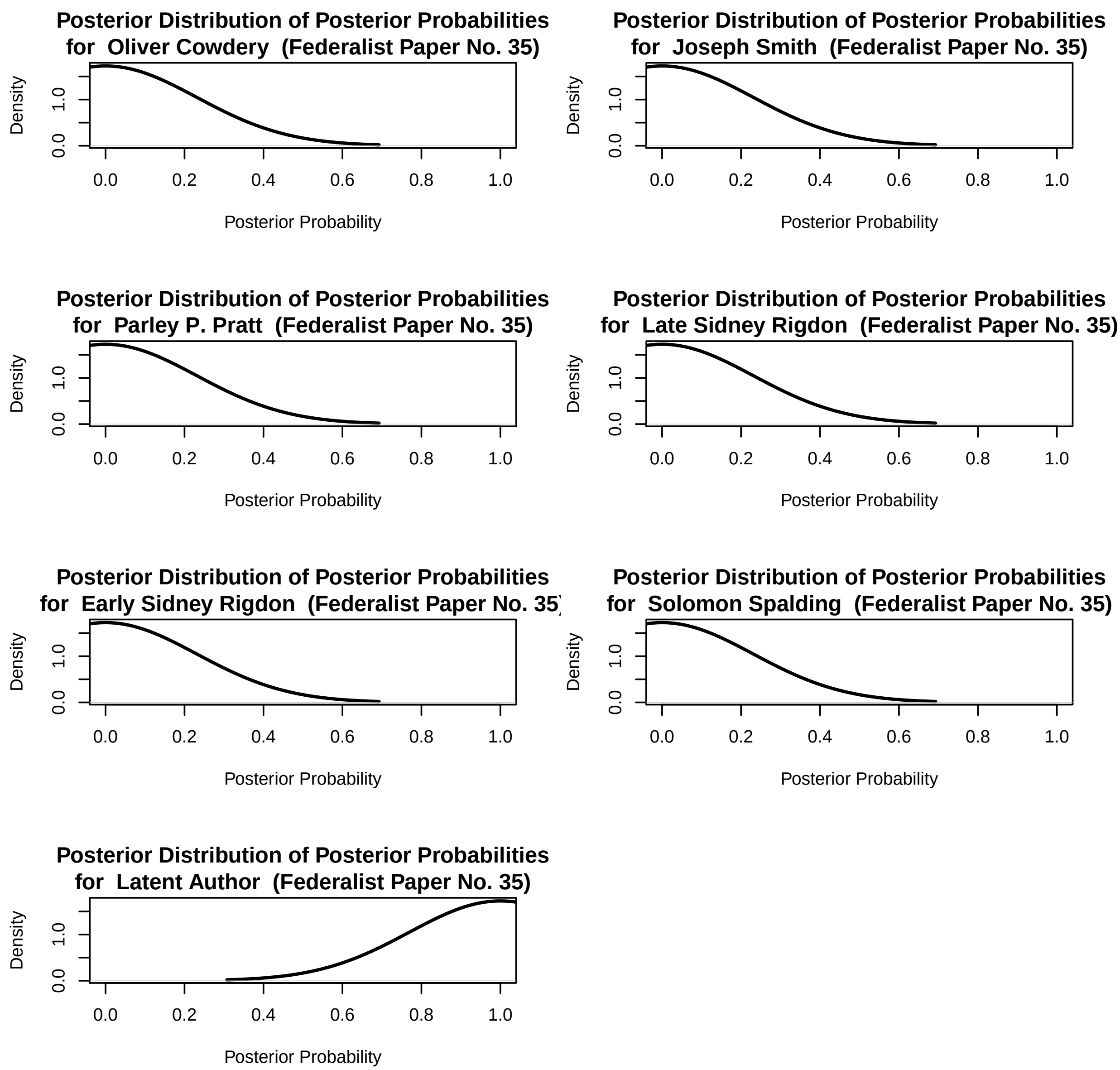




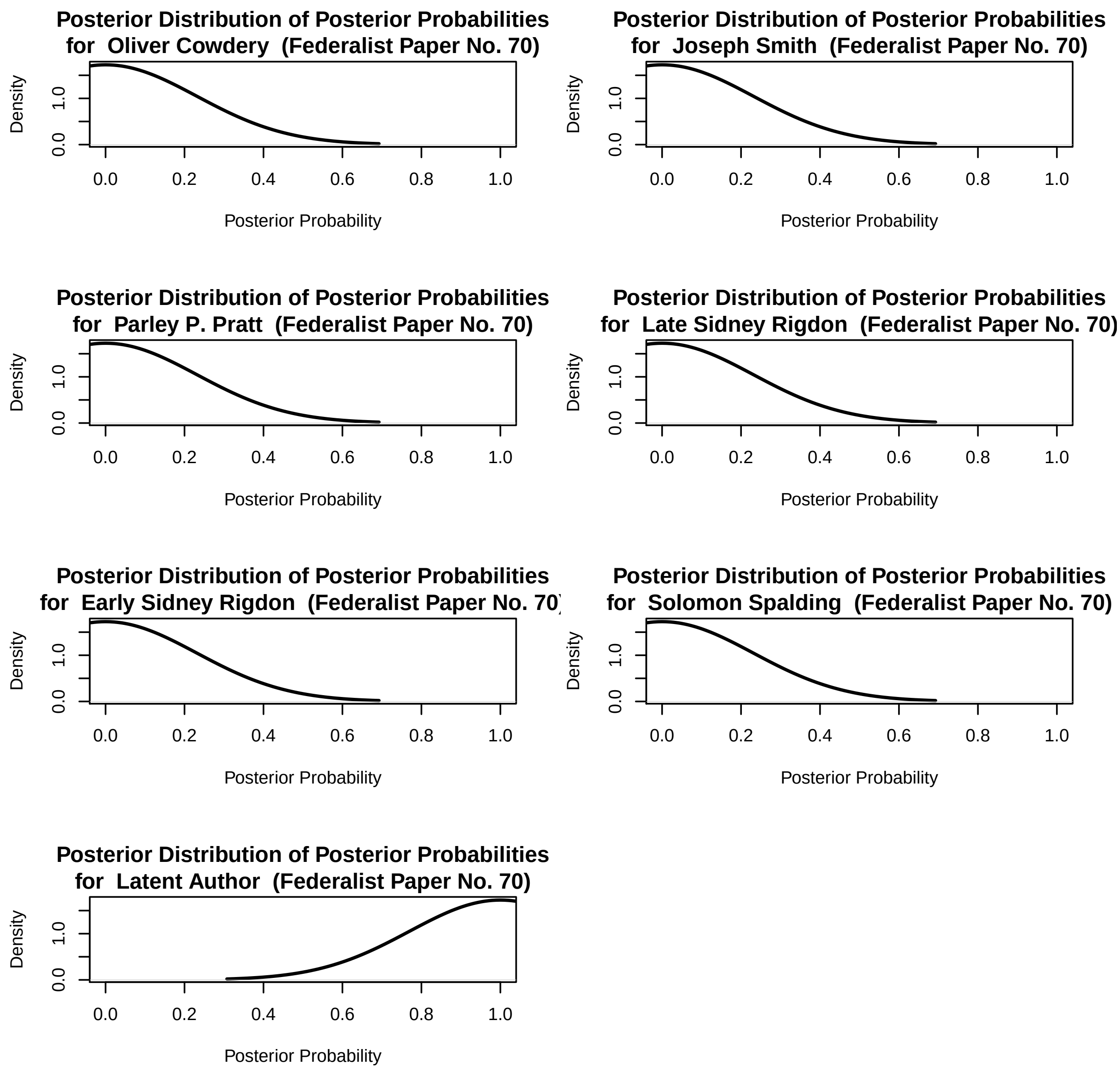
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 22



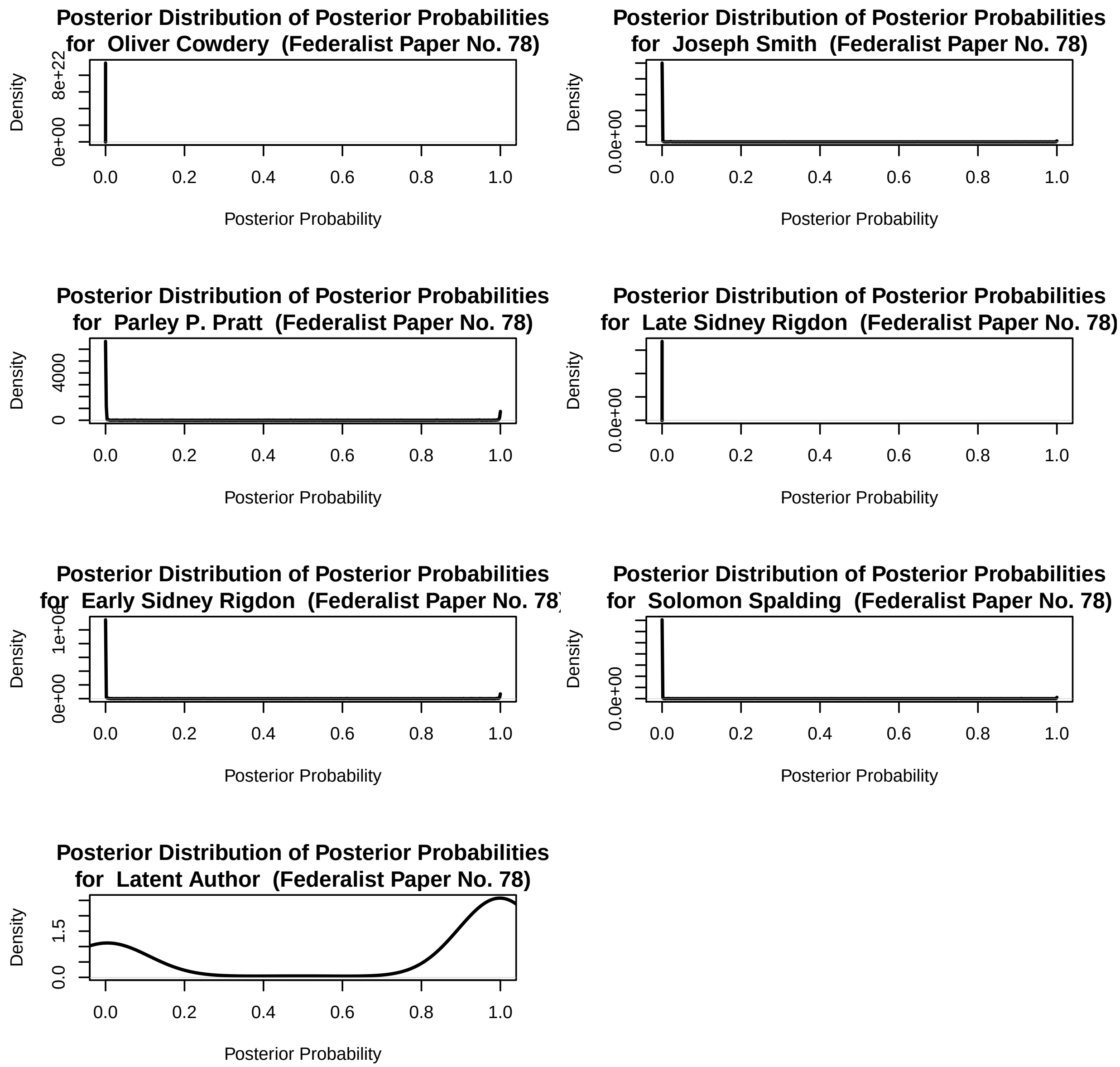
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 35



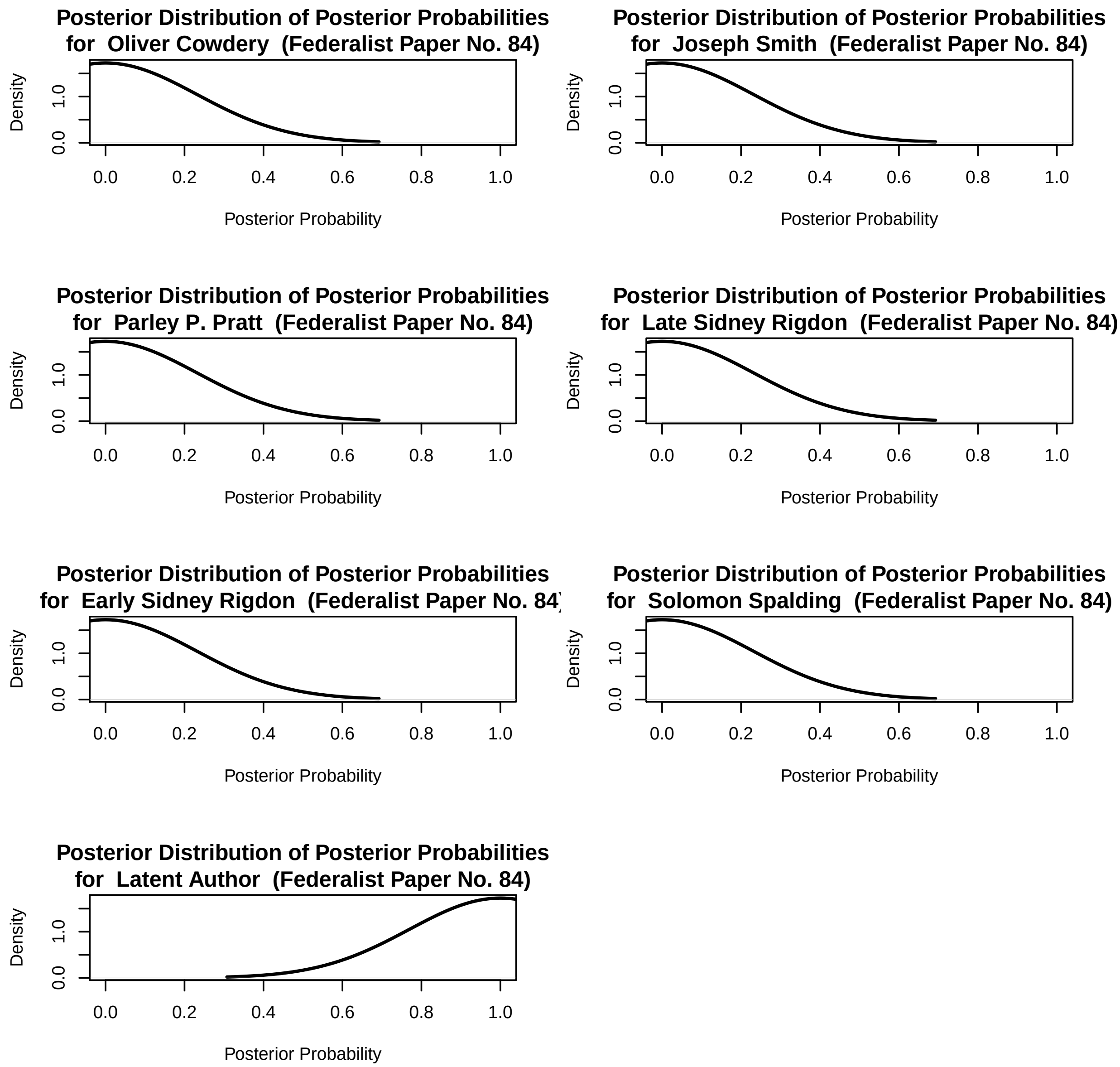
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 70



Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 78

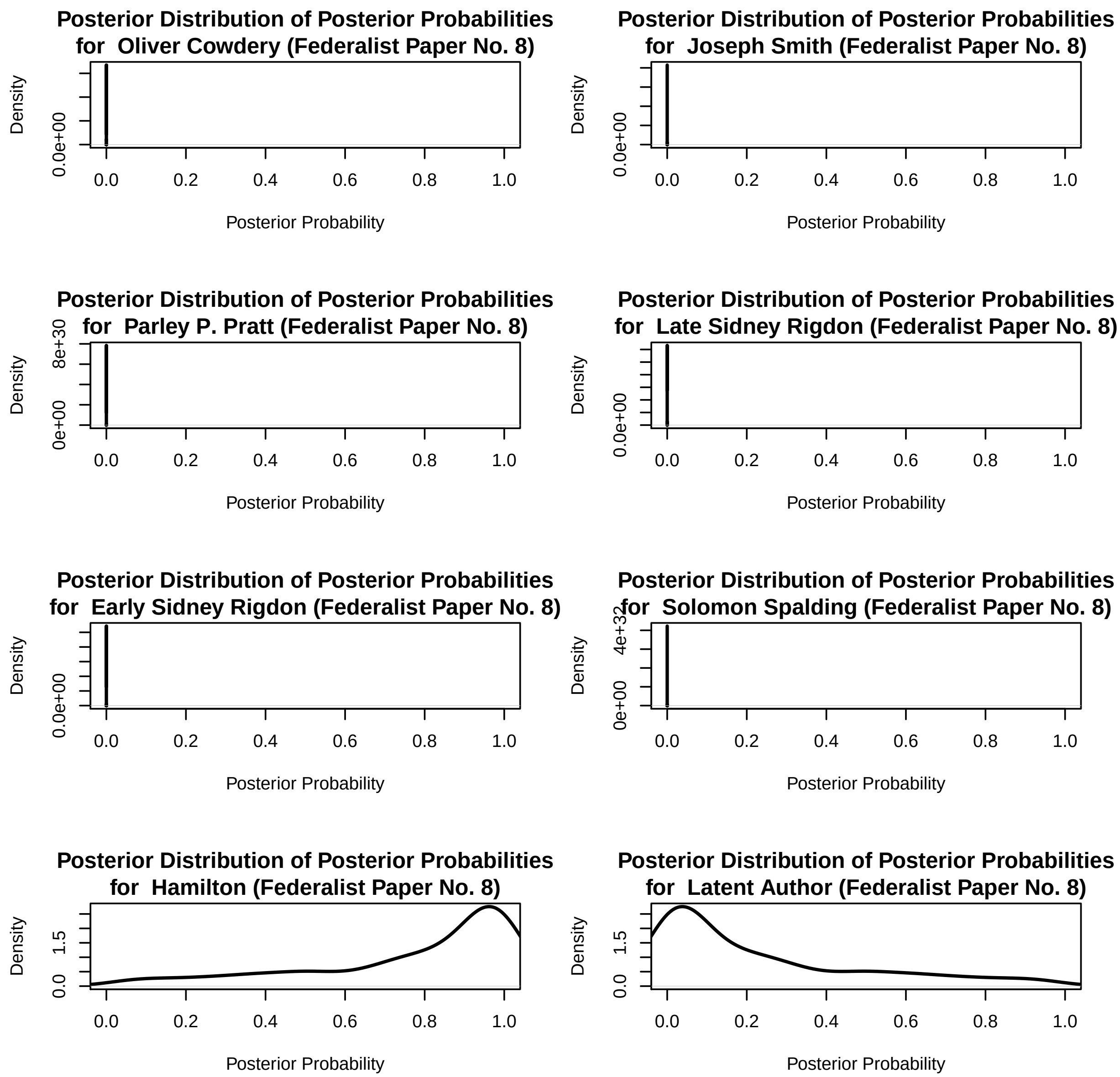


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 84

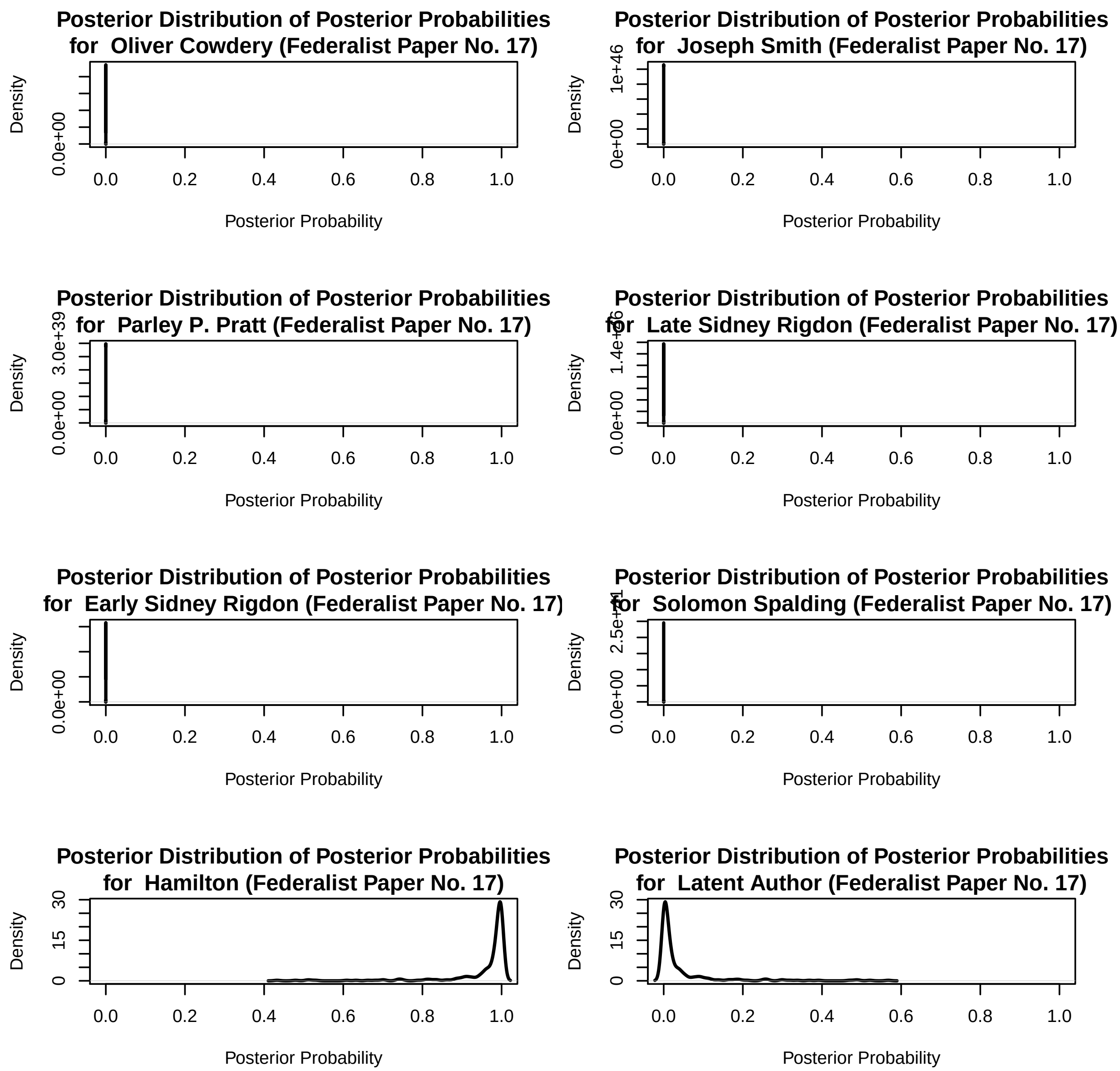


B.2 Hamilton Federalist Papers that were misclassified by extended NSC (with Hamilton in the training set)

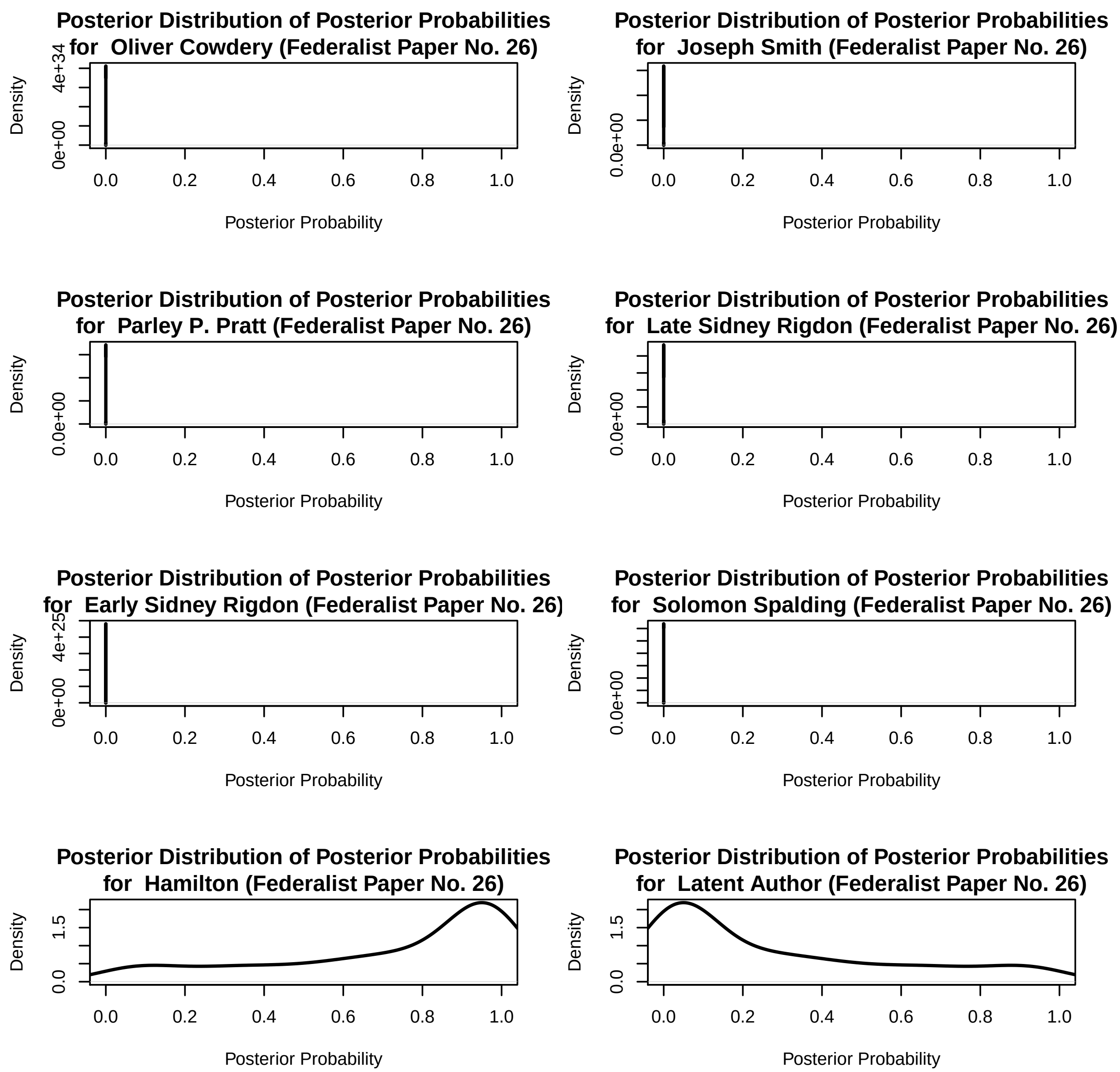
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 8



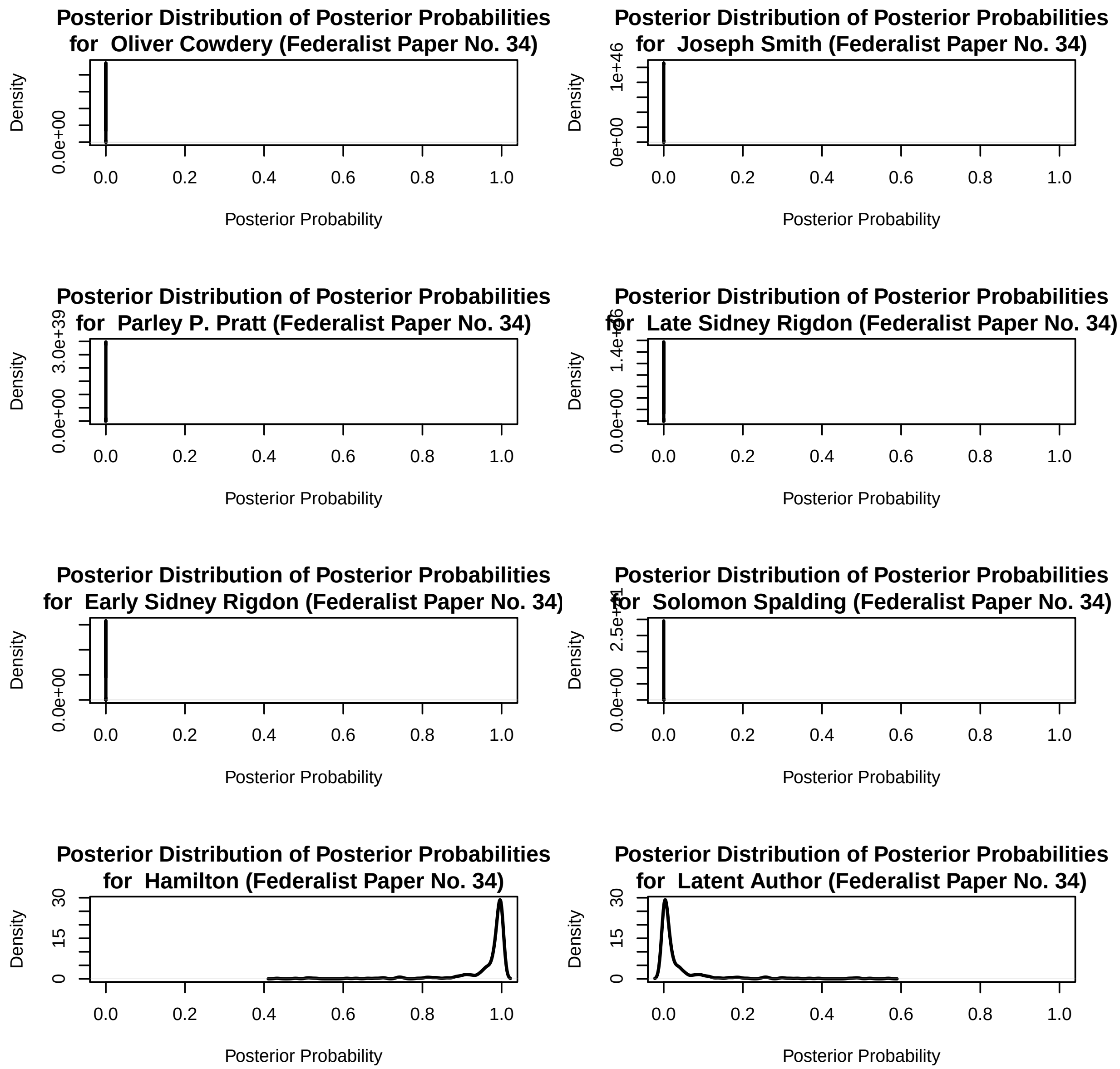
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 17



Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 26

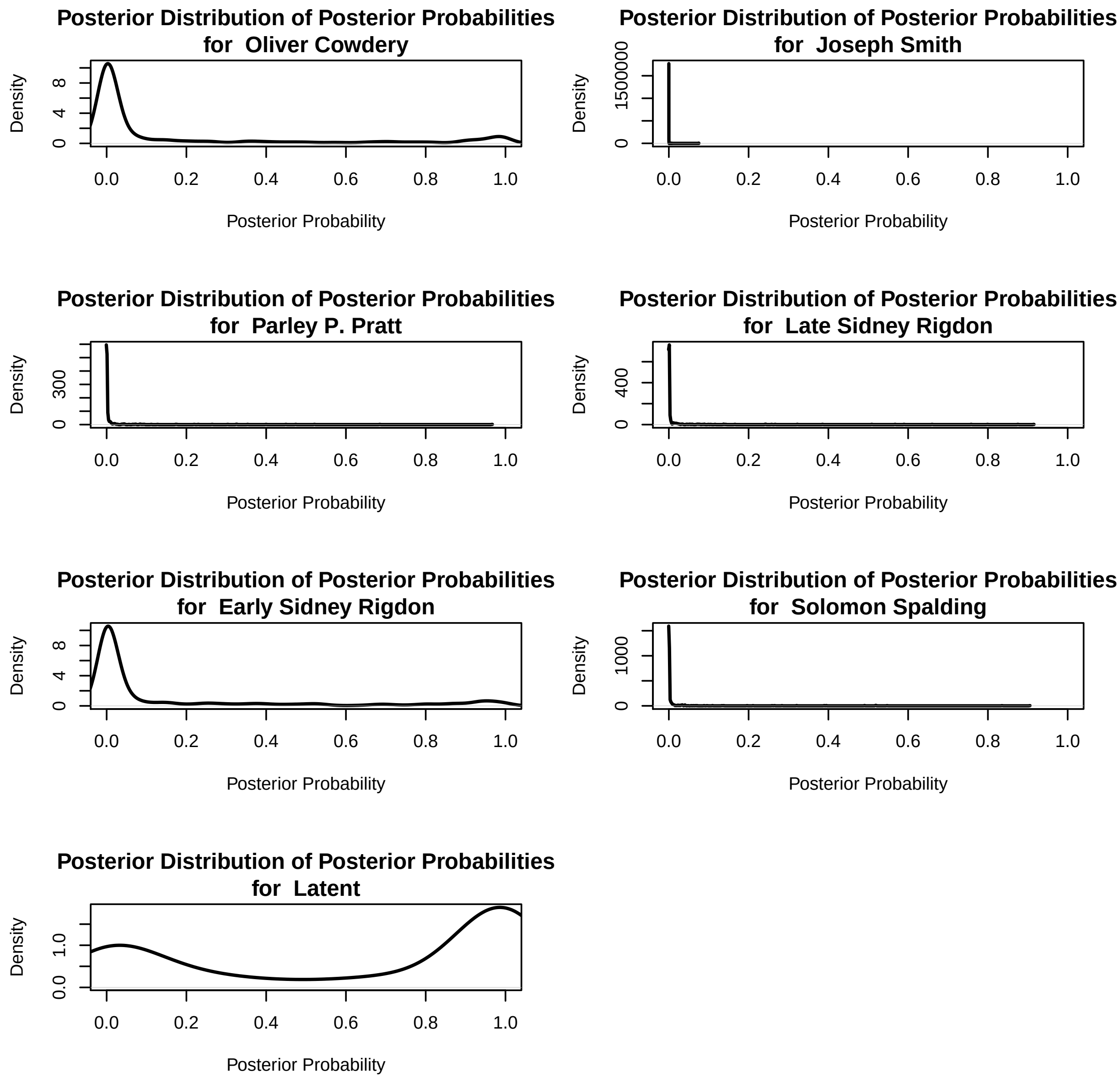


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 34

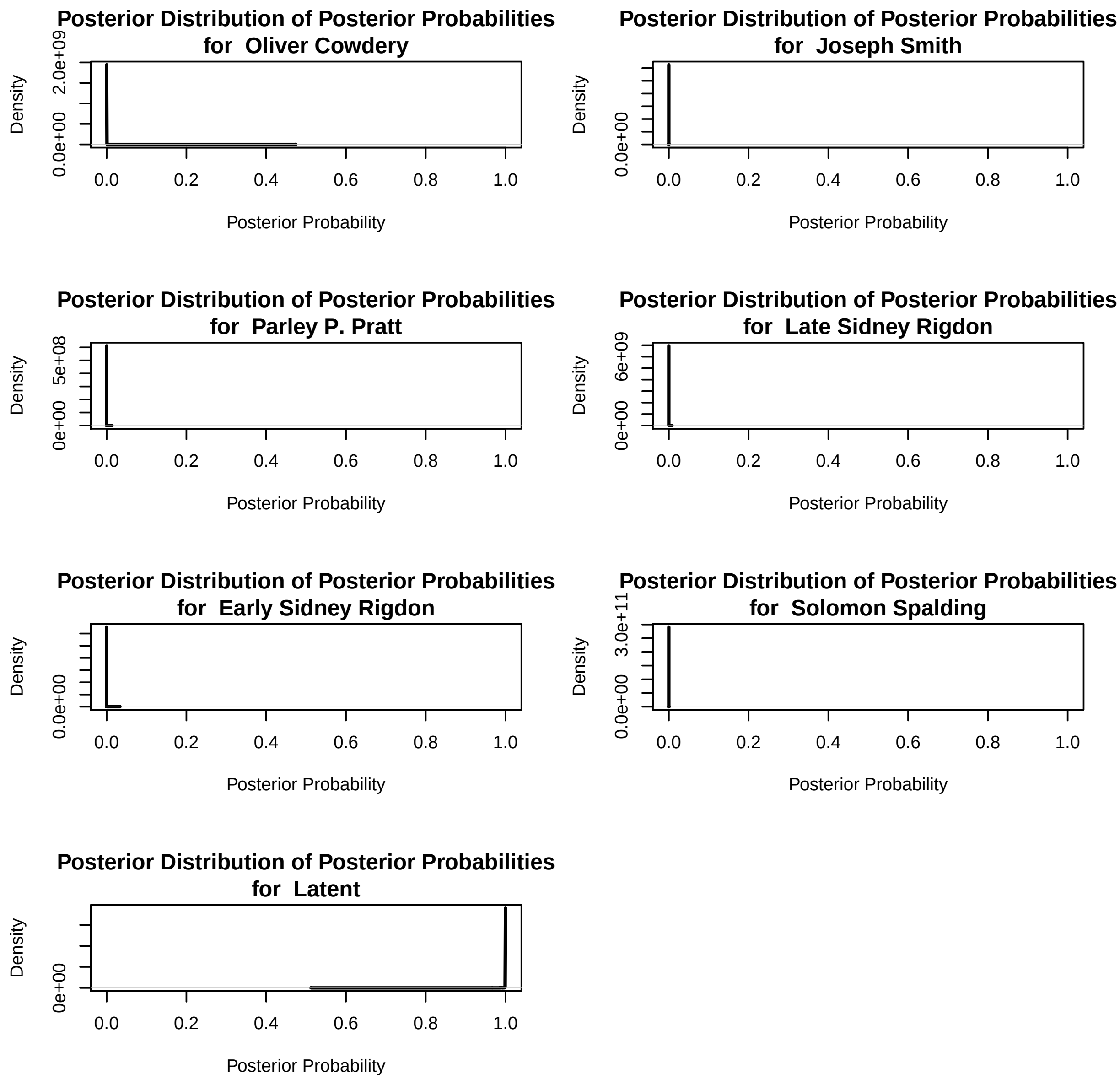


B.3 5000 words block of texts from Book of Mormon

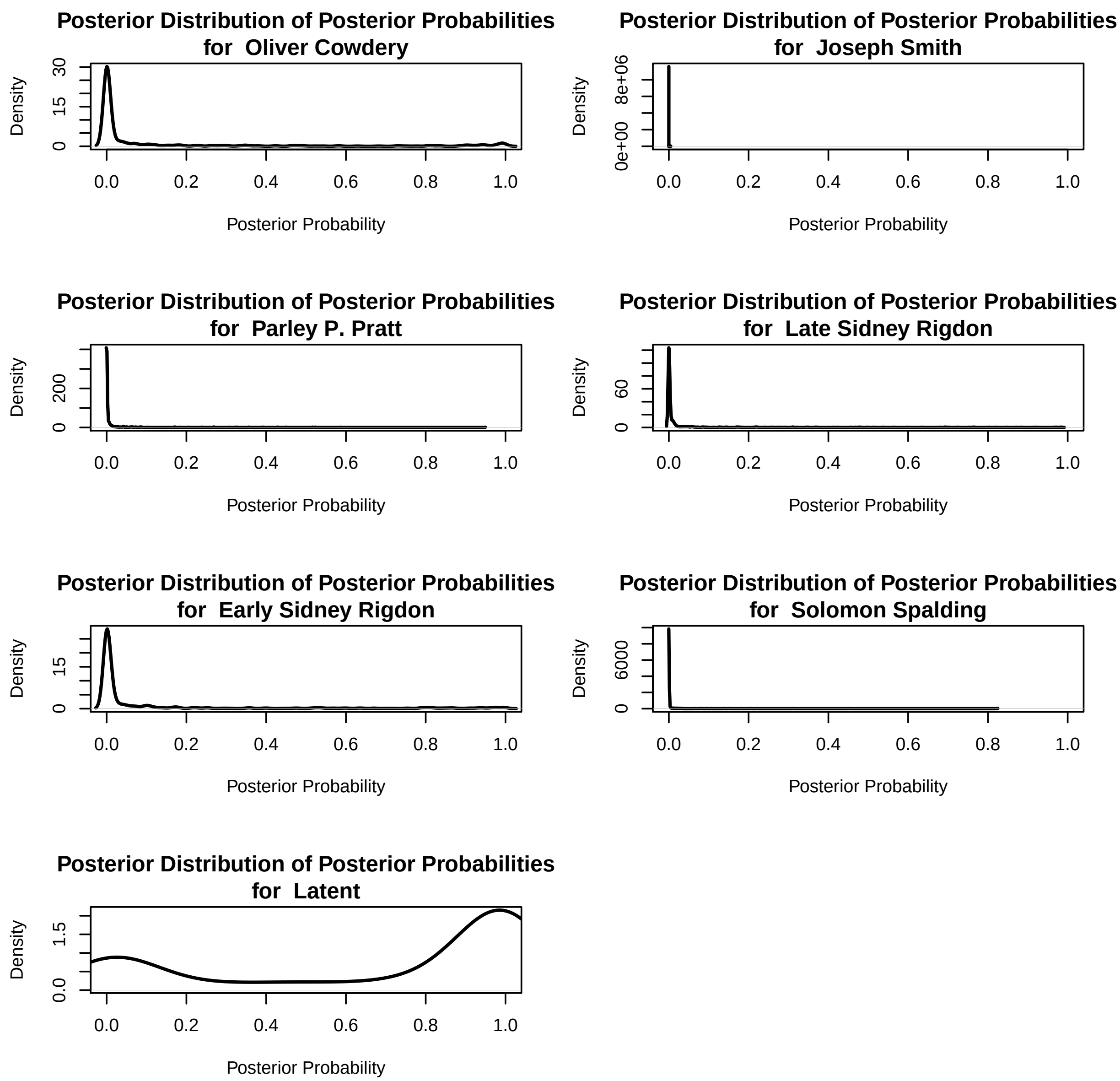
Posterior distribution of Posterior Probabilities for each authors; block no. 1



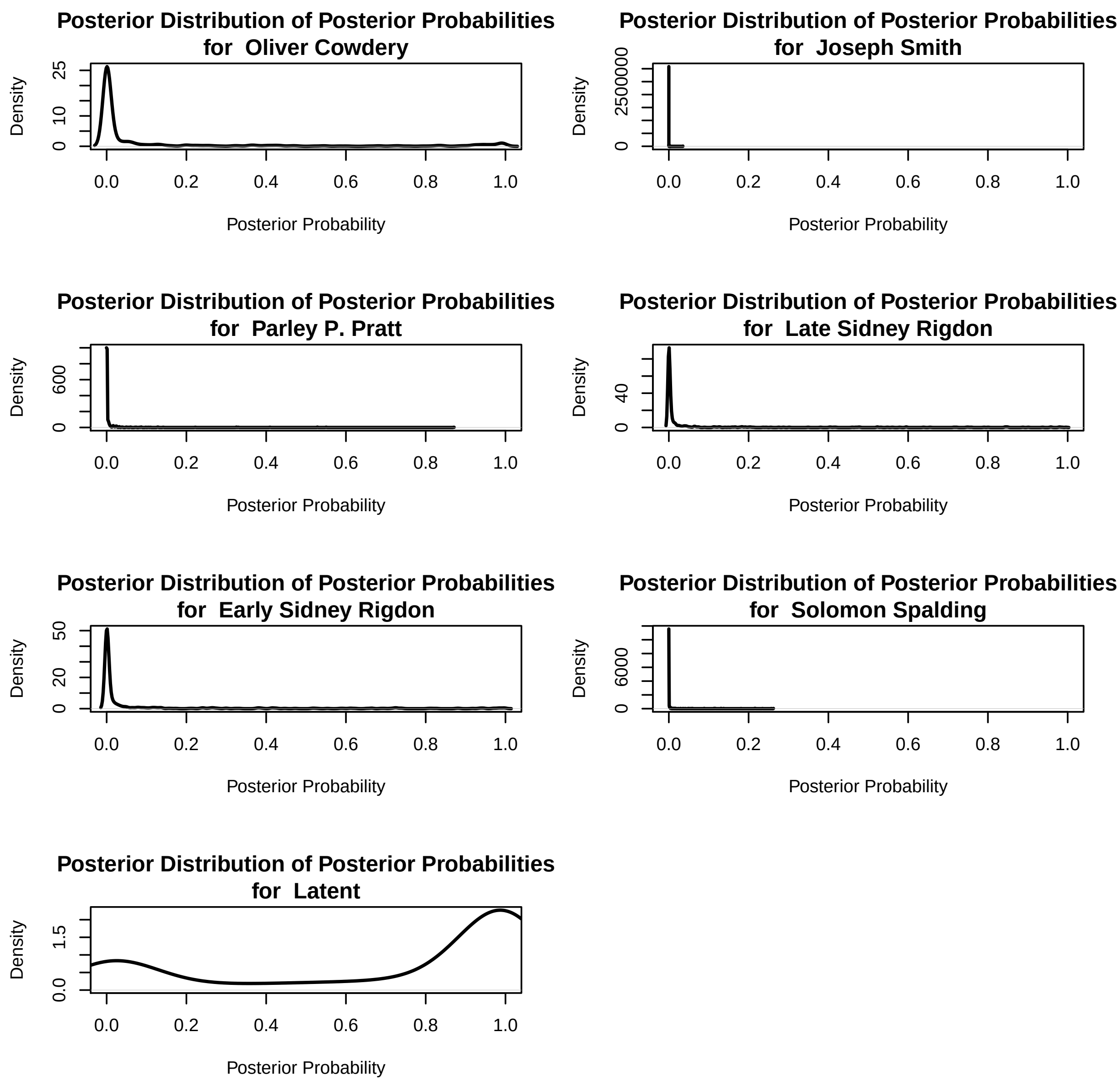
Posterior distribution of Posterior Probabilities for each authors; block no. 2



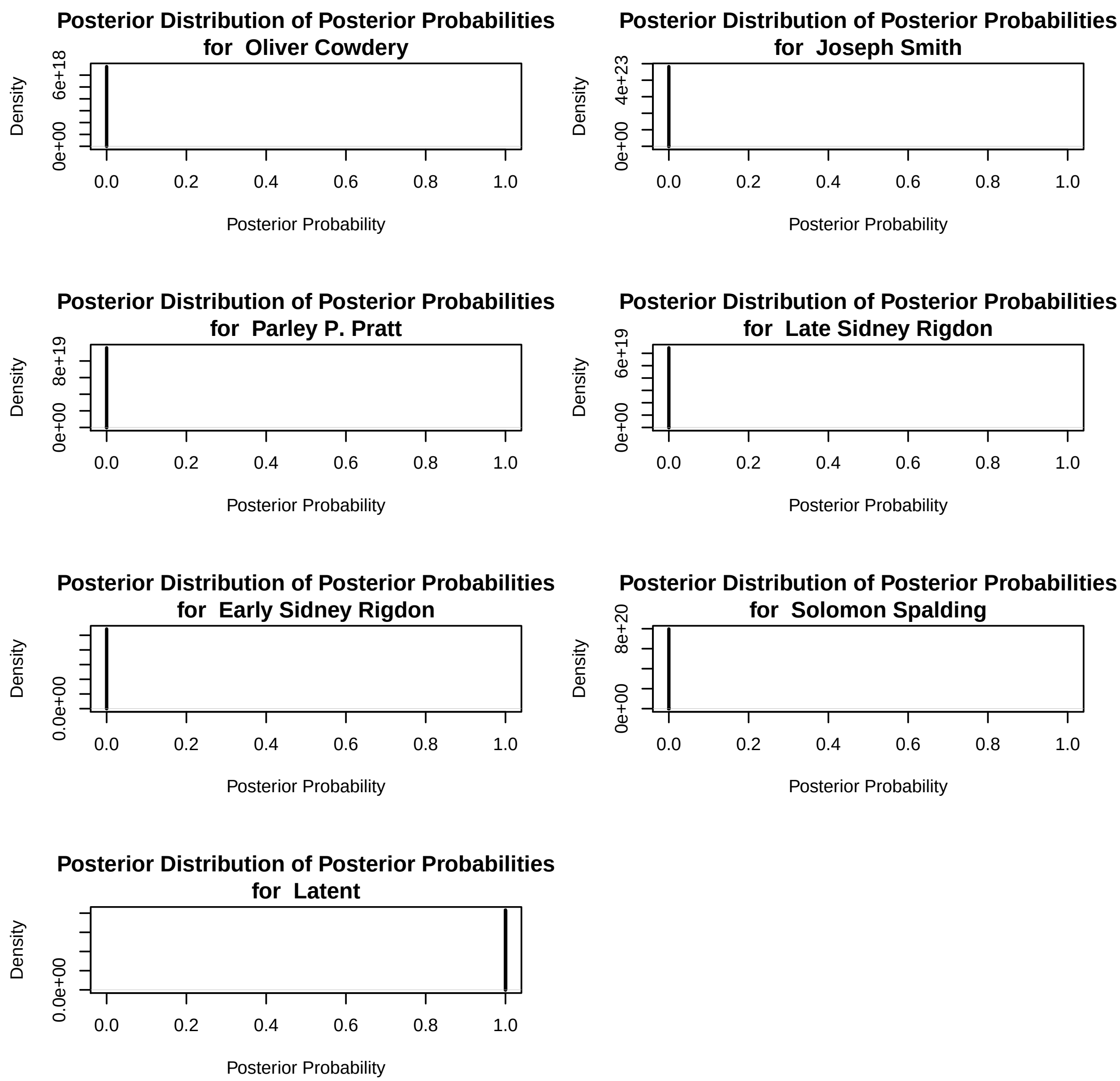
Posterior distribution of Posterior Probabilities for each authors; block no. 3



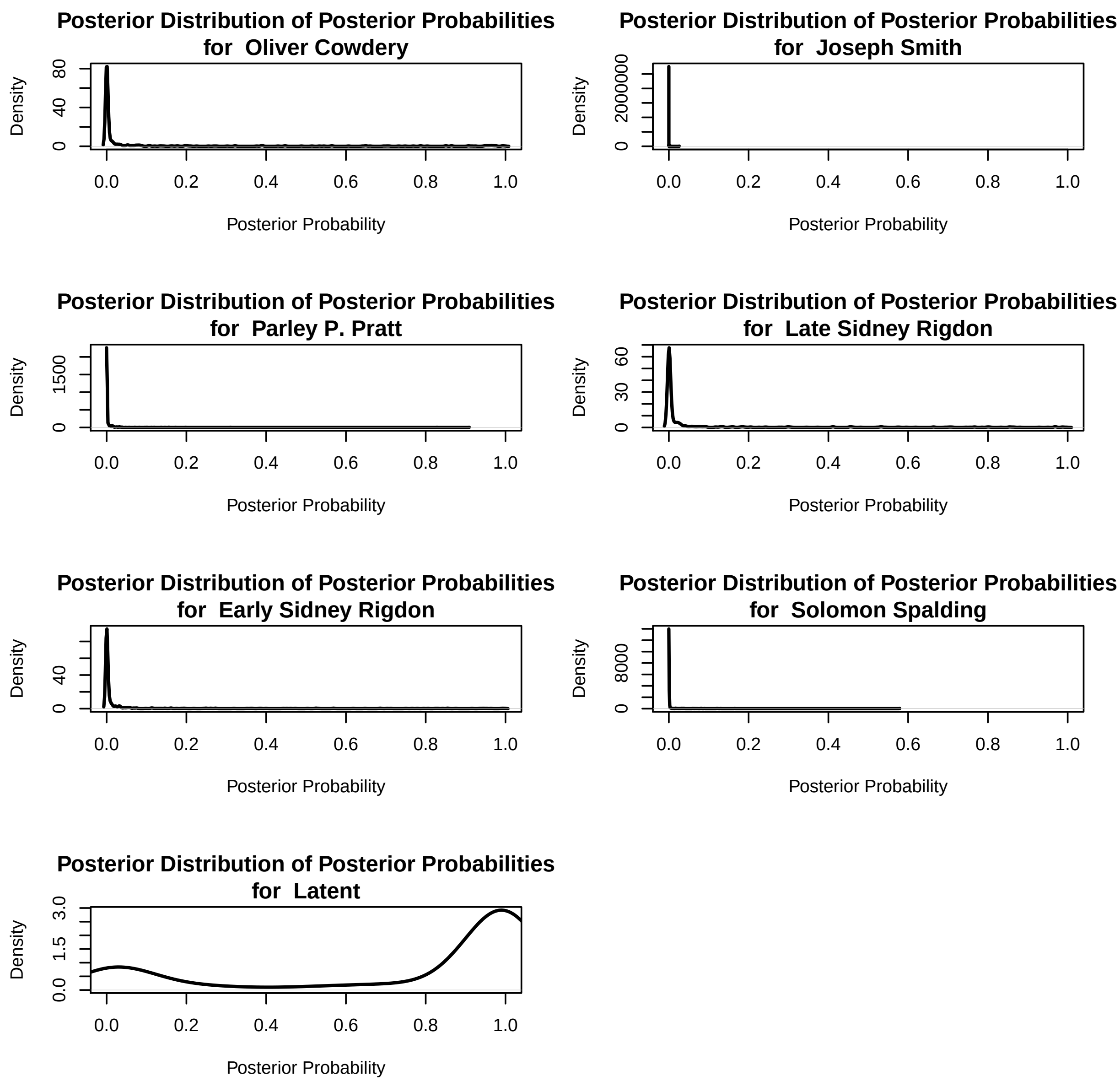
Posterior distribution of Posterior Probabilities for each authors; block no. 4



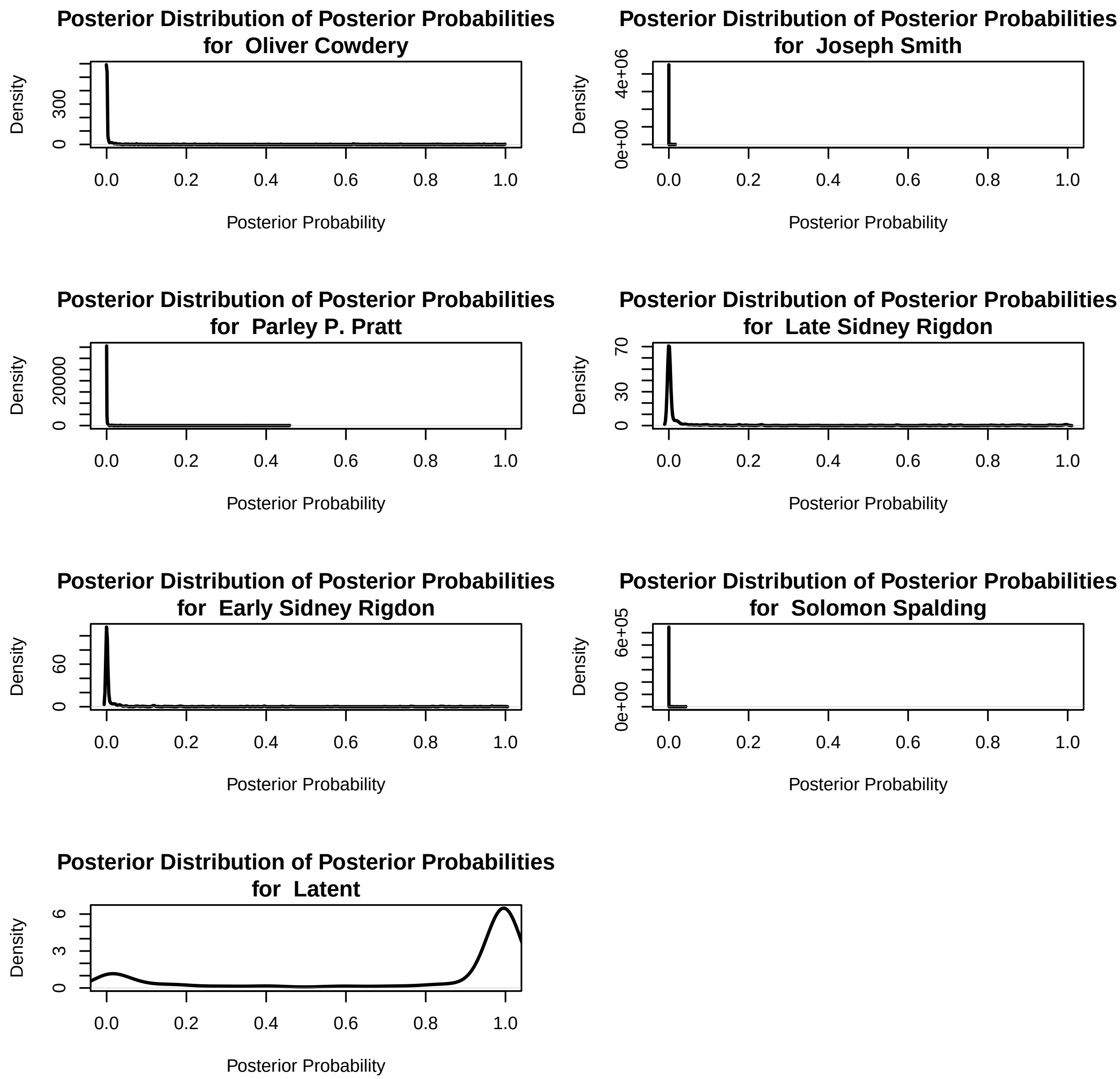
Posterior distribution of Posterior Probabilities for each authors; block no. 5



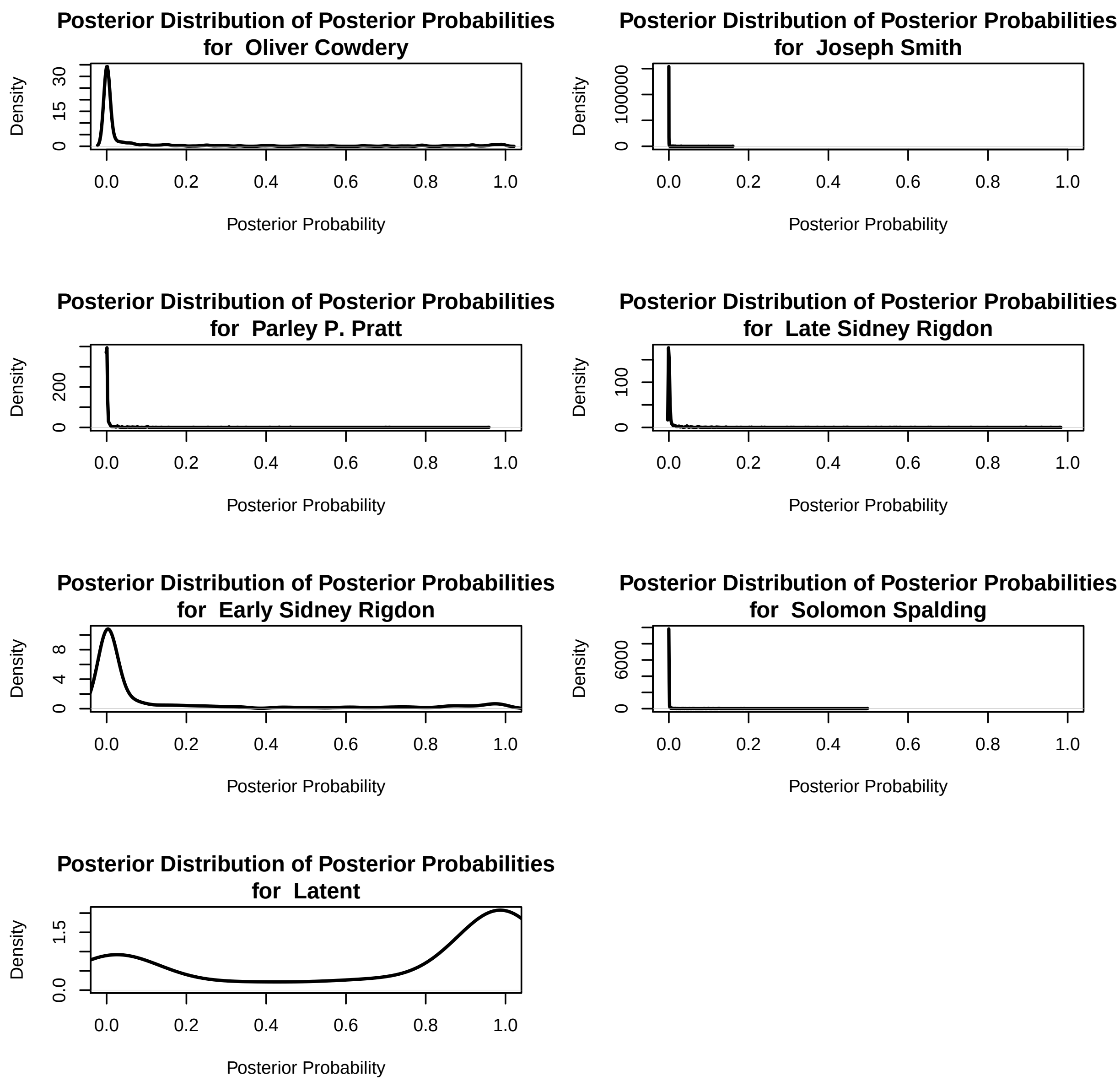
Posterior distribution of Posterior Probabilities for each authors; block no. 6



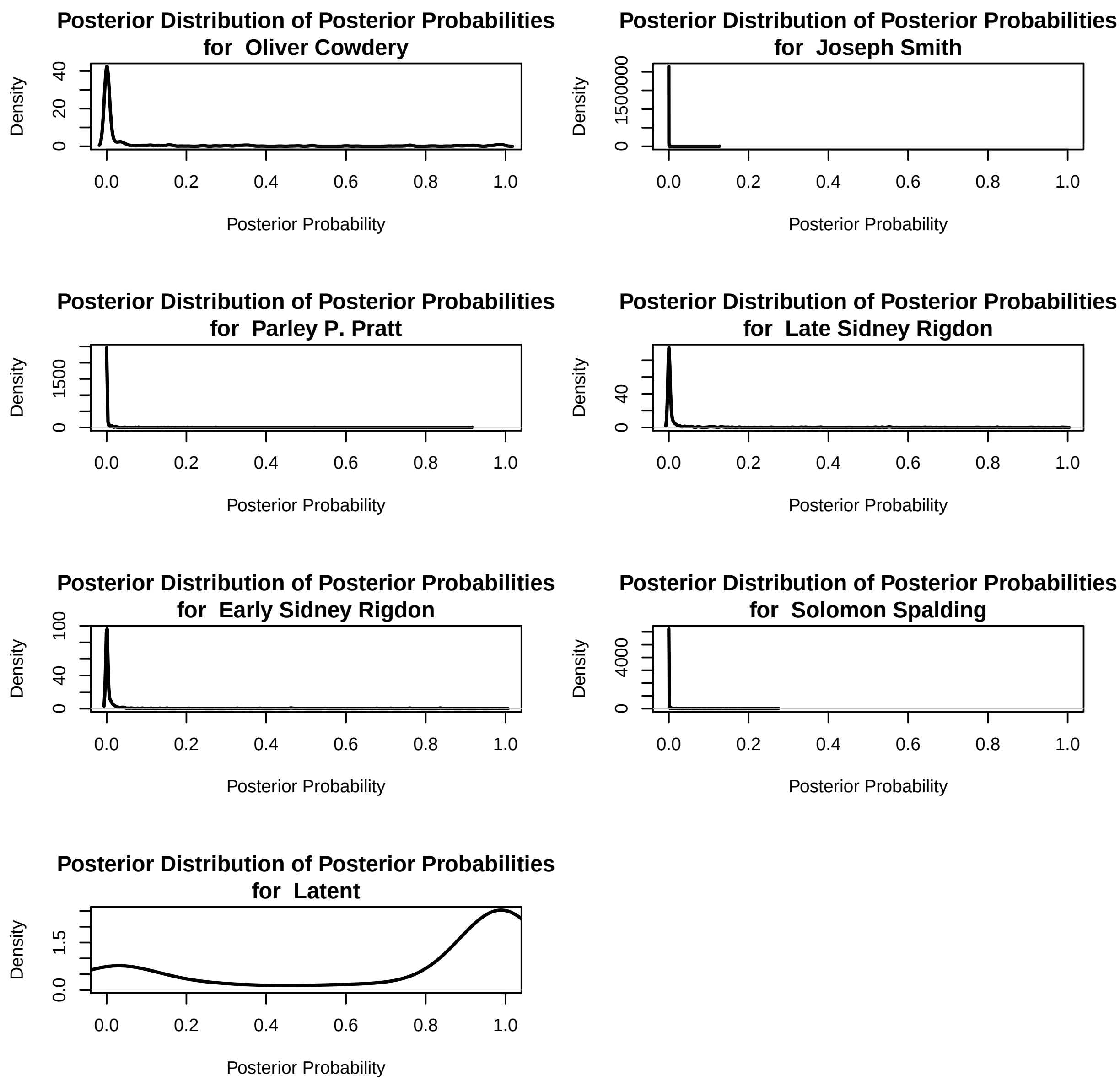
Posterior distribution of Posterior Probabilities for each authors; block no. 7



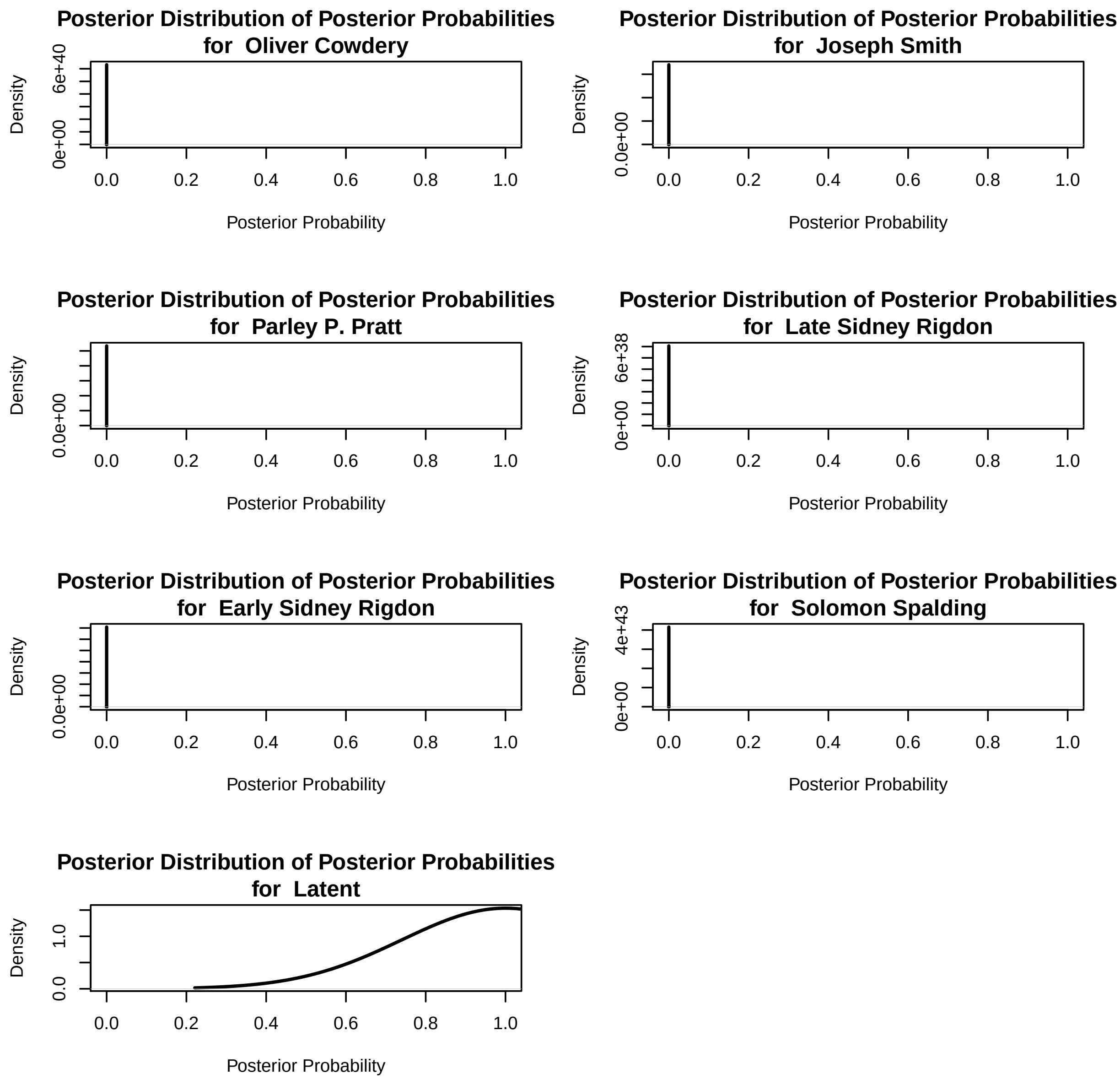
Posterior distribution of Posterior Probabilities for each authors; block no. 8



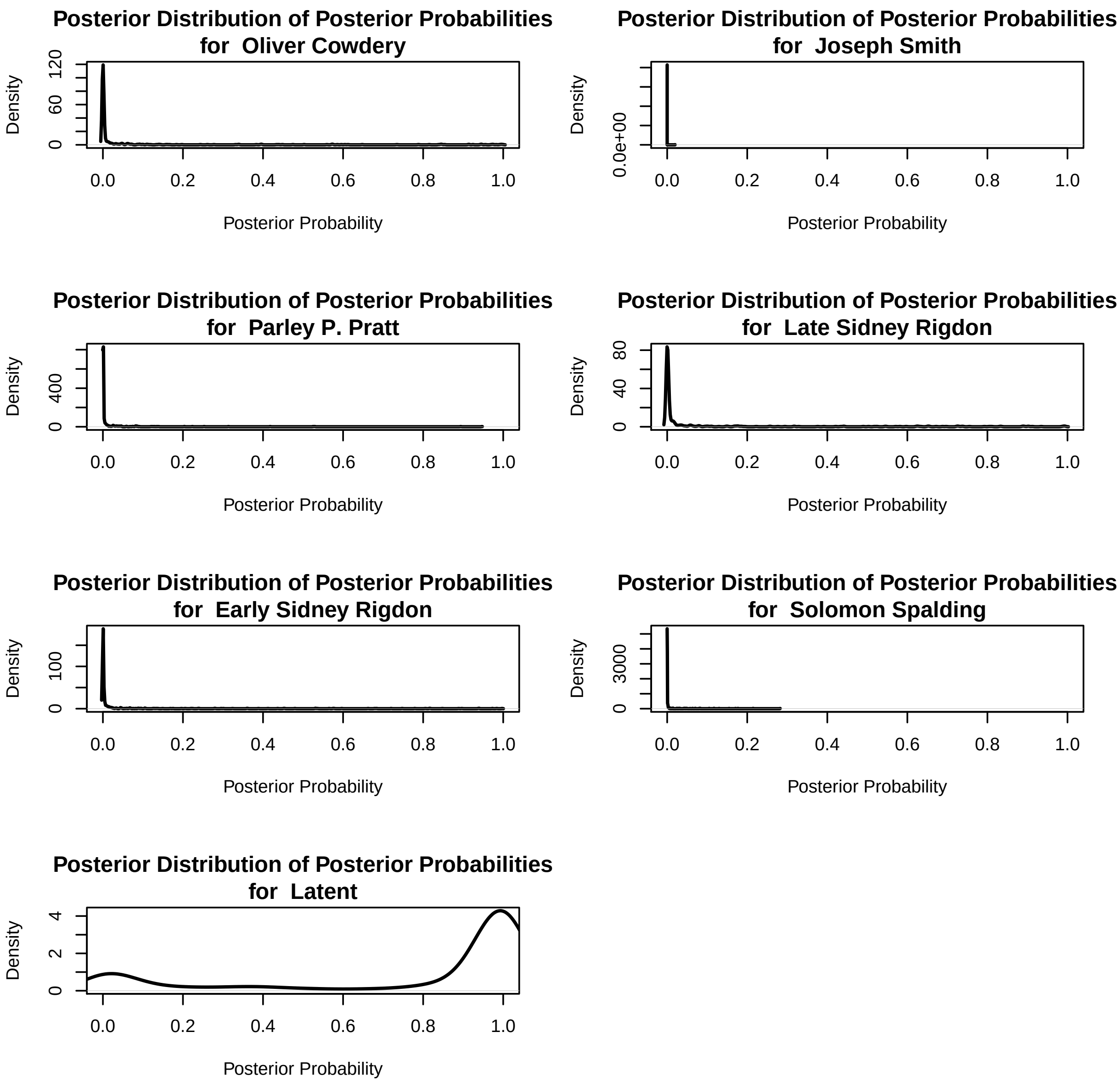
Posterior distribution of Posterior Probabilities for each authors; block no. 9



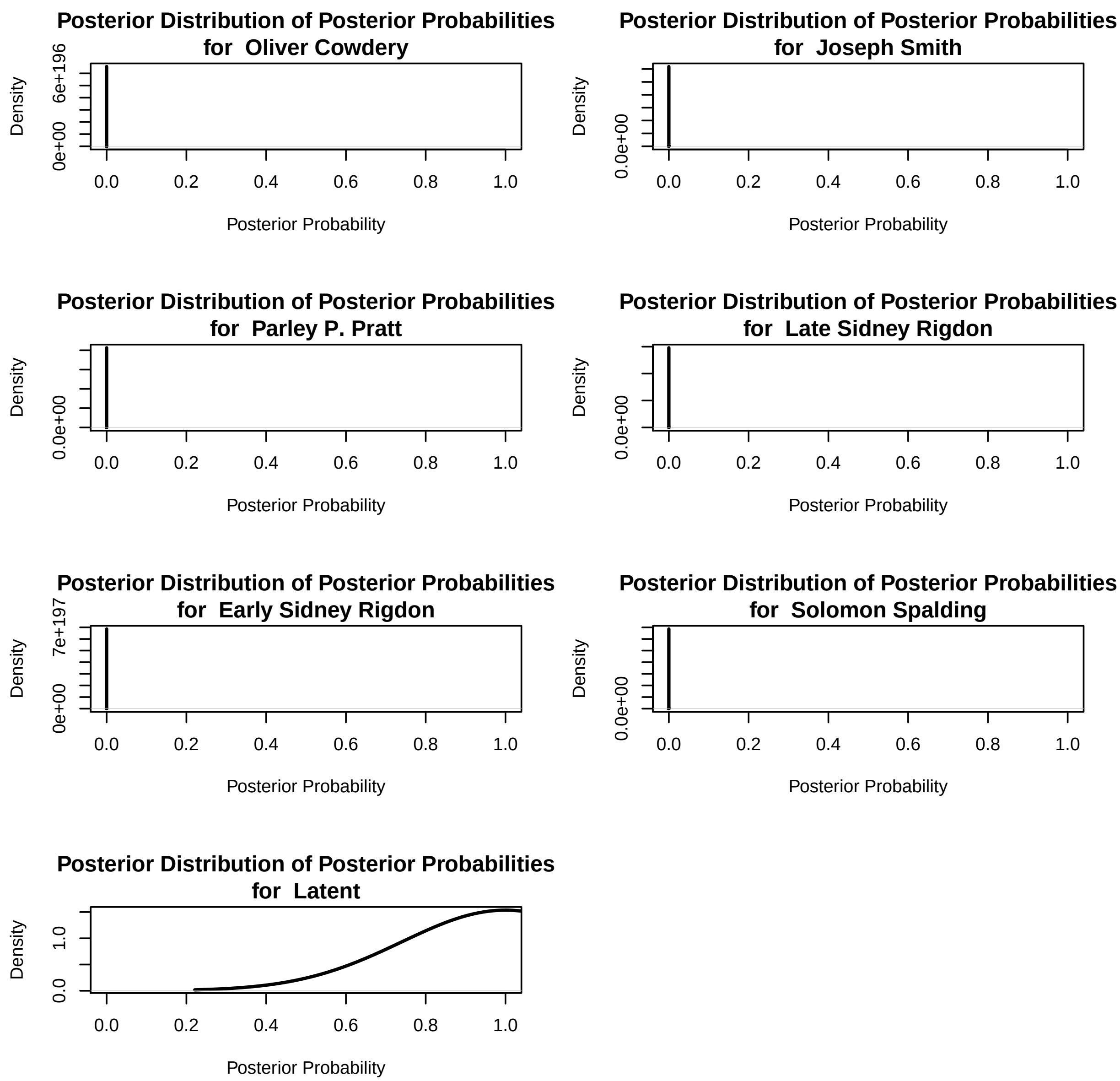
Posterior distribution of Posterior Probabilities for each authors; block no. 10



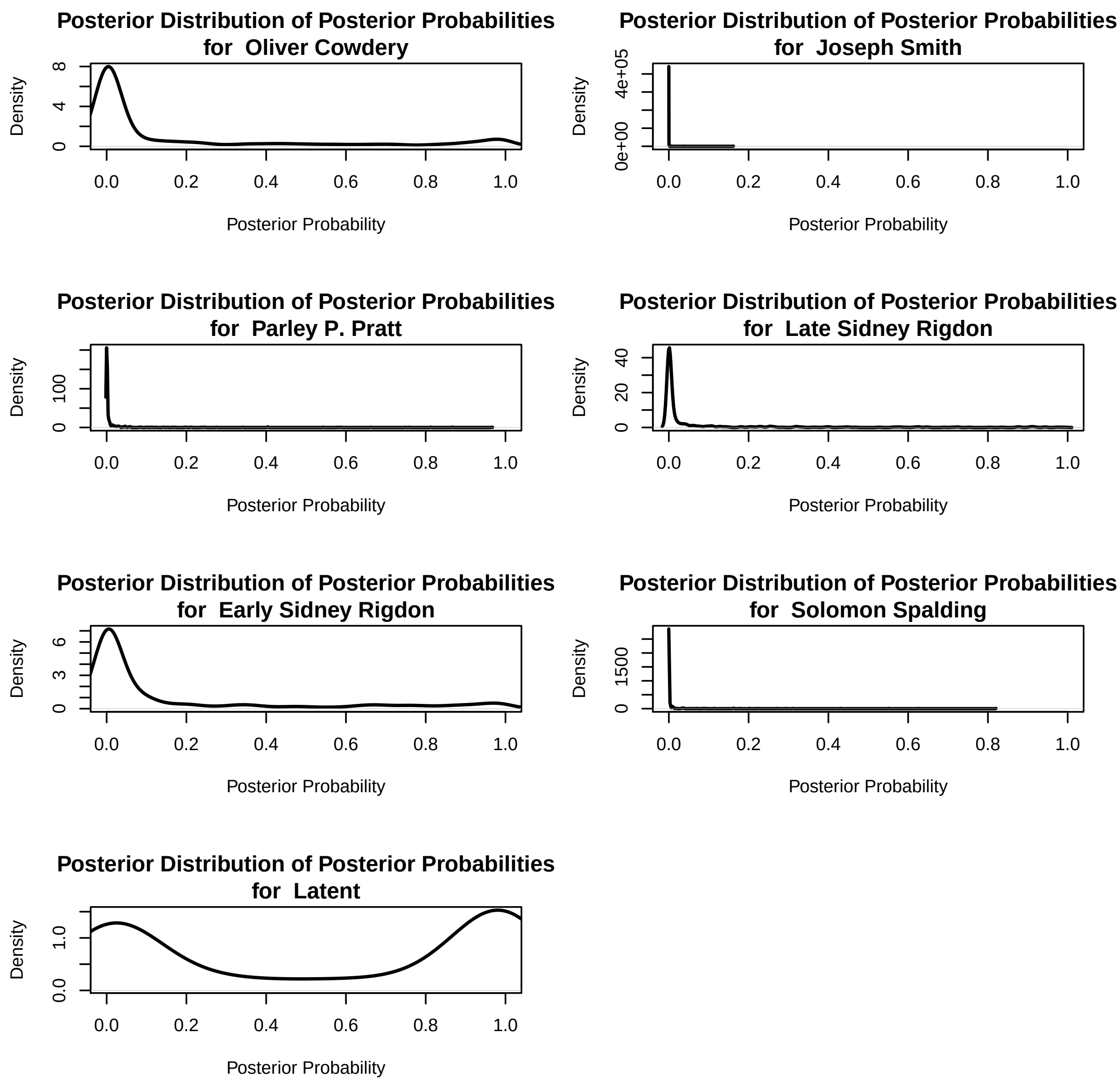
Posterior distribution of Posterior Probabilities for each authors; block no. 11



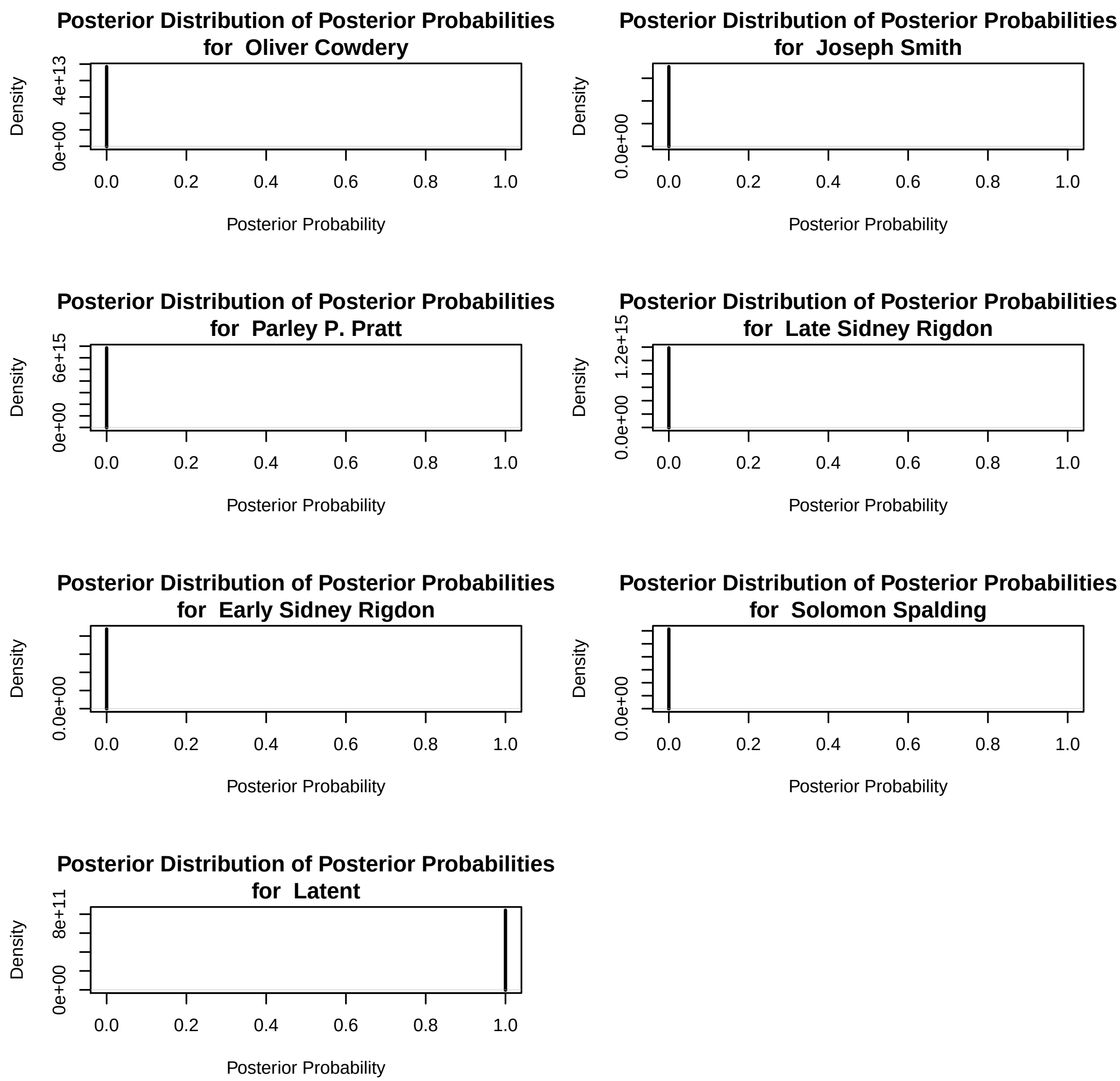
Posterior distribution of Posterior Probabilities for each authors; block no. 12



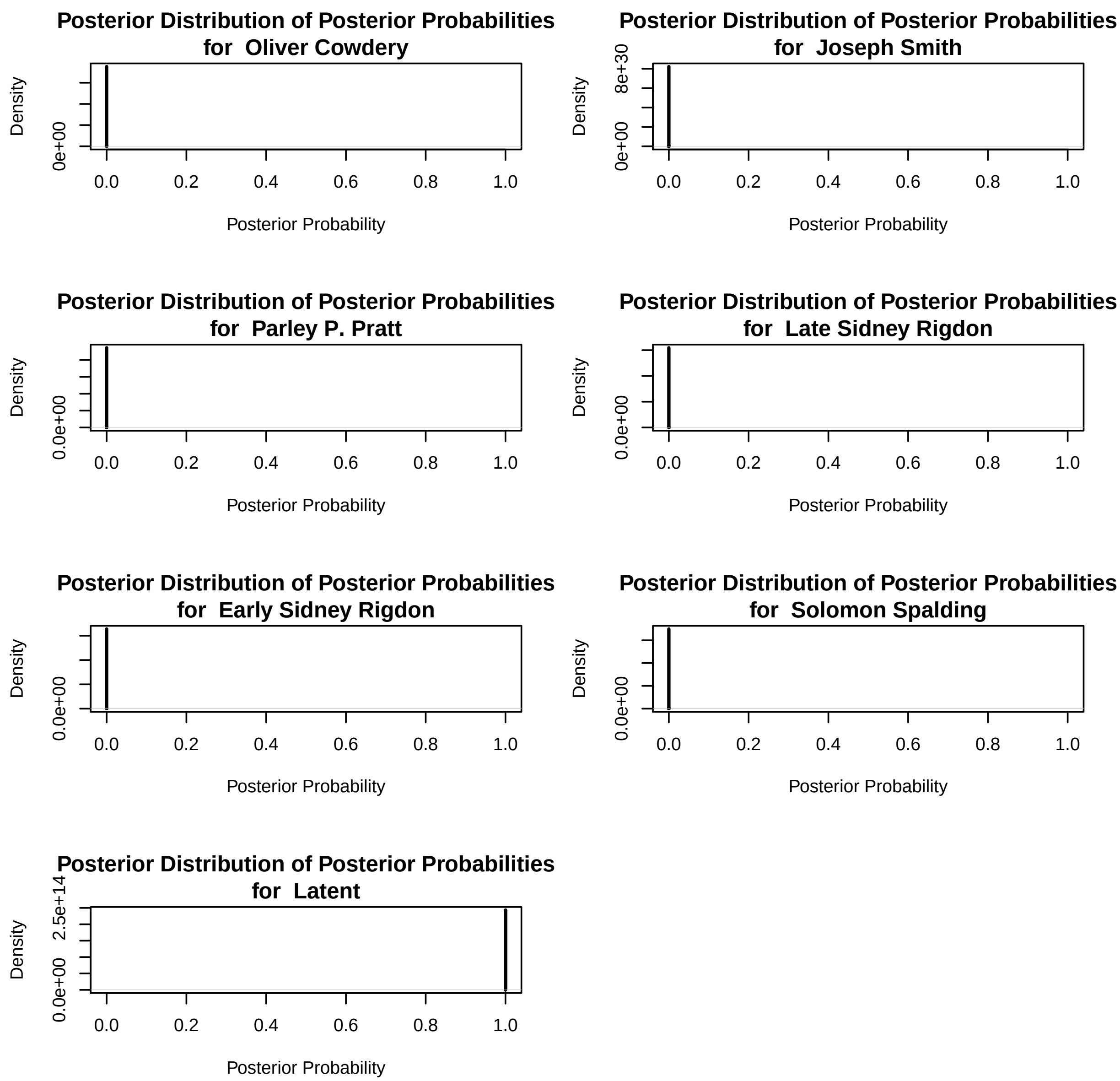
Posterior distribution of Posterior Probabilities for each authors; block no. 13



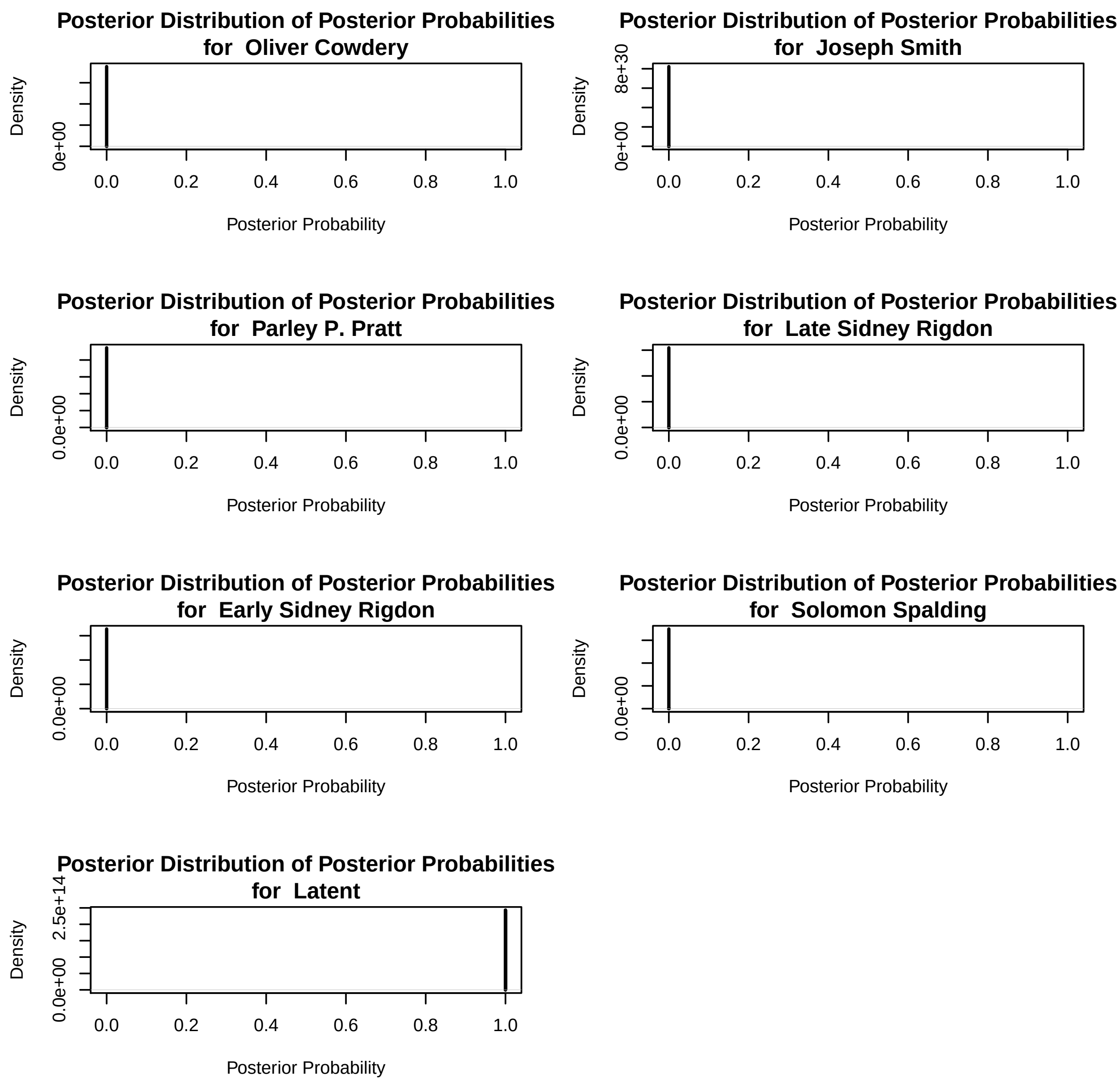
Posterior distribution of Posterior Probabilities for each authors; block no. 14



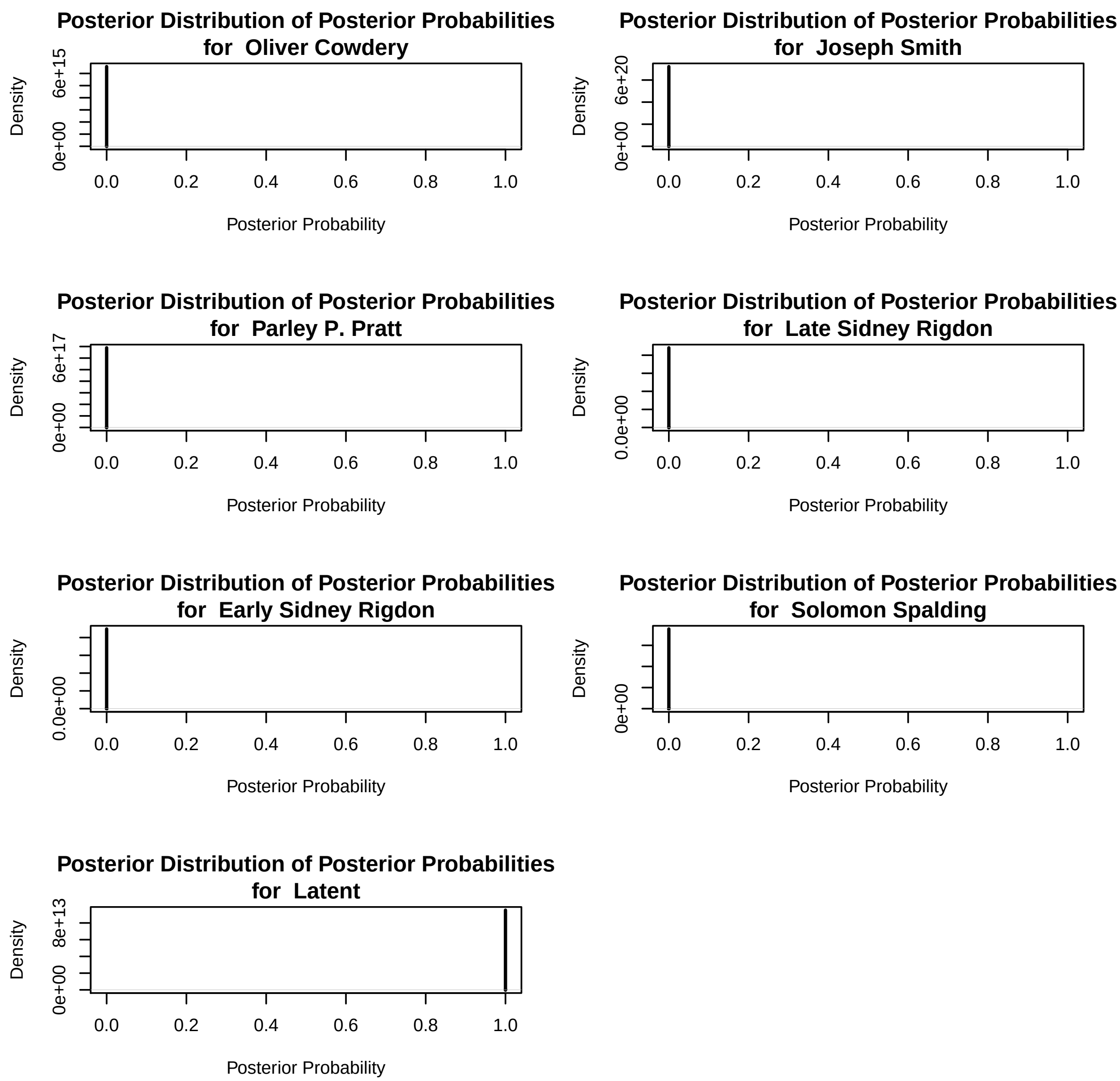
Posterior distribution of Posterior Probabilities for each authors; block no. 15



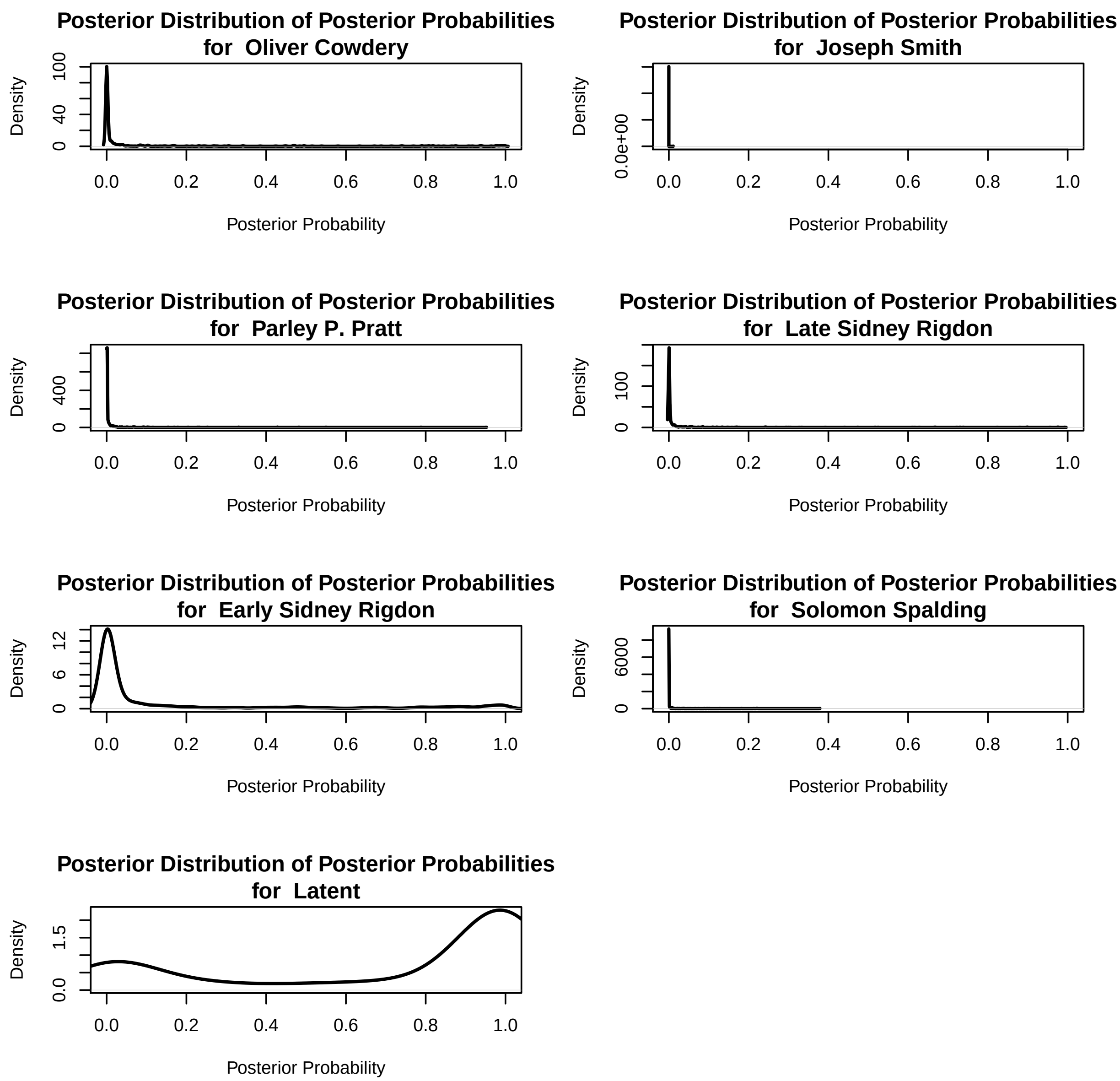
Posterior distribution of Posterior Probabilities for each authors; block no. 16



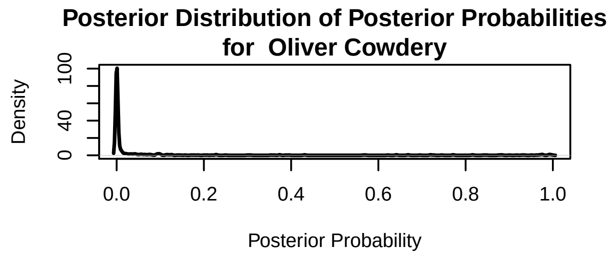
Posterior distribution of Posterior Probabilities for each authors; block no. 17

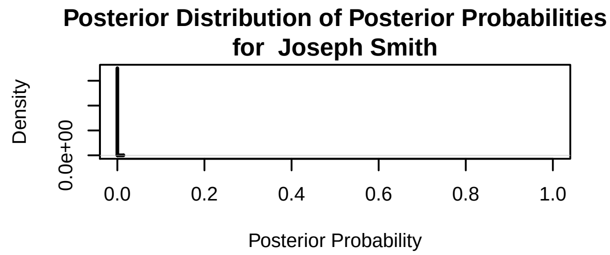


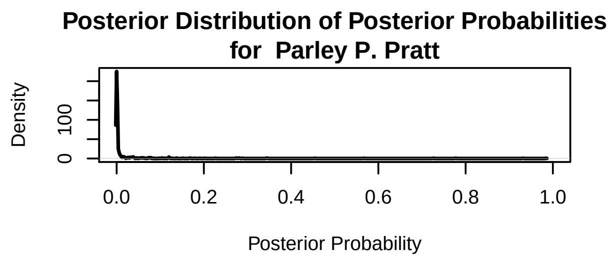
Posterior distribution of Posterior Probabilities for each authors; block no. 18

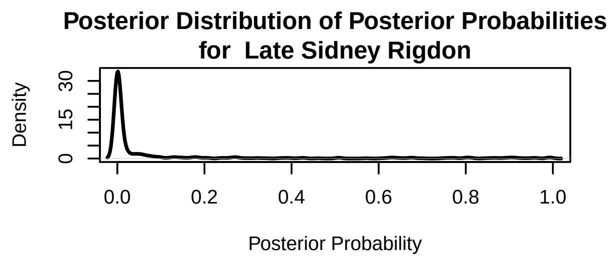


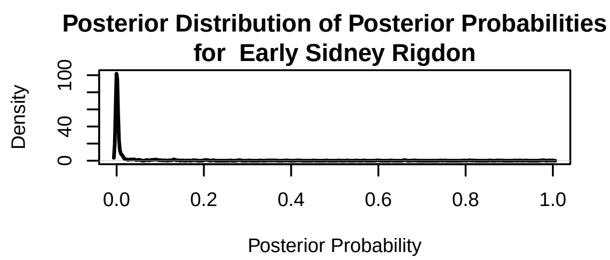
Posterior distribution of Posterior Probabilities for each authors; block no. 19

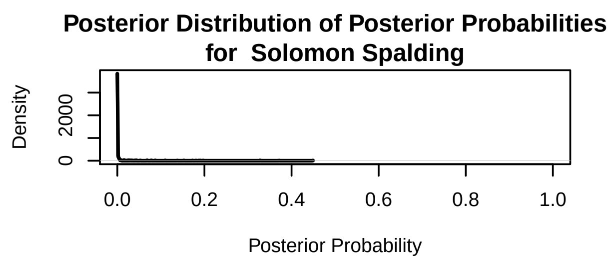


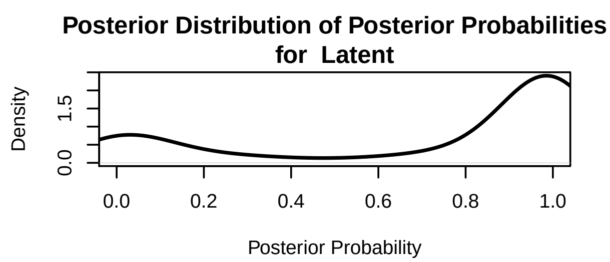




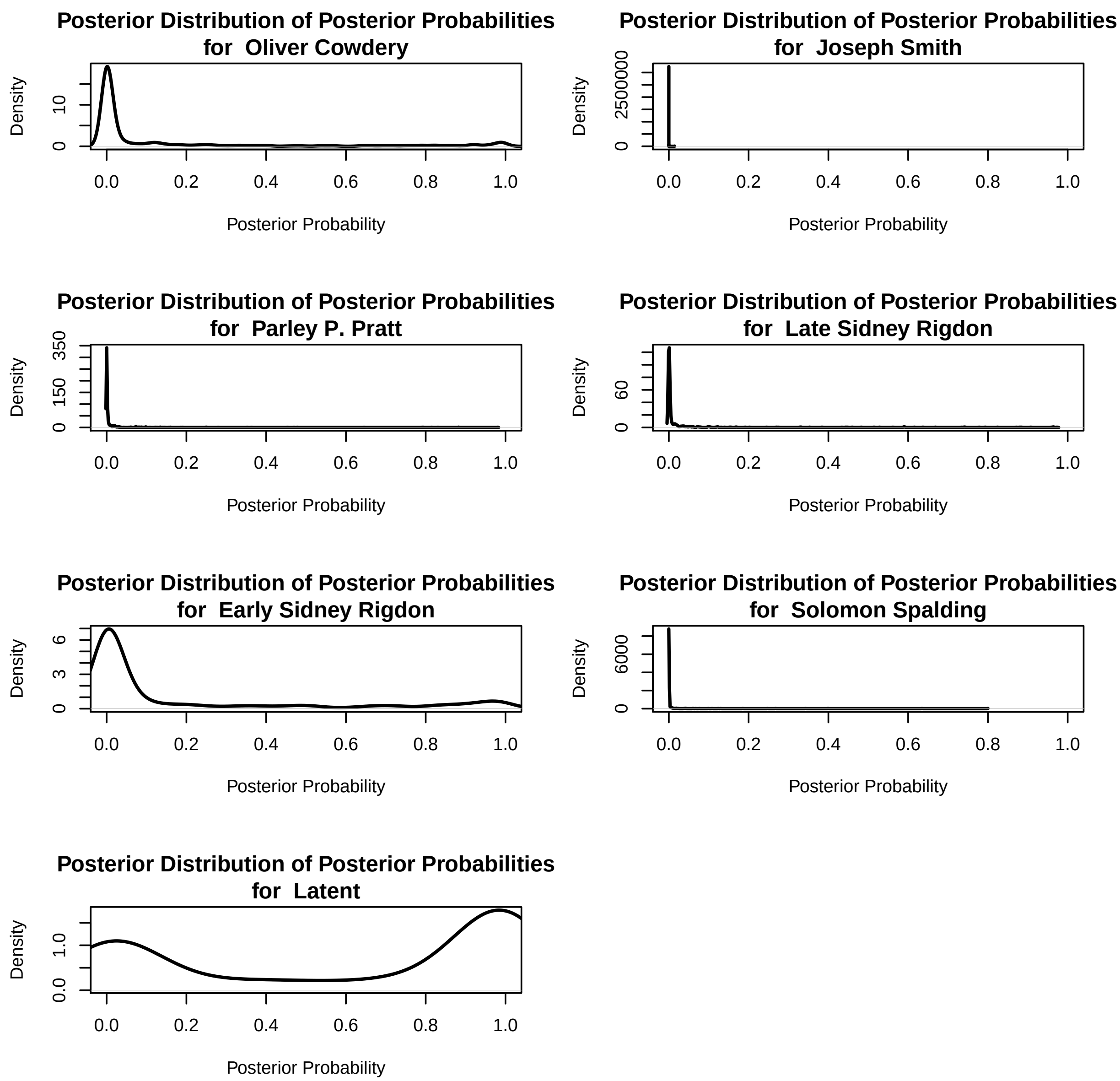




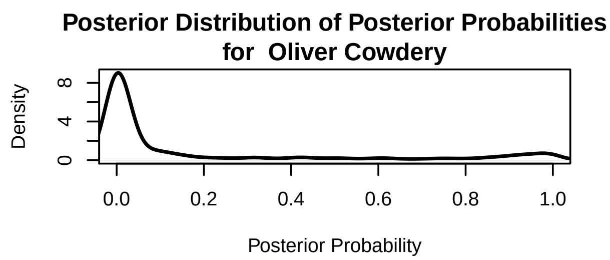


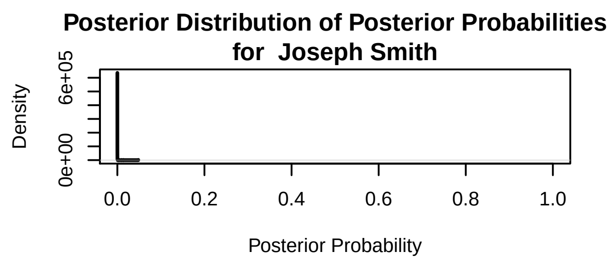


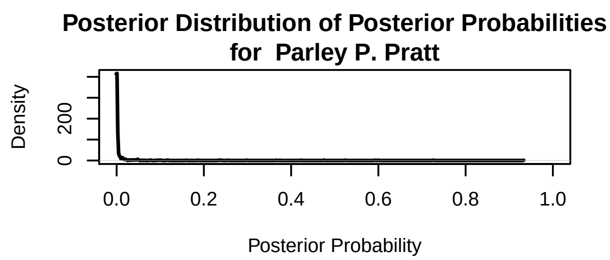
Posterior distribution of Posterior Probabilities for each authors; block no. 20

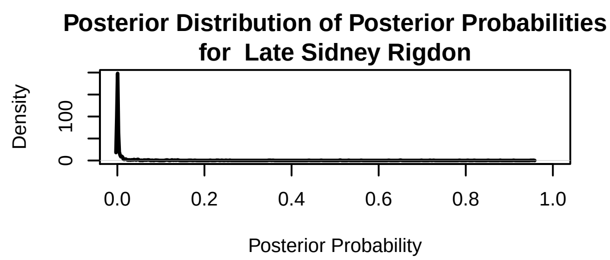


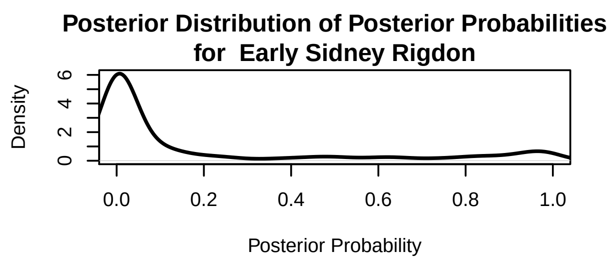
Posterior distribution of Posterior Probabilities for each authors; block no. 21

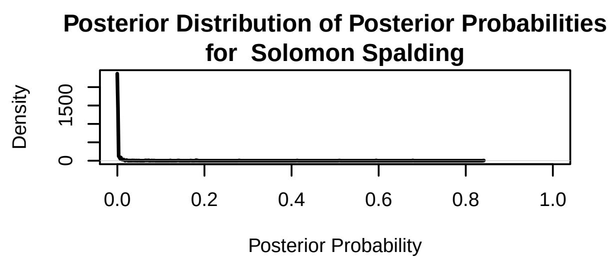


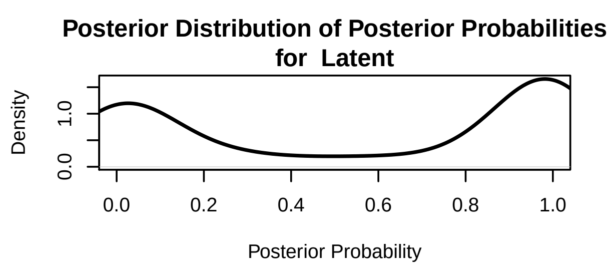




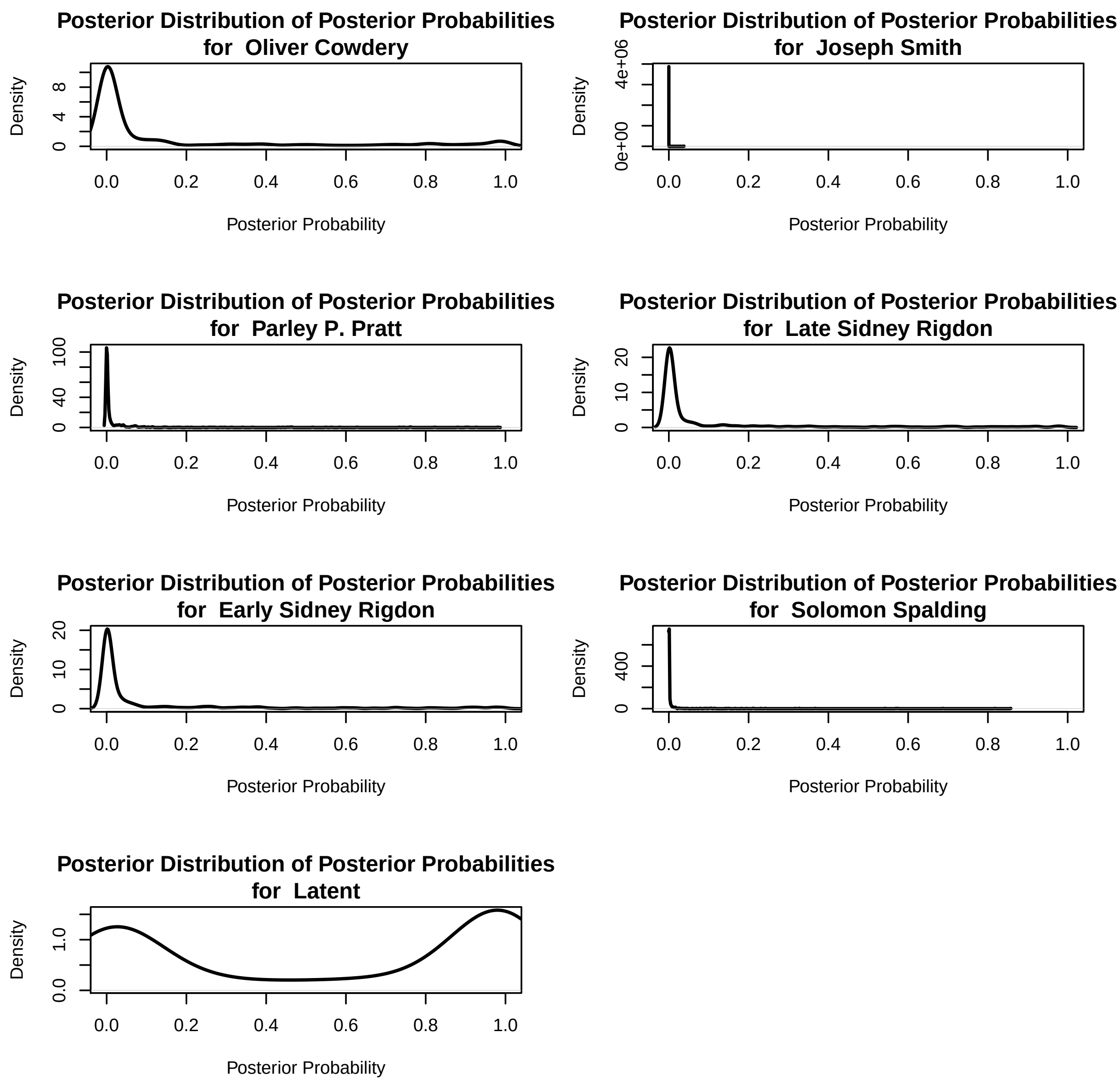




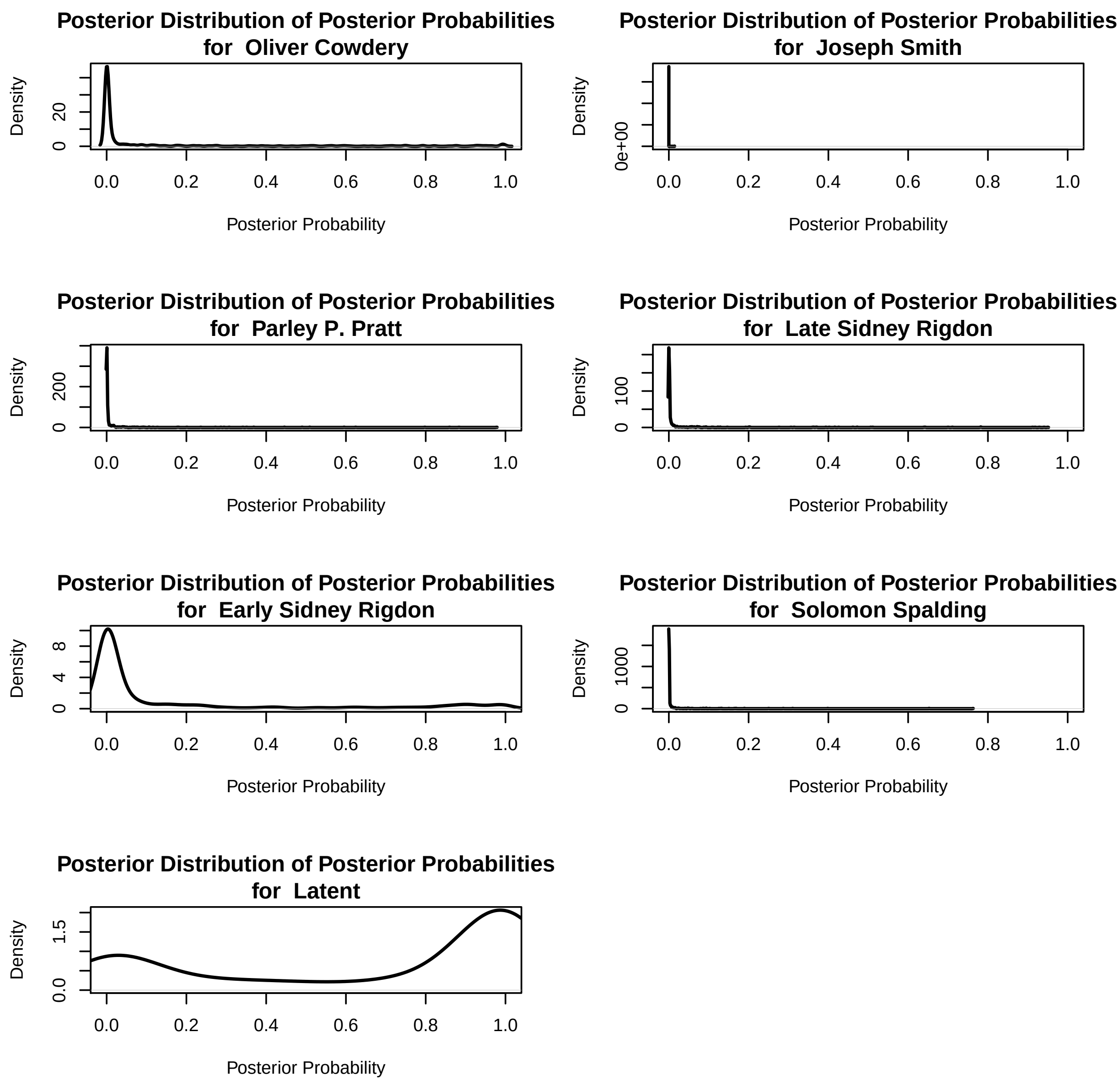




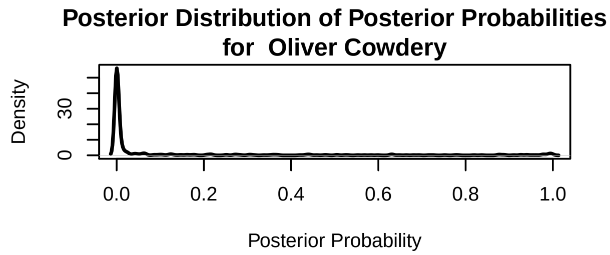
Posterior distribution of Posterior Probabilities for each authors; block no. 22

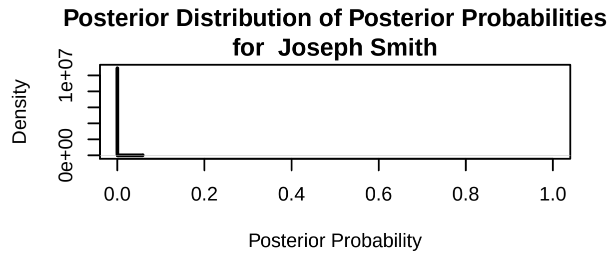


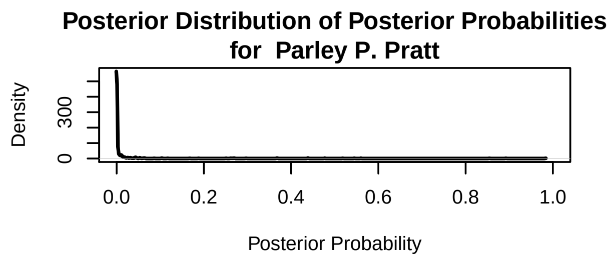
Posterior distribution of Posterior Probabilities for each authors; block no. 23

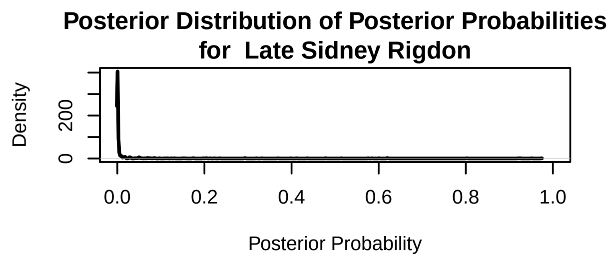


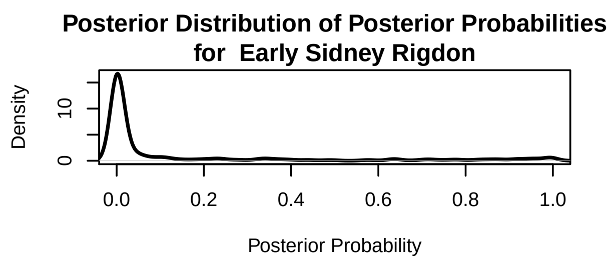
Posterior distribution of Posterior Probabilities for each authors; block no. 24

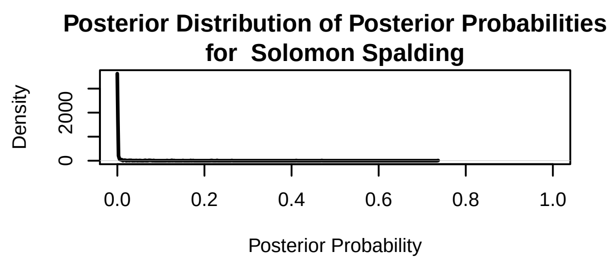


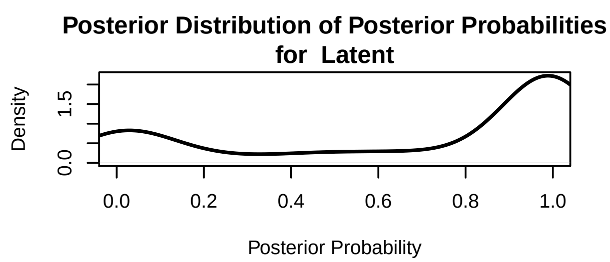




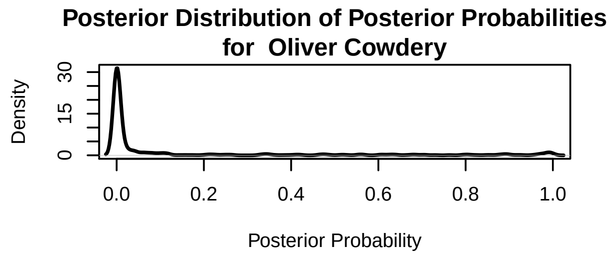


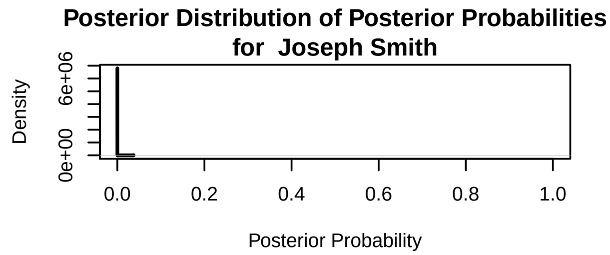


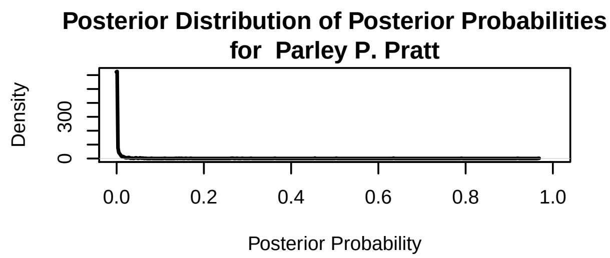


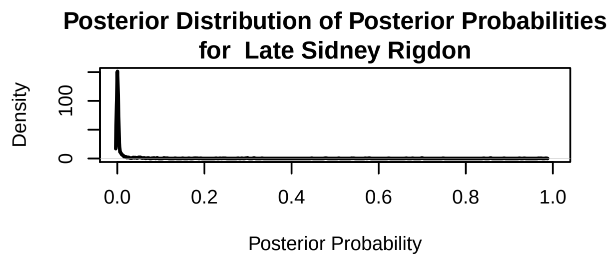


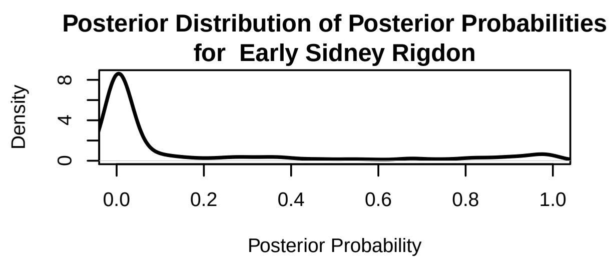
Posterior distribution of Posterior Probabilities for each authors; block no. 25

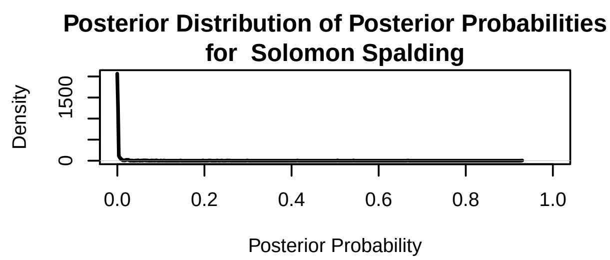


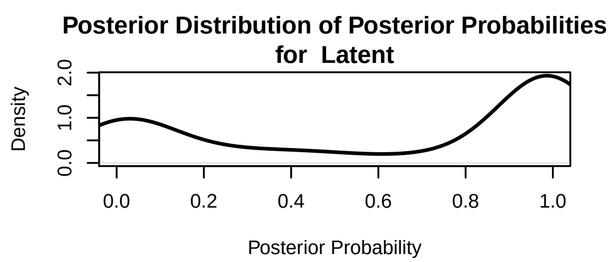




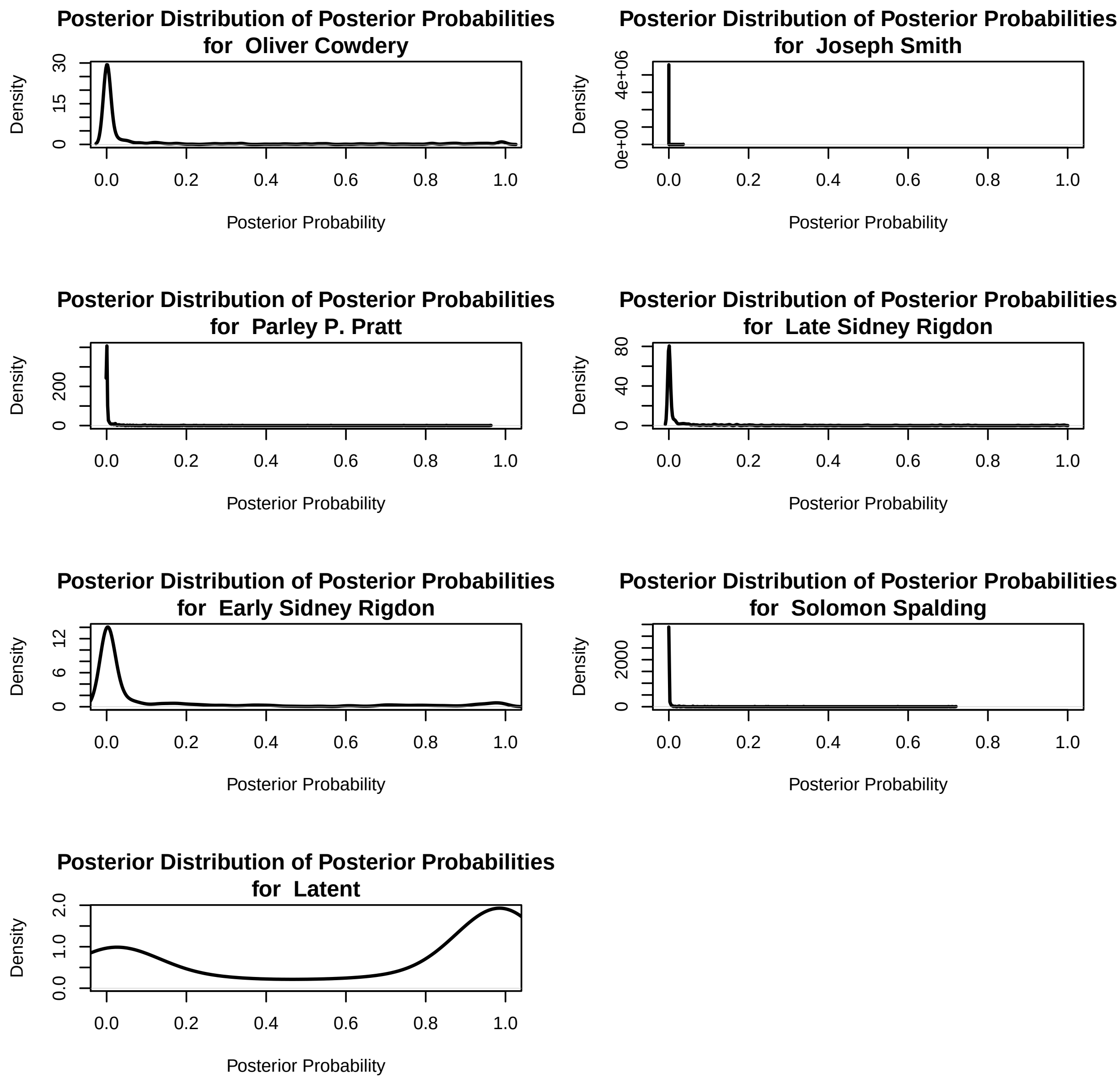




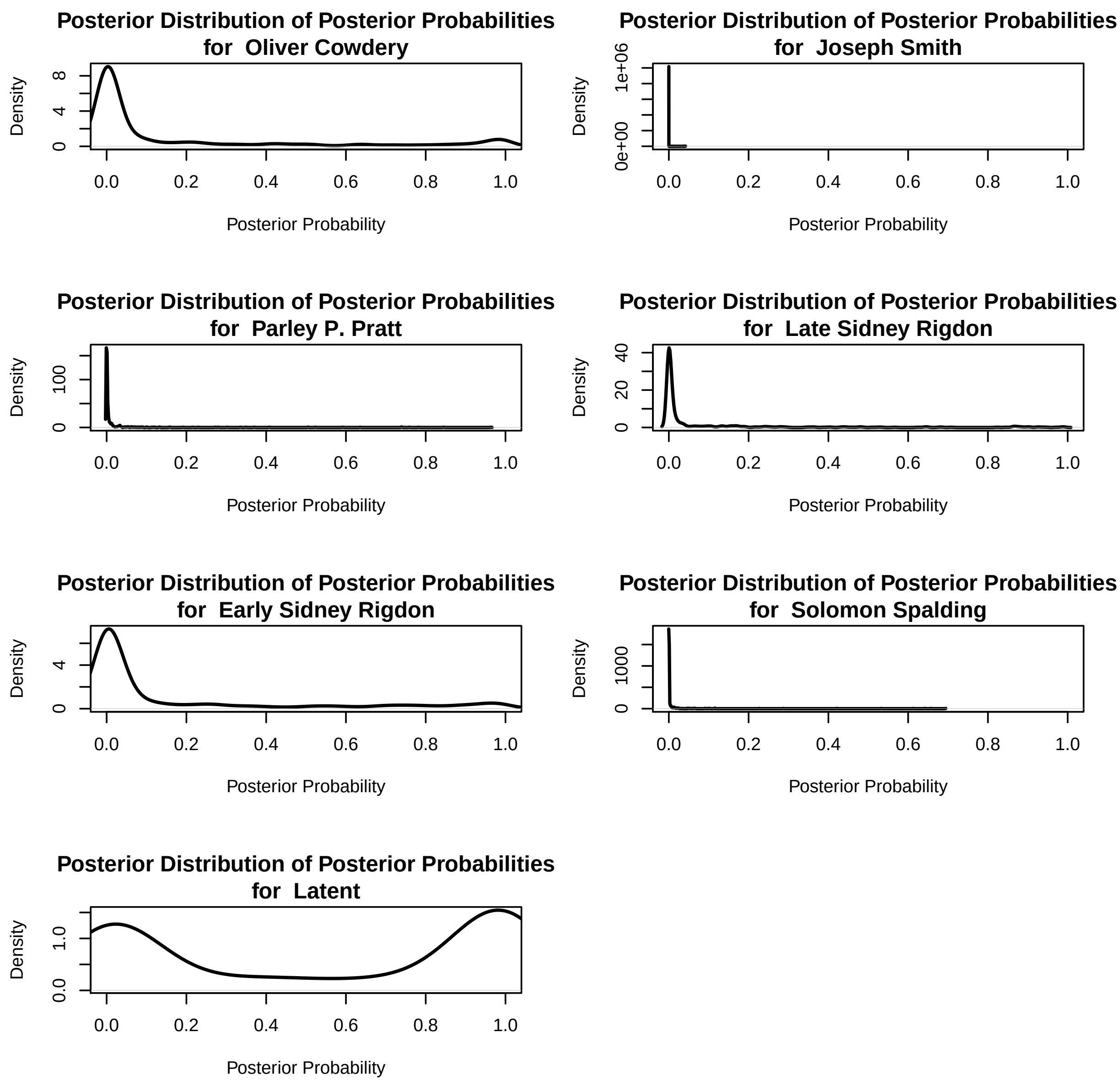




Posterior distribution of Posterior Probabilities for each authors; block no. 26

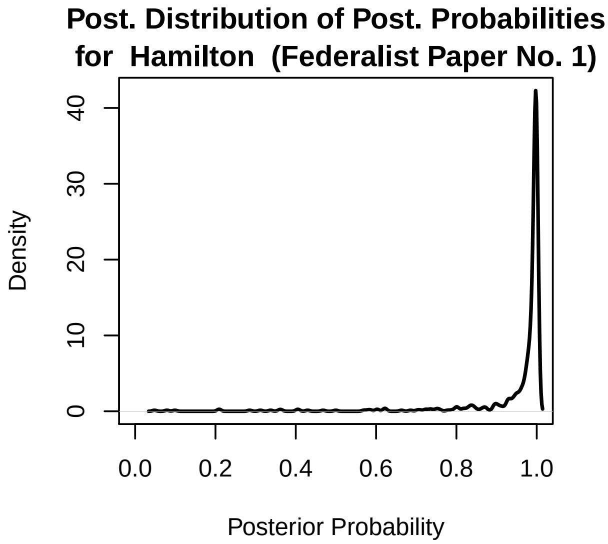


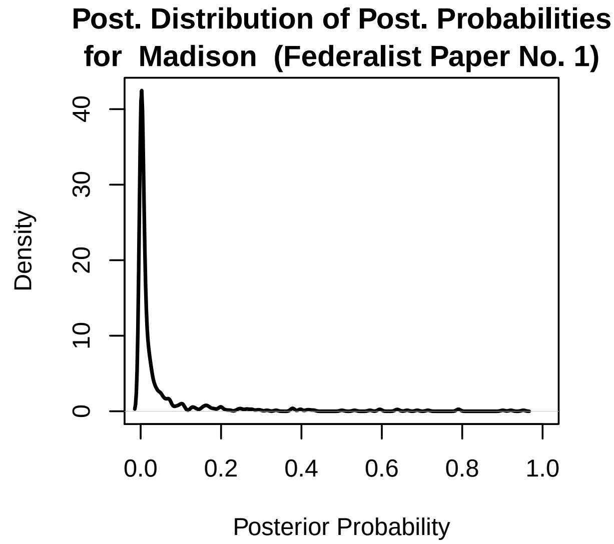
Posterior distribution of Posterior Probabilities for each authors; block no. 27

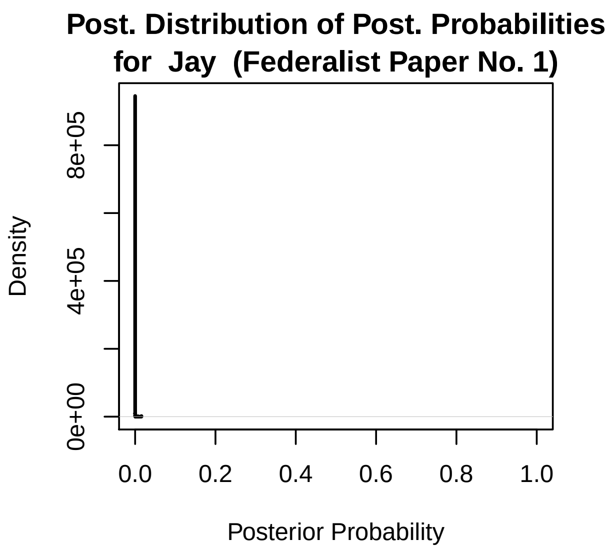


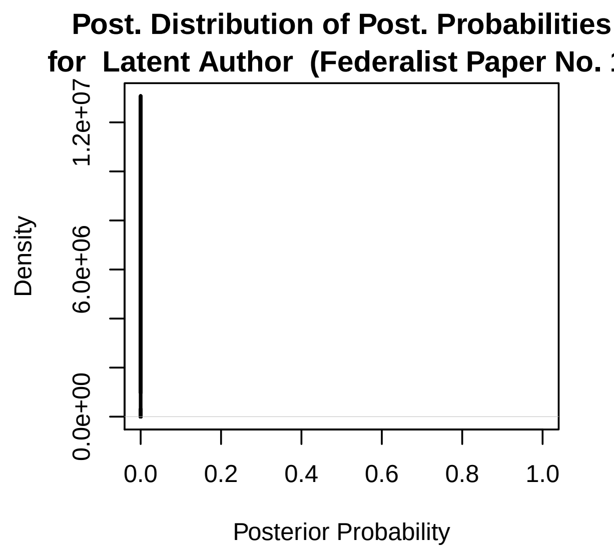
B.4 Hamilton Federalist Papers with Hamilton, Madison and Jay in the training set

Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 1

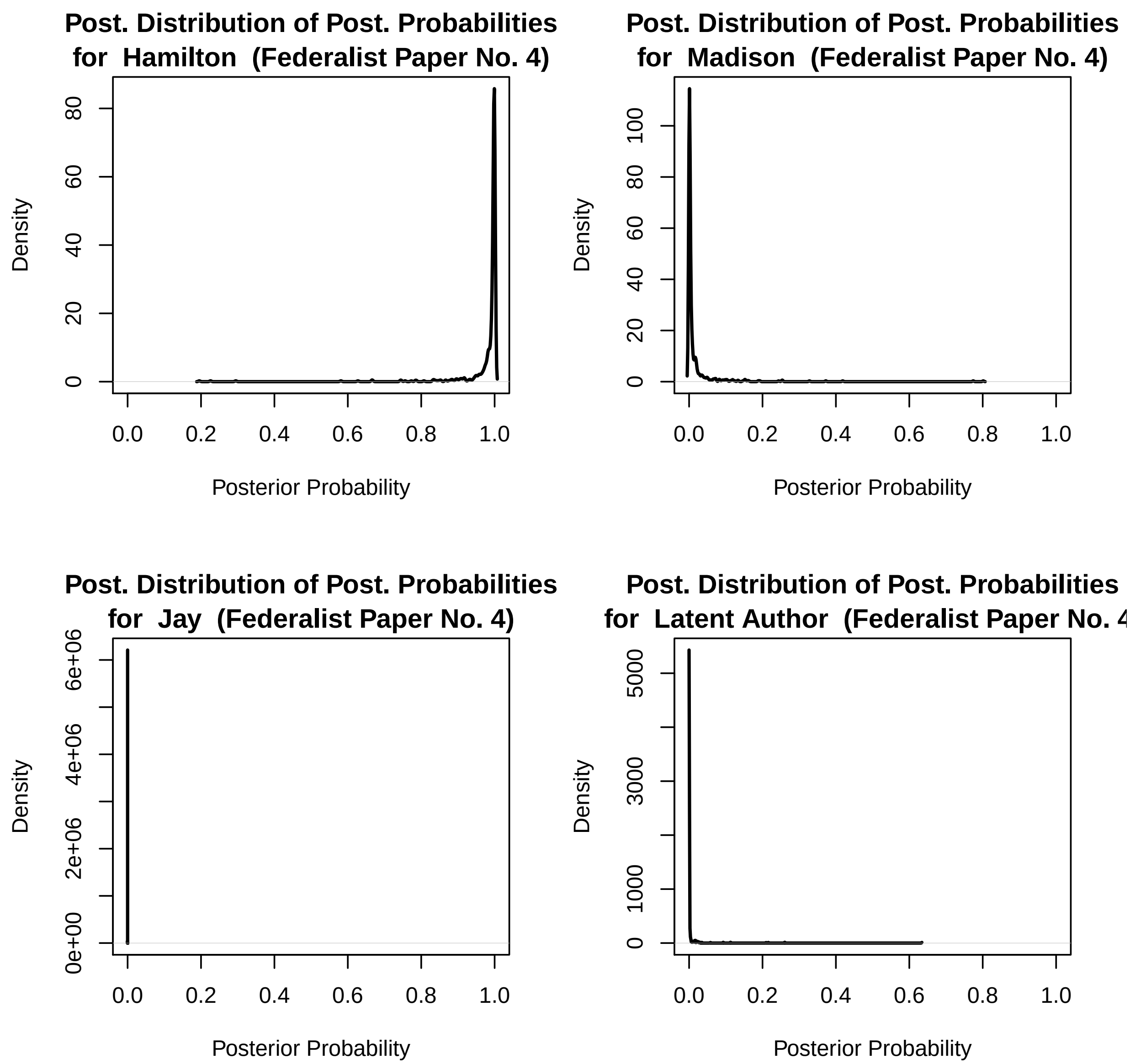




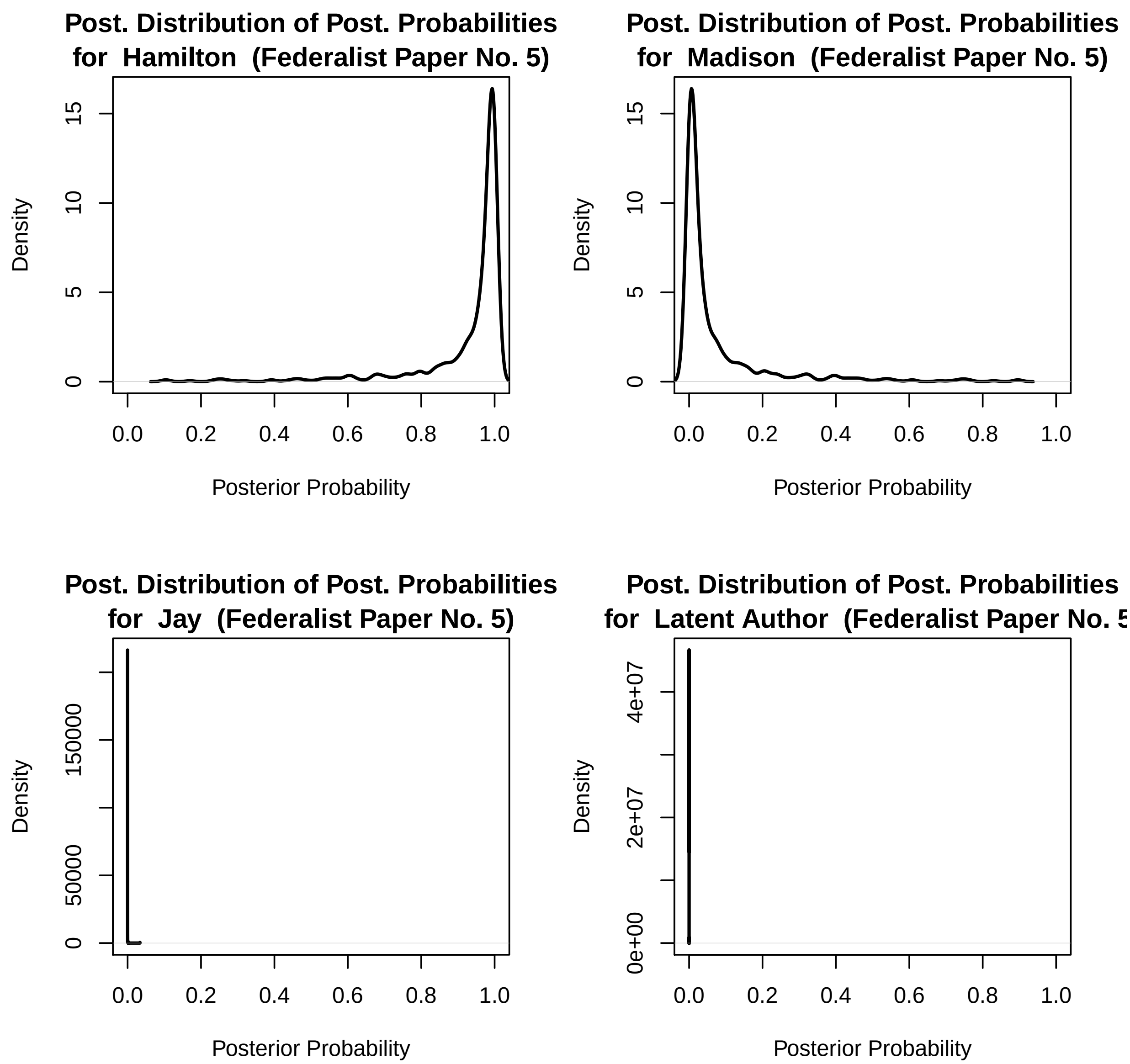




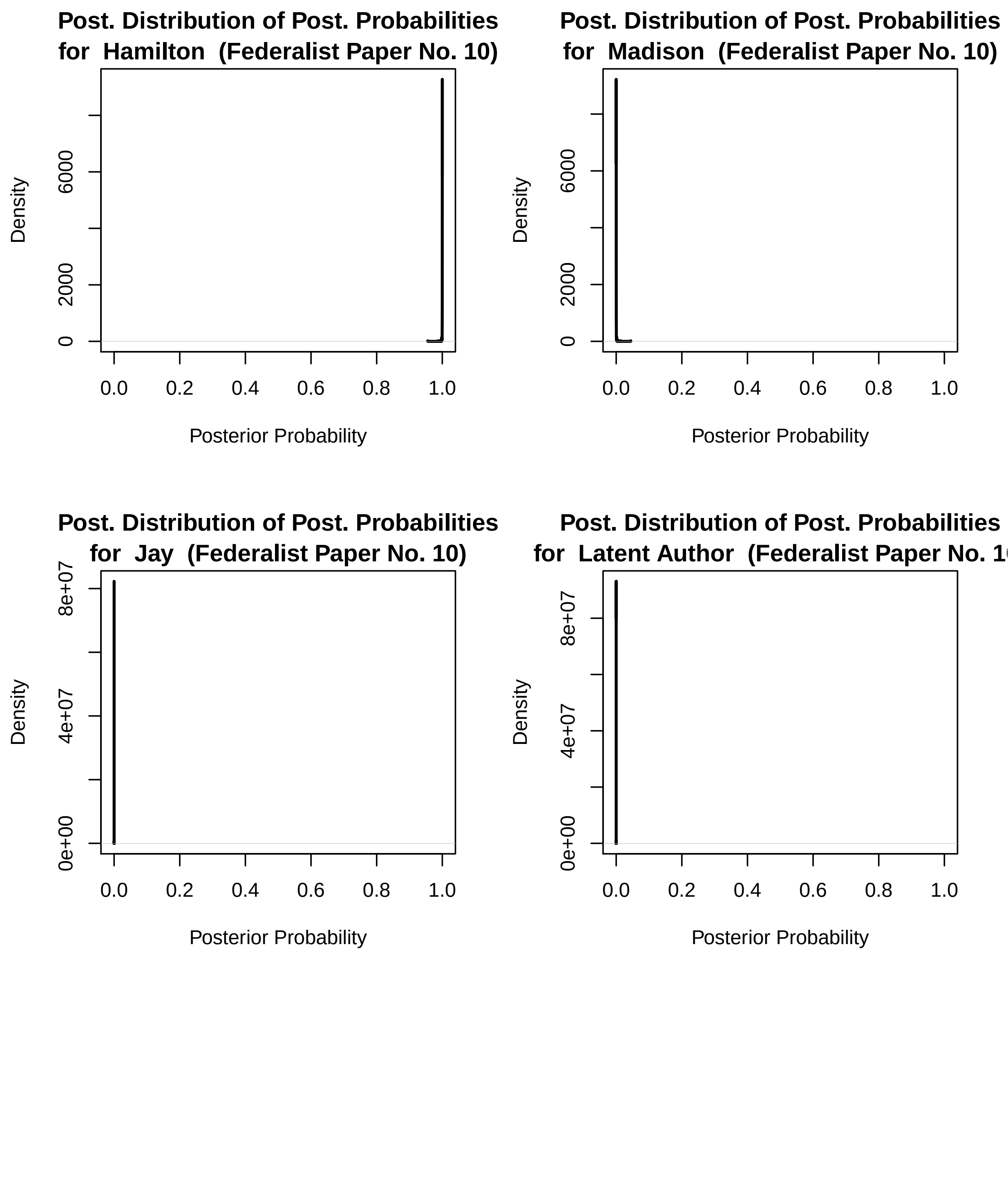
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 4



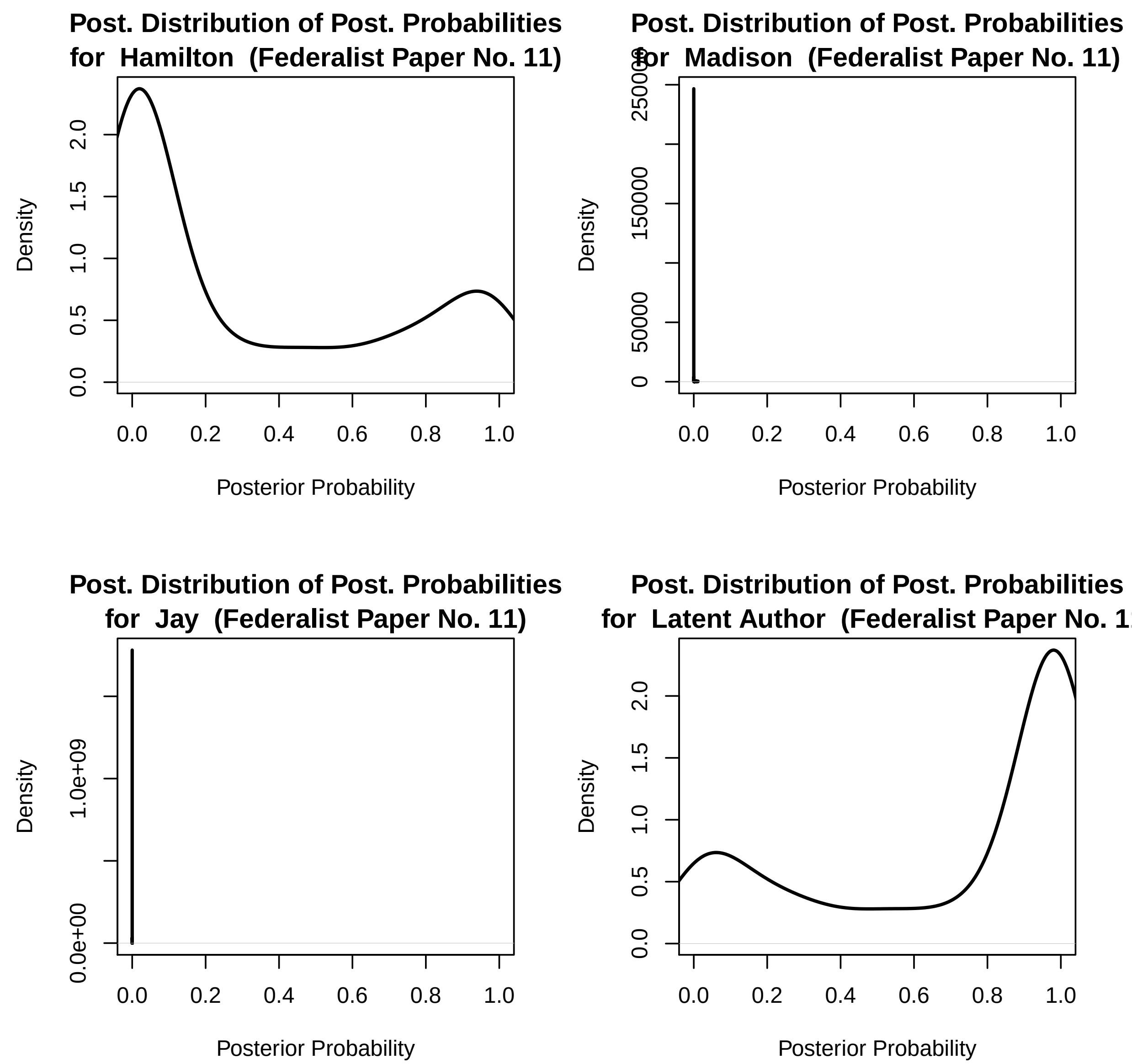
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 5



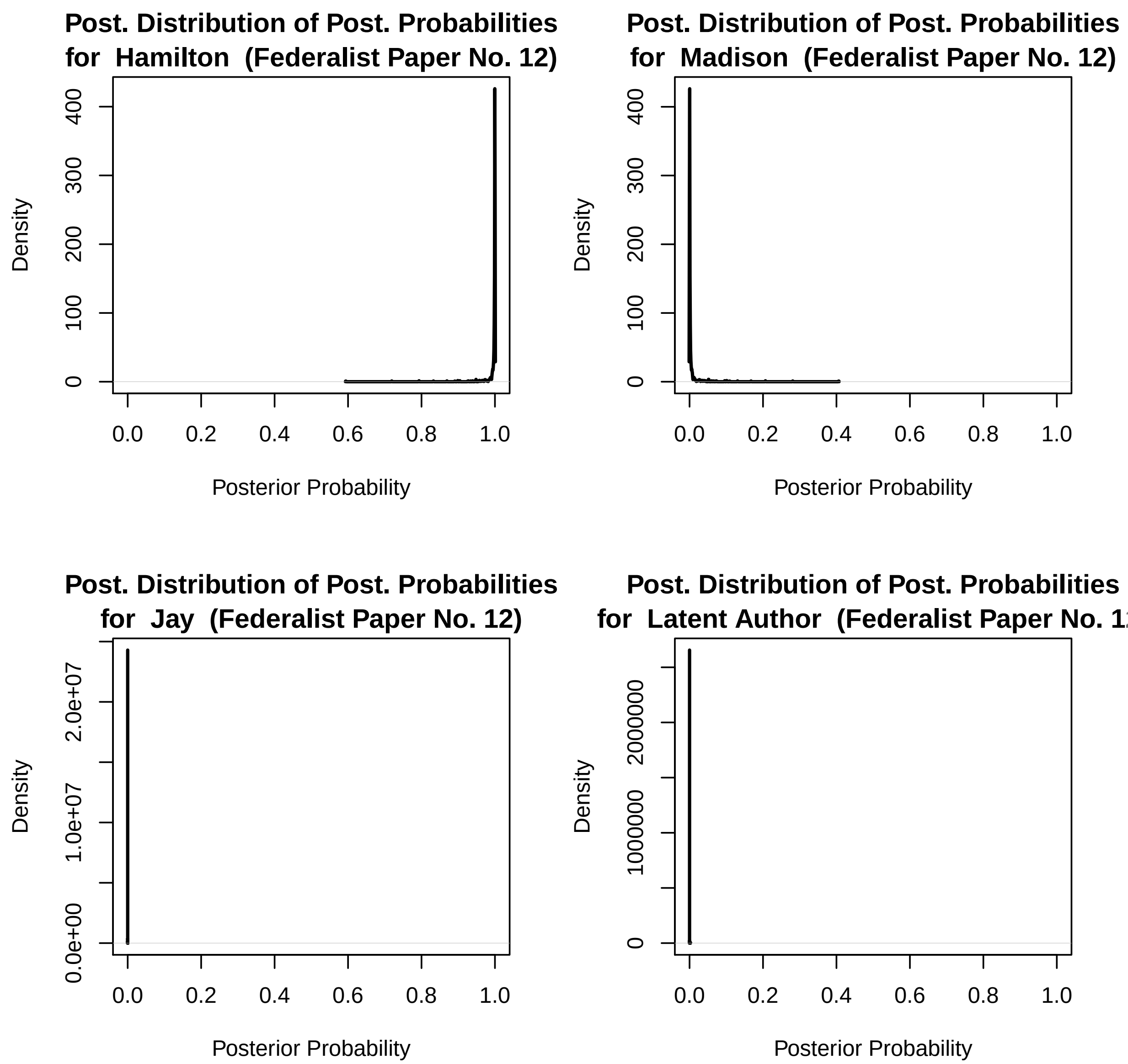
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 10



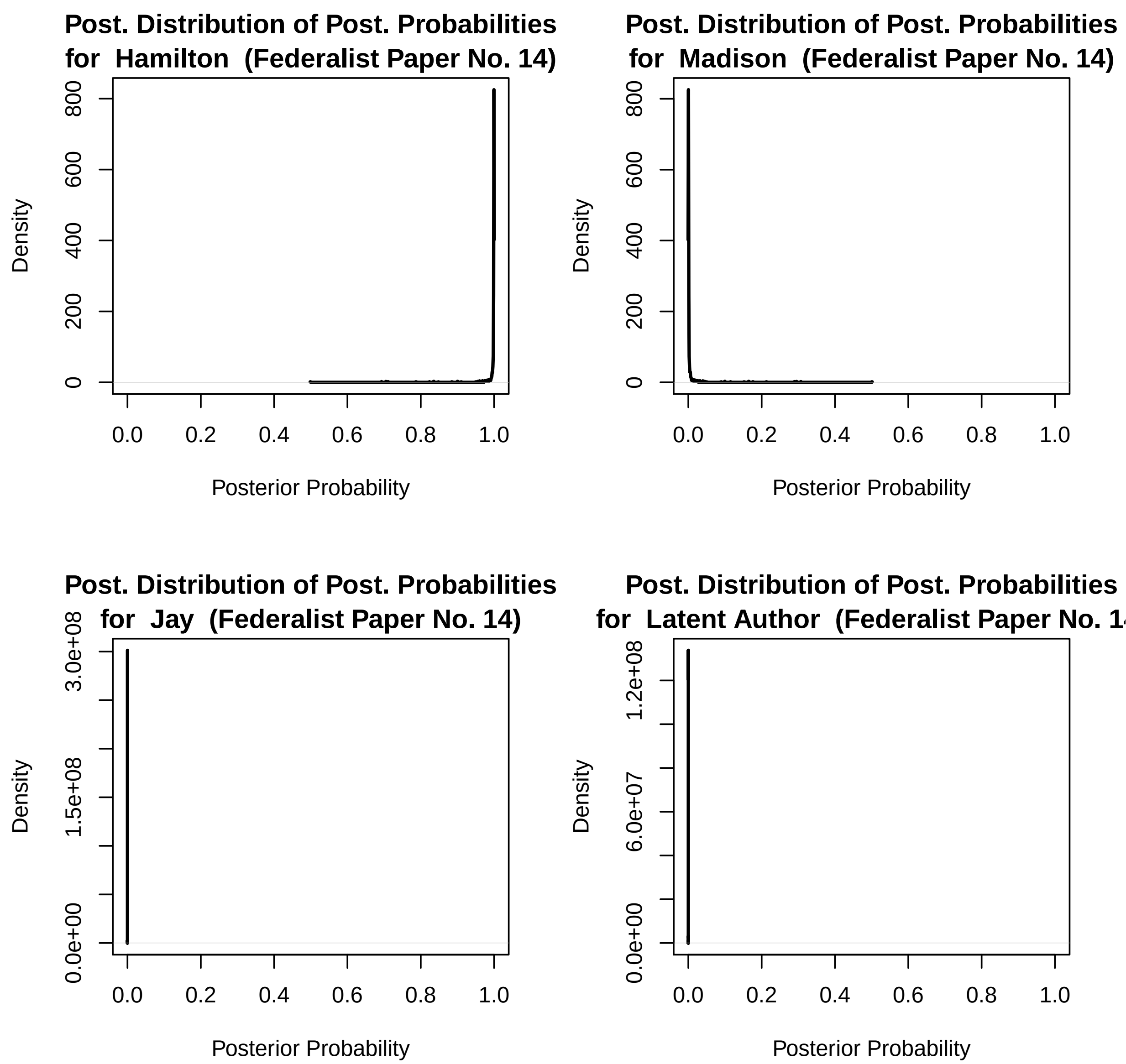
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 11



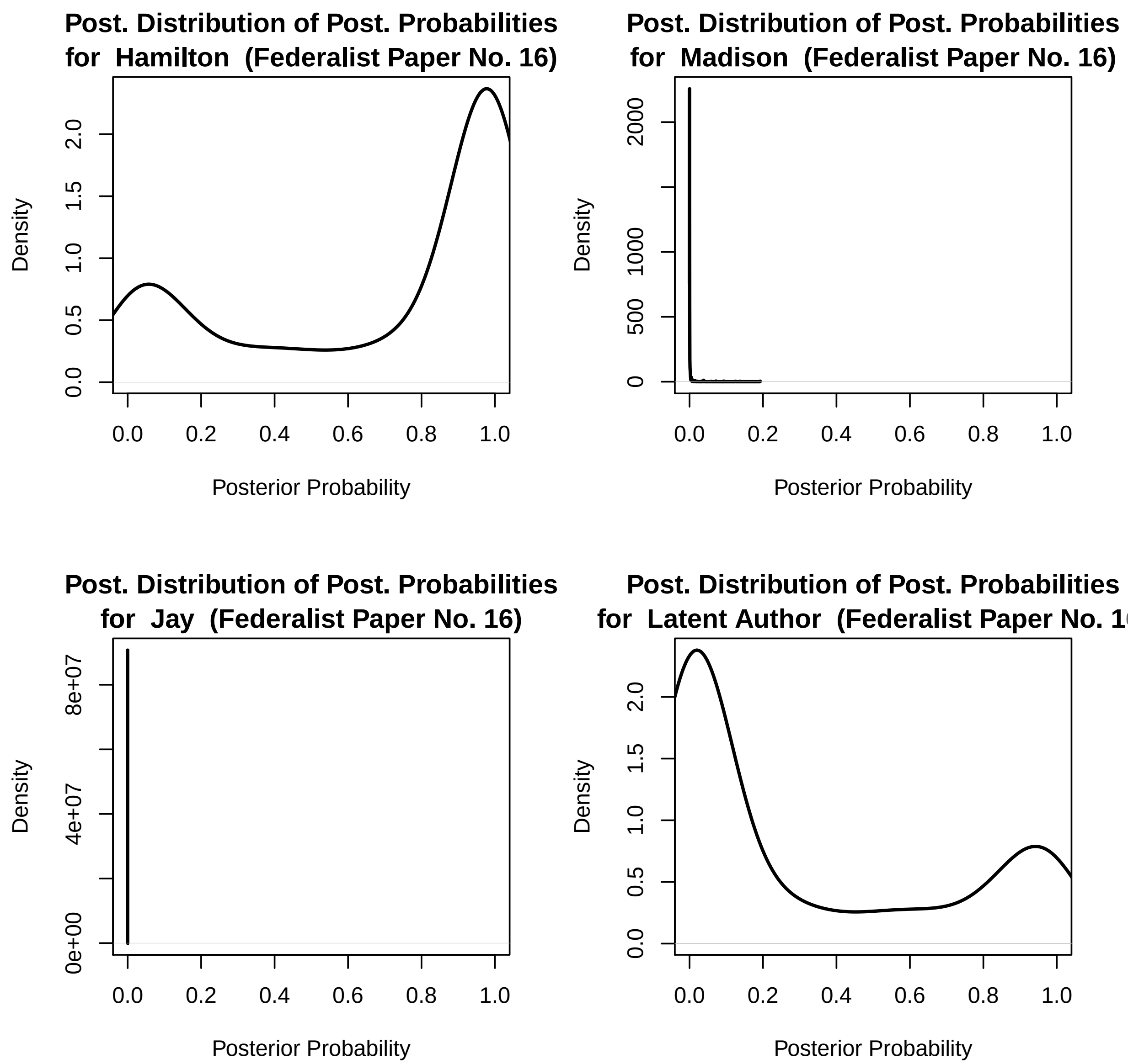
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 12



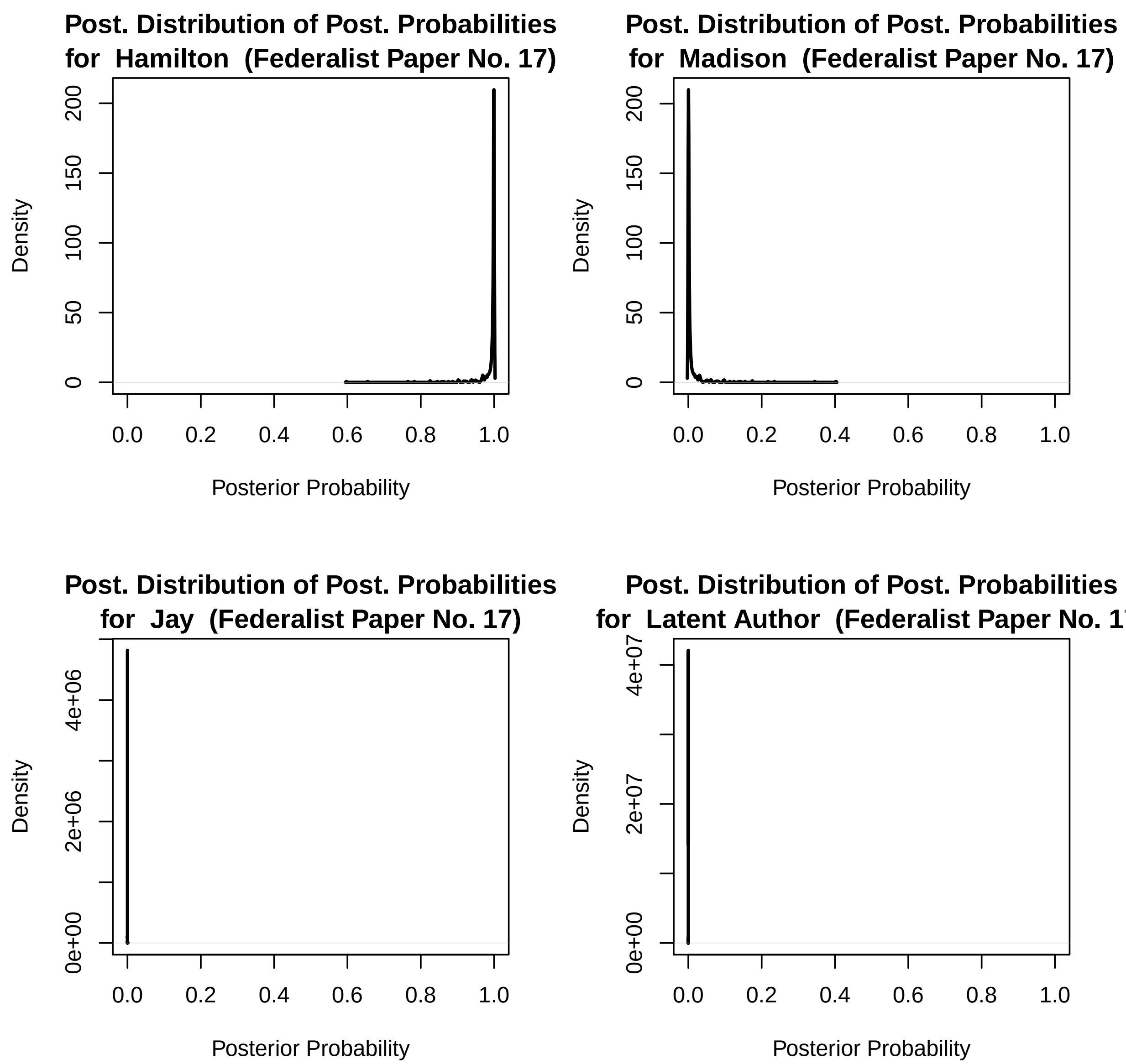
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 14



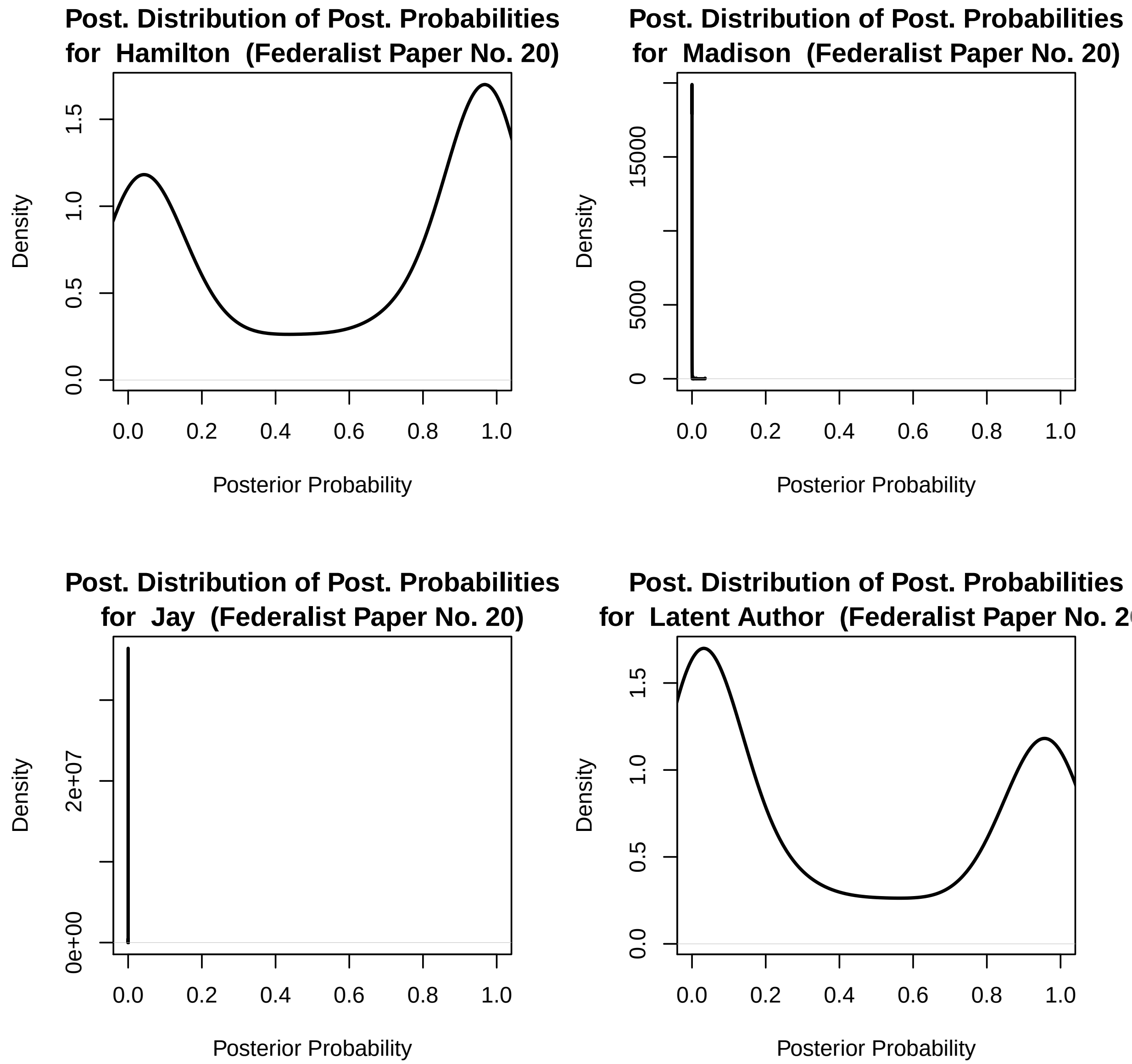
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 16



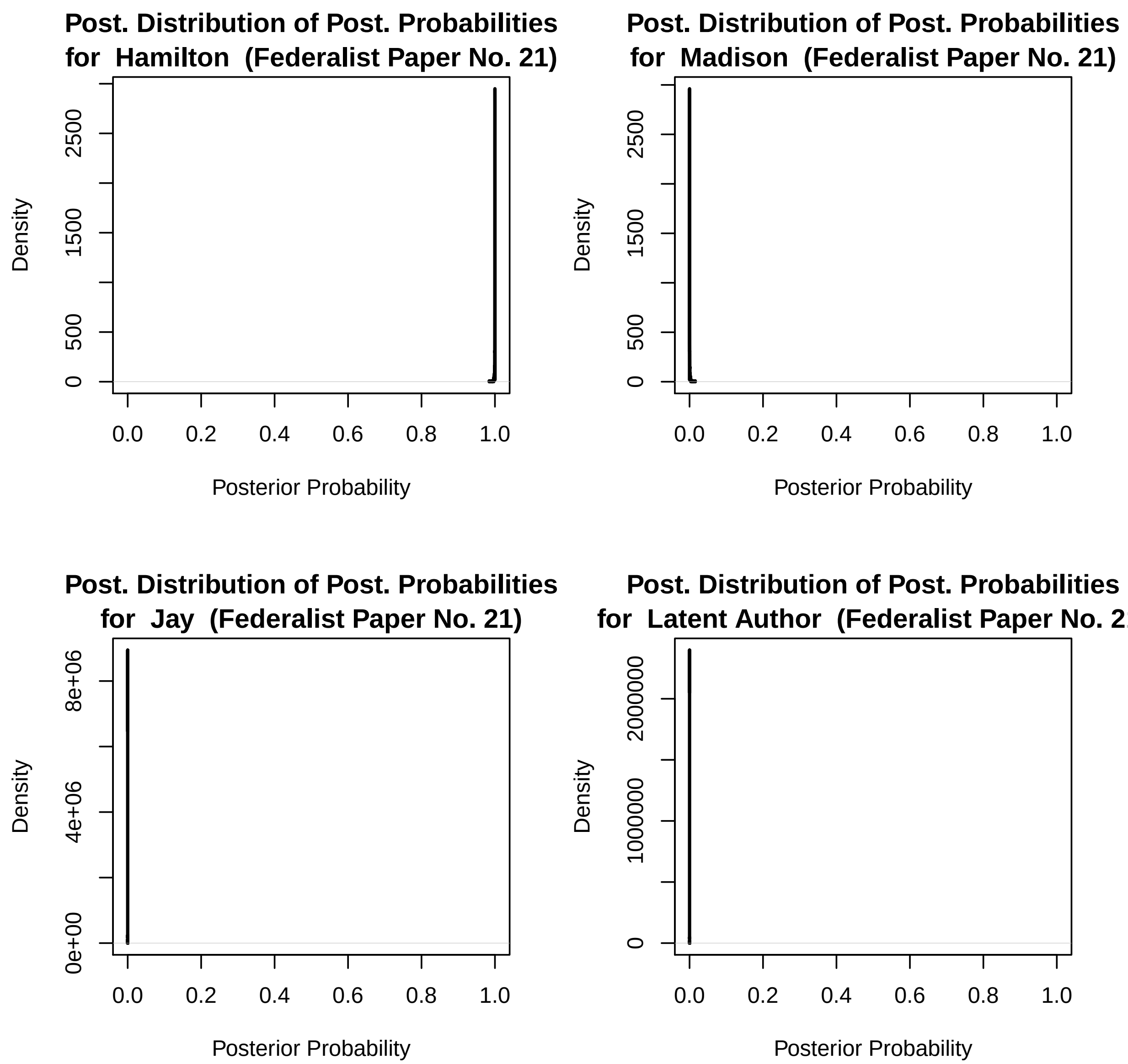
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 17



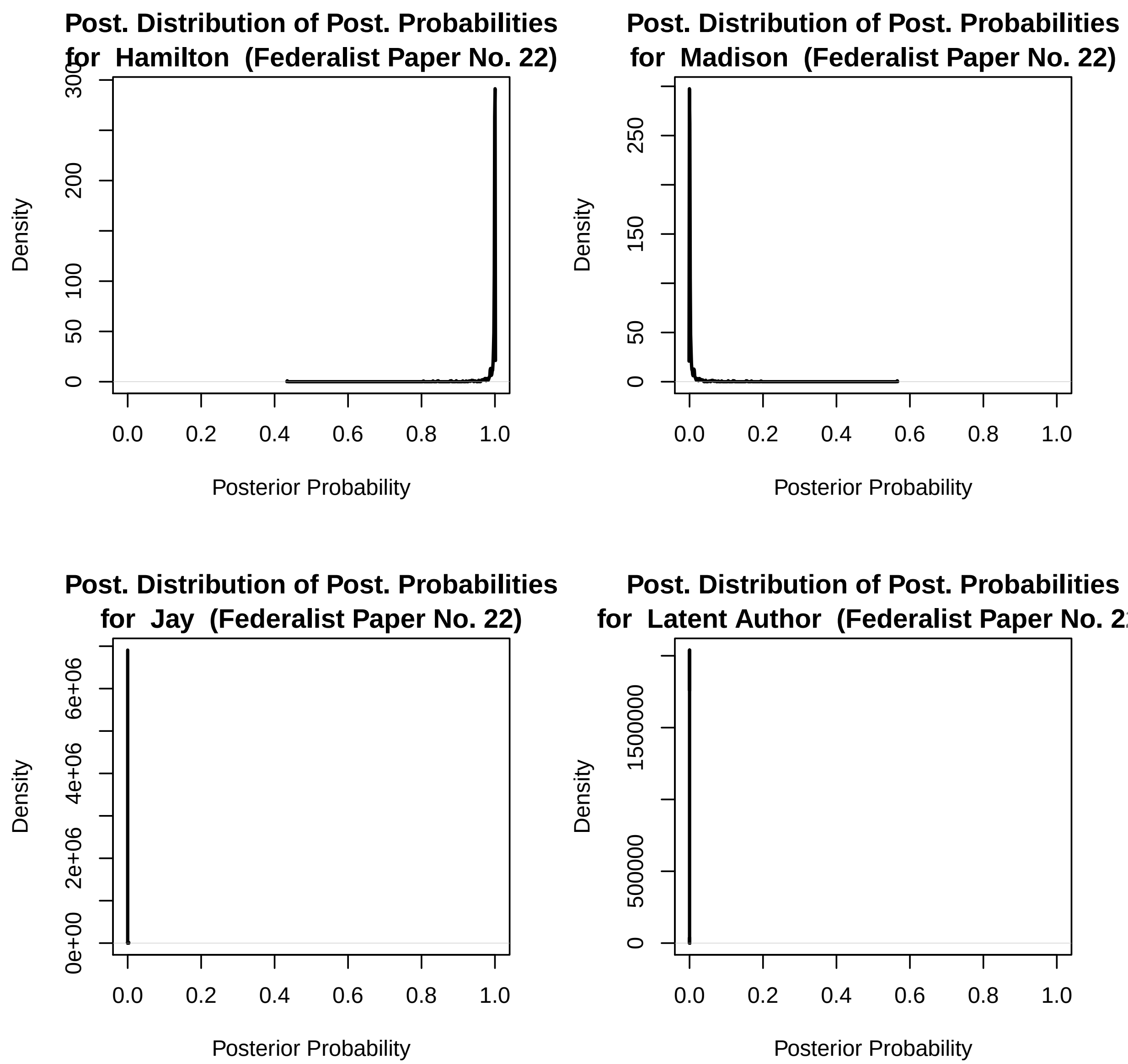
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 20



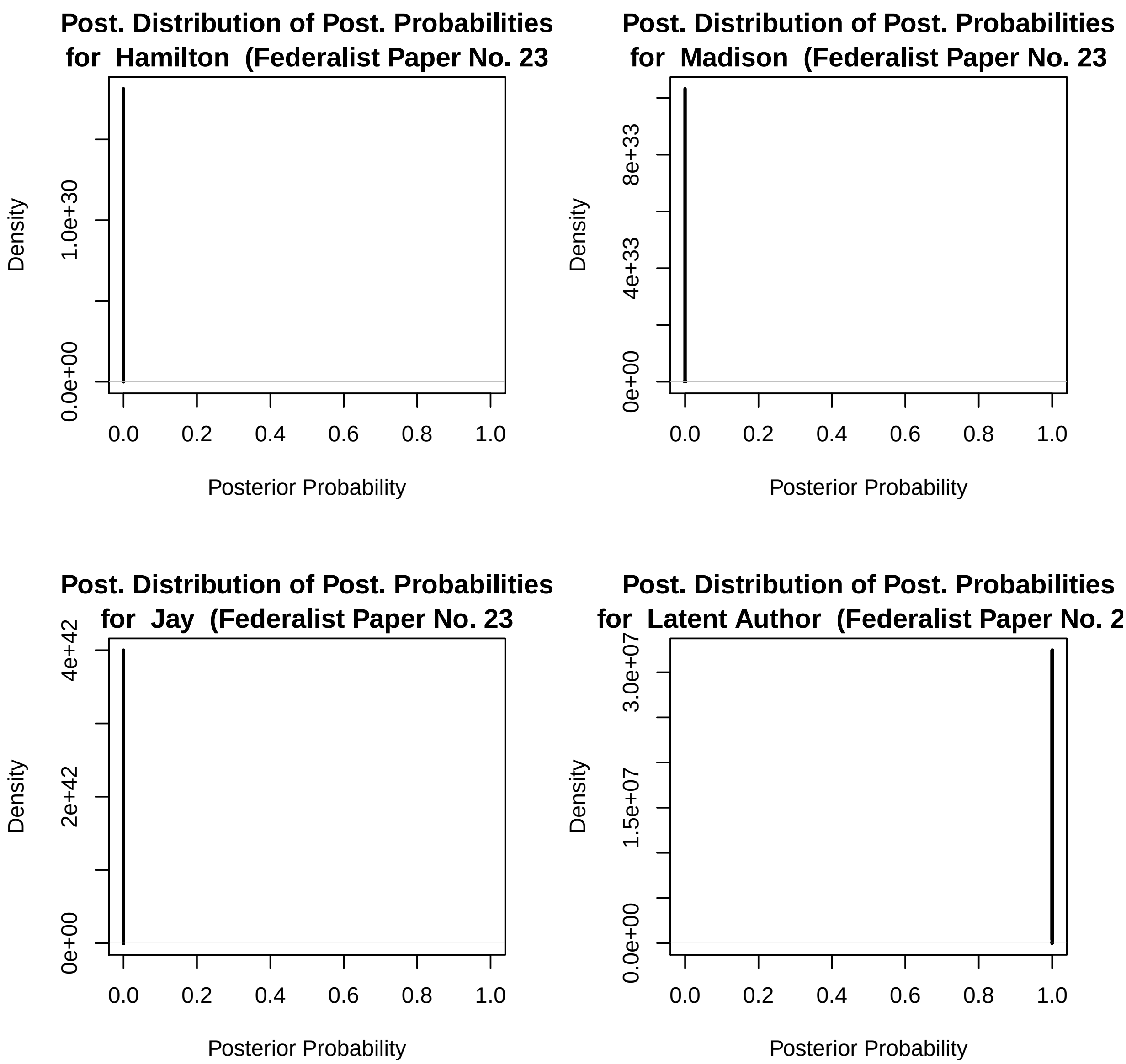
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 21



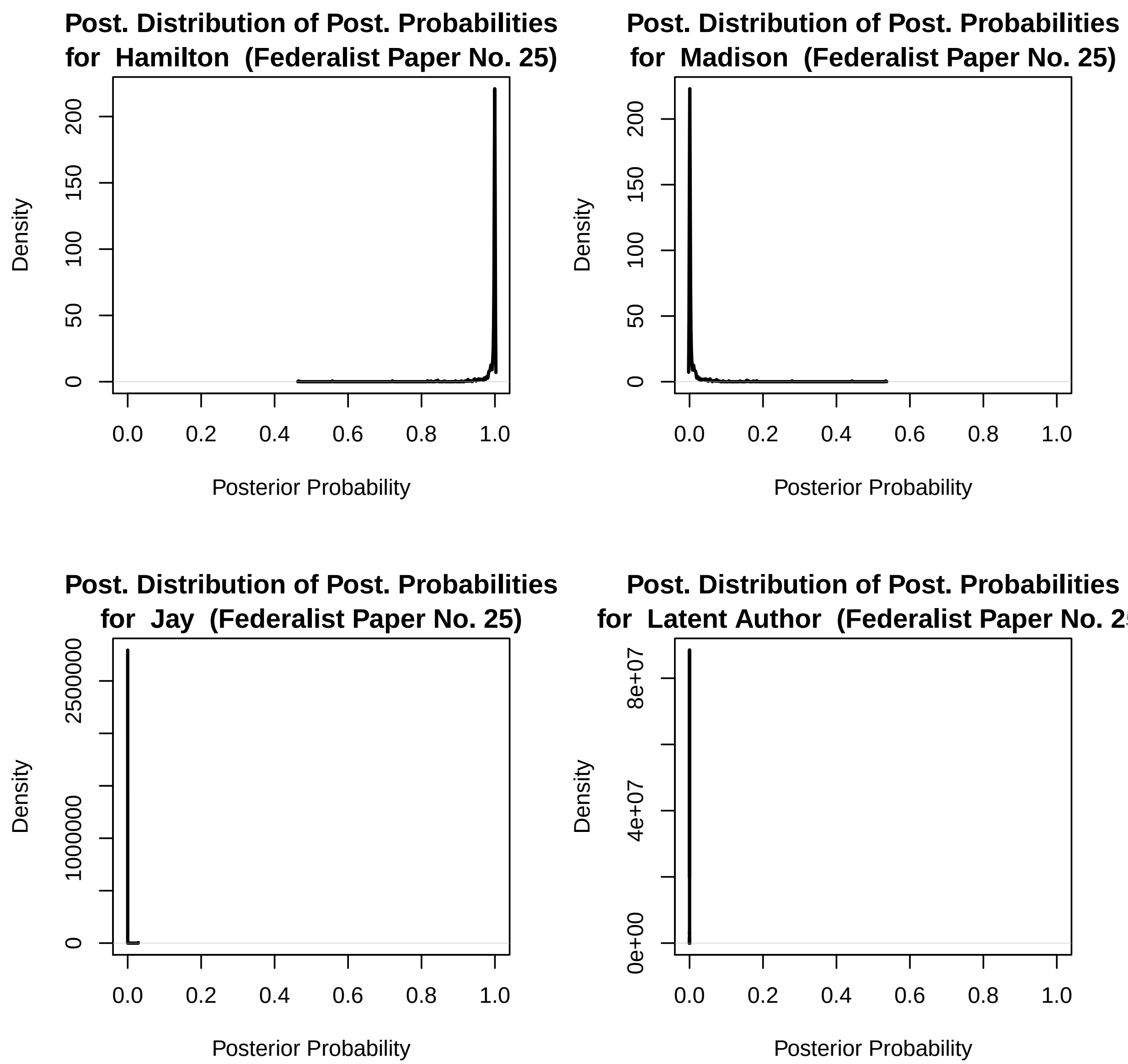
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 22



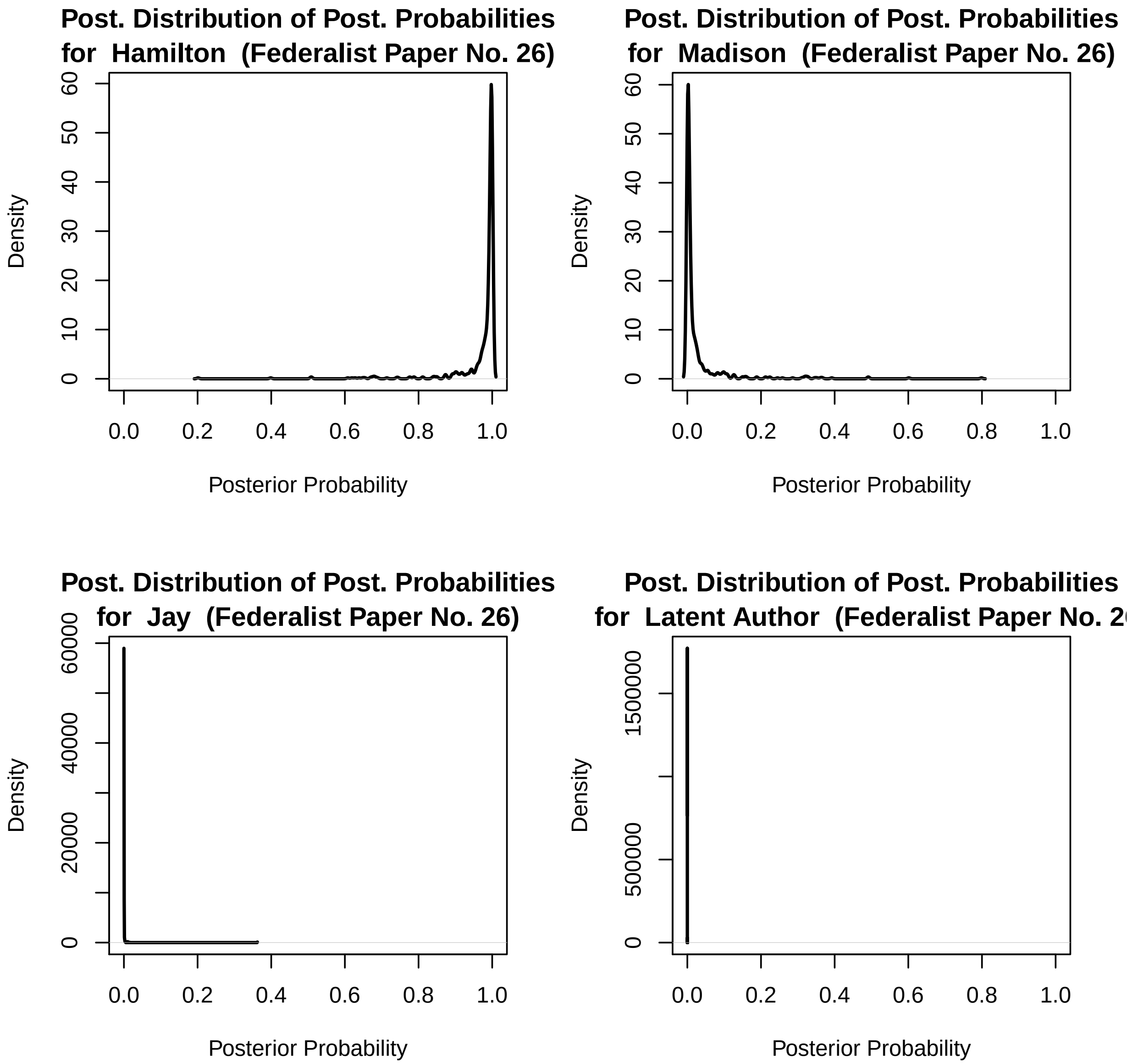
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 23



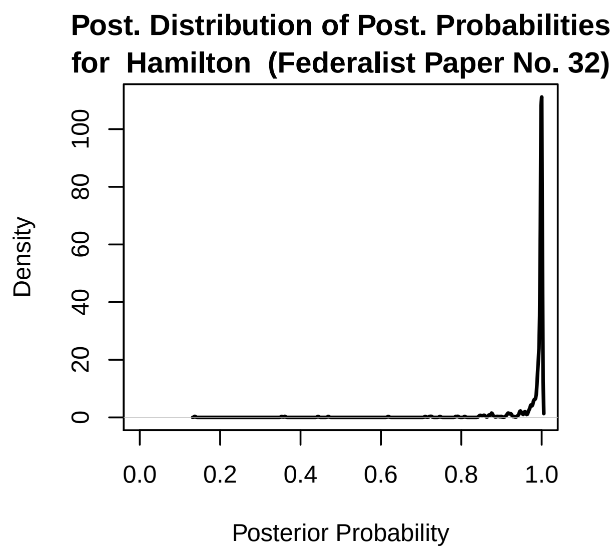
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 25

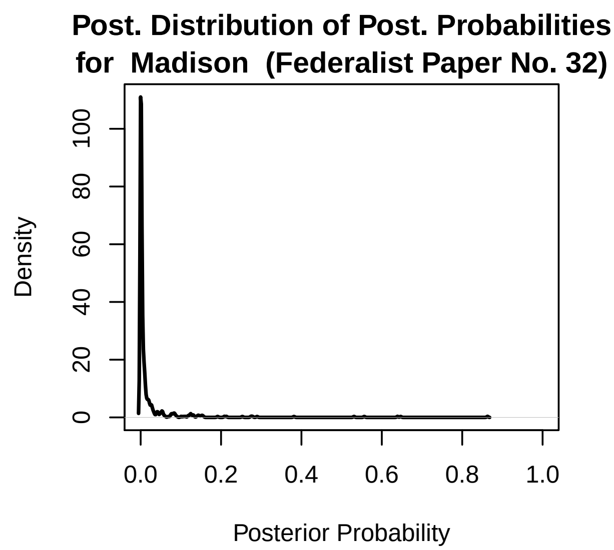


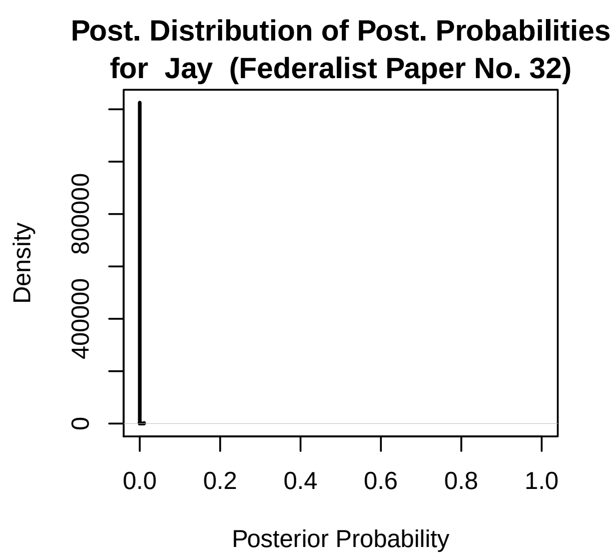
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 26

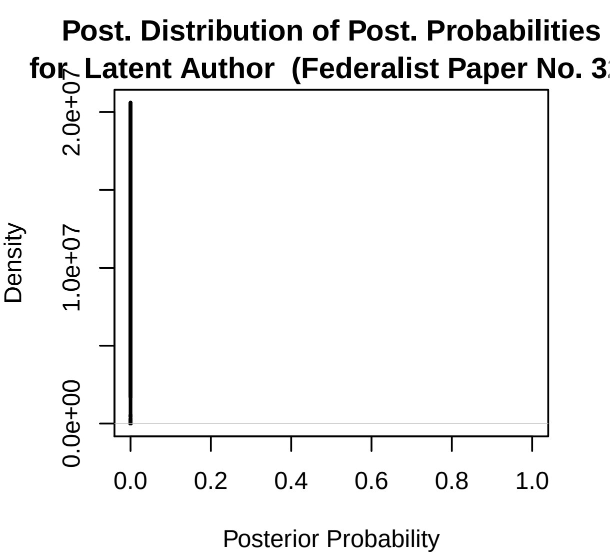


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 32

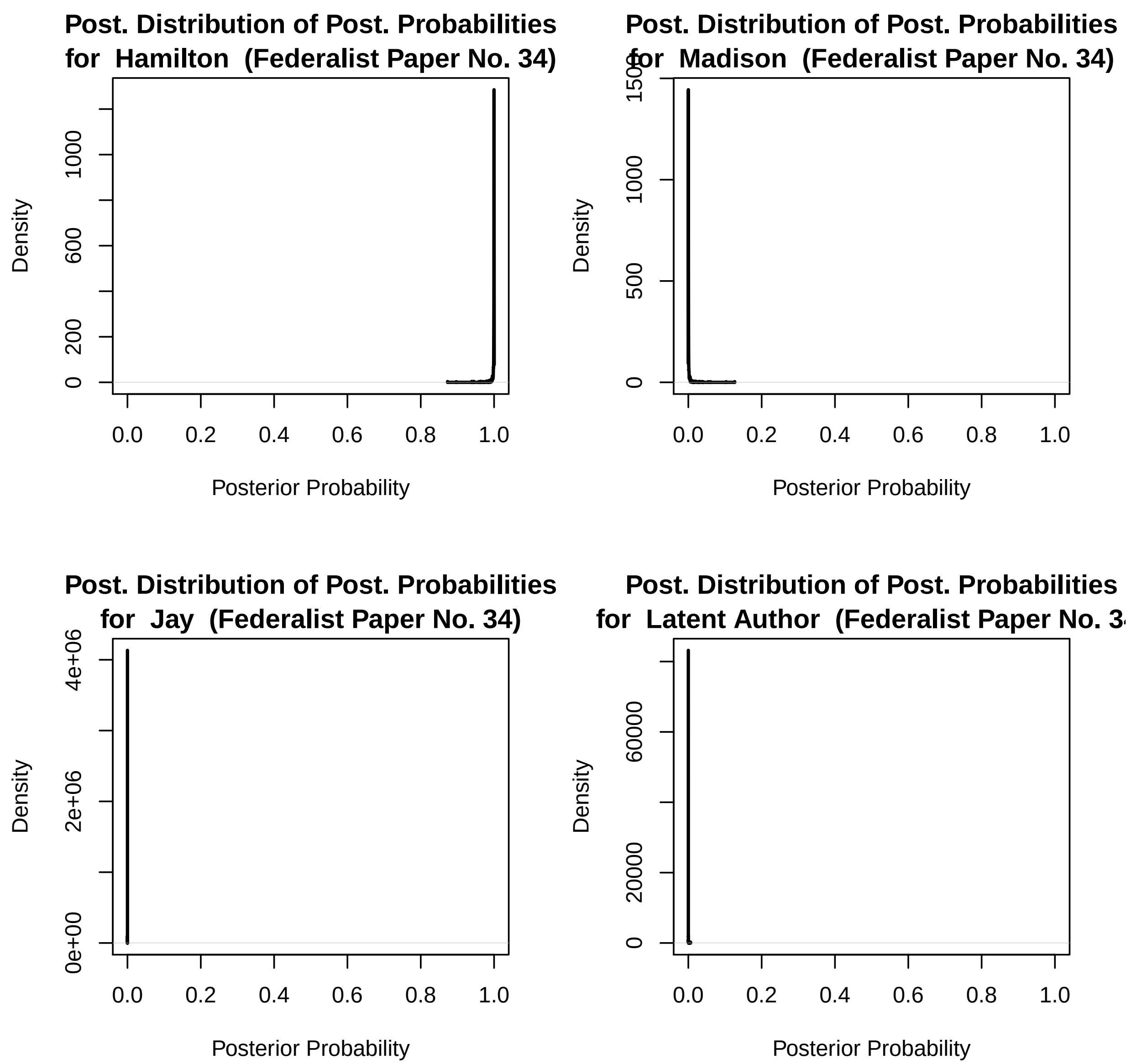




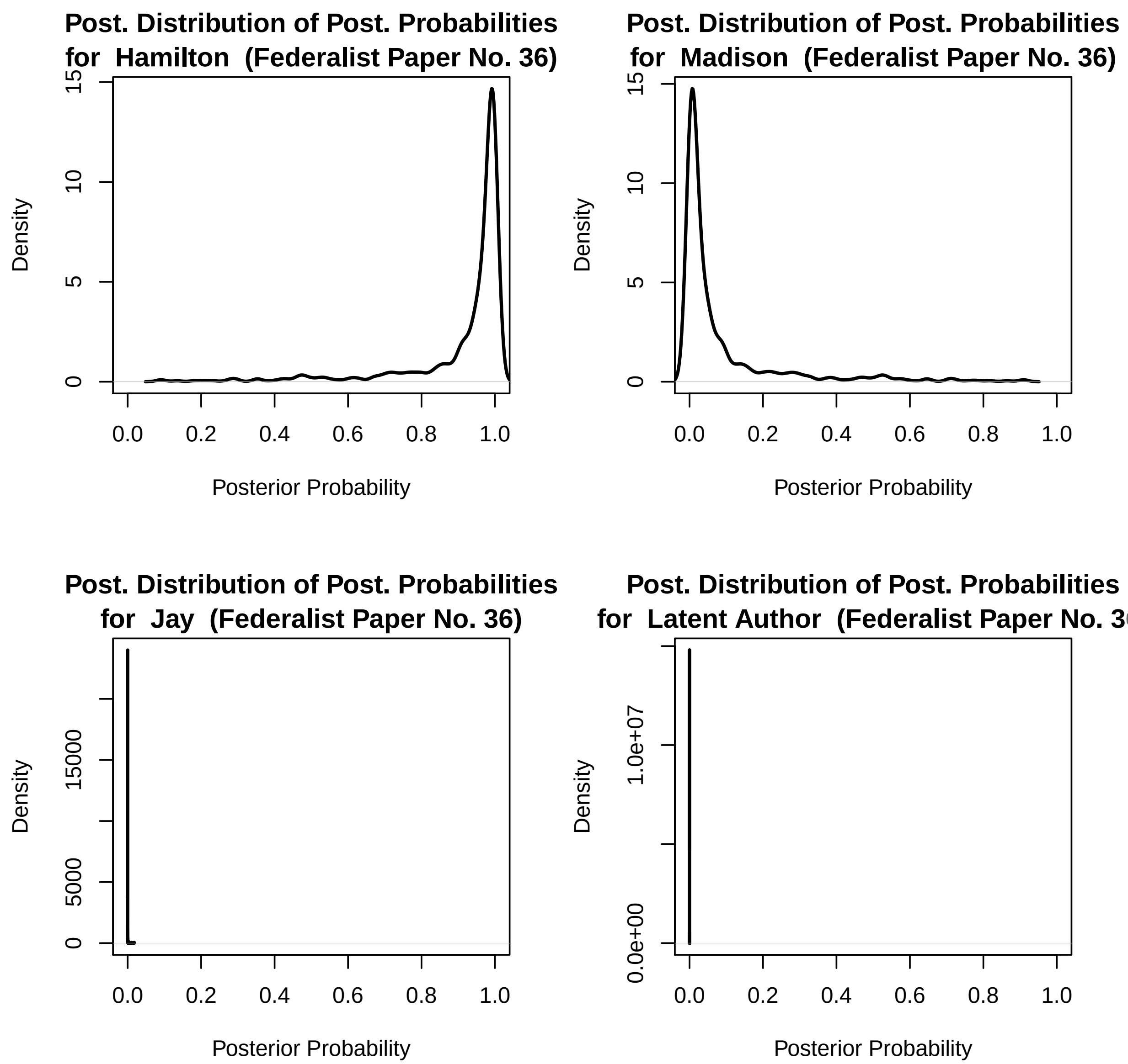




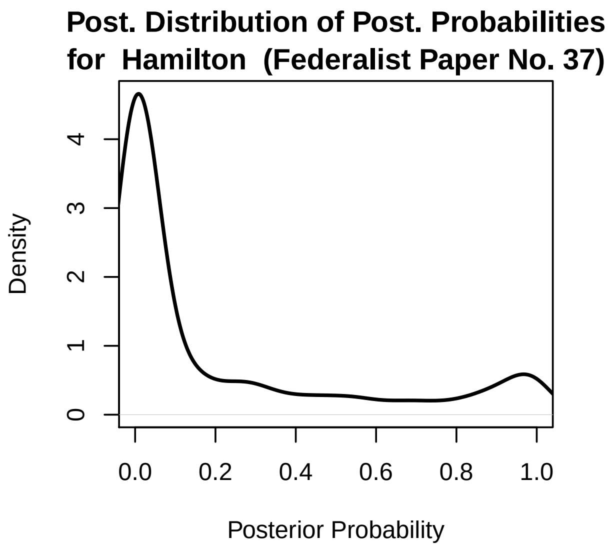
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 34

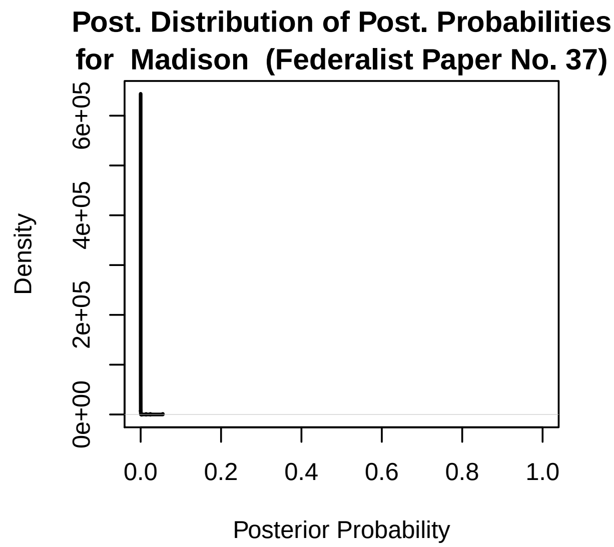


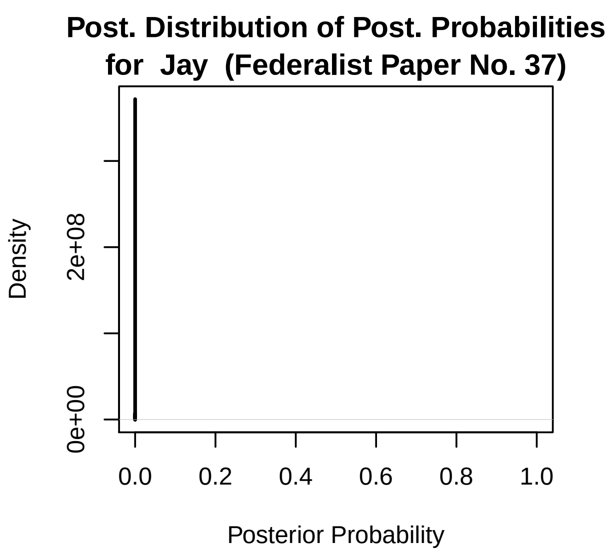
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 36

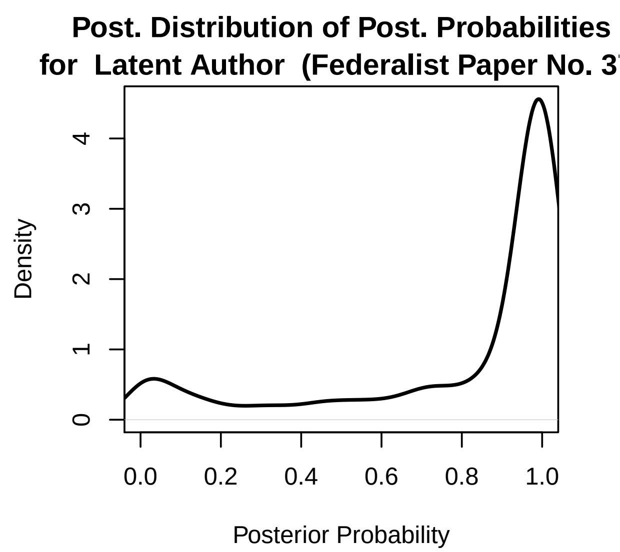


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 37

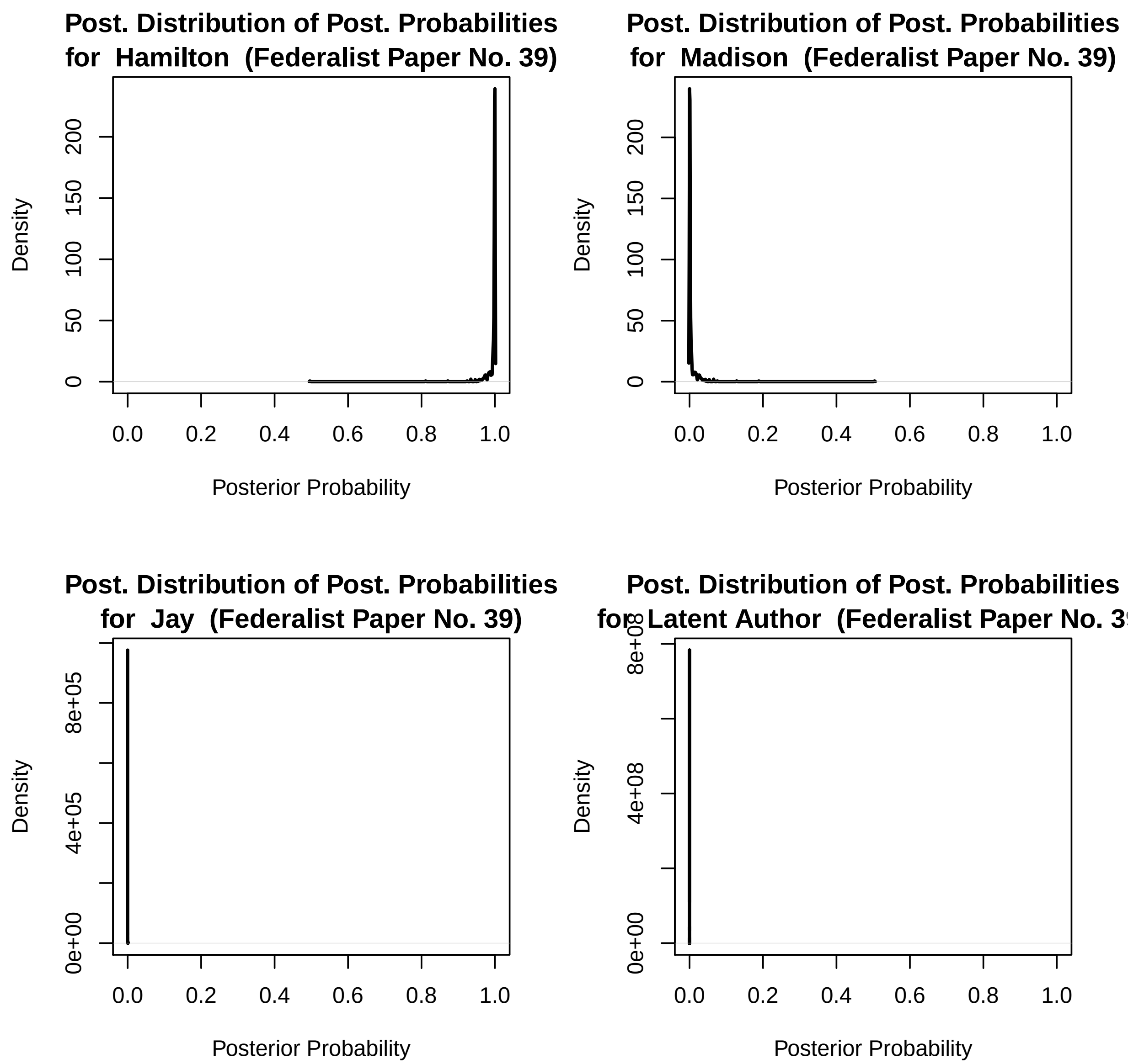




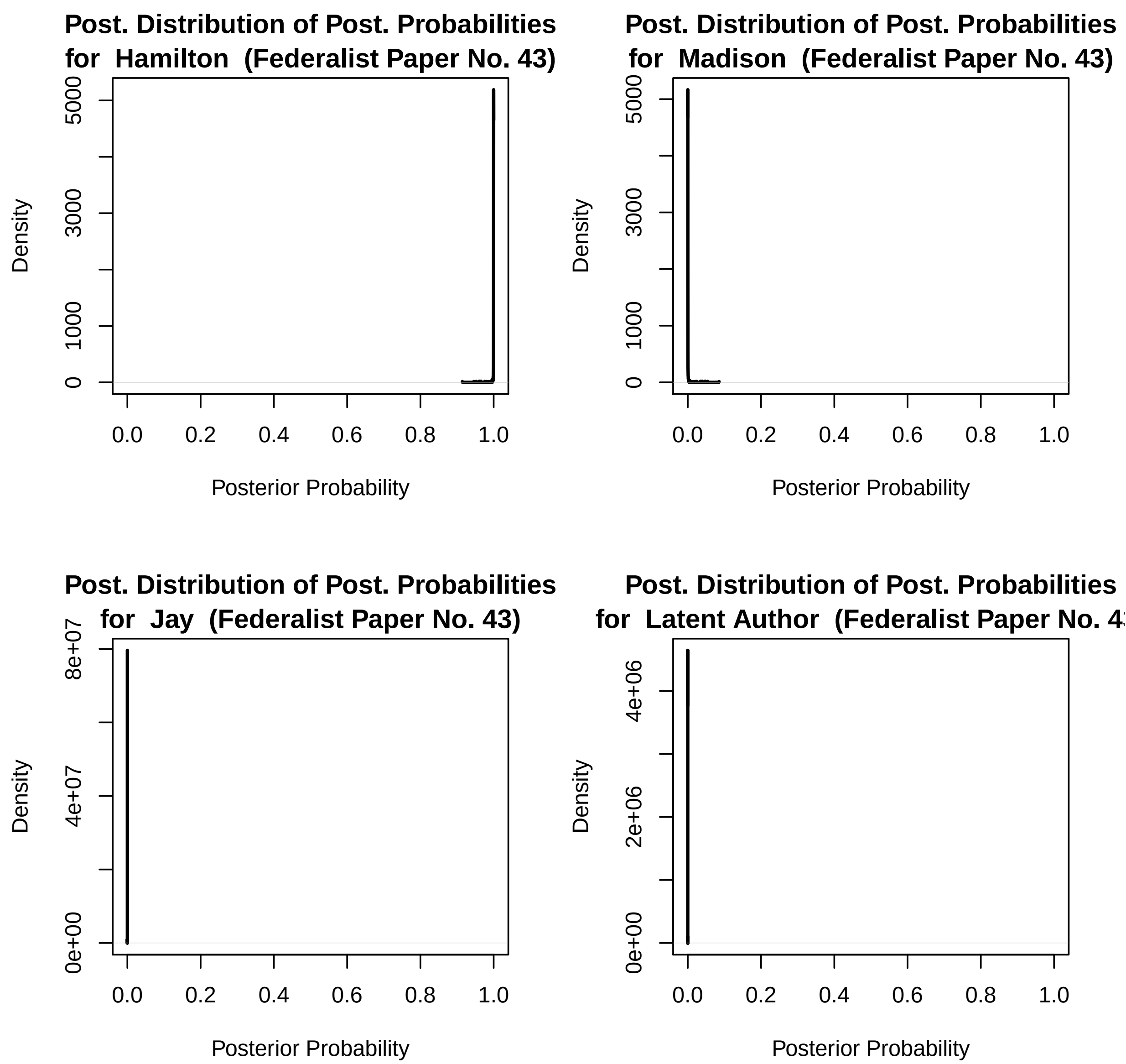




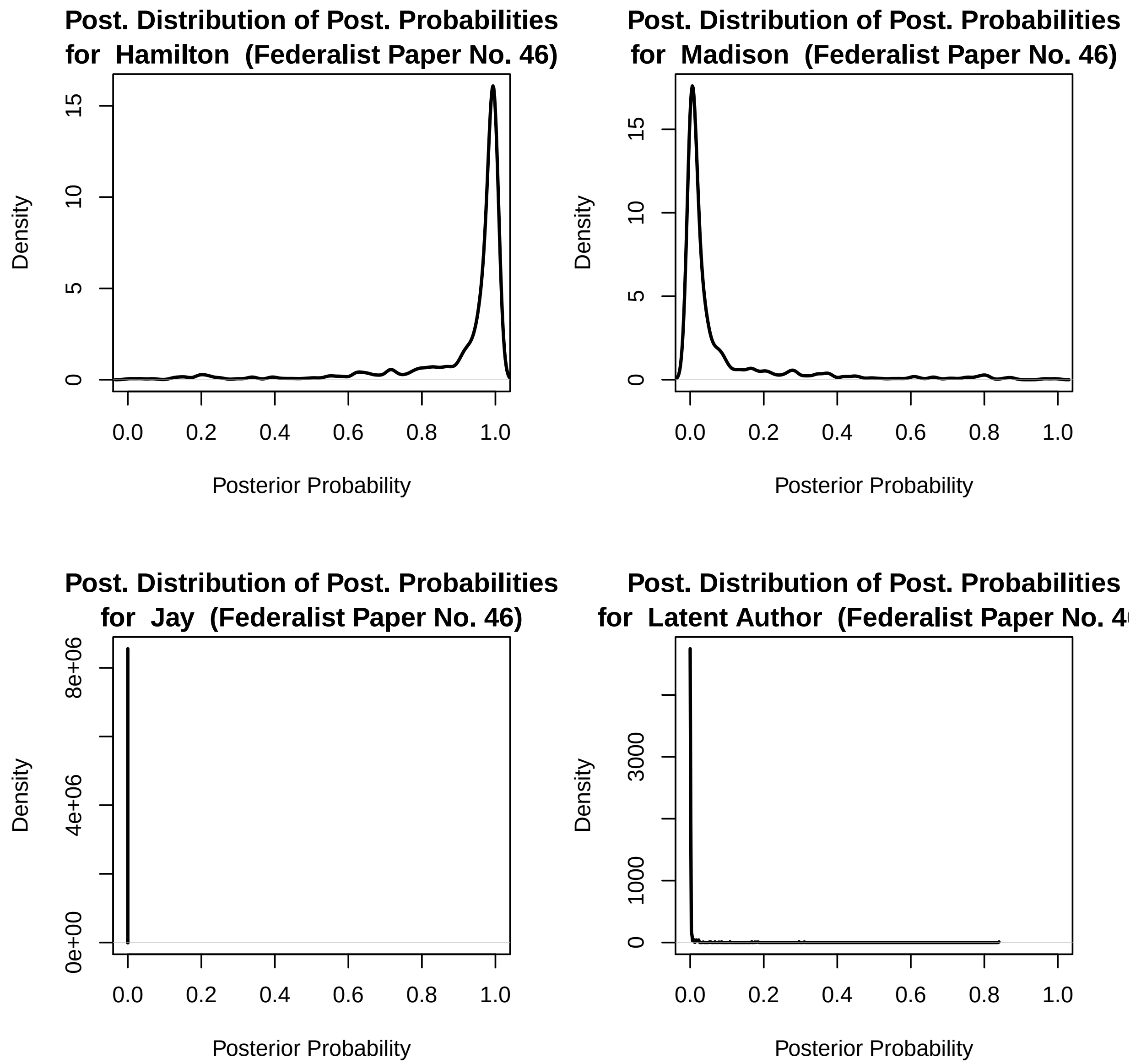
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 39



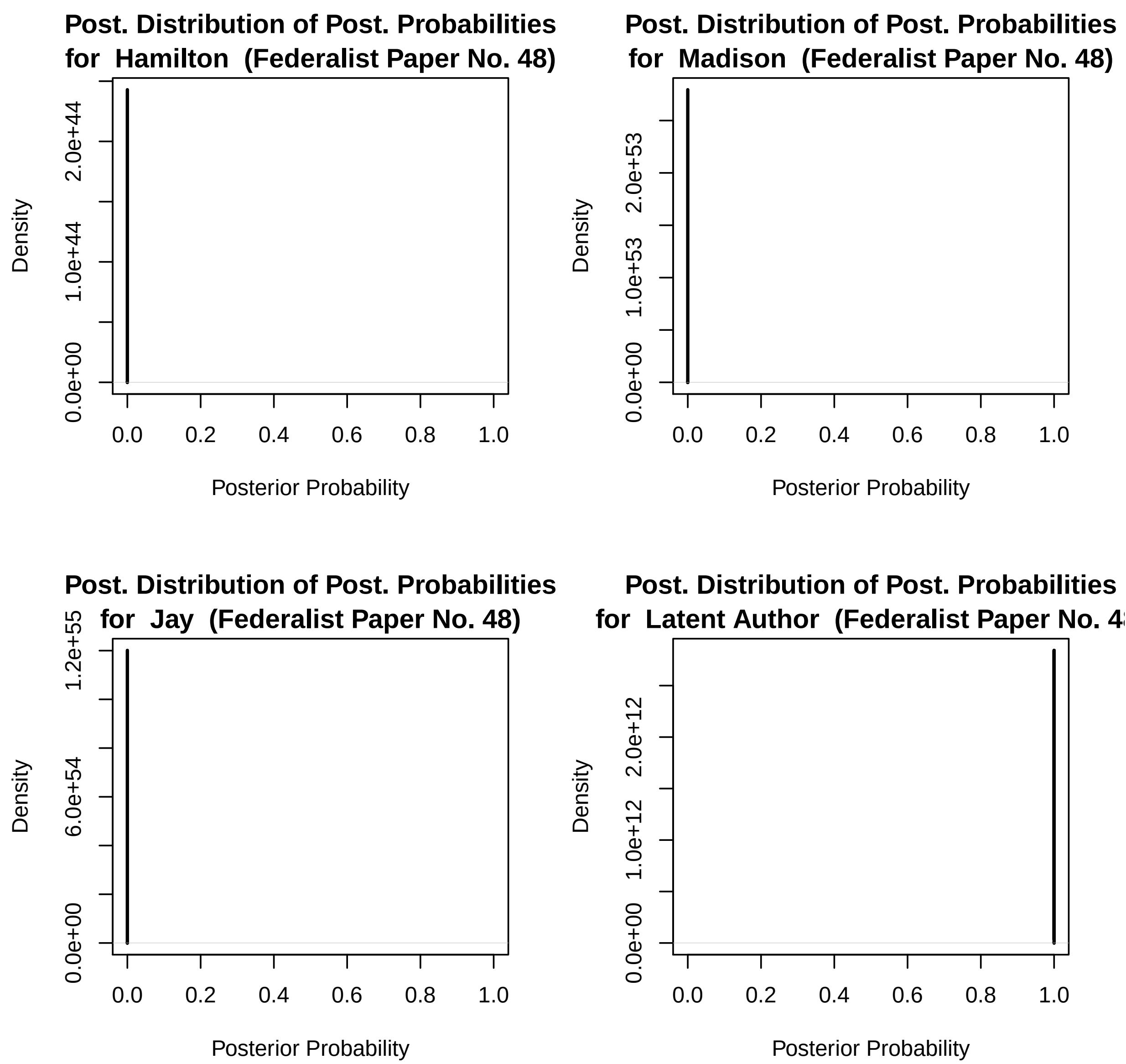
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 43



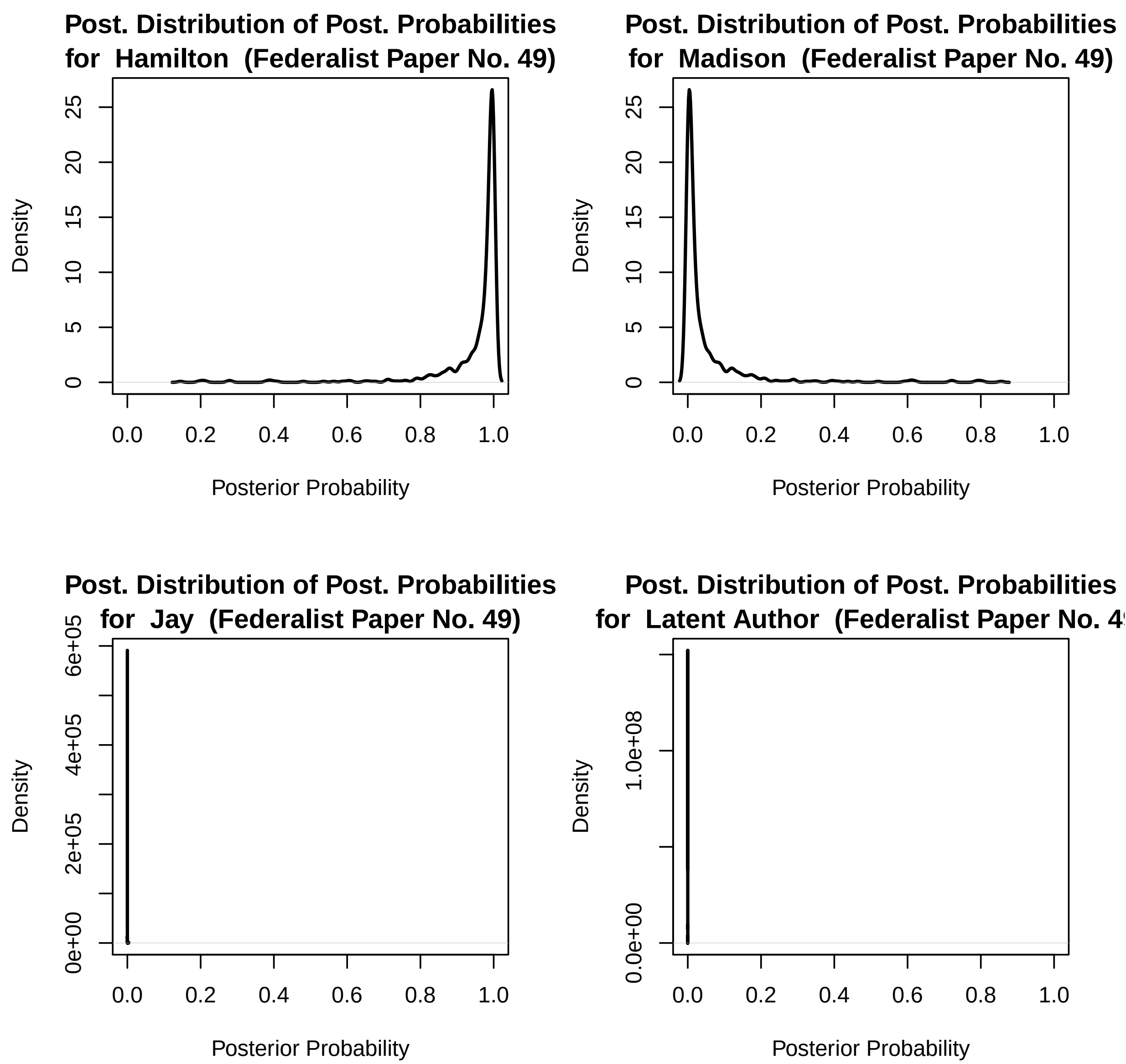
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 46



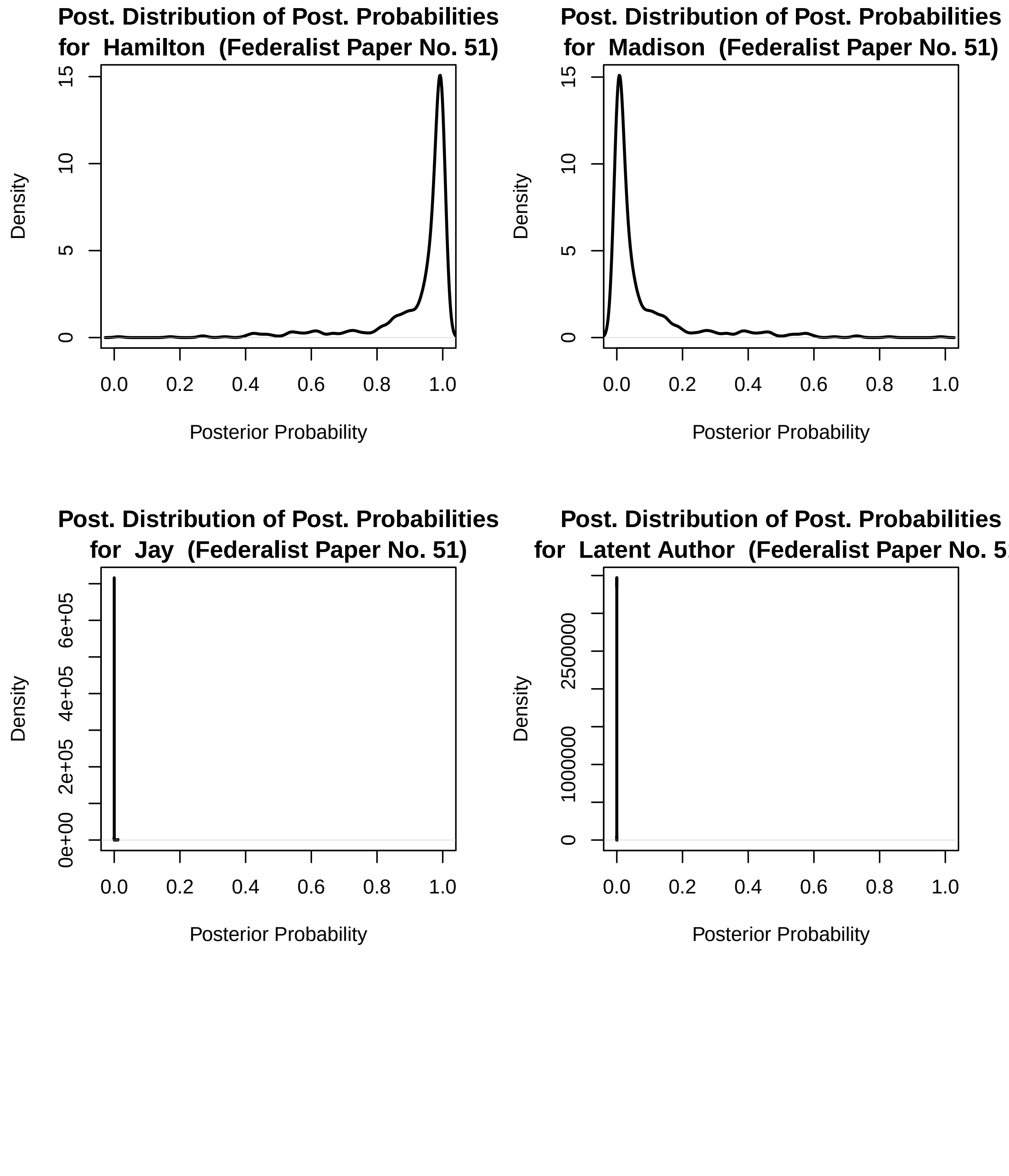
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 48



Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 49

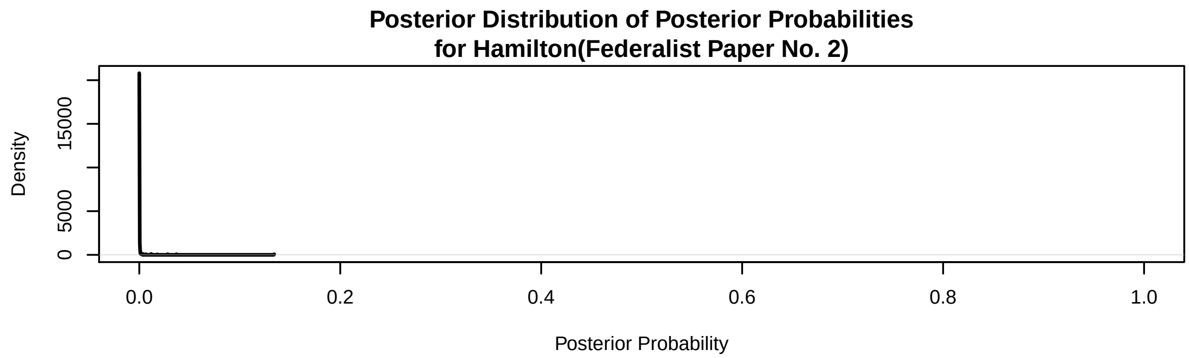


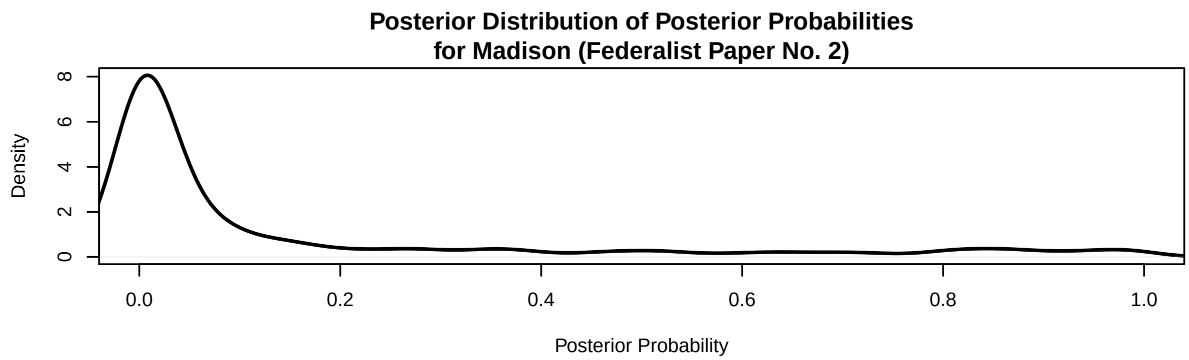
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 51

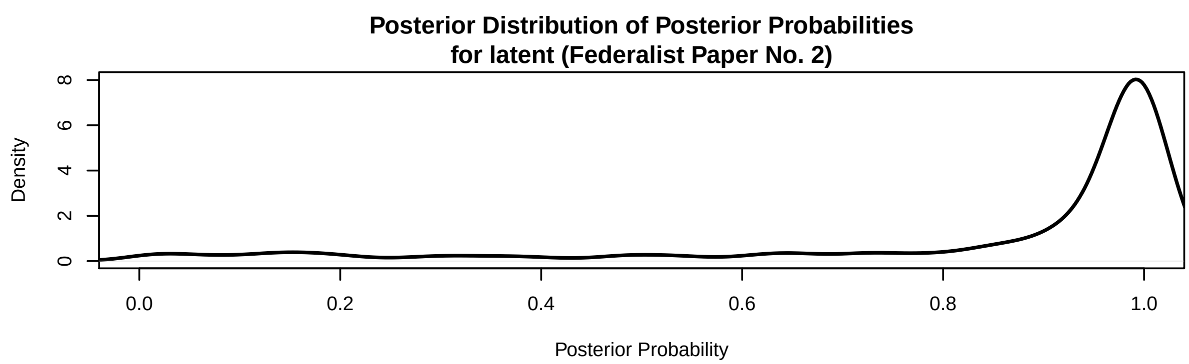


B.5 Jay Federalist Papers with Hamilton and Madison in the training set

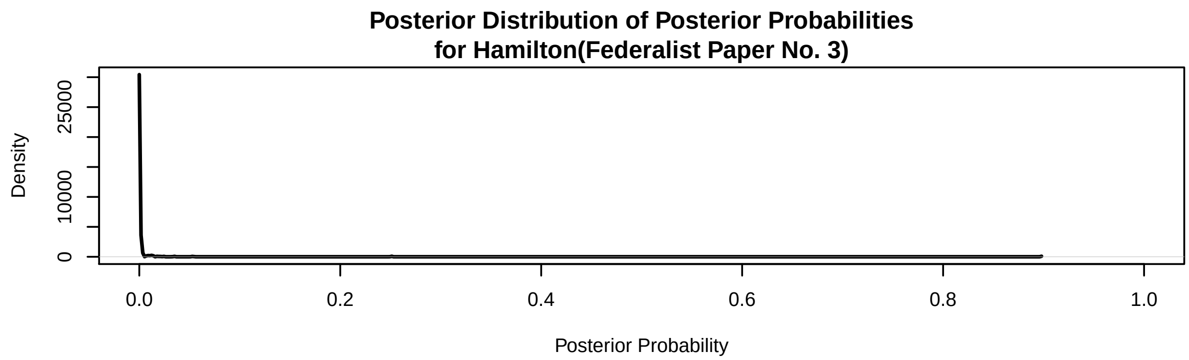
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 2

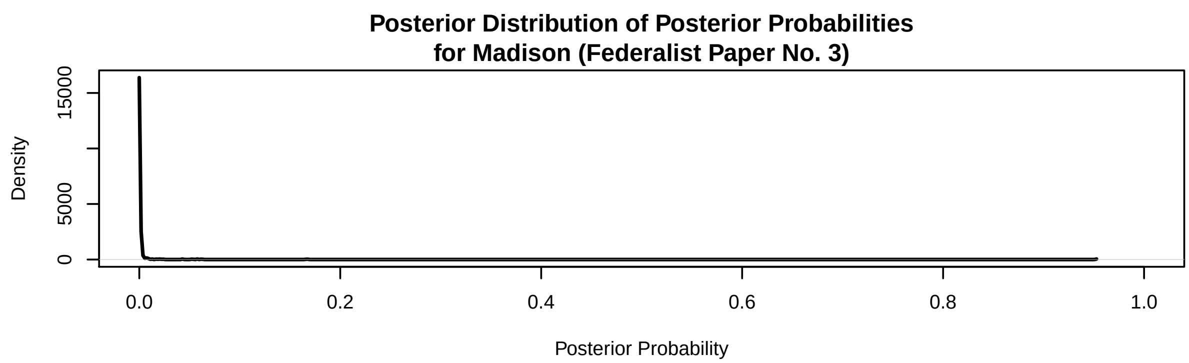


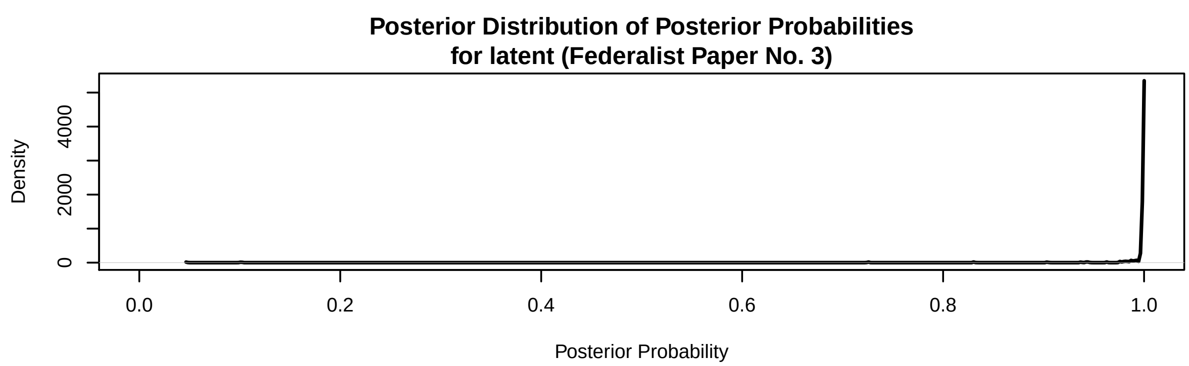




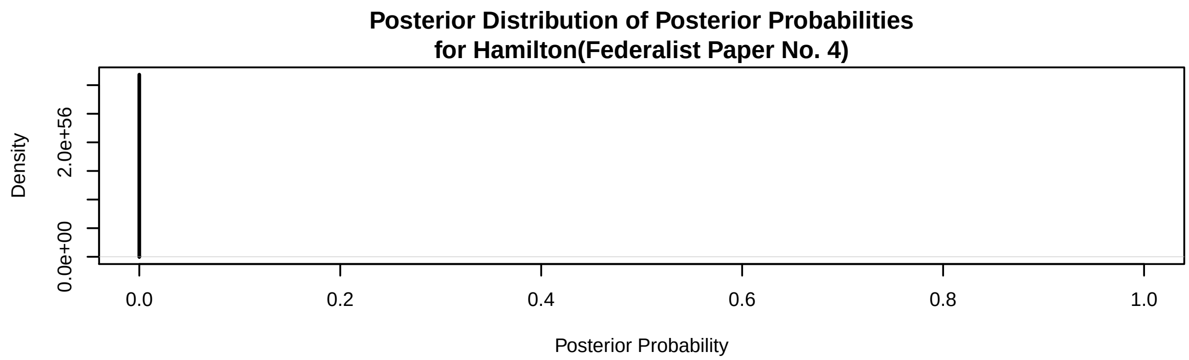
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 3

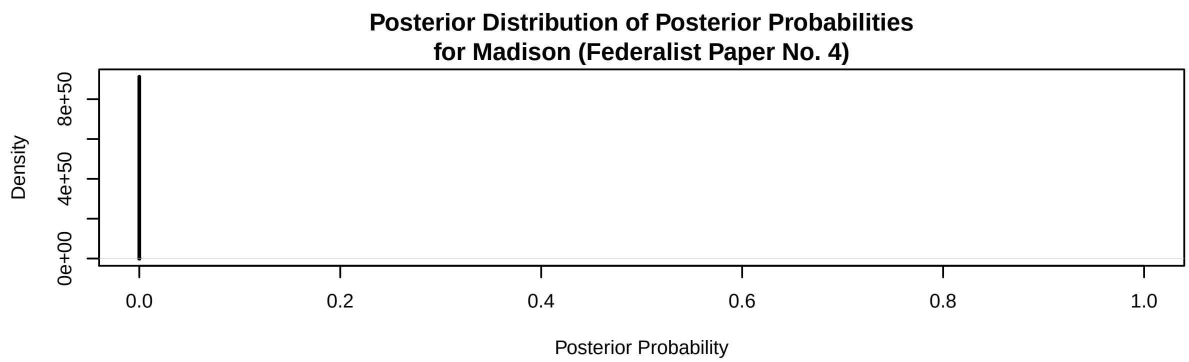


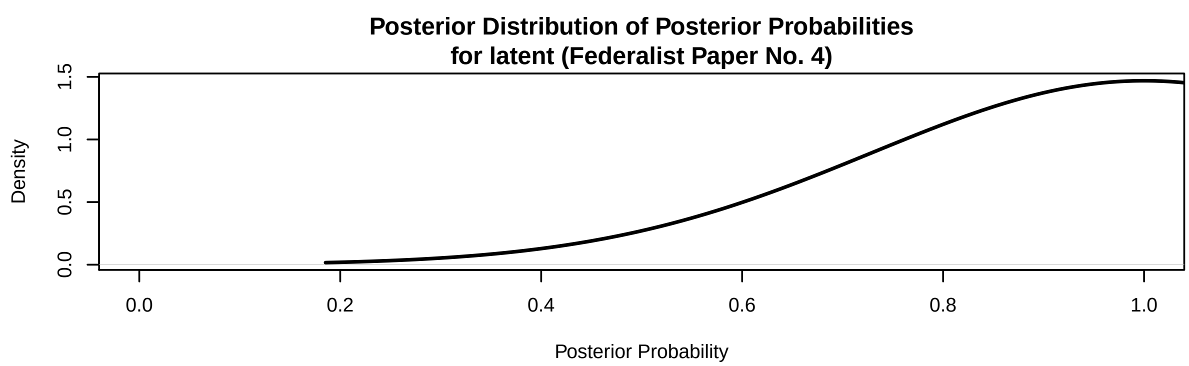




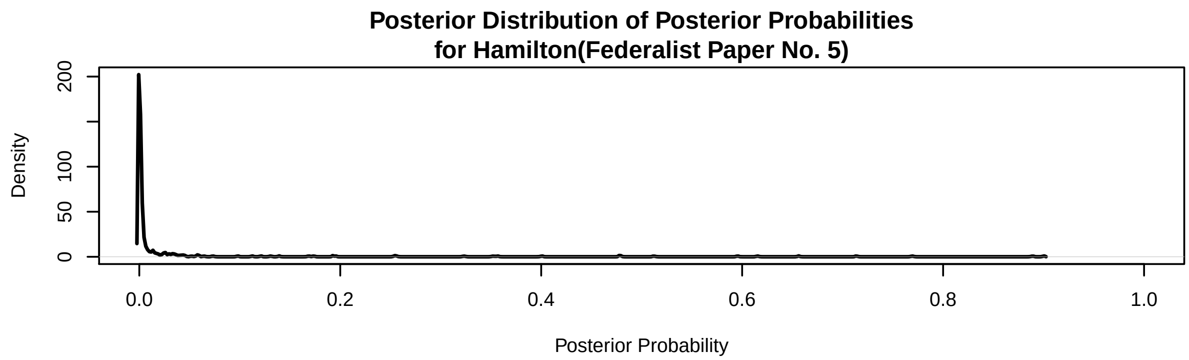
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 4

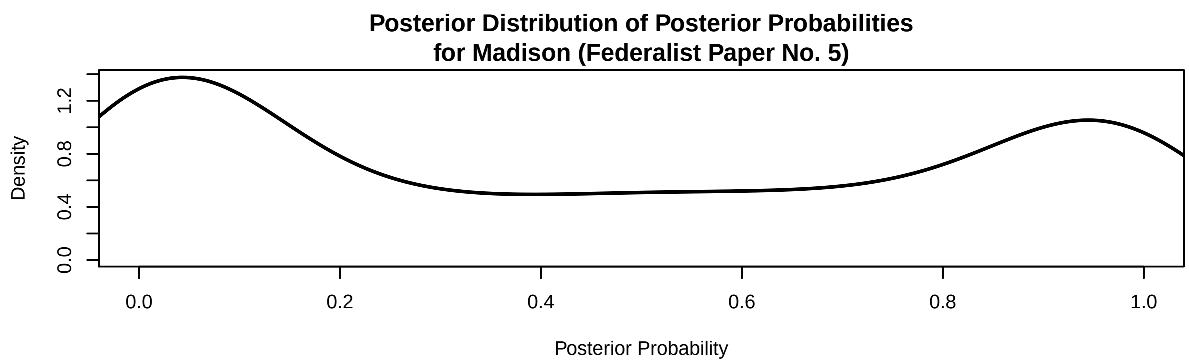


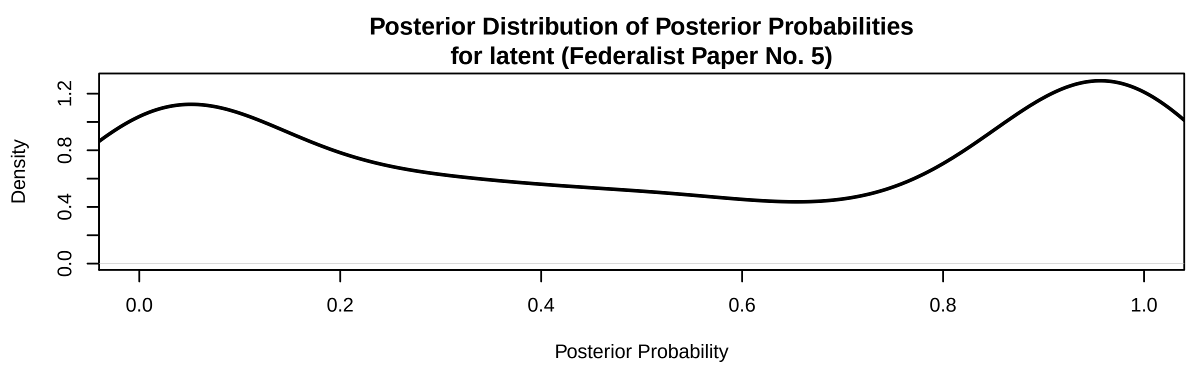




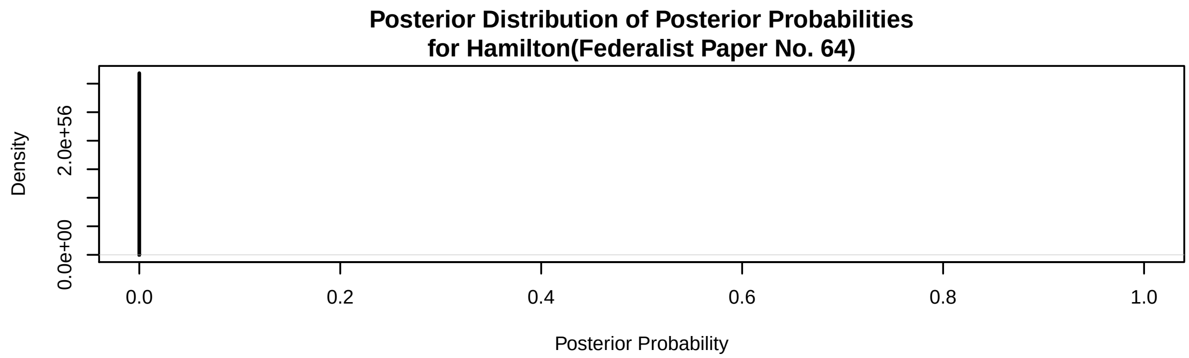
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 5

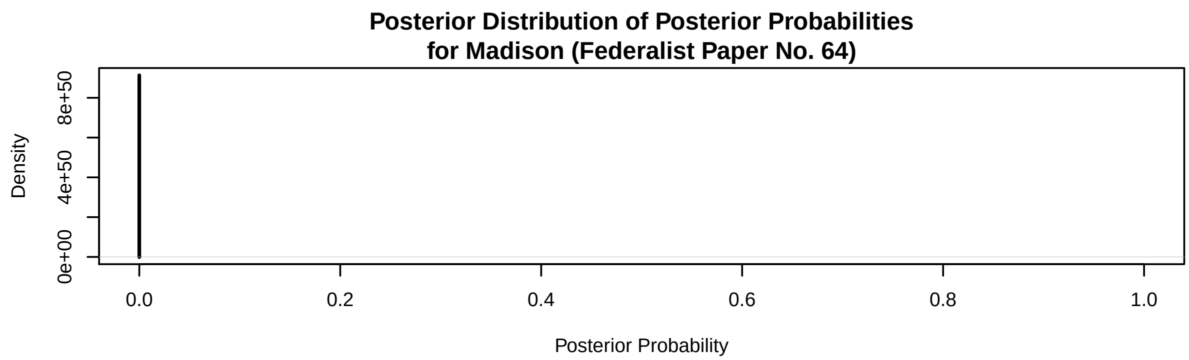


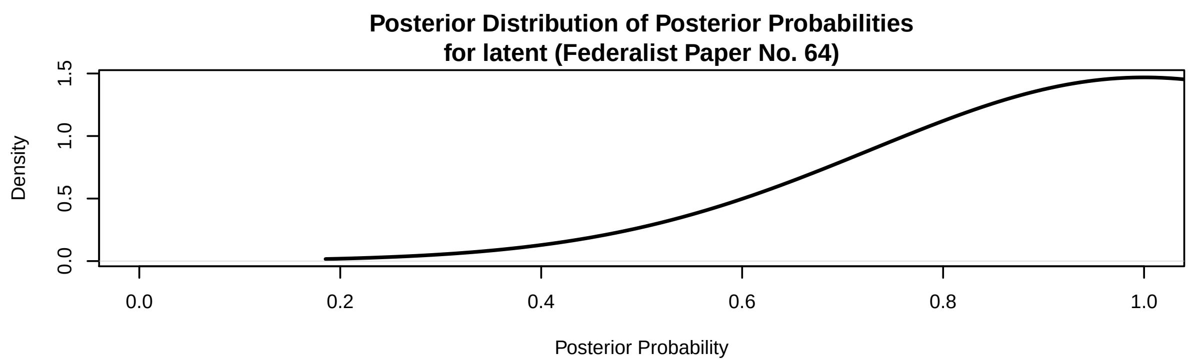




Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 64

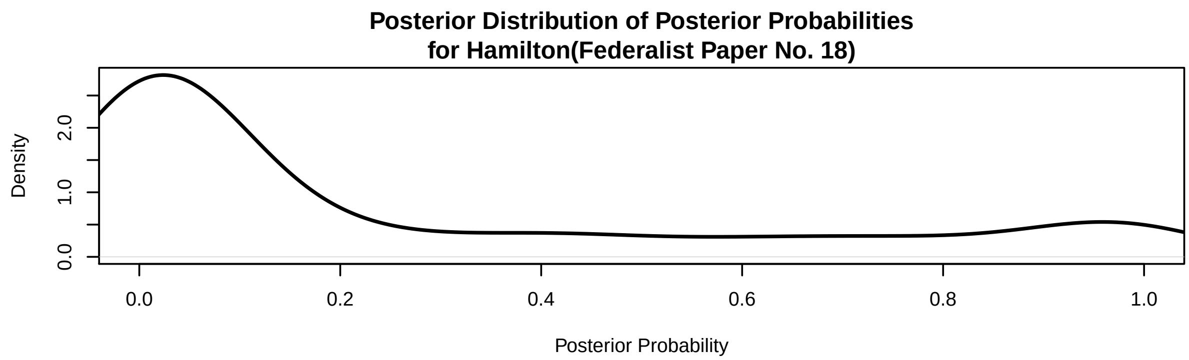


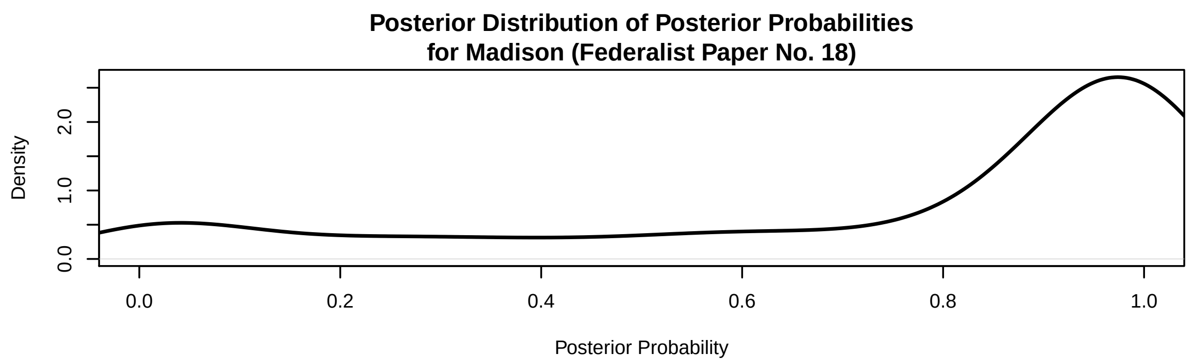


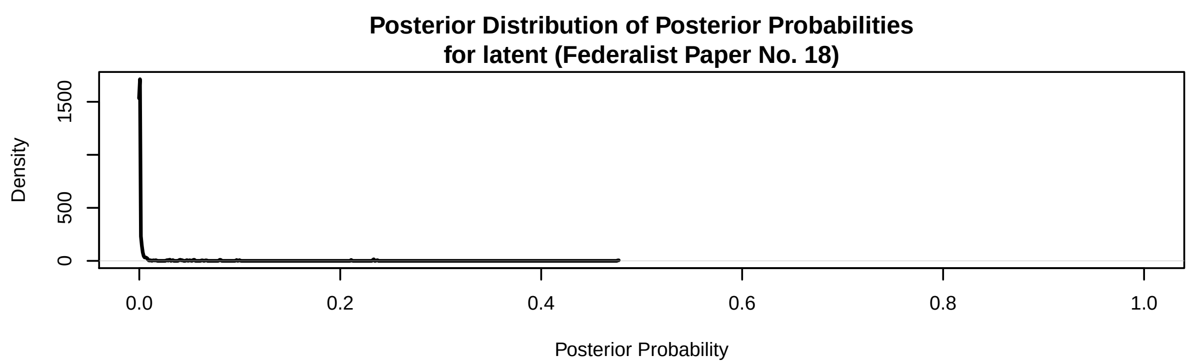


B.6 Joint Federalist Papers with Hamilton and Madison in the training set

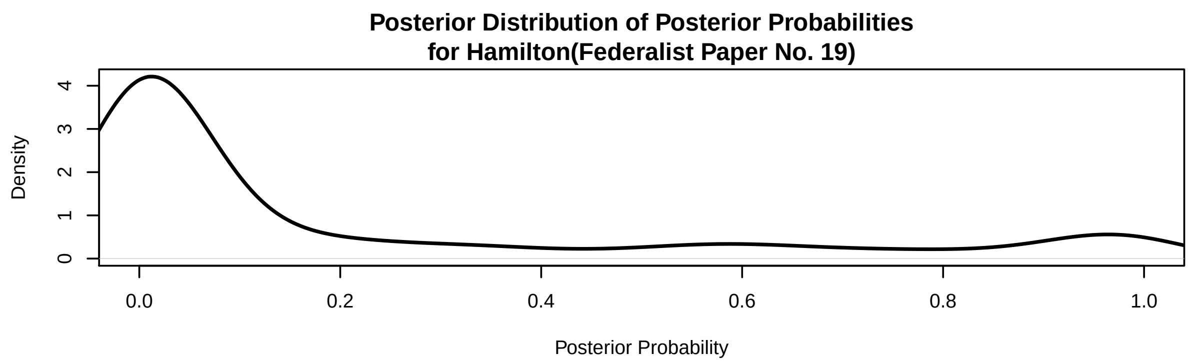
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 18

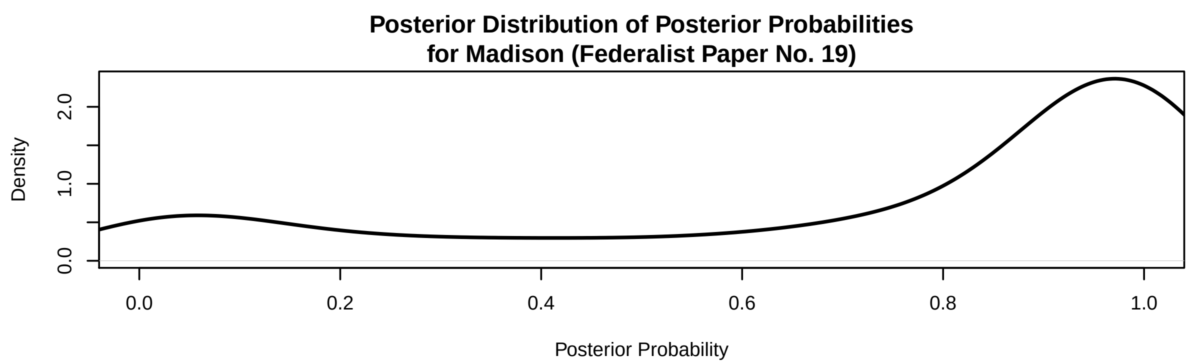


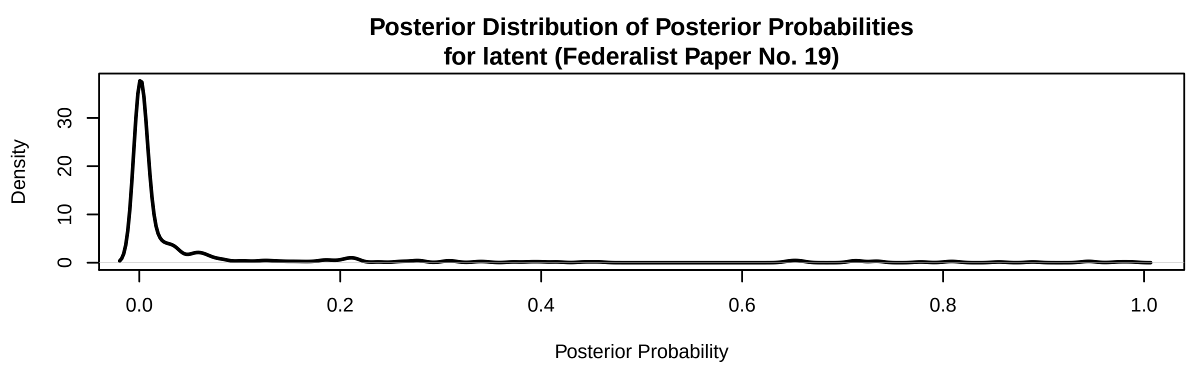




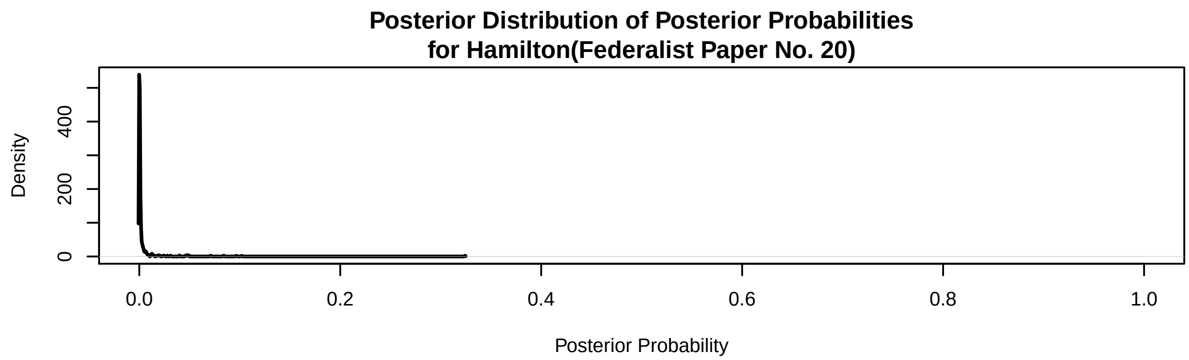
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 19

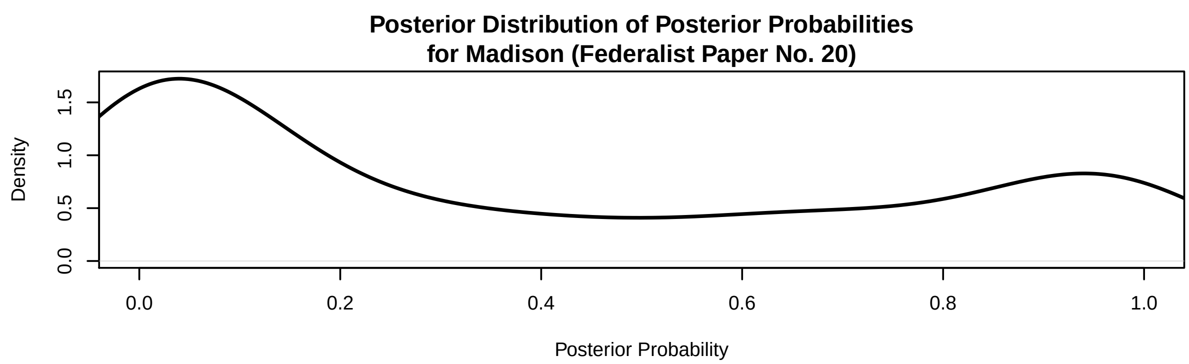


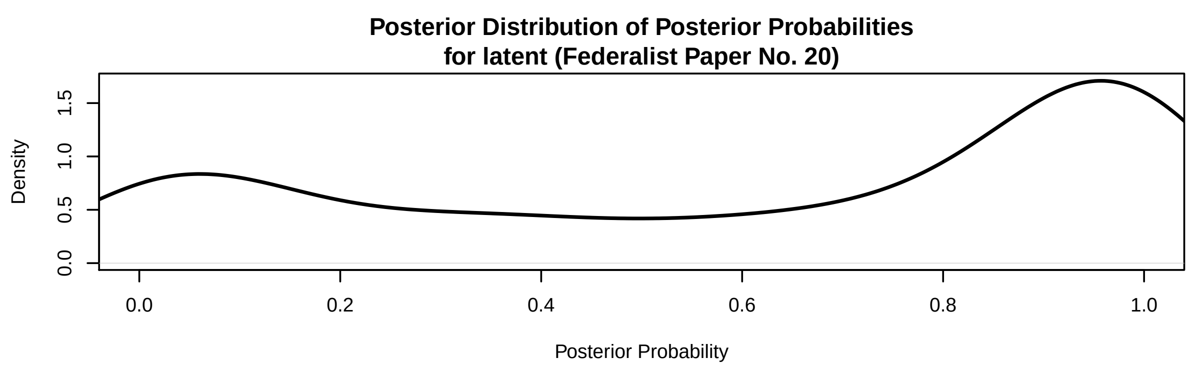




Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 20

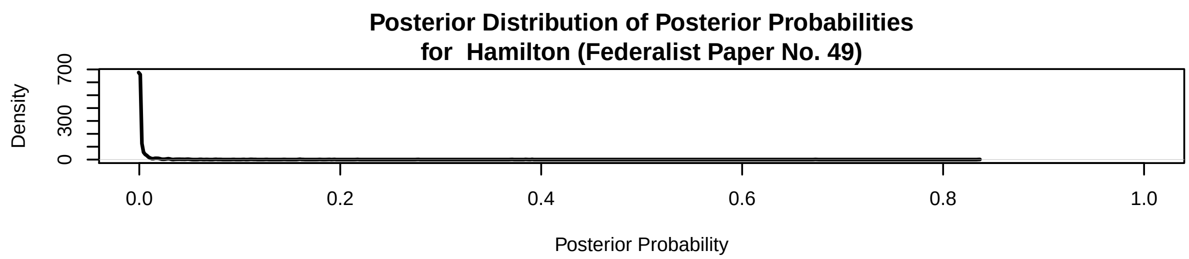


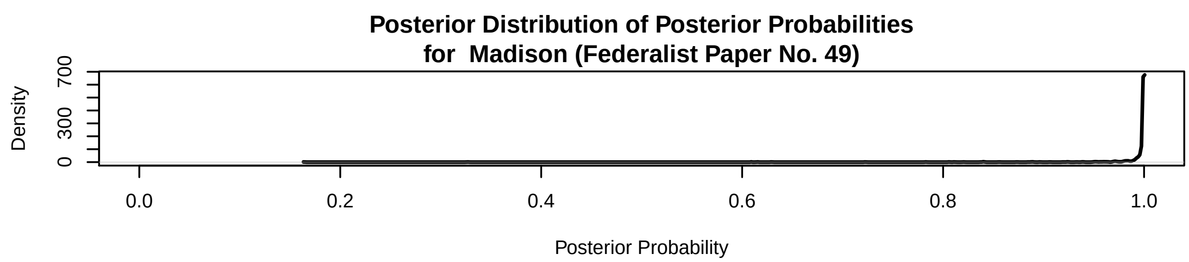


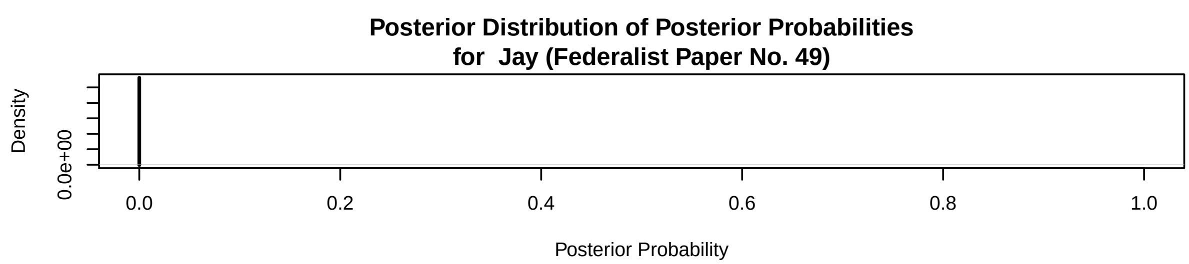


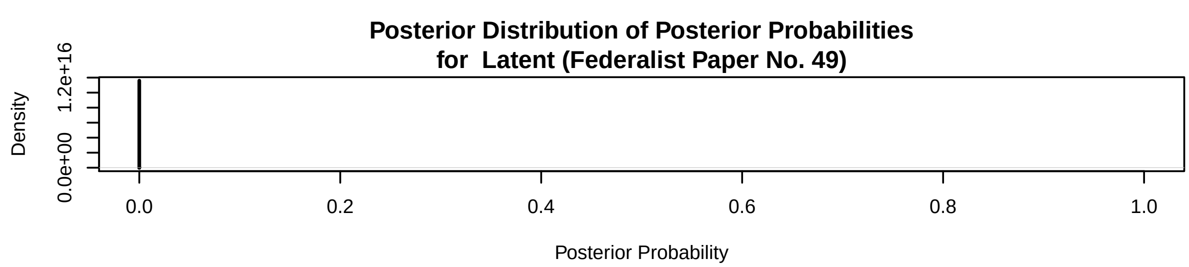
B.7 14 disputed Federalist Papers with Hamilton, Madison and Jay in the training set

Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 49

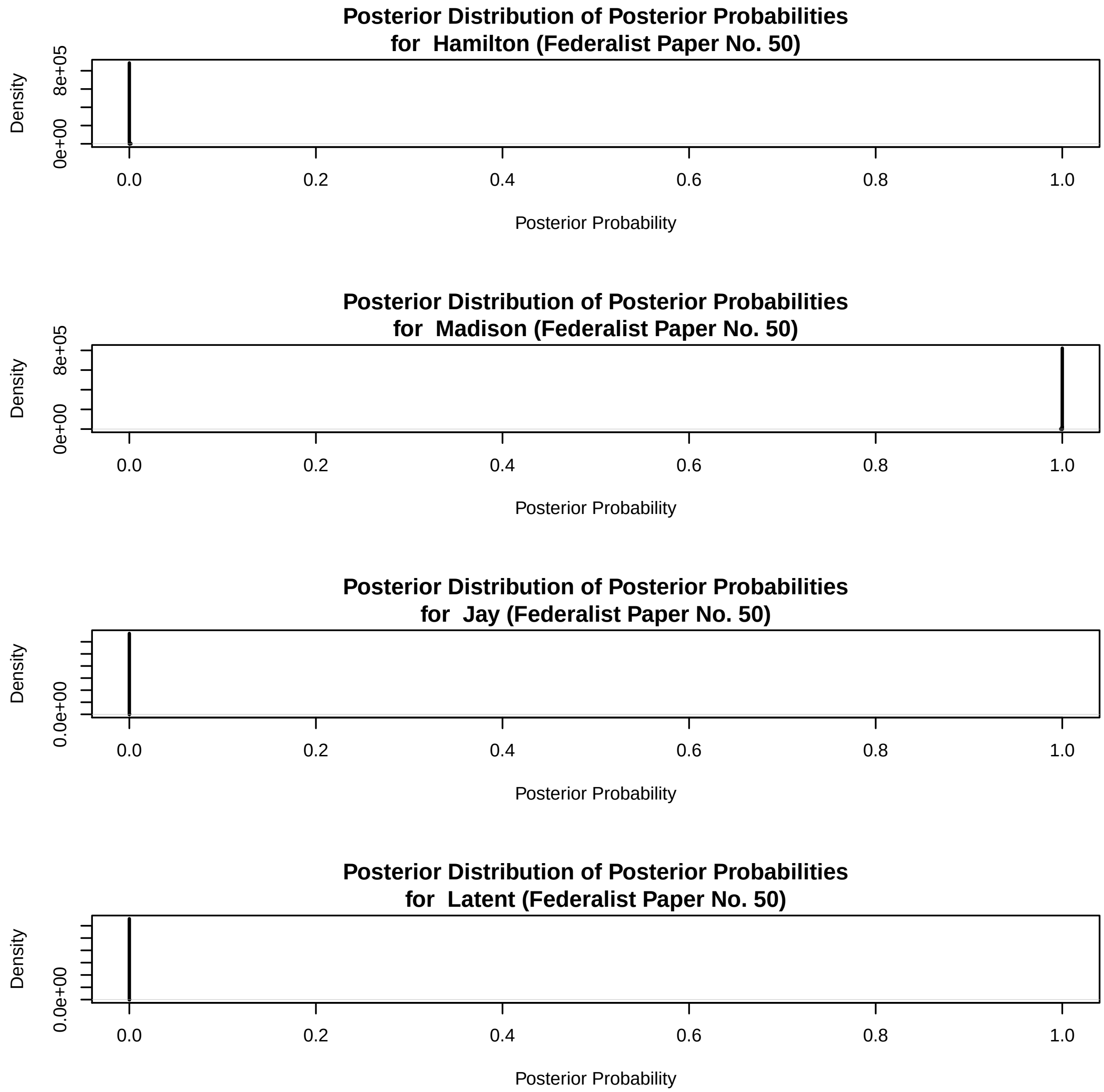




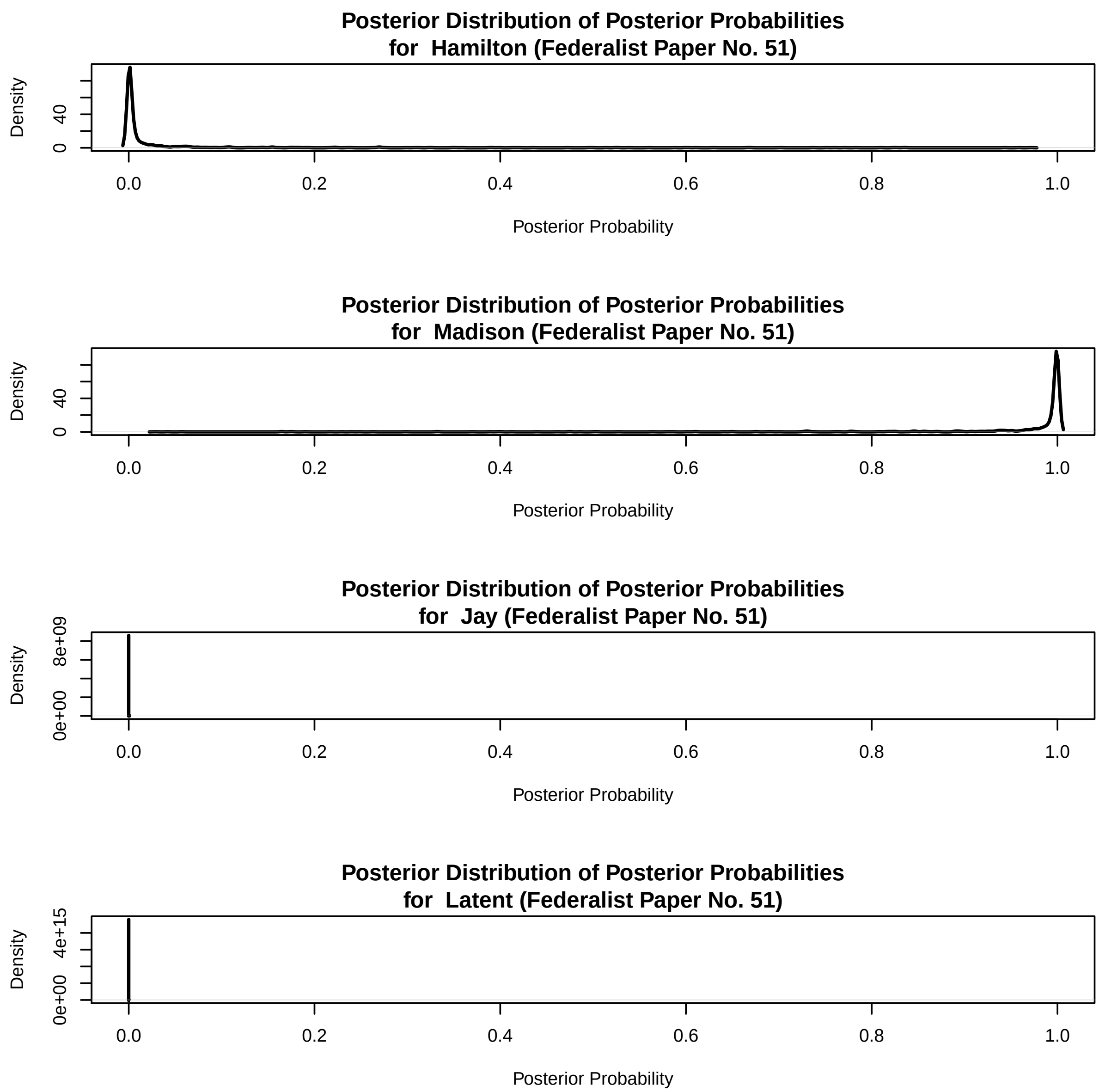




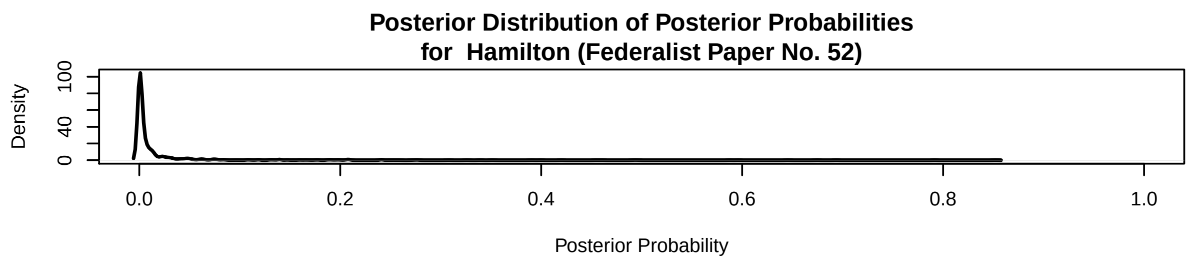
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 50

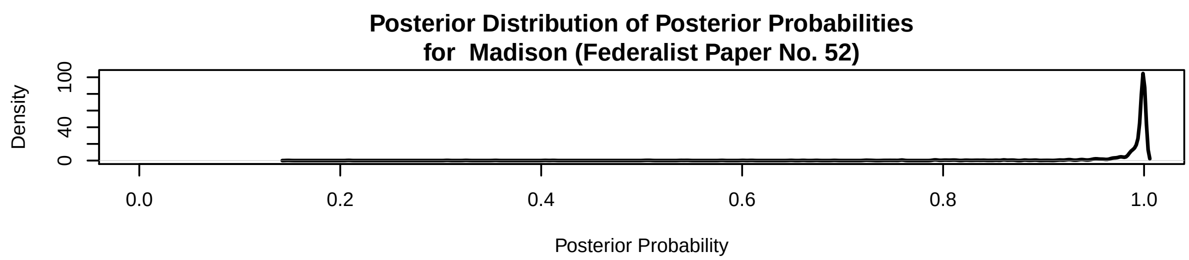


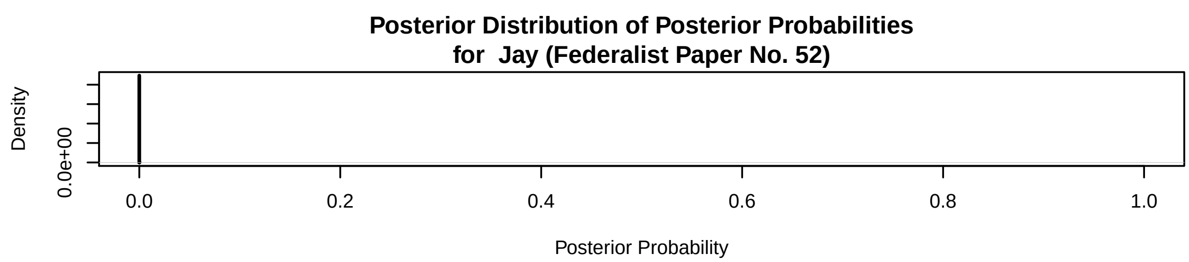
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 51

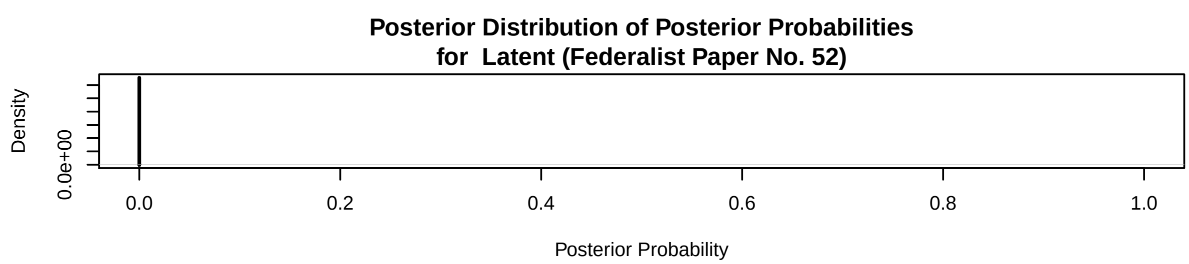


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 52

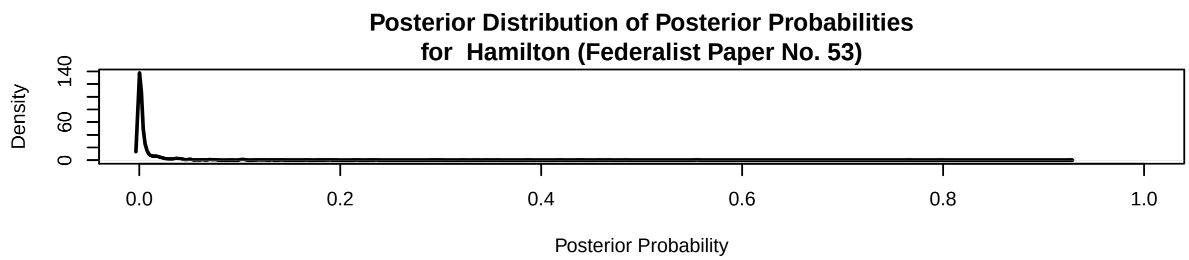


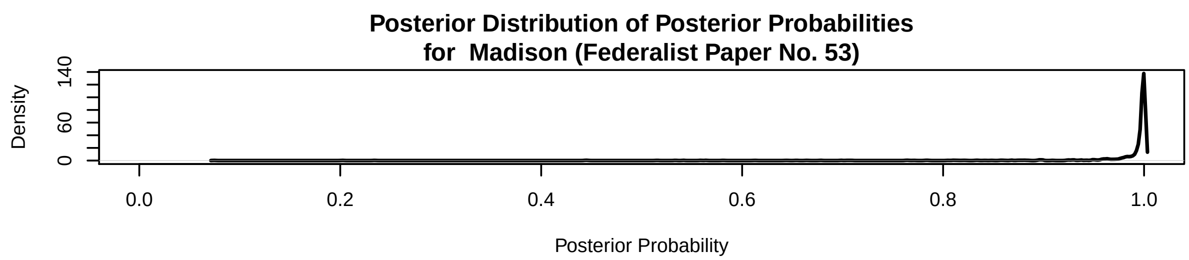


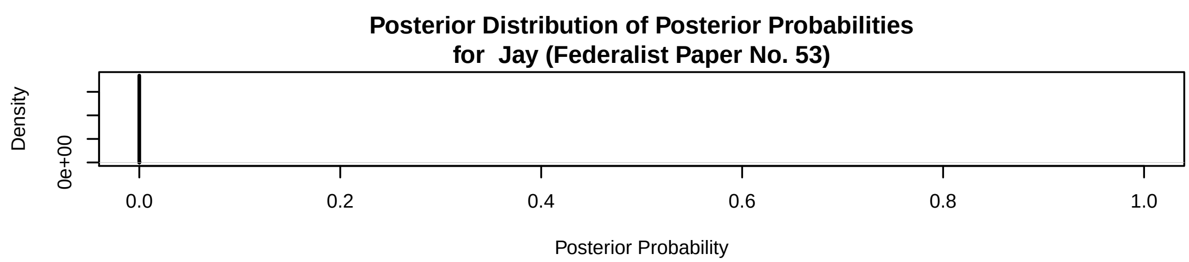


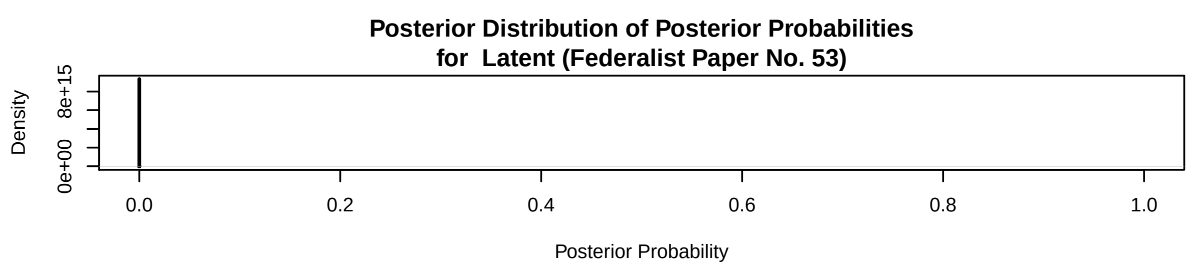


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 53

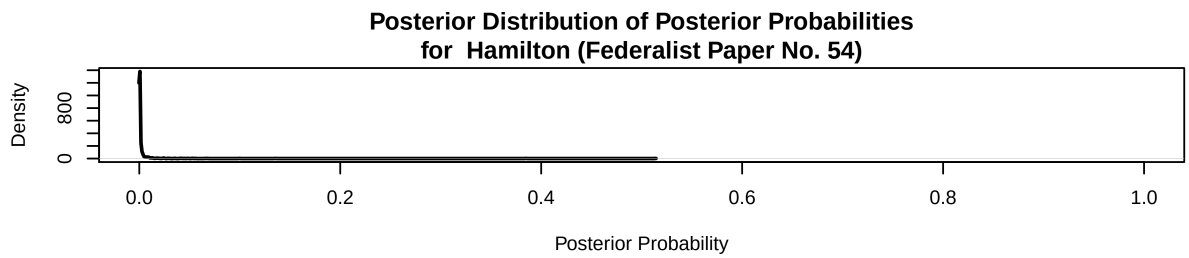


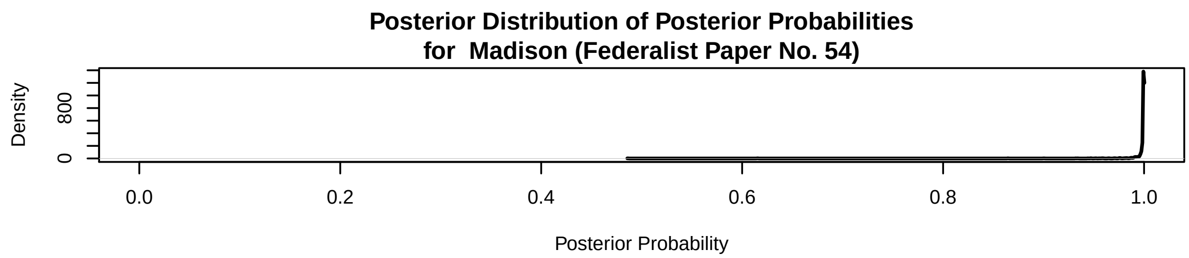


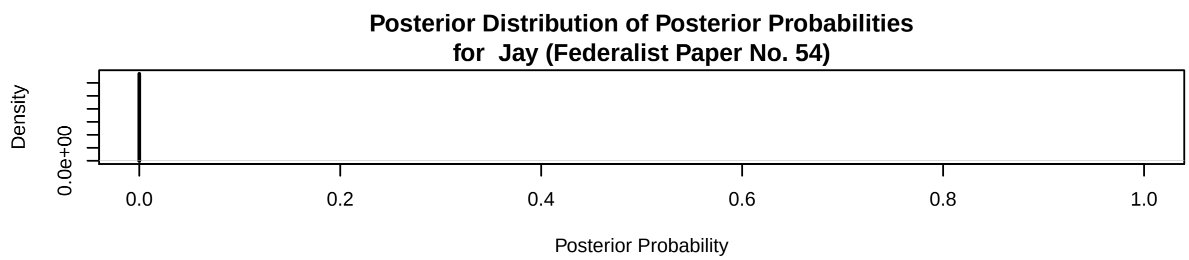


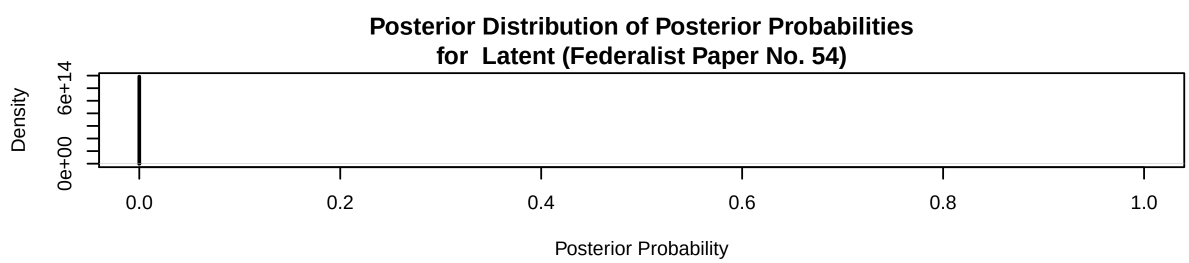


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 54

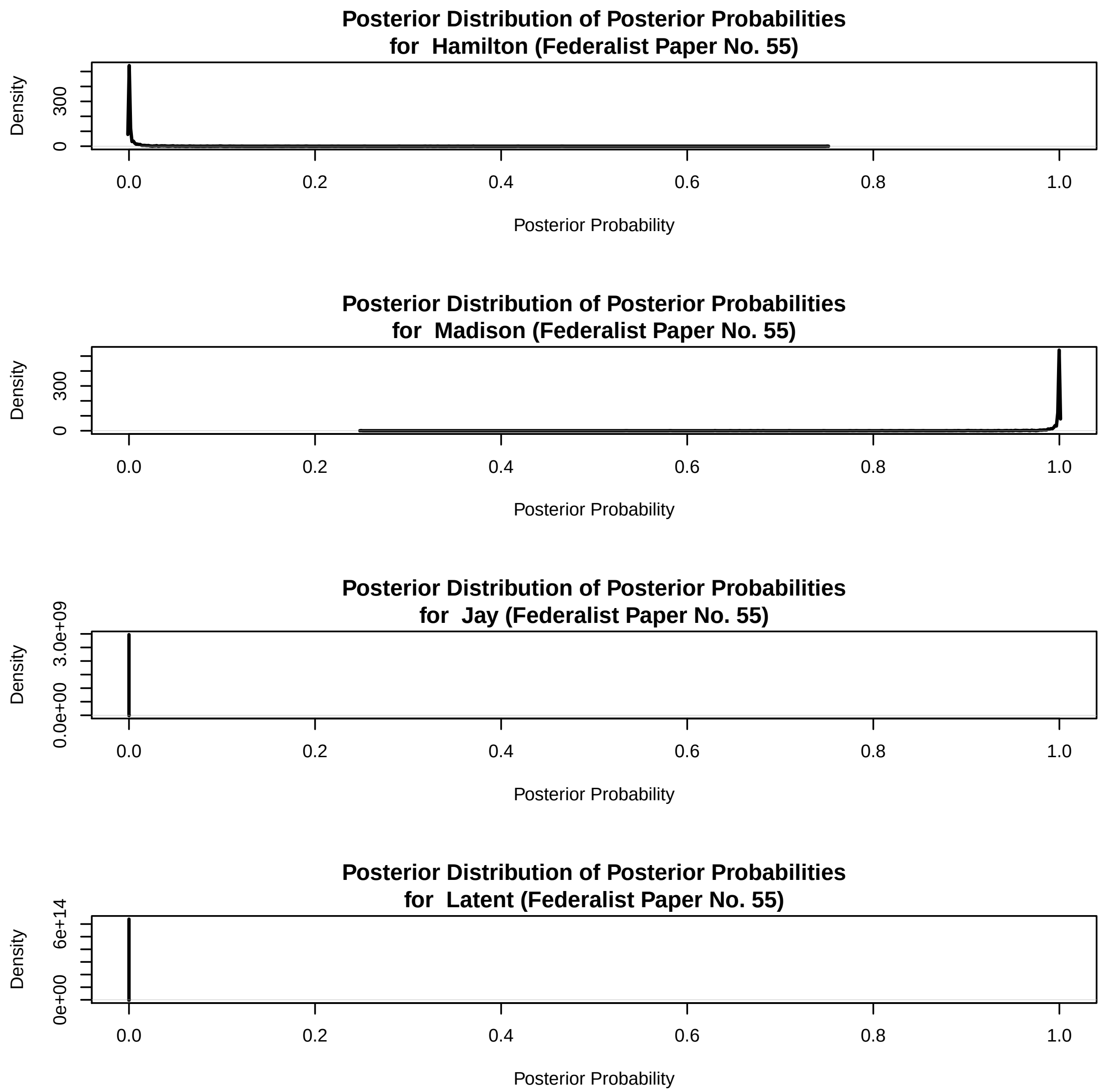




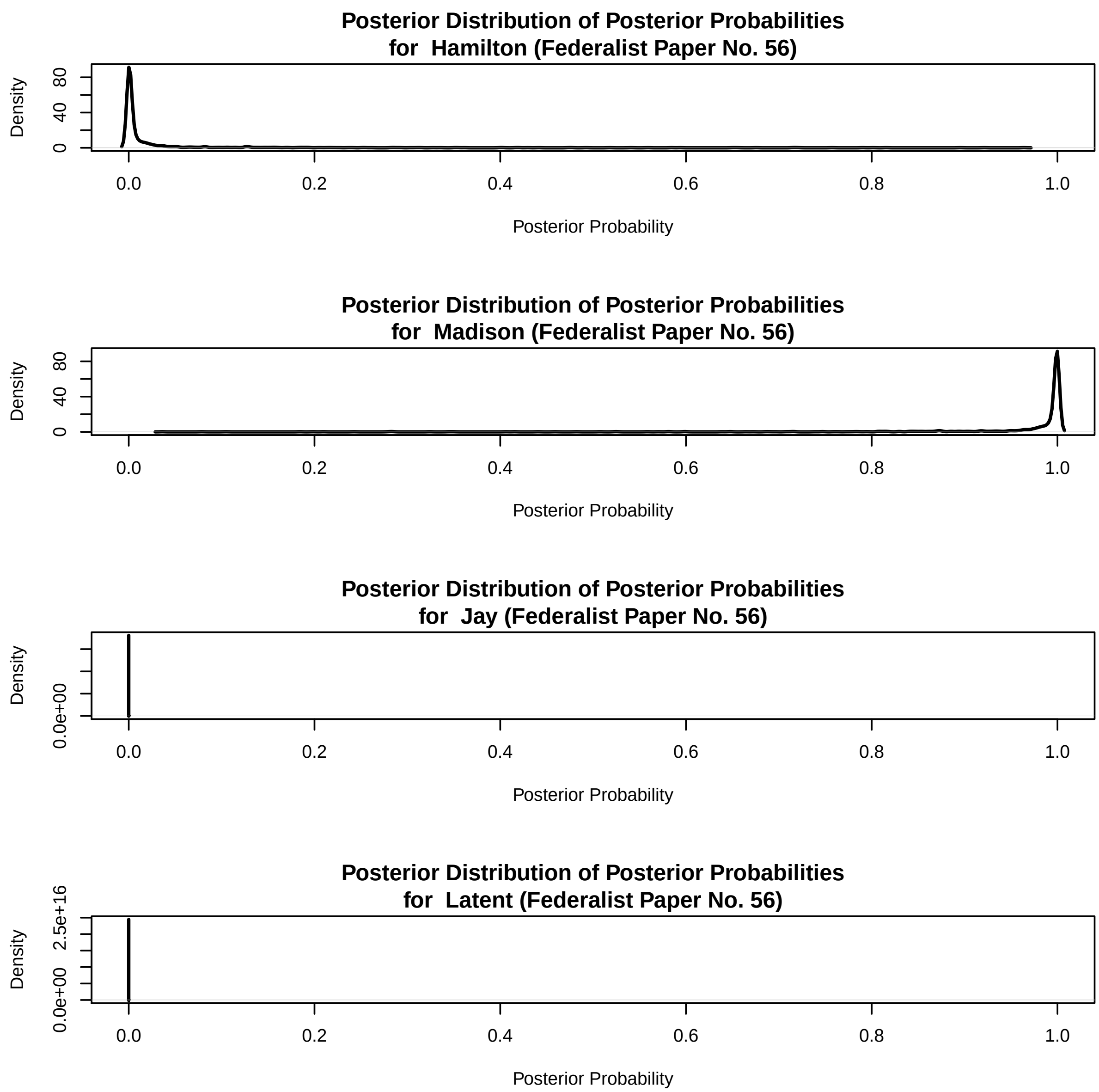




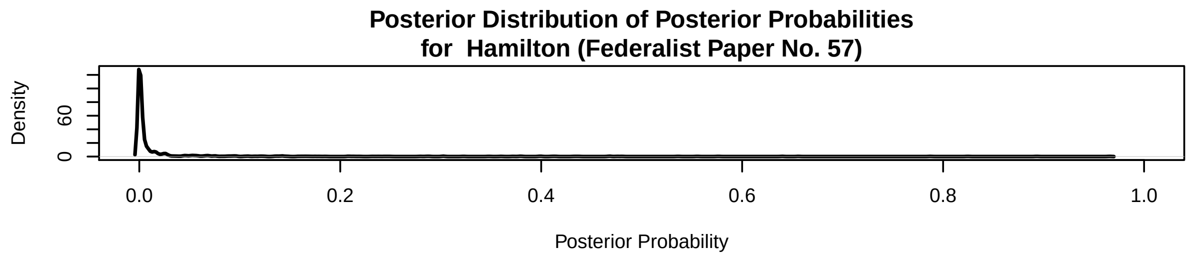
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 55

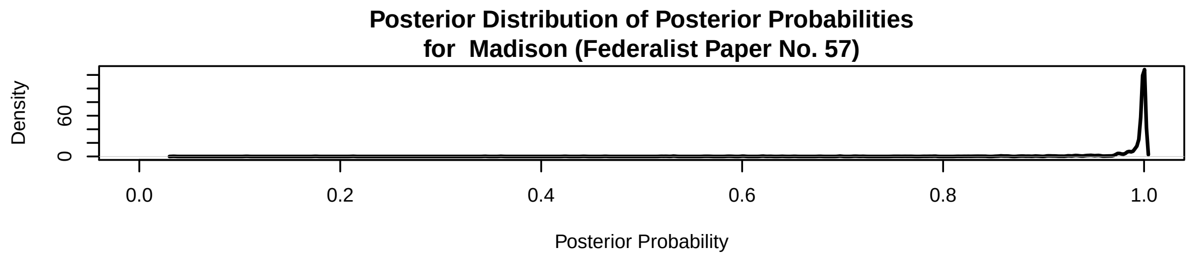


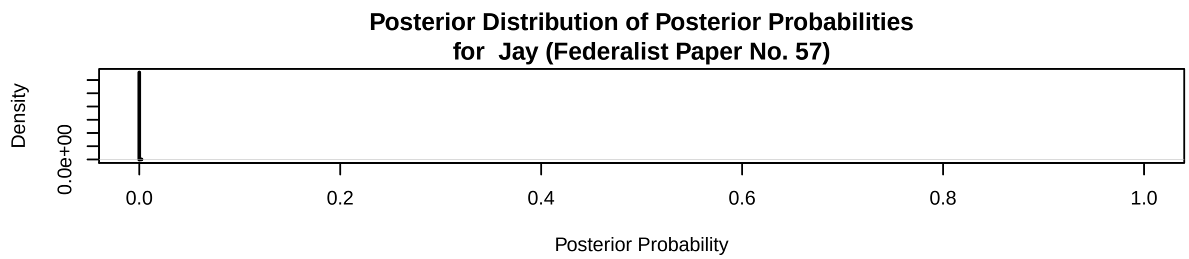
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 56

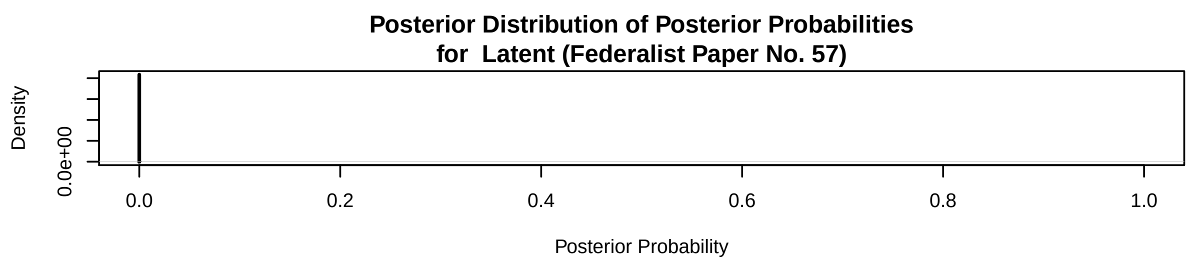


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 57

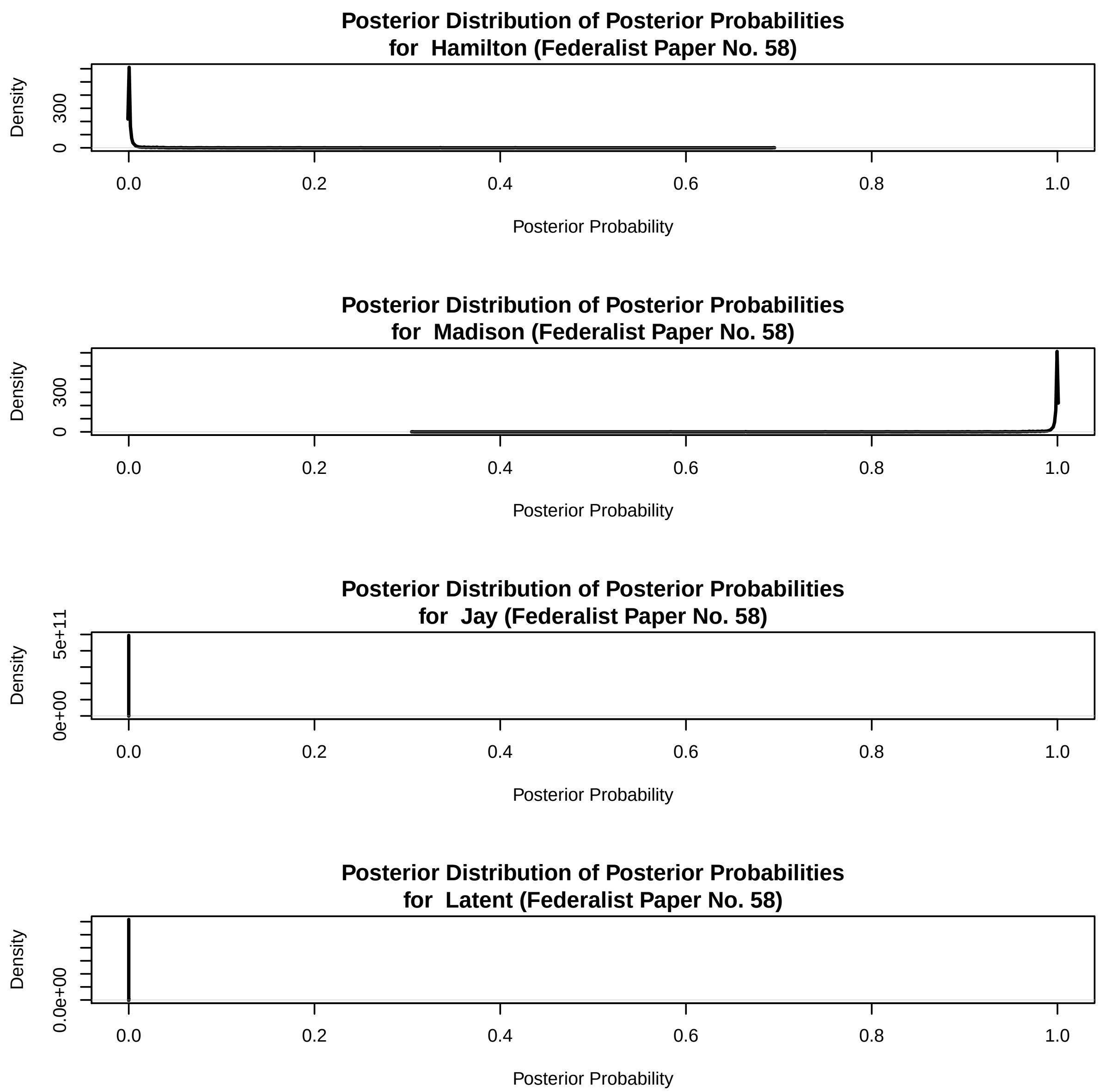




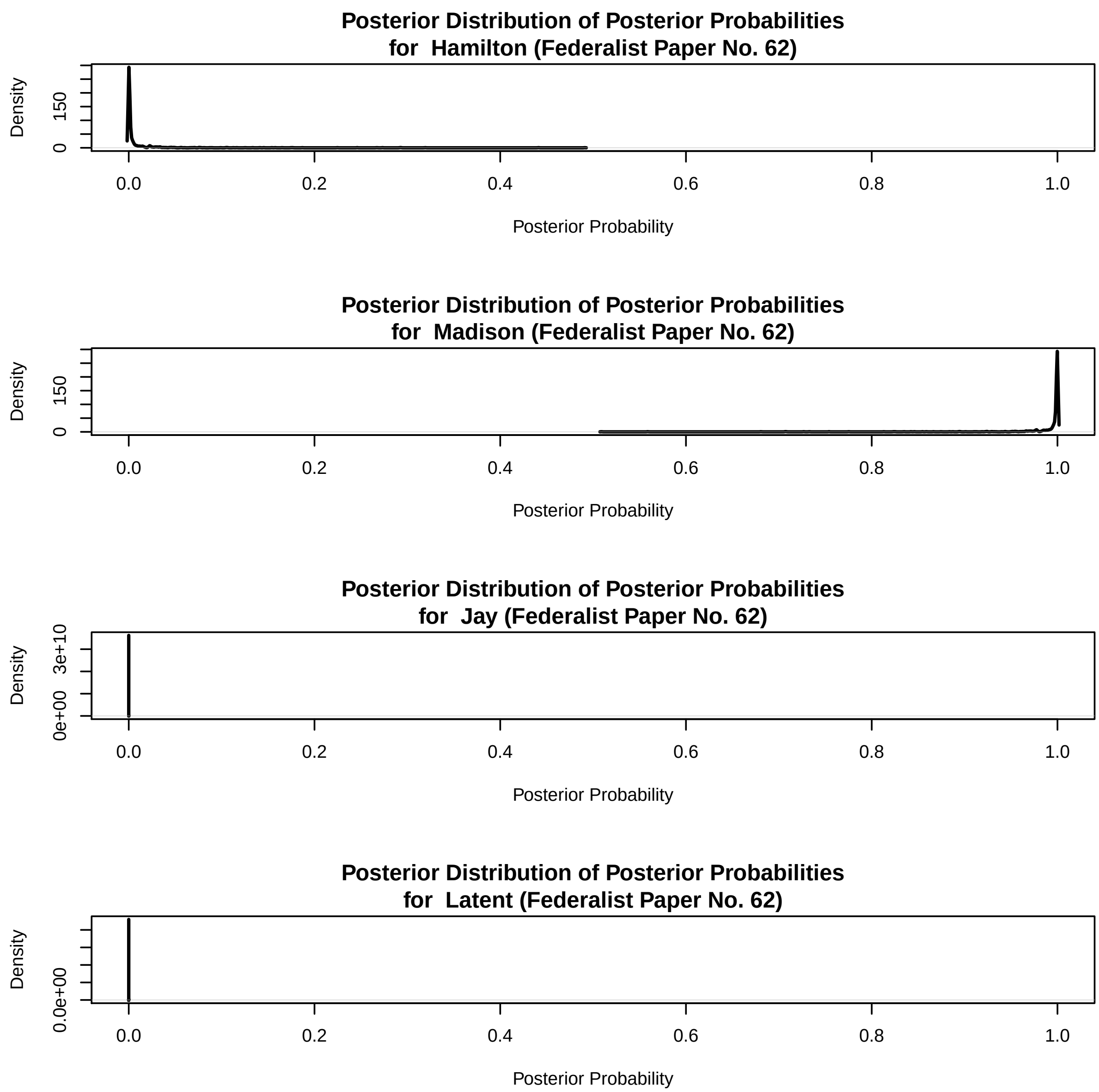




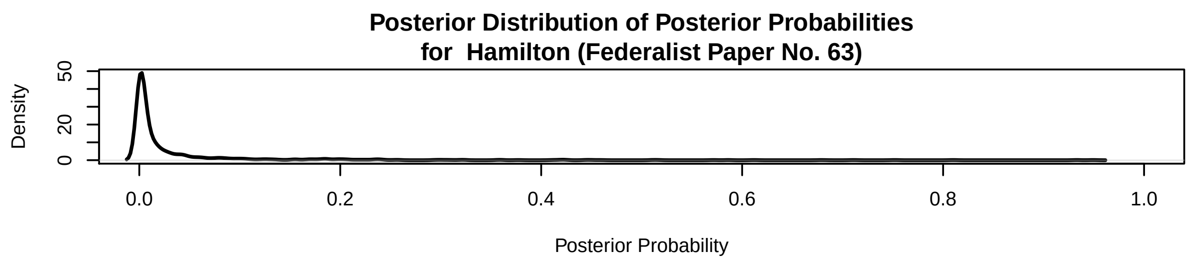
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 58

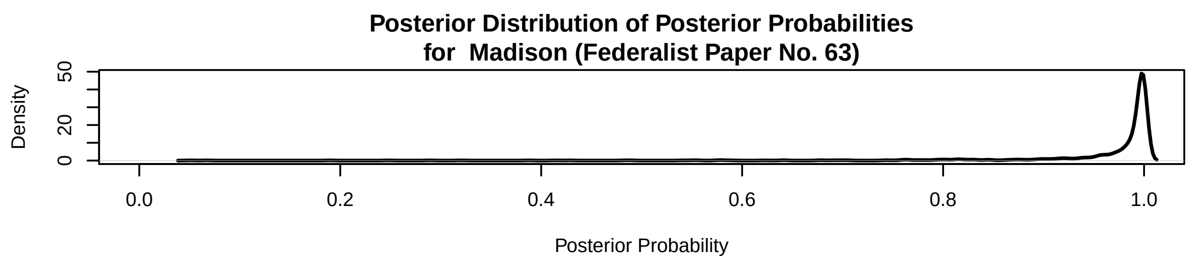


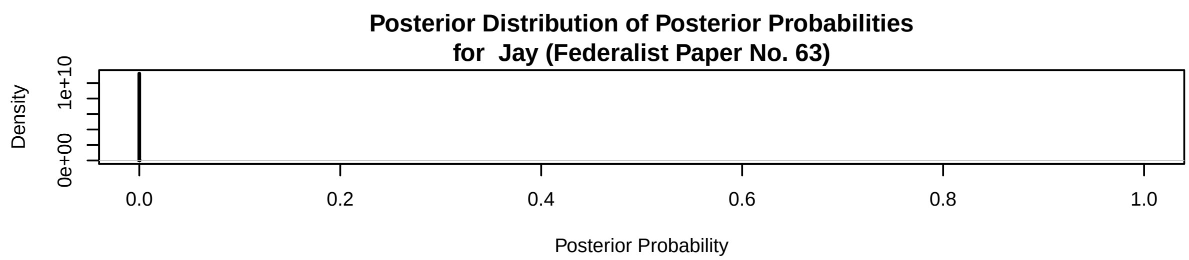
Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 62

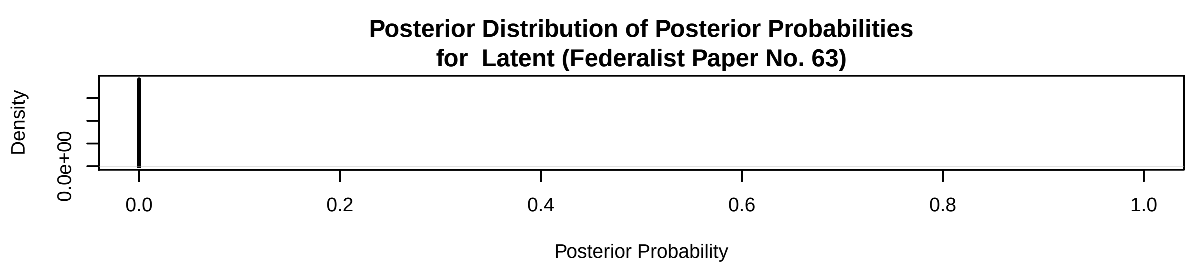


Posterior distribution of Posterior Probabilities for each authors for Federalist Paper No. 63









APPENDIX C

R CODES

```
########################
####  Extended NSC  ####
########################
### Data set up and Configurations
rm(list=ls())
setwd("C:/Users/Tomo/Documents/STATS 698 Research/My Codes/Fed Paper 1")
data<-read.table("completed federalist.txt")

names(data)<-c("file", "author","total","different","total_sentences","R","V2/V",
"K","a_fws","an_fws","and_fws","in_fws","it_fws","it_lws","of_fws","of_2_lws",
"the_fws","the_2_lws","with_2_lws","a_2_lws","a_adj","a_x_and","a_x_of","a_x_a",
"a_x_x_a","and_adj","and_the","and_x_of","and_x_and","and_x_x_and","as_x_as",
"as_x_x_as","be_a","be_to","but_a","by_the","i_am","i_have","i_x_i","i_x_x_i",
"in_a","in_the","of_a","of_the","of_x_and","the_and","the_of","the_in","the_to",
"the_x_and","the_x_the","the_x_x_the","to_be","to_the","to_x_to","to_x_x_to",
"you_x_you","you_x_x_you","verb_to_verb","an_a","any_all","no_not","up_upon",
"with_without","a","about","according","after","again","against","all","also",
"although","always","am","among","an","and","any","apt","are","as","at","away",
"back","be","because","been","before","behold","between","both","but","by",
"came","can","commonly","consequently","considerably","could","direction","do",
"don't","down","enough","even","ever","every","exceeding","for","forth","from",
"had","has","have","he","her","here","herself","him","his","how","i","if","in",
"innovation","into","is","it","its","just","kind","know","language","lest",
"little","matter","may","me","might","more","must","my","nay","nevertheless",
"no","not","notwithstanding","now","O","of","on","one","only","or","out","over",
"particularly","pass","probability","said","she","should","so","some","such",
"than","that","the","their","them","then","there","thereby","therefore",
"thereof","these","they","this","those","though","through","thus","to",
"towards","until","unto","up","upon","verily","very","vigorous","was","well",
"were","what","whatsoever","when","where","whereby","wherefore","whether",
"which","while","whilst","who","why","will","with","without","work","would",
"yea","you","your")

all<-cbind(data$file,data$author,data$total,data$a,data$after,
data$again,data$against,data$all,data$among,
data$an,data$and,data$are,data$as,data$at,data$away,data$be,
data$because,data$been,data$before,data$but,data$by,data$came,
data$children,data$come,data$day,data$did,data$do,data$down,
data$earth,data$even,data$every,data$father,data$for.,data$forth,
data$from,data$go,data$great,data$had,data$hand,data$have,
data$he,data$her,data$him,data$his,data$i,data$if.,data$in.,
data$into,data$is,data$it,data$king,data$know,data$land,data$made,
data$man,data$many,data$may,data$me,data$men,data$might,data$more,
data$my,data$no,data$not,data$now,data$O,data$of,data$on,
data$one,data$or,data$our,data$out,data$over,data$pass,data$people,
data$power,data$said,data$say,data$shall,data$should,data$so,
data$son,data$that,data$the,data$their,data$them,data$then,
data$there,data$therefore,data$these,data$they,data$things,
data$this,data$those,data$thus,data$time,data$to,data$up,data$upon,
data$us,data$was,data$we,data$were,data$which,data$who,data$will,
data$with,data$words,data$would,data$you,data$your)

all[,-c(1,2,3)]<-all[,-c(1,2,3)]/all[,3]

### (1)Disputed, (2)Hamilton, (3)Jay, (4)Joint, (5)Madison.
```

```
Madison.train<-all[all[,2]==5,]
Jay.train<-all[all[,2]==3,]
Hamilton<-all[all[,2]==2,]
set.seed(698)
test.index<-sample(1:51,51/2)
Hamilton.test<-Hamilton[test.index,]
Hamilton.train<-Hamilton[-test.index,]

xx<-rbind(Hamilton.train,Madison.train,Jay.train)[,-c(1,3)]
col<-c(rep(1,sum(xx[,1]==2)),rep(2,sum(xx[,1]==5)),rep(3,sum(xx[,1]==3)))
x<-cbind(col,xx)[,-2]
y<-Hamilton.test[,-c(1,2,3)]
#y<-cbind(1,yy)[,-2]
m<-length(unique(x[,1]))
prior<-rep(1/(m+1),(m+1))
thresh<-1
lambda<-3
r<-ncol(x)-1
n<-nrow(x)

### Obtain the Shrunken Centroids
xbarij<-ssij<-ni<-matrix(0,nrow=m,ncol=r)
row<-0

for(i in 1:m){
  while((x[row+1])==i){
    row<-row+1
    for(jj in 2:(r+1)){
      j<-jj-1
      xbarij[i,j]=xbarij[i,j]+x[row,jj]
      ssij[i,j]=ssij[i,j]+x[row,jj]^2
      ni[i,j]= ni[i,j]+1
    }
  }
}

xbarij<-xbarij/ni
xbarj<-t(t(xbarij)%*%rep(1,m))/rep(m,r)
s2j<-t(t(ssij-(xbarij^2)*ni)%*%(rep(1,m)))/rep((n-m),r)
sj<-sqrt(s2j)

sjsort<-sj
sjsort[,rank(sj)]<-sj
mid<-r/2+.5
if(round(mid)==mid){delta=sjsort[1,mid]
}else{delta=.5*(sjsort[1,round(mid)]+sjsort[1,round(mid)+1])}

xtwidij<-xbarij;

for(j in 1:r){
  for(i in 1:m){
    qij<-(sj[j]+delta)*sqrt(1/ni[i,1]+1/n)
    dumt<-(abs((xtwidij[i,j]-xbarj[1,j])/qij)-thresh)
    dumt=max(dumt,0)
    if((xtwidij[i,j]-xbarj[1,j])==0){dumt<-0
    }else{dumt<-dumt*qij*abs(xtwidij[i,j]-xbarj[1,j])/(xtwidij[i,j]-xbarj[1,j])}
    xtwidij[i,j]<-xbarj[1,j]+dumt
  }
}
ntest<-nrow(y)
postprob<-matrix(0,nrow=ntest,ncol=(m+1))

for(j in 1:r){
  if(sj[1,j]==0){sj[1,j]<-0.000001}
```

```
}

ghostdiff<-rep(0,r)
for(i in 1:ntest){
  use<-matrix(1,nrow=m,ncol=1)
  uu<-matrix(y[i,],nrow=1)
  sdiff<-(kronecker(use,uu)-xtwidij)/(kronecker(sj,rep(1,m)))
  for(j in 1:r){
    ghostdiff[j]<-min(max(abs(sdiff[,j])),rep(lambda,m))
  }
  sdiff<-rbind(sdiff,ghostdiff)
  sumdiff=-diag(sdiff%*%t(sdiff))/2

  pptemp<-exp(sumdiff+500)
  denom<-t(pptemp)%*%prior
  pptemp<-(pptemp*prior)/(kronecker(denom,matrix(1,nrow=1,ncol=(m+1))))
  postprob[i,]<-pptemp
}
postprob ## Posterior Probability

maxprob<-NA
author<-NA
for(i in 1:dim(postprob)[1]){
maxprob[i]<-max(postprob[i,])
author[i]<-which.max(postprob[i,])
}
maxprob
author

chapter<-1:ntest
wrong<-1-(c(sum(author==1)/ntest))
wrong

cbind(postprob,maxprob,author,chapter)
write.table(author,"m1.txt")

####################################
####  NSC without fudge factor  ####
####################################
### Data set up and Configurations
rm(list=ls())
setwd("C:/Users/Tomo/Documents/STATS 698 Research/My Codes/Fed Paper 1")
data<-read.table("completed federalist.txt")

names(data)<-c("file", "author","total","different","total_sentences","R","V2/V",
"K","a_fws","an_fws","and_fws","in_fws","it_fws","it_lws","of_fws","of_2_lws",
"the_fws","the_2_lws","with_2_lws","a_2_lws","a_adj","a_x_and","a_x_of","a_x_a",
"a_x_x_a","and_adj","and_the","and_x_of","and_x_and","and_x_x_and","as_x_as",
"as_x_x_as","be_a","be_to","but_a","by_the","i_am","i_have","i_x_i","i_x_x_i",
"in_a","in_the","of_a","of_the","of_x_and","the_and","the_of","the_in","the_to",
"the_x_and","the_x_the","the_x_x_the","to_be","to_the","to_x_to","to_x_x_to",
"you_x_you","you_x_x_you","verb_to_verb","an_a","any_all","no_not","up_upon",
"with_without","a","about","according","after","again","against","all","also",
"although","always","am","among","an","and","any","apt","are","as","at","away",
"back","be","because","been","before","behold","between","both","but","by",
"came","can","commonly","consequently","considerably","could","direction","do",
"don't","down","enough","even","ever","every","exceeding","for","forth","from",
"had","has","have","he","her","here","herself","him","his","how","i","if","in",
"innovation","into","is","it","its","just","kind","know","language","lest",
"little","matter","may","me","might","more","must","my","nay","nevertheless",
"no","not","notwithstanding","now","O","of","on","one","only","or","out","over",
"particularly","pass","probability","said","she","should","so","some","such",
"than","that","the","their","them","then","there","thereby","therefore",
"thereof","these","they","this","those","though","through","thus","to",
```

```
"towards","until","unto","up","upon","verily","very","vigorous","was","well",
"were","what","whatsoever","when","where","whereby","wherefore","whether",
"which","while","whilst","who","why","will","with","without","work","would",
"yea","you","your")

all<-cbind(data$file,data$author,data$total,data$a,data$after,
data$again,data$against,data$all,data$among,
data$an,data$and,data$are,data$as,data$at,data$away,data$be,
data$because,data$been,data$before,data$but,data$by,data$came,
data$children,data$come,data$day,data$did,data$do,data$down,
data$earth,data$even,data$every,data$father,data$for.,data$forth,
data$from,data$go,data$great,data$had,data$hand,data$have,
data$he,data$her,data$him,data$his,data$i,data$if.,data$in.,
data$into,data$is,data$it,data$king,data$know,data$land,data$made,
data$man,data$many,data$may,data$me,data$men,data$might,data$more,
data$my,data$no,data$not,data$now,data$O,data$of,data$on,
data$one,data$or,data$our,data$out,data$over,data$pass,data$people,
data$power,data$said,data$say,data$shall,data$should,data$so,
data$son,data$that,data$the,data$their,data$them,data$then,
data$there,data$therefore,data$these,data$they,data$things,
data$this,data$those,data$thus,data$time,data$to,data$up,data$upon,
data$us,data$was,data$we,data$were,data$which,data$who,data$will,
data$with,data$words,data$would,data$you,data$your)

all[,-c(1,2,3)]<-all[,-c(1,2,3)]/all[,3]

### (1)Disputed, (2)Hamilton, (3)Jay, (4)Joint, (5)Madison.
Madison.train<-all[all[,2]==5,]
Jay.train<-all[all[,2]==3,]
Hamilton<-all[all[,2]==2,]
set.seed(698)
test.index<-sample(1:51,51/2)
Hamilton.test<-Hamilton[test.index,]
Hamilton.train<-Hamilton[-test.index,]

xx<-rbind(Hamilton.train,Madison.train,Jay.train)[,-c(1,3)]
col<-c(rep(1,sum(xx[,1]==2)),rep(2,sum(xx[,1]==5)),rep(3,sum(xx[,1]==3)))
x<-cbind(col,xx)[,-2]
y<-Hamilton.test[,-c(1,2,3)]
#y<-cbind(1,yy)[,-2]
m<-length(unique(x[,1]))
prior<-rep(1/(m+1),(m+1))
thresh<-1.2
lambda<-2
r<-ncol(x)-1
n<-nrow(x)



xbarij<-ssij<-ni<-matrix(0,nrow=m,ncol=r)
row<-0

for(i in 1:m){
  while((x[row+1])==i){
    row<-row+1
    for(jj in 2:(r+1)){
      j<-jj-1
      xbarij[i,j]=xbarij[i,j]+x[row,jj]
      ssij[i,j]=ssij[i,j]+x[row,jj]^2
      ni[i,j]= ni[i,j]+1
    }
  }
}
```

```
xbarij<-xbarij/ni
xbarj<-t(t(xbarij)%*%rep(1,m))/rep(m,r)
s2j<-t(t(ssij-(xbarij^2)*ni)%*%(rep(1,m)))/rep((n-m),r)
sj<-sqrt(s2j)

xtwidij<-xbarij;

for(j in 1:r){
  for(i in 1:m){
    qij<-(sj[j])*sqrt(1/ni[i,1]+1/n)
    dumt<-(abs((xtwidij[i,j]-xbarj[1,j])/qij)-thresh)
    dumt=max(dumt,0)
    if((xtwidij[i,j]-xbarj[1,j])==0){dumt<-0
    }else{dumt<-dumt*qij*abs(xtwidij[i,j]-xbarj[1,j])/(xtwidij[i,j]-xbarj[1,j])}
    xtwidij[i,j]<-xbarj[1,j]+dumt
  }
}
ntest<-nrow(y)
postprob<-matrix(0,nrow=ntest,ncol=(m+1))

for(j in 1:r){
  if(sj[1,j]==0){sj[1,j]<-0.000001}
}

ghostdiff<-rep(0,r)
for(i in 1:ntest){
  use<-matrix(1,nrow=m,ncol=1)
  uu<-matrix(y[i,],nrow=1)
  sdiff<-(kronecker(use,uu)-xtwidij)/(kronecker(sj,rep(1,m)))
  for(j in 1:r){
    ghostdiff[j]<-min(max(abs(sdiff[,j])),rep(lambda,m))
  }
  sdiff<-rbind(sdiff,ghostdiff)
  sumdiff=-diag(sdiff%*%t(sdiff))/2

  pptemp<-exp(sumdiff+500)
  denom<-t(pptemp)%*%prior
  pptemp<-(pptemp*prior)/(kronecker(denom,matrix(1,nrow=1,ncol=(m+1))))
  postprob[i,]<-pptemp
}
postprob

maxprob<-NA
author<-NA
for(i in 1:dim(postprob)[1]){
maxprob[i]<-max(postprob[i,])
author[i]<-which.max(postprob[i,])
}
maxprob
author

chapter<-1:ntest
wrong<-1-(c(sum(author==1)/ntest))
wrong

cbind(postprob,maxprob,author,chapter)
write.table(author,"m1.txt")

##################################
####  NSC withoutt shrinkage  ####
##################################
### Data set up and Configurations
rm(list=ls())
setwd("C:/Users/Tomo/Documents/STATS 698 Research/My Codes/Fed Paper 1")
```

```
data<-read.table("completed federalist.txt")

names(data)<-c("file", "author","total","different","total_sentences","R","V2/V",
"K","a_fws","an_fws","and_fws","in_fws","it_fws","it_lws","of_fws","of_2_lws",
"the_fws","the_2_lws","with_2_lws","a_2_lws","a_adj","a_x_and","a_x_of","a_x_a",
"a_x_x_a","and_adj","and_the","and_x_of","and_x_and","and_x_x_and","as_x_as",
"as_x_x_as","be_a","be_to","but_a","by_the","i_am","i_have","i_x_i","i_x_x_i",
"in_a","in_the","of_a","of_the","of_x_and","the_and","the_of","the_in","the_to",
"the_x_and","the_x_the","the_x_x_the","to_be","to_the","to_x_to","to_x_x_to",
"you_x_you","you_x_x_you","verb_to_verb","an_a","any_all","no_not","up_upon",
"with_without","a","about","according","after","again","against","all","also",
"although","always","am","among","an","and","any","apt","are","as","at","away",
"back","be","because","been","before","behold","between","both","but","by",
"came","can","commonly","consequently","considerably","could","direction","do",
"don't","down","enough","even","ever","every","exceeding","for","forth","from",
"had","has","have","he","her","here","herself","him","his","how","i","if","in",
"innovation","into","is","it","its","just","kind","know","language","lest",
"little","matter","may","me","might","more","must","my","nay","nevertheless",
"no","not","notwithstanding","now","O","of","on","one","only","or","out","over",
"particularly","pass","probability","said","she","should","so","some","such",
"than","that","the","their","them","then","there","thereby","therefore",
"thereof","these","they","this","those","though","through","thus","to",
"towards","until","unto","up","upon","verily","very","vigorous","was","well",
"were","what","whatsoever","when","where","whereby","wherefore","whether",
"which","while","whilst","who","why","will","with","without","work","would",
"yea","you","your")

all<-cbind(data$file,data$author,data$total,data$a,data$after,
data$again,data$against,data$all,data$among,
data$an,data$and,data$are,data$as,data$at,data$away,data$be,
data$because,data$been,data$before,data$but,data$by,data$came,
data$children,data$come,data$day,data$did,data$do,data$down,
data$earth,data$even,data$every,data$father,data$for.,data$forth,
data$from,data$go,data$great,data$had,data$hand,data$have,
data$he,data$her,data$him,data$his,data$i,data$if.,data$in.,
data$into,data$is,data$it,data$king,data$know,data$land,data$made,
data$man,data$many,data$may,data$me,data$men,data$might,data$more,
data$my,data$no,data$not,data$now,data$O,data$of,data$on,
data$one,data$or,data$our,data$out,data$over,data$pass,data$people,
data$power,data$said,data$say,data$shall,data$should,data$so,
data$son,data$that,data$the,data$their,data$them,data$then,
data$there,data$therefore,data$these,data$they,data$things,
data$this,data$those,data$thus,data$time,data$to,data$up,data$upon,
data$us,data$was,data$we,data$were,data$which,data$who,data$will,
data$with,data$words,data$would,data$you,data$your)

all[,-c(1,2,3)]<-all[,-c(1,2,3)]/all[,3]

### (1)Disputed, (2)Hamilton, (3)Jay, (4)Joint, (5)Madison.
Madison.train<-all[all[,2]==5,]
Jay.train<-all[all[,2]==3,]
Hamilton<-all[all[,2]==2,]
set.seed(698)
test.index<-sample(1:51,51/2)
Hamilton.test<-Hamilton[test.index,]
Hamilton.train<-Hamilton[-test.index,]

xx<-rbind(Hamilton.train,Madison.train,Jay.train)[,-c(1,3)]
col<-c(rep(1,sum(xx[,1]==2)),rep(2,sum(xx[,1]==5)),rep(3,sum(xx[,1]==3)))
x<-cbind(col,xx)[,-2]
y<-Hamilton.test[,-c(1,2,3)]
#y<-cbind(1,yy)[,-2]
m<-length(unique(x[,1]))
prior<-rep(1/(m+1),(m+1))
```

```
thresh<-3
lambda<-1.5
r<-ncol(x)-1
n<-nrow(x)

xbarij<-ssij<-ni<-matrix(0,nrow=m,ncol=r)
row<-0

for(i in 1:m){
  while((x[row+1])==i){
    row<-row+1
    for(jj in 2:(r+1)){
      j<-jj-1
      xbarij[i,j]=xbarij[i,j]+x[row,jj]
      ssij[i,j]=ssij[i,j]+x[row,jj]^2
      ni[i,j]= ni[i,j]+1
    }
  }
}

xbarij<-xbarij/ni
xbarj<-t(t(xbarij)%*%rep(1,m))/rep(m,r)
s2j<-t(t(ssij-(xbarij^2)*ni)%*%(rep(1,m)))/rep((n-m),r)
sj<-sqrt(s2j)

sjsort<-sj
sjsort[,rank(sj)]<-sj
mid<-r/2+.5
if(round(mid)==mid){delta=sjsort[1,mid]
}else{delta=.5*(sjsort[1,round(mid)]+sjsort[1,round(mid)+1])}

xtwidij<-xbarij;

for(j in 1:r){
  for(i in 1:m){
    qij<-(sj[j]+delta)*sqrt(1/ni[i,1]+1/n)
    dumt<-(abs((xtwidij[i,j]-xbarj[1,j])/qij))
    dumt=max(dumt,0)
    if((xtwidij[i,j]-xbarj[1,j])==0){dumt<-0
    }else{dumt<-dumt*qij*abs(xtwidij[i,j]-xbarj[1,j])/(xtwidij[i,j]-xbarj[1,j])}
    xtwidij[i,j]<-xbarj[1,j]+dumt
  }
}
ntest<-nrow(y)
postprob<-matrix(0,nrow=ntest,ncol=(m+1))

for(j in 1:r){
  if(sj[1,j]==0){sj[1,j]<-0.000001}
}

ghostdiff<-rep(0,r)
for(i in 1:ntest){
  use<-matrix(1,nrow=m,ncol=1)
  uu<-matrix(y[i,],nrow=1)
  sdiff<-(kronecker(use,uu)-xtwidij)/(kronecker(sj,rep(1,m)))
  for(j in 1:r){
    ghostdiff[j]<-min(max(abs(sdiff[,j])),rep(lambda,m))
  }
  sdiff<-rbind(sdiff,ghostdiff)
  sumdiff=-diag(sdiff%*%t(sdiff))/2

  pptemp<-exp(sumdiff+500)
  denom<-t(pptemp)%*%prior
  pptemp<-(pptemp*prior)/(kronecker(denom,matrix(1,nrow=1,ncol=(m+1))))
```

```
  postprob[i,]<-pptemp
}
postprob

maxprob<-NA
author<-NA
for(i in 1:dim(postprob)[1]){
maxprob[i]<-max(postprob[i,])
author[i]<-which.max(postprob[i,])
}
maxprob
author

wrong<-1-(c(sum(author==1)/ntest))
wrong

chapter<-1:ntest

#cbind(postprob,maxprob,author,chapter)
write.table(author,"m2.txt")

############################################
####  NSC with class-specific threshold  ####
############################################
### Data set up and Configurations
rm(list=ls())
setwd("C:/Users/Tomo/Documents/STATS 698 Research/My Codes/Fed Paper 1")
data<-read.table("completed federalist.txt")

names(data)<-c("file", "author","total","different","total_sentences","R","V2/V",
"K","a_fws","an_fws","and_fws","in_fws","it_fws","it_lws","of_fws","of_2_lws",
"the_fws","the_2_lws","with_2_lws","a_2_lws","a_adj","a_x_and","a_x_of","a_x_a",
"a_x_x_a","and_adj","and_the","and_x_of","and_x_and","and_x_x_and","as_x_as",
"as_x_x_as","be_a","be_to","but_a","by_the","i_am","i_have","i_x_i","i_x_x_i",
"in_a","in_the","of_a","of_the","of_x_and","the_and","the_of","the_in","the_to",
"the_x_and","the_x_the","the_x_x_the","to_be","to_the","to_x_to","to_x_x_to",
"you_x_you","you_x_x_you","verb_to_verb","an_a","any_all","no_not","up_upon",
"with_without","a","about","according","after","again","against","all","also",
"although","always","am","among","an","and","any","apt","are","as","at","away",
"back","be","because","been","before","behold","between","both","but","by",
"came","can","commonly","consequently","considerably","could","direction","do",
"don't","down","enough","even","ever","every","exceeding","for","forth","from",
"had","has","have","he","her","here","herself","him","his","how","i","if","in",
"innovation","into","is","it","its","just","kind","know","language","lest",
"little","matter","may","me","might","more","must","my","nay","nevertheless",
"no","not","notwithstanding","now","O","of","on","one","only","or","out","over",
"particularly","pass","probability","said","she","should","so","some","such",
"than","that","the","their","them","then","there","thereby","therefore",
"thereof","these","they","this","those","though","through","thus","to",
"towards","until","unto","up","upon","verily","very","vigorous","was","well",
"were","what","whatsoever","when","where","whereby","wherefore","whether",
"which","while","whilst","who","why","will","with","without","work","would",
"yea","you","your")

all<-cbind(data$file,data$author,data$total,data$a,data$after,
data$again,data$against,data$all,data$among,
data$an,data$and,data$are,data$as,data$at,data$away,data$be,
data$because,data$been,data$before,data$but,data$by,data$came,
data$children,data$come,data$day,data$did,data$do,data$down,
data$earth,data$even,data$every,data$father,data$for.,data$forth,
data$from,data$go,data$great,data$had,data$hand,data$have,
data$he,data$her,data$him,data$his,data$i,data$if.,data$in.,
data$into,data$is,data$it,data$king,data$know,data$land,data$made,
data$man,data$many,data$may,data$me,data$men,data$might,data$more,
```

```
data$my,data$no,data$not,data$now,data$O,data$of,data$on,
data$one,data$or,data$our,data$out,data$over,data$pass,data$people,
data$power,data$said,data$say,data$shall,data$should,data$so,
data$son,data$that,data$the,data$their,data$them,data$then,
data$there,data$therefore,data$these,data$they,data$things,
data$this,data$those,data$thus,data$time,data$to,data$up,data$upon,
data$us,data$was,data$we,data$were,data$which,data$who,data$will,
data$with,data$words,data$would,data$you,data$your)

all[,-c(1,2,3)]<-all[,-c(1,2,3)]/all[,3]

### (1)Disputed, (2)Hamilton, (3)Jay, (4)Joint, (5)Madison.
Madison.train<-all[all[,2]==5,]
Jay.train<-all[all[,2]==3,]
Hamilton<-all[all[,2]==2,]
set.seed(698)
test.index<-sample(1:51,51/2)
Hamilton.test<-Hamilton[test.index,]
Hamilton.train<-Hamilton[-test.index,]

xx<-rbind(Hamilton.train,Madison.train,Jay.train)[,-c(1,3)]
col<-c(rep(1,sum(xx[,1]==2)),rep(2,sum(xx[,1]==5)),rep(3,sum(xx[,1]==3)))
x<-cbind(col,xx)[,-2]
y<-Hamilton.test[,-c(1,2,3)]
#y<-cbind(1,yy)[,-2]
m<-length(unique(x[,1]))
prior<-rep(1/(m+1),(m+1))
thresh<-3
lambda<-1.5
r<-ncol(x)-1
n<-nrow(x)

nsc<-function(la,x=x,y=x,b){

prior<-rep(1/(m+1),(m+1))
thresh<-la
lambda<-2
xbarij<-ssij<-ni<-matrix(0,nrow=m,ncol=r)
row<-0

for(i in 1:m){
  while((x[row+1])==i){
    row<-row+1
    for(jj in 2:(r+1)){
      j<-jj-1
      xbarij[i,j]=xbarij[i,j]+x[row,jj]
      ssij[i,j]=ssij[i,j]+x[row,jj]^2
      ni[i,j]= ni[i,j]+1
    }
  }
}

xbarij<-xbarij/ni
xbarj<-t(t(xbarij)%*%rep(1,m))/rep(m,r)
s2j<-t(t(ssij-(xbarij^2)*ni)%*%(rep(1,m)))/rep((n-m),r)
sj<-sqrt(s2j)

sjsort<-sj
sjsort[,rank(sj)]<-sj
mid<-r/2+.5
if(round(mid)==mid){delta=sjsort[1,mid]
}else{delta=.5*(sjsort[1,round(mid)]+sjsort[1,round(mid)+1])}

xtwidij<-xbarij;
```

```
for(j in 1:r){
  for(i in 1:m){
    qij<-(sj[j]+delta)*sqrt(1/ni[i,1]+1/n)
    dumt<-(abs((xtwidij[i,j]-xbarj[1,j])/qij)-thresh[i])
    dumt=max(dumt,0)
    if((xtwidij[i,j]-xbarj[1,j])==0){dumt<-0
    }else{dumt<-dumt*qij*abs(xtwidij[i,j]-xbarj[1,j])/(xtwidij[i,j]-xbarj[1,j])}
    xtwidij[i,j]<-xbarj[1,j]+dumt
  }
}
ntest<-nrow(y)
postprob<-matrix(0,nrow=ntest,ncol=(m+1))

for(j in 1:r){
  if(sj[1,j]==0){sj[1,j]<-0.000001}
}

ghostdiff<-rep(0,r)
for(i in 1:ntest){
  use<-matrix(1,nrow=m,ncol=1)
  uu<-matrix(y[i,],nrow=1)
  sdiff<-(kronecker(use,uu)-xtwidij)/(kronecker(sj,rep(1,m)))
  for(j in 1:r){
    ghostdiff[j]<-min(max(abs(sdiff[,j])),rep(lambda,m))
  }
  sdiff<-rbind(sdiff,ghostdiff)
  sumdiff=-diag(sdiff%*%t(sdiff))/2

  pptemp<-exp(sumdiff+500)
  denom<-t(pptemp)%*%prior
  pptemp<-(pptemp*prior)/(kronecker(denom,matrix(1,nrow=1,ncol=(m+1))))
  postprob[i,]<-pptemp
}
postprob

maxprob<-NA
author<-NA
for(i in 1:dim(postprob)[1]){
maxprob[i]<-max(postprob[i,])
author[i]<-which.max(postprob[i,])
}
maxprob
wrong<-1-(c(sum(author==1)/ntest))

chapter<-1:ntest

postprob<-cbind(postprob,maxprob,author,chapter)

#par(mfrow=c(3,3))
#plot(chapter,postprob[,1],main="Posterior Prob (Author 1)",ylab="Probability",type='h')
#plot(chapter,postprob[,2],main="Posterior Prob (Author 2)",ylab="Probability",type='h')
#plot(chapter,postprob[,3],main="Posterior Prob (Author 3)",ylab="Probability",type='h')
#plot(chapter,postprob[,4],main="Posterior Prob (Author 4)",ylab="Probability",type='h')
#plot(chapter,postprob[,5],main="Posterior Prob (Author 5)",ylab="Probability",type='h')
#plot(chapter,postprob[,6],main="Posterior Prob (Author 6)",ylab="Probability",type='h')
#plot(chapter,postprob[,7],main="Posterior Prob (Author 7)",ylab="Probability",type='h')
return(wrong)
}

### Determine which combination yields the lowest classification
source("http://statistics.byu.edu/faculty/wfc/stat435/COMBINAT.Q")
aa<-seq(0,1,by=0.1)
v<-combn(aa,m)
```

```
dim(v)[2]->ll
ll
misclass<-NA

for(i in 1:ll){
out<-nsc(v[,i],x,y,v[,i])
misclass[i]<-out
cat(i," misclass=",misclass[i],"\n")
}
misclass

select<-which(misclass==min(misclass))
v[,select]


########################
#### The actual results
prior<-rep(1/(m+1),m+1)
thresh<-v[,64]
lambda<-2
xbarij<-ssij<-ni<-matrix(0,nrow=m,ncol=r)
row<-0

for(i in 1:m){
  while((x[row+1])==i){
    row<-row+1
    for(jj in 2:(r+1)){
      j<-jj-1
      xbarij[i,j]=xbarij[i,j]+x[row,jj]
      ssij[i,j]=ssij[i,j]+x[row,jj]^2
      ni[i,j]= ni[i,j]+1
    }
  }
}

xbarij<-xbarij/ni
xbarj<-t(t(xbarij)%*%rep(1,m))/rep(m,r)
s2j<-t(t(ssij-(xbarij^2)*ni)%*%(rep(1,m)))/rep((n-m),r)
sj<-sqrt(s2j)

sjsort<-sj
sjsort[,rank(sj)]<-sj
mid<-r/2+.5
if(round(mid)==mid){delta=sjsort[1,mid]
}else{delta=.5*(sjsort[1,round(mid)]+sjsort[1,round(mid)+1])}

xtwidij<-xbarij;

for(j in 1:r){
  for(i in 1:m){
    qij<-(sj[j]+delta)*sqrt(1/ni[i,1]+1/n)
    dumt<-(abs((xtwidij[i,j]-xbarj[1,j])/qij)-thresh[i])
    dumt=max(dumt,0)
    if((xtwidij[i,j]-xbarj[1,j])==0){dumt<-0
    }else{dumt<-dumt*qij*abs(xtwidij[i,j]-xbarj[1,j])/(xtwidij[i,j]-xbarj[1,j])}
    xtwidij[i,j]<-xbarj[1,j]+dumt
  }
}
ntest<-nrow(y)
postprob<-matrix(0,nrow=ntest,ncol=(m+1))

for(j in 1:r){
  if(sj[1,j]==0){sj[1,j]<-0.000001}
}
```

```
ghostdiff<-rep(0,r)
for(i in 1:ntest){
  use<-matrix(1,nrow=m,ncol=1)
  uu<-matrix(y[i,],nrow=1)
  sdiff<-(kronecker(use,uu)-xtwidij)/(kronecker(sj,rep(1,m)))
  for(j in 1:r){
    ghostdiff[j]<-min(max(abs(sdiff[,j])),rep(lambda,m))
  }
  sdiff<-rbind(sdiff,ghostdiff)
  sumdiff=-diag(sdiff%*%t(sdiff))/2

  pptemp<-exp(sumdiff+500)
  denom<-t(pptemp)%*%prior
  pptemp<-(pptemp*prior)/(kronecker(denom,matrix(1,nrow=1,ncol=(m+1))))
  postprob[i,]<-pptemp
}
postprob

maxprob<-NA
author<-NA
for(i in 1:dim(postprob)[1]){
maxprob[i]<-max(postprob[i,])
author[i]<-which.max(postprob[i,])
}
maxprob
wrong<-1-(c(sum(author==1)/ntest))
wrong

postprob<-cbind(postprob,maxprob,author,chapter)

author
wrong
write.table(author,"m3.txt")

#############################################
####  NSC shrunk in different direction  ####
#############################################
### Data set up and Configurations
rm(list=ls())
setwd("C:/Users/Tomo/Documents/STATS 698 Research/My Codes/Fed Paper 1")
data<-read.table("completed federalist.txt")

names(data)<-c("file", "author","total","different","total_sentences","R","V2/V",
"K","a_fws","an_fws","and_fws","in_fws","it_fws","it_lws","of_fws","of_2_lws",
"the_fws","the_2_lws","with_2_lws","a_2_lws","a_adj","a_x_and","a_x_of","a_x_a",
"a_x_x_a","and_adj","and_the","and_x_of","and_x_and","and_x_x_and","as_x_as",
"as_x_x_as","be_a","be_to","but_a","by_the","i_am","i_have","i_x_i","i_x_x_i",
"in_a","in_the","of_a","of_the","of_x_and","the_and","the_of","the_in","the_to",
"the_x_and","the_x_the","the_x_x_the","to_be","to_the","to_x_to","to_x_x_to",
"you_x_you","you_x_x_you","verb_to_verb","an_a","any_all","no_not","up_upon",
"with_without","a","about","according","after","again","against","all","also",
"although","always","am","among","an","and","any","apt","are","as","at","away",
"back","be","because","been","before","behold","between","both","but","by",
"came","can","commonly","consequently","considerably","could","direction","do",
"don't","down","enough","even","ever","every","exceeding","for","forth","from",
"had","has","have","he","her","here","herself","him","his","how","i","if","in",
"innovation","into","is","it","its","just","kind","know","language","lest",
"little","matter","may","me","might","more","must","my","nay","nevertheless",
"no","not","notwithstanding","now","O","of","on","one","only","or","out","over",
"particularly","pass","probability","said","she","should","so","some","such",
"than","that","the","their","them","then","there","thereby","therefore",
"thereof","these","they","this","those","though","through","thus","to",
"towards","until","unto","up","upon","verily","very","vigorous","was","well",
```

```
"were","what","whatsoever","when","where","whereby","wherefore","whether",
"which","while","whilst","who","why","will","with","without","work","would",
"yea","you","your")

all<-cbind(data$file,data$author,data$total,data$a,data$after,
data$again,data$against,data$all,data$among,
data$an,data$and,data$are,data$as,data$at,data$away,data$be,
data$because,data$been,data$before,data$but,data$by,data$came,
data$children,data$come,data$day,data$did,data$do,data$down,
data$earth,data$even,data$every,data$father,data$for.,data$forth,
data$from,data$go,data$great,data$had,data$hand,data$have,
data$he,data$her,data$him,data$his,data$i,data$if.,data$in.,
data$into,data$is,data$it,data$king,data$know,data$land,data$made,
data$man,data$many,data$may,data$me,data$men,data$might,data$more,
data$my,data$no,data$not,data$now,data$O,data$of,data$on,
data$one,data$or,data$our,data$out,data$over,data$pass,data$people,
data$power,data$said,data$say,data$shall,data$should,data$so,
data$son,data$that,data$the,data$their,data$them,data$then,
data$there,data$therefore,data$these,data$they,data$things,
data$this,data$those,data$thus,data$time,data$to,data$up,data$upon,
data$us,data$was,data$we,data$were,data$which,data$who,data$will,
data$with,data$words,data$would,data$you,data$your)

all[,-c(1,2,3)]<-all[,-c(1,2,3)]/all[,3]

### (1)Disputed, (2)Hamilton, (3)Jay, (4)Joint, (5)Madison.
Madison.train<-all[all[,2]==5,]
Jay.train<-all[all[,2]==3,]
Hamilton<-all[all[,2]==2,]
set.seed(698)
test.index<-sample(1:51,51/2)
Hamilton.test<-Hamilton[test.index,]
Hamilton.train<-Hamilton[-test.index,]

xx<-rbind(Hamilton.train,Madison.train,Jay.train)[,-c(1,3)]
col<-c(rep(1,sum(xx[,1]==2)),rep(2,sum(xx[,1]==5)),rep(3,sum(xx[,1]==3)))
x<-cbind(col,xx)[,-2]
y<-Hamilton.test[,-c(1,2,3)]
#y<-cbind(1,yy)[,-2]
m<-length(unique(x[,1]))
prior<-rep(1/(m+1),(m+1))
thresh<-0.0001
lambda<-1.5
r<-ncol(x)-1
n<-nrow(x)


xbarij<-ssij<-ni<-matrix(0,nrow=m,ncol=r)
row<-0

for(i in 1:m){
  while((x[row+1])==i){
    row<-row+1
    for(jj in 2:(r+1)){
      j<-jj-1
      xbarij[i,j]=xbarij[i,j]+x[row,jj]
      ssij[i,j]=ssij[i,j]+x[row,jj]^2
      ni[i,j]= ni[i,j]+1
    }
  }
}

xbarij<-xbarij/ni # 6 by 130
```

```
xbari<-xbarij%*%matrix(1,nrow=r,ncol=1)/matrix(r,nrow=m,ncol=1)
#xbari<-apply(xbarij,1,mean)

# 130 by 1
s2j<-t(ssij-(xbarij^2)*ni)%*%(rep(1,m))/matrix(n-m,nrow=r,ncol=1)
sj<-sqrt(s2j)

sjsort<-sj
sjsort[rank(sj),]<-sj
mid<-r/2+.5
if(round(mid)==mid){delta=sjsort[1,mid]
}else{delta=.5*(sjsort[round(mid),1]+sjsort[round(mid)+1,1])}

xtwidij<-xbarij;

for(i in 1:m){  #6
  for(j in 1:r){ #130
    qij<-sqrt((1-2/r)/ni[i]*s2j[j] + sum(s2j)/((r^2)*ni[i]))
    dumt<-(abs((xtwidij[i,j]-xbari[i])/qij)-thresh)
    dumt=max(dumt,0)
    if((xtwidij[i,j]-xbari[i])==0){dumt<-0
    }else{dumt<-dumt*qij*abs(xtwidij[i,j]-xbari[i])/(xtwidij[i,j]-xbari[i])}
    xtwidij[i,j]<-xbari[i]+dumt
  }
}
xtwidij
ntest<-nrow(y)
postprob<-matrix(0,nrow=ntest,ncol=(m+1))

for(j in 1:r){
  if(sj[j]==0){sj[j]<-0.000001}
}

ghostdiff<-rep(0,r)
for(i in 1:ntest){
  use<-matrix(1,nrow=m,ncol=1) # 6 by 1
  uu<-matrix(y[i,],nrow=1)     # 1 by 130
  sdiff<-(kronecker(use,uu)-xtwidij)/(kronecker(t(sj),rep(1,m)))
  for(j in 1:r){
    ghostdiff[j]<-min(max(abs(sdiff[,j])),rep(lambda,m))
  }
  sdiff<-rbind(sdiff,ghostdiff)
  sumdiff=-diag(sdiff%*%t(sdiff))/2

  pptemp<-exp(sumdiff)
  denom<-t(pptemp)%*%prior
  pptemp<-(pptemp*prior)/(kronecker(denom,matrix(1,nrow=(m+1),ncol=1)))
  postprob[i,]<-pptemp
}
postprob

maxprob<-NA
author<-NA
for(i in 1:dim(postprob)[1]){
maxprob[i]<-max(postprob[i,])
author[i]<-which.max(postprob[i,])
}
maxprob
author
wrong<-1-(c(sum(author==1)/ntest))
wrong

write.table(author,"m4.txt")
```

```
######################################
####  Full Independent Bayesian  ####
######################################
### Data set up and Configurations
rm(list=ls())
library(MSBVAR)
setwd("C:/Users/Tomo/Documents/STATS 698 Research/My Codes/Fed Paper 1")
data<-read.table("completed federalist.txt")

names(data)<-c("file", "author","total","different","total_sentences","R","V2/V",
"K","a_fws","an_fws","and_fws","in_fws","it_fws","it_lws","of_fws","of_2_lws",
"the_fws","the_2_lws","with_2_lws","a_2_lws","a_adj","a_x_and","a_x_of","a_x_a",
"a_x_x_a","and_adj","and_the","and_x_of","and_x_and","and_x_x_and","as_x_as",
"as_x_x_as","be_a","be_to","but_a","by_the","i_am","i_have","i_x_i","i_x_x_i",
"in_a","in_the","of_a","of_the","of_x_and","the_and","the_of","the_in","the_to",
"the_x_and","the_x_the","the_x_x_the","to_be","to_the","to_x_to","to_x_x_to",
"you_x_you","you_x_x_you","verb_to_verb","an_a","any_all","no_not","up_upon",
"with_without","a","about","according","after","again","against","all","also",
"although","always","am","among","an","and","any","apt","are","as","at","away",
"back","be","because","been","before","behold","between","both","but","by",
"came","can","commonly","consequently","considerably","could","direction","do",
"don't","down","enough","even","ever","every","exceeding","for","forth","from",
"had","has","have","he","her","here","herself","him","his","how","i","if","in",
"innovation","into","is","it","its","just","kind","know","language","lest",
"little","matter","may","me","might","more","must","my","nay","nevertheless",
"no","not","notwithstanding","now","O","of","on","one","only","or","out","over",
"particularly","pass","probability","said","she","should","so","some","such",
"than","that","the","their","them","then","there","thereby","therefore",
"thereof","these","they","this","those","though","through","thus","to",
"towards","until","unto","up","upon","verily","very","vigorous","was","well",
"were","what","whatsoever","when","where","whereby","wherefore","whether",
"which","while","whilst","who","why","will","with","without","work","would",
"yea","you","your")

all<-cbind(data$file,data$author,data$total,data$a,data$after,
data$again,data$against,data$all,data$among,
data$an,data$and,data$are,data$as,data$at,data$away,data$be,
data$because,data$been,data$before,data$but,data$by,data$came,
data$children,data$come,data$day,data$did,data$do,data$down,
data$earth,data$even,data$every,data$father,data$for.,data$forth,
data$from,data$go,data$great,data$had,data$hand,data$have,
data$he,data$her,data$him,data$his,data$i,data$if.,data$in.,
data$into,data$is,data$it,data$king,data$know,data$land,data$made,
data$man,data$many,data$may,data$me,data$men,data$might,data$more,
data$my,data$no,data$not,data$now,data$O,data$of,data$on,
data$one,data$or,data$our,data$out,data$over,data$pass,data$people,
data$power,data$said,data$say,data$shall,data$should,data$so,
data$son,data$that,data$the,data$their,data$them,data$then,
data$there,data$therefore,data$these,data$they,data$things,
data$this,data$those,data$thus,data$time,data$to,data$up,data$upon,
data$us,data$was,data$we,data$were,data$which,data$who,data$will,
data$with,data$words,data$would,data$you,data$your)

all[,-c(1,2,3)]<-all[,-c(1,2,3)]/all[,3]

### (1)Disputed, (2)Hamilton, (3)Jay, (4)Joint, (5)Madison.
Madison.train<-all[all[,2]==5,]
Jay.train<-all[all[,2]==3,]
Hamilton<-all[all[,2]==2,]
set.seed(698)
test.index<-sample(1:51,51/2)
Hamilton.test<-Hamilton[test.index,]
Hamilton.train<-Hamilton[-test.index,]
```

```
xx<-rbind(Hamilton.train,Madison.train,Jay.train)[,-c(1,3)]
col<-c(rep(1,sum(xx[,1]==2)),rep(2,sum(xx[,1]==5)),rep(3,sum(xx[,1]==3)))
x<-cbind(col,xx)[,-2]
y<-Hamilton.test[,-c(1,2,3)]
#y<-cbind(1,yy)[,-2]
m<-length(unique(x[,1]))
prior<-rep(1/(m+1),(m+1))
thresh<-0.0001
lambda<-1.5
r<-ncol(x)-1
n<-NA
for(i in 1:m){
  n[i]<-length(x[x[,1]==i,1])
}
n


## Logit transform
logit<-function(x){
  x<-log(x/(1-x))
}
x[,-1]<-logit(x[,-1])
y<-logit(y)

### replace any 0's with very small number in training and test set
x[x==-Inf]<- -20
y[y==-Inf]<- -20
sum(x==-Inf)
sum(y==-Inf)
sum(x==Inf)
sum(y==Inf)

thresh<-2.5
lambda<-1.5


### Unlogit
unlogit<-function(x){
  x<-exp(x)/(1+exp(x))
}

### Priors
mean(c(logit(0.1),logit(0.0001)))  ## -5.703732
((logit(0.1)--5.703732)/3)^2
mu<--5.703732
tau2<-1.366177
c(unlogit(mu-3*sqrt(tau2)),unlogit(mu+3*sqrt(tau2)))
c(mu-3*sqrt(tau2),mu+3*sqrt(tau2))

MU<-mu*rep(1,r)
TAU2<-tau2*diag(r)

alpha<-3
beta<-0.1
1/(beta*(alpha-1))
1/((beta^2)*((alpha-1)^2)*(alpha-2))
#print(unlogit(1/(beta*(alpha-1))))
#print(unlogit(1/((beta^2)*((alpha-1)^2)*(alpha-2))))

#pdf("priorsig.pdf")
#plot(density(unlogit(prior.sigma)),main=expression(paste("Prior Predictive for ",sigma[j]^2)),
#xlab=expression(sigma[j]))
#dev.off()
```

```
#pdf("priortheta.pdf")
#plot(density(unlogit(prior.theta)),main=expression(paste("Prior Predictive for ",theta[i])),
#xlab=expression(theta[i]),xlim=c(-0.01,0.08))
#dev.off()

### Lay out the prior distributions
prior.theta<-rmultnorm(10000,MU,TAU2,tol=1e-10)
prior.sigma<-1/rgamma(10000,alpha,scale=beta)
Sigma<-prior.sigma*diag(r)
prior.y<-rnorm(10000,prior.theta,sqrt(prior.sigma))

### Prior Predictive
#plot(density(unlogit(prior.theta)),main=expression(paste("Prior Predictive for ",theta)),
#xlab=expression(theta),xlim=c(-0.01,0.1))
#plot(density(unlogit(prior.sigma)),
main=expression(paste("Prior Predictive for ",sigma)),xlab=expression(sigma))
#plot(density(unlogit(prior.y)),main=expression(paste("Prior Predictive for ",y)),
xlab=expression(y))


### Starting Points
nsim<-1000
theta<-array(mu,dim=c(r,m,nsim))
sig2<-array(TAU2,dim=c(r,r,nsim))

N<-dim(x)[1]
N
sample<-rep(1,n[1])
for(i in 2:m){
  sample<-c(sample,rep(i,n[i]))
}
sample
rate<-NA

### Start Gibbs Sampling
for(ii in 2:nsim){

  ### generate theta's
  for(i in 1:m){
    x.sum<-NA
    for(j in 2:(r+1)){
      x.sum[j]<-sum(x[x[,1]==i,j])
    }
    x.sum<-x.sum[-c(1)]
    x.sum

    ## mustar = r by 1
    mustar<-(solve(tau2*n[i]*solve(sig2[,,ii-1])+diag(r)))%*%(tau2*solve(sig2[,,(ii-1)])%*%
    x.sum+MU)
    mustar

    ## sigstar = r by r
    #sigstar<-(solve(tau2*n[i]*solve(sig2[,,ii-1])+diag(r)))
    sigstar<-(solve(tau2*n[i]*sig2[,,ii-1]+diag(r)))*tau2
    sigstar

    ## theta = r by m by nsim
    theta[,i,ii]<-rmultnorm(1,mustar,sqrt(sigstar))
  }

  ### generate sig2
  astar<-alpha+N/2
  for(j in 1:r){
```

```
    theta.x<-0
    sum.xsq<-0
    theta.sq<-0

    for(l in 1:N){
      sum.xsq<-sum.xsq+(x[l,(j+1)]^2)
      theta.x<-theta.x+(theta[j,(sample[l]),ii]*x[l,(j+1)])
    }
    for(i in 1:m){

      ni<-n[i]
      theta.sq<-theta.sq+(ni*(theta[j,i,ii]^2))
    }
    bstar<-(((sum.xsq-2*theta.x+theta.sq)/2)+1/beta)^(-1)
    sig2[j,j,ii]<-1/rgamma(1,astar,scale=bstar)
  }
  cat("iter =",ii,"\n")
}
### Determine burn period
pdf("convergence4.pdf")
plot(1:300,theta[i,1,1:300],type='l',main="Convergence plot",
xlab="Gibbs Samples",ylab=expression(paste(theta[i]," (logit scale)")))
title(sub="(first 300 Gibbs Samples)",line=-27.3)
abline(v=100,col='red',lty=2)
dev.off()

burn<-100
length<-nsim-burn
good<-burn:nsim


### Posterior Predictive of theta
pdf("post_theta4.pdf")
par(mfrow=c(3,1))
plot(density(unlogit(theta[,1,100])),xlim=c(-0.0001,0.075),
    main=expression(paste("Prior and Posterior Predictive for ",theta[i])),
    xlab=expression(paste(theta[i])),ylim=c(0,300),ylab="Density")
    lines(density(unlogit(prior.theta)),col='red')
    legend(0.02,250,legend=c("Prior predictive","Posterior predictive"),
    col=c("red","black"),lty=1)
    title(sub="(Alexander Hamilton)",line=-10)

plot(density(unlogit(theta[,2,100])),xlim=c(-0.0001,0.075),
    main=expression(paste("Prior and Posterior Predictive for ",theta[i])),
    xlab=expression(paste(theta[i])),ylim=c(0,300),ylab="Density")
    lines(density(unlogit(prior.theta)),col='red')
    legend(0.02,250,legend=c("Prior predictive","Posterior predictive"),
    col=c("red","black"),lty=1)
    title(sub="(James Madison)",line=-10)

plot(density(unlogit(theta[,3,100])),xlim=c(-0.0001,0.075),
    main=expression(paste("Prior and Posterior Predictive for ",theta[i])),
    xlab=expression(paste(theta[i])),ylim=c(0,300),ylab="Density")
    lines(density(unlogit(prior.theta)),col='red')
    legend(0.02,250,legend=c("Prior predictive","Posterior predictive"),
    col=c("red","black"),lty=1)
    title(sub="(John Jay)",line=-10)

dev.off()

### Posterior Predictive of theta (logit scale)
pdf("post_thetalogit4.pdf")
par(mfrow=c(3,1))
plot(density(theta[,1,100]),xlim=c(-12,0),
```

```
    main=expression(paste("Prior and Posterior Predictive for ",theta[i], " (logit scale)")),
    xlab=expression(paste(italic(logit),"(",theta[i],")")),ylim=c(0,0.5),ylab="Density")
    lines(density(prior.theta),col='red')
    legend(-12,0.47,legend=c("Prior","Posterior"),col=c("red","black"),lty=1)
    title(sub="(Alexander Hamilton)",line=-10)

plot(density(theta[,2,100]),xlim=c(-12,0),
    main=expression(paste("Prior and Posterior Predictive for ",theta[i], " (logit scale)")),
    xlab=expression(paste(italic(logit),"(",theta[i],")")),ylim=c(0,0.5),ylab="Density")
    lines(density(prior.theta),col='red')
    legend(-12,0.47,legend=c("Prior","Posterior"),col=c("red","black"),lty=1)
    title(sub="(James Madison)",line=-10)

plot(density(theta[,3,100]),xlim=c(-12,0),
    main=expression(paste("Prior and Posterior Predictive for ",theta[i], " (logit scale)")),
    xlab=expression(paste(italic(logit),"(",theta[i],")")),ylim=c(0,0.5),ylab="Density")
    lines(density(prior.theta),col='red')
    legend(-12,0.47,legend=c("Prior","Posterior"),col=c("red","black"),lty=1)
    title(sub="(John Jay)",line=-10)

dev.off()

### Posterior Predictive of y (logit scale)
pdf("post_y4.pdf")
par(mfrow=c(3,1))
  a<-rmultnorm(10000,theta[,1,100],sqrt(sig2[,,100]))
  plot(density(a),xlim=c(-16,2),
    main=expression(paste("Prior and Posterior Predictive for ",X[il], " (logit scale)")),
    xlab=expression(paste(italic(logit),"(",X[il],")")),ylim=c(0,0.25),ylab="Probability")
    lines(density(prior.y),col='red')
    legend(-16,0.234,legend=c("Prior","Posterior"),col=c("red","black"),lty=1)
    title(sub="(Alexander Hamilton)",line=-10)

  a<-rmultnorm(10000,theta[,2,100],sqrt(sig2[,,100]))
  plot(density(a),xlim=c(-16,2),
    main=expression(paste("Prior and Posterior Predictive for ",X[il], " (logit scale)")),
    xlab=expression(paste(italic(logit),"(",X[il],")")),ylim=c(0,0.25),ylab="Density")
    lines(density(prior.y),col='red')
    legend(-16,0.234,legend=c("Prior","Posterior"),col=c("red","black"),lty=1)
    title(sub="(James Madison)",line=-10)

  a<-rmultnorm(10000,theta[,3,100],sqrt(sig2[,,100]))
  plot(density(a),xlim=c(-16,2),
    main=expression(paste("Prior and Posterior Predictive for ",X[il], " (logit scale)")),
    xlab=expression(paste(italic(logit),"(",X[il],")")),ylim=c(0,0.25),ylab="Density")
    lines(density(prior.y),col='red')
    legend(-16,0.234,legend=c("Prior","Posterior"),col=c("red","black"),lty=1)
    title(sub="(John Jay)",line=-10)

dev.off()

### Posterior Predictive of y
pdf("post_ylogit4.pdf")
par(mfrow=c(3,1))
  a<-rmultnorm(10000,theta[,1,100],sqrt(sig2[,,100]))
  plot(density(unlogit(a)),xlim=c(0.001,0.07),
    main=expression(paste("Prior and Posterior Predictive for ",X[il])),
    xlab=expression(X[il]),ylim=c(0,430),ylab="Density")
    lines(density(unlogit(prior.y)),col='red')
    legend(0.02,350,legend=c("Prior predictive","Posterior predictive"),
    col=c("red","black"),lty=1)
    title(sub="(Alexander Hamilton)",line=-10)

  a<-rmultnorm(10000,theta[,2,100],sqrt(sig2[,,100]))
```

```
  plot(density(unlogit(a)),xlim=c(0.001,0.07),
    main=expression(paste("Prior and Posterior Predictive for ",X[il])),
    xlab=expression(X[il]),ylim=c(0,430),ylab="Density")
    lines(density(unlogit(prior.y)),col='red')
    legend(0.02,350,legend=c("Prior predictive","Posterior predictive"),
    col=c("red","black"),lty=1)
    title(sub="(James Madison)",line=-10)

  a<-rmultnorm(10000,theta[,3,100],sqrt(sig2[,,100]))
  plot(density(unlogit(a)),xlim=c(0.001,0.07),
    main=expression(paste("Prior and Posterior Predictive for ",X[il])),
    xlab=expression(X[il]),ylim=c(0,430),ylab="Density")
    lines(density(unlogit(prior.y)),col='red')
    legend(0.02,350,legend=c("Prior predictive","Posterior predictive"),
    col=c("red","black"),lty=1)
    title(sub="(John Jay)",line=-10)
dev.off()
```

```
ten.s<-tw.s<-median.s<-NA
ntest<-dim(y)[1]
final.select<-NA
select<-matrix(NA,nrow=ntest,ncol=(length-0))
rate<-NA
prob<-array(NA,dim=c(length,(m+1),ntest))
library(nnet)
lambda<-2.5
for(ns in 1:ntest){
  L<-as.matrix(rep(lambda,r),ncol=r)
  prior<-rep(1/(m+1),m+1)
  for(gs in 1:(length-0)){

    ### Obtain Sum of all m authors' contribution
    eachauthor<-NA
    for(i in 1:m){
      inner<-NA
      for(j in 1:r){
        innermost<-(y[ns,(j)]-theta[j,i,(gs+burn)])/sqrt(sig2[j,j,(gs+burn)])
        inner[j]<-innermost^2
      }
      eachauthor[i]<-prior[i]*exp(-0.5*sum(inner))
    }
    allauthor<-sum(eachauthor)

    ### Obtain the (m+1)th author's contribution
    inside<-matrix(NA,nrow=r,ncol=m)
    for(i in 1:m){
      for(j in 1:r){
        #centroid<-theta[j,i,(gs+burn)]
        #s.star<-0.259*(centroid^1.1147)/(y[ns,3]^0.661)
        #s.star
        #inside[j,i]<-abs((y[ns,(j+1)]-theta[j,i,(gs+burn)])/s.star)
        inside[j,i]<-abs((y[ns,(j)]-theta[j,i,(gs+burn)])/sqrt(sig2[j,j,(gs+burn)]))
      }
    }
    left<-apply(inside,1,max)
    a.j<-apply(cbind(left,L),1,min)
    print(a.j)

    extra<-prior[m+1]*exp(-0.5*sum(a.j^2))
```

```
    if(extra+allauthor==0){
      blog<-1*10^(-260)
      }else{
      blog<-extra+allauthor
    }
    pr<-eachauthor/(blog)
    pr<-c(pr,extra/(blog))
    pr

    ### Plug into the Probability matrix
    for(i in 1:(m+1)){
      prob[gs,i,ns]<-pr[i]
      print(pr[i])
    }
    cat("Gibbs Sample =",gs,"chapters=",ns,"\n")
  }
  par(mfrow=c(3,3))
  #for(i in 1:(m+1)){
  #  #plot(density(prob[,i,ns]),main=i,xlim=c(0,1))
  #  hist(prob[,i,ns],main=i,xlim=c(0,1))
  #  #abline(v=qnt[i,-2],col='red',lty=2)
  #  title(paste("Chapter",ns),outer=T,line=-1)
  #}
  for(i in 1:(m+1)){
    #plot(prob[,i,ns],type='h',main=i,ylim=c(0,max(prob)))
    plot(prob[,i,ns],type='h',main=i,ylim=c(0,1))
  }
  title(paste("Chapter",ns),outer=T,line=-1)


  ### Quantile
  quant.ten<-quant.tw<-med<-NA
  for(i in 1:(m+1)){
    quant.ten[i]<-quantile(prob[,i,ns],probs=0.1)
    quant.tw[i]<-quantile(prob[,i,ns],probs=0.25)
    med[i]<-quantile(prob[,i,ns],probs=0.5)
  }
  ten.s[ns]<-which(max(quant.ten)==quant.ten)
  tw.s[ns]<-which(max(quant.tw)==quant.tw)
  median.s[ns]<-which(max(med)==med)



  #select<-NA
  for(i in 1:dim(prob)[1]){
    select[ns,i]<-which.is.max(prob[i,,ns])
  }
  rate[ns]<-sum(select[ns,]==y[ns,1])/dim(prob)[1]

  cnt<-NA
  for(i in 1:length(unique(select[ns,]))){
    cnt[i]<-sum(select[ns,]==unique(select[ns,])[i])
  }
  final.select[ns]<-unique(select[ns,])[cnt==max(cnt)]

  cat("Lambda =",lambda,"rate =",rate[ns],"author =",y[ns,1],
  "\nselection =",select[ns,],"\n")
  cat("Final selection =",final.select[ns],"\n")
}
#final.select
ten.s
tw.s
median.s
```

```
1-sum(ten.s==1)/ntest
1-sum(tw.s==1)/ntest
1-sum(median.s==1)/ntest
#1-sum(final.select==7)/ntest


write.table(ten.s,"m5q10.txt")
write.table(tw.s,"m5q25.txt")
write.table(median.s,"m5q50.txt")

############################
####  Plotting Results  ####
############################
rm(list=ls())
setwd("C:/Users/Tomo/Documents/STATS 698 Research/My Codes/Fed Paper 1/")
results1<-read.table("m0.txt")
results2<-read.table("m1.txt")
results3<-read.table("m2.txt")
results4<-read.table("m3.txt")
results5<-read.table("m4.txt")
results6a<-read.table("m5q10.txt")
results6b<-read.table("m5q25.txt")
results6c<-read.table("m5q50.txt")

chap<-25
m1<-1-sum(results1==1)/chap
m2<-1-sum(results2==1)/chap
m3<-1-sum(results3==1)/chap
m4<-1-sum(results4==1)/chap
m5<-1-sum(results5==1)/chap
m6a<-1-sum(results6a==1)/chap
m6b<-1-sum(results6b==1)/chap
m6c<-1-sum(results6c==1)/chap
c(m1,m2,m3,m4,m5,m6a,m6b,m6c)
m<-4
colmat1<-colmat2<-colmat3<-colmat4<-colmat5<-colmat6a<-colmat6b<-
colmat6c<-matrix(0,ncol=m,nrow=chap)
for(i in 1:chap){
  for(j in 1:m){
    if(as.vector(results1)[i,1]==j){colmat1[i,j]<-1}
    if(as.vector(results2)[i,1]==j){colmat2[i,j]<-1}
    if(as.vector(results3)[i,1]==j){colmat3[i,j]<-1}
    if(as.vector(results4)[i,1]==j){colmat4[i,j]<-1}
    if(as.vector(results5)[i,1]==j){colmat5[i,j]<-1}
    if(as.vector(results6a)[i,1]==j){colmat6a[i,j]<-1}
    if(as.vector(results6b)[i,1]==j){colmat6b[i,j]<-1}
    if(as.vector(results6c)[i,1]==j){colmat6c[i,j]<-1}
  }
}

library(TeachingDemos)


pdf("fed1.pdf")
par(mfrow=c(3,1),oma=c(0,0.01,0.01,0.01),cex.main=2.2,cex.axis=1.5,cex.lab=1.6)
plot(1:chap,colmat6a[1:chap,1],type='h',lwd=25,axes=FALSE,ylim=c(0.3,0.9),
xlab="25 Alexander Hamilton Federalist Papers",ylab="",
main=expression(paste("Full Independent Bayesian (50th percentile): Misclassified 8%")),
col='blue')
axis(1,at=c(seq(0,chap,by=5)),lab=c(seq(0,chap,by=5)))
lines(1:chap,colmat6a[1:chap,2],type='h',col='red',lwd=25)
lines(1:chap,colmat6a[1:chap,3],type='h',col='purple',lwd=25)
lines(1:chap,colmat6a[1:chap,4],type='h',col='darkgreen',lwd=25)
```

```
plot(1:chap,colmat6b[1:chap,1],type='h',lwd=25,axes=FALSE,ylim=c(0.3,0.9),
xlab="25 Alexander Hamilton Federalist Papers",ylab="",
main=expression(paste("Full Independent Bayesian (25th percentile): Misclassified 8%")),
col='blue')
axis(1,at=c(seq(0,chap,by=5)),lab=c(seq(0,chap,by=5)))
lines(1:chap,colmat6b[1:chap,2],type='h',col='red',lwd=25)
lines(1:chap,colmat6b[1:chap,3],type='h',col='purple',lwd=25)
lines(1:chap,colmat6b[1:chap,4],type='h',col='darkgreen',lwd=25)

plot(1:chap,colmat6c[1:chap,1],type='h',lwd=25,axes=FALSE,ylim=c(0.3,0.9),
xlab="25 Alexander Hamilton Federalist Papers",ylab="",
main=expression(paste("Full Independent Bayesian (50th percentile): Misclassified 8%")),
cex=1.3,col='blue')
axis(1,at=c(seq(0,chap,by=5)),lab=c(seq(0,chap,by=5)))
lines(1:chap,colmat6c[1:chap,2],type='h',col='red',lwd=25)
lines(1:chap,colmat6c[1:chap,3],type='h',col='purple',lwd=25)
lines(1:chap,colmat6c[1:chap,4],type='h',col='darkgreen',lwd=25)
dev.off()

plot(1:chap,colmat6c[1:chap,1],type='n',axes=F)
legend(1,1,legend=c("Hamilton","Madison","Jay","Latent"),
lty=1,lwd=10,col=c("blue","red","purple","darkgreen"),ncol=2,cex=1.4)
```