**Detecting Multiple Authorship of United States Supreme Court
Legal Decisions Using Function Words**

by

Jeffrey S. Rosenthal
Department of Statistics
University of Toronto

and

Albert H. Yoon
Faculty of Law
University of Toronto

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Detecting Multiple Authorship of United States Supreme Court Legal Decisions Using Function Words

by

Jeffrey S. Rosenthal and Albert H. Yoon

Department of Statistics and Faculty of Law

University of Toronto

(December 3, 2009.)

1   Introduction

This paper describes various statistical analyses performed on the texts of legal judgments written by United States Supreme Court (USSC) justices.

Our primary motivation is the following. It is believed that certain USSC justices (e.g. Scalia, see [8] p. 271) primarily write their own legal decisions, while others (e.g. Kennedy, see [8] p. 274) rely heavily on their law clerks to do much of the writing, though there are few hard facts about this and it is mostly a matter of speculation. We attempt to verify this hypothesis by measuring the *variability* of writing style of the majority opinions written by the various justices. We conclude that, indeed, majority opinions written by Kennedy have significantly greater variability than do those by Scalia (measured using various statistics involving the frequencies of various *function words*, as described below). Furthermore, using a bootstrap approach, we confirm that this observation is statistically significant at the 5% level. So, under the (reasonable, but unproven) assumption that multiple authorship leads to greater variability of writing style, this conclusion would appear to provide convincing evidence that Kennedy does indeed get more writing assistance from law clerks than does Scalia.

Our secondary motivation concerns the question of whether it is possible to determine which justice is the (recorded) author of a given decision, solely by use of word frequencies. Informal enquiries with USSC legal experts indicate that they do not believe they are able to do this. Nevertheless, in this paper we consider various approaches (naive Bayes classifier, linear classifier), and show using a cross-validation approach that such algorithms can indeed predict authorship with accuracy approaching 90%. (Of course, in this case the recorded authorship is already known, so identifying it is not useful *per se*, but this serves as a measure of the extent to which different justices have identifiably different writing styles.)

Our methodology thus appears to provide useful methods both for determining multiple authorship and for identifying the recorded authorship, solely using function words – at least for USSC decisions (and perhaps beyond). Our analysis required writing extensive software, all of which will be made freely available [14] for purposes of reproducing or extending our results. Further details are given herein.

1.1 Background on the United States Supreme Court

The United States Supreme Court is the highest court in the United States. The USSC has a predominantly discretionary docket, granting certiorari only for "cases involving unsettled questions of federal constitutional or statutory law of general interest" ([13], p. 238). While the USSC is not unique in issuing opinions or creating precedent, its position at the apex of the judicial hierarchy ensures that practitioners, legal scholars, and law students closely scrutinize its opinions.

Unlike the executive and legislative branches of the federal government, the USSC is administratively lean. The court itself consists of nine justices. The chambers of each justice typically consist of one secretary and four law clerks. The workload, given the small number of justices and staff, is considerable. Each year, the justices, with the assistance of only their law clerks, determines which among roughly 8,000 cases to grant certiorari (i.e., grant to hear the case), and writes opinions – majority, concurring, dissenting – for roughly 80 cases. Justices typically serve on the USSC well past typical retirement age or vesting in their pension (both typically occuring at age 65), often until death or prevented by physical or mental incapacity. Perhaps in part due to the Court's tradition of longevity and hard work, Americans consistently rank the USSC as the most trusted branch of government [6]. Justice Louis Brandeis once commented, "[t]he reason the public thinks so much of the Justices of the Supreme Court is that they are almost the only people in Washington who do their own work" ([11], pp. 12–13).

At the same time, the USSC remains one of the least understood branches. Unlike Congress, the USSC deliberates in private. The deliberations result in a single public document: the opinion itself. Accordingly, the process by which the USSC produces each opinion remains largely unknown. Prior to the 1950s, justices performed most of the substantive requirements of the job, including the content of opinions ([12], p. 208). Law clerks performed mostly administrative tasks. When the job demands increased over time, however, without a corresponding increase in judgeships, justices relied more on law clerks to prepare certiorari and oral argument memos, as well as draft and edit opinions ([12], p. 151). While the prestige of a Supreme Court law clerkship is well accepted within legal circles [7], the

clerks' contribution to their respective justices remains largely anecdotal [12, 17, 8, 19].

The anecdotal evidence that justices vary in their reliance on law clerks in the drafting and editing of opinions suggests that justices would vary in their level of distinctiveness in authorship. At the extremes, a justice who wrote his own opinions would presumptively have a more distinct writing style than another justice who relied heavily on his law clerks. The intuition here is that the greater the number of participants in the writing process, the more heterogeneous the writing style.

The institutional design of USSC clerkships provides an additional exogenous mechanism to test this hypothesis: USSC clerkships are usually for a single term, from October through August of the subsequent year. Accordingly, the cohort for law clerks changes in predictable fashion, allowing an examination of justices' writing style within and across term years.

Although this paper focuses on statistical methodology, the question of judicial authorship is an important one in both political science and law. More than the current scholarship acknowledges, the output of the USSC reflects a principal-agent relationship between the justices and their clerks. As with any principal-agent relationship, the degrees to which the clerks' interests overlap with the justices' depend on their incentives and degree of oversight. But to even approach this question requires first a more complete understanding of judicial authorship. While it is possible to tackle this by reading every Supreme Court opinion, our discussions with current USSC scholars suggest that differences in writing variability across justices may be too subtle to discern manually. Our paper provides a more systematic approach.

1.2 Statistical analysis via function words

Statistical analysis of texts has a long history, for example related to the famous Shakespeare authorship question (see e.g. [15, 2]), though many such questions have been investigated primarily by historical and other non-statistical means (see e.g. [18]).

One challenge with statistical textual analysis is separating those writing features pertaining to writing style, from those pertaining to specific subject matter content. To deal with this, a number of authors (e.g. [5, 9, 10, 20, 1]) have made extensive use of so-called *function words*, i.e. words such as *all*, *have*, *not*, and *than*, whose usage frequencies are thought to be largely independent of the subject matter under discussion. Previous studies [20, 1] have found these function words to be of some use, albeit limited, in determining authorship of disputed writings. In any case, such function words appear to be a useful starting point for content-independent statistical analysis.

In particular, in their study of the Federalist Papers, Mosteller and Wallace ([10], p. 38,

Table 2.5-2) produce a list of 70 function words, culled for their purposes from certain earlier studies [5, 9]. This list provides the basis for our statistical analysis, though to improve stability we eliminated the seven function words (*every, my, shall, should, upon, will, your*) that occur with frequency less than 0.001 in the USSC judgments, leaving us with 63 function words (Table 1). (We also considered adding *while* and *whilst*, which [10] also found to be very useful, but they too had frequency less than 0.001 in the USSC judgments. In any case, our results changed very little upon adding our removing these few words.)

Of course, it is also possible to consider larger-scale features (e.g. sentences), and smaller-scale features (e.g. individual characters). But in the present study, we stick to the function words only.

> *a, all, also, an, and, any, are, as, at, be, been, but, by, can, do, down, even, for, from, had, has, have, her, his, if, in, into, is, it, its, may, more, must, no, not, now, of, on, one, only, or, our, so, some, such, than, that, the, their, then, there, things, this, to, up, was, were, what, when, which, who, with, would*

Table 1: the 63 function words used in the present study.

1.3   Data Acquisition

Our data consisted of the complete text of judgments of the USSC from 1991–2009, as provided by the Cornell Law School web site [16]. For consistency, we primarily considered the majority opinions written by the various USSC justices, though we do briefly consider dissenting opinions below as well. (While expansive, the data in [16] occasionally introduces transcription errors, and furthermore apparently does not contain quite every USSC opinion, e.g. those by Justice O'Connor are apparently missing. We assume, however, that such minor limitations in the data do not significantly bias our results.)

Although the judgment texts were publicly available [16], it was still necessary to download all the data files from the web, convert them to simple plain-text format, remove extraneous header and footer text, sort the judgments by authoring justice and by court session, and index all the judgments by date written. Given the large number (well over 1,000) of judgments to consider, we wrote extensive software [14] in C and Unix to quickly and automatically perform this task.

Using this software, we downloaded and processed and sorted all of the majority opinion judgments (and also separately the dissenting opinions) of various USSC justices. To avoid trivialities, judgments containing fewer than 250 words were systematically excluded. The resulting files were then used as data for all of our statistical work below.

2   Statistical analyses of word counts

We suppose that our function words are numbered from $j = 1$ to $j = 63$. We further suppose that a given justice has written judgments numbered from $i = 1$ to $i = K$. Let w_i be the total number of words in judgment i, and let c_{ij} be the number of times that function word j appears in judgment i.

2.1   Word fractions

Since judgments can differ considerably in their length, the raw counts c_{ij} by themselves are not particularly meaningful. One approach is to instead consider the quantities

$$f_{ij} \;=\; c_{ij} \,/\, w_i \,,$$

representing the fraction of words in judgment i which are function word j. If the sample standard deviation $sd(f_{1j}, f_{2j}, \ldots, f_{Kj})$ is much larger for one justice than for another, this suggests that the justice has a much more variable writing style.

Unfortunately, this analysis is not entirely independent of such factors as the length of judgments, the different justices' different propensities to use different words, etc.

For example, for a given function word j, consider the null hypothesis that a given justice has some fixed unknown propensity p_j for using the function word j, independently for each word of each of their judgments. That is, assume that each word of each judgment has the same (unknown) probability p_j of being the function word j. In this case, the distribution of the count c_{ij} of reference word j in judgment i would be binomial, corresponding to w_i independent experiments each having probability of success p_j. We can write this formally as:

$$c_{ij} \;\sim\; \mathrm{Binomial}(w_i, p_j) \,,$$

so that c_{ij} would have mean $w_i p_j$ and variance $w_i p_j (1 - p_j)$.

If we then considered the fraction of function word j in each judgment i, i.e. $f_{ij} = c_{ij} \,/\, w_i$, then f_{ij} would have mean p_j, and variance $p_j(1-p_j)/w_i$. In particular, the mean and variance of f_{ij} would depend on the individual propensities p_j and the judgment lengths w_i, making statistical inferences difficult. Specifically, while we could consider calculating the sum of sample standard deviations

$$V_1 \;=\; \sum_{j=1}^{63} sd(f_{1j}, f_{2j}, \ldots, f_{Kj}) \,,$$

and using that as a measure of the variability of writing style across judgments of a given justice, such a comparison would not be entirely satisfactory since it would be influenced

by such factors as p_j and w_i, so it would e.g. tend to unfairly assign smaller variability to justices who tend to write shorter decisions.

2.2 Correcting the word fraction values

One approach to overcoming this obstacle is to adjust the fractions f_{ij} to make them less sensitive to p_j and w_i. Specifically, letting

$$\mu_j \;=\; \frac{c_{1j} + c_{2j} + \ldots + c_{Kj}}{w_1 + w_2 + \ldots + w_K}$$

be our best estimate of p_j, we would then have that roughly speaking $f_{ij} - \mu_j$ has mean 0 and variance $p_j(1 - p_j)/w_i$, whence

$$r_{ij} \;=\; w_i^{1/2}(f_{ij} - \mu_j)$$

has mean 0 and variance $p_j(1 - p_j)$, which is independent of the judgment length w_i. Hence, another option would be to consider the sum of standard deviations

$$V_2 \;=\; \sum_{j=1}^{63} sd(r_{1j}, r_{2j}, \ldots, r_{Kj})\,.$$

On the other hand, this null variance $p_j(1 - p_j)$ still depends on the unknown propensity p_j. Now, if our estimate μ_j were exact, i.e. if $\mu_j = p_j$, then this would mean that

$$q_{ij} \;=\; \frac{w_i^{1/2}(f_{ij} - \mu_j)}{(\mu_j(1 - \mu_j))^{1/2}}$$

would have mean 0 and variance approximately 1.

So, in principle, these quantities q_{ij} can be used as approximately scale-independent versions of the original function word fractions f_{ij}. If so, then the sample standard deviations

$$V_3 \;=\; \sum_{j=1}^{63} sd(q_{1j}, q_{2j}, \ldots, q_{Kj})$$

would provide an estimate of the writing variability of a given justice in a way that would be largely independent of the individual values p_j and w_i.

However, even this is not entirely satisfactory, since the above analysis relies on the assumption that the observed estimates μ_j are perfect estimates of the propensities p_j, which in general they would not quite be, and it is difficult to accurately take into account the additional variability from the uncertainty in the μ_j. In particular, if p_j is quite close to zero (as it often will be), then dividing by μ_j might be quite an unstable operation, leading to unreliable results. (In the most extreme case, if $\mu_j = 0$, then dividing by μ_j would lead to "not a number", i.e. NaN. We fix this by simply omiting all terms with $\mu_j = 0$ from the sum; nevertheless, the issue illustrates another form of instability inherent in V_3.)

2.3 A chi-squared approach

Since in our case the counts c_{ij} are exact, while estimates such as μ_j are inexact, this suggests that we instead use the chi-squared statistic. Specifically, we consider the value

$$chisq \;=\; \sum_{i=1}^{K} \sum_{j=0}^{63} \frac{(c_{ij} - e_{ij})^2}{e_{ij}} \,,$$

where again w_i is the total number of words in judgment i, and c_{ij} is the number of times that function word j appears in judgment i, and now $c_{i0} = w_i - c_{i1} - \ldots - c_{iK}$ is the number of words in judgment i which are *not* function words, and

$$e_{ij} \;=\; (c_{1j} + c_{2j} + \ldots + c_{Kj}) \left(\frac{w_i}{w_1 + w_2 + \ldots + w_K} \right) \;=\; w_i \left(\frac{c_{1j} + c_{2j} + \ldots + c_{Kj}}{w_1 + w_2 + \ldots + w_K} \right)$$

is the *expected number* of times that function word j *would* have appeared in judgment i, under the null hypothesis that the total number $c_{1j} + c_{2j} + \ldots + c_{Kj}$ of appearances of reference word j were each equally likely to occur in any of the total number $w_1 + w_2 + \ldots + w_K$ of words in all of the justice's K judgments combined.

Under the null hypothesis, according to the standard chi-squared statistical test (see e.g. [4], Theorem 9.1.1, or any other standard introductory statistics textbook), the *chisq* statistic should follow a chi-squared distribution with $(63 + 1 - 1)(K - 1) = 63(K - 1)$ degrees of freedom, hence with mean $63(K - 1)$. (The "+1" arises because of the c_{i0} terms, i.e. since we also count the leftover non-function words too.) So, dividing this statistic by its null mean, we obtain the new statistic

$$V_4 \;=\; chisq/df \;=\; chisq \,/\, 63(K - 1) \,.$$

The value of chisq/df should be approximately 1 under the null hypothesis, and larger than 1 for writing collections which exhibit greater writing style variability. In particular, the extent to which chisq/df is larger than 1 appears to be a fairly reliable and robust way to estimate writing style variability, and we use it below.

As an aside, we note that such *chisq* values are less stable/useful when the expected cell counts are very close to zero. It may be possible to correct for this in various ways (e.g., Yates' correction), but this is not entirely clear. However, since we already eliminated from consideration those function words which have very low frequency in the USSC judgments, and those USSC judgments which are extremely short, overall we do not expect this small-cell issue to be a significant problem, and we do not consider it further.

2.4 Variability results

We developed software [14] to compute each of the above variability statistics V_1, V_2, V_3, and V_4 (among other statistics). We then applied our software to a variety of USSC justices' judgments. The results were as follows. (Unless otherwise specified, the results are for *majority* judgments written by that justice.)

	# judgments	V_1	V_2	V_3	V_4
Kennedy	147	0.198694	14.070624	2075.810061	8.257910
Scalia	156	0.188897	11.907545	1772.956673	6.059062
Stevens	148	0.185262	13.470465	2005.212104	8.387597
Souter	143	0.193926	13.579813	1909.395861	8.295534
Thomas	140	0.235498	14.189851	2032.116409	7.984608
Ginsburg	130	0.208412	13.748464	2034.408488	7.753709
Breyer	121	0.209246	12.712241	1915.215702	6.961088
Rehnquist	127	0.210449	12.352877	1774.805458	5.963600
Stevens dissent	214	0.277549	11.396058	1669.284886	5.116318
Scalia dissent	114	0.267204	10.941911	1559.111187	4.883980
Kennedy dissent	47	0.436531	10.453099	1627.703912	4.858435

Looking at these results, a number of facts become clear. Firstly, we see that Kennedy does indeed have higher writing-style variability than does Scalia, by each of the four measures, thus apparently confirming our original hypothesis (see also the next section re statistical significance). The other justices mostly fall inbetween these extremes, though Souter and Stevens also have very high variability, while Breyer and especially Rehnquist have lower variability.

As for the dissenting judgments, we might expect them to have much smaller writing variability since they tend to be more focused and also more likely to be written by the justice alone. This is indeed confirmed by the measures V_2, V_3, and V_4, but not by V_1 which gets tripped up by the fact that dissents tend to be shorter and V_1 does not correct for this. So, this provides confirmation that dissent judgments tend to have more consistent writing style, and also illustrates why V_1 is not an appropriate measure of variability.

As for the other variability measures, V_2 and V_3 each have some merit, but as discussed above V_2 is too dependent on the extraneous factors p_j, while V_3 is too unstable due to variability of the estimates μ_j. Thus, overall we feel that V_4 is the most stable measure, so we concentrate on V_4 below.

Remark: Of course, the different V_i are each on a different scale, so it is meaningless to e.g. compare the value of V_1 directly with the value of V_2. It is only meaningful to compare the

same variability statistic (e.g. V_4) when computed for different collections of judgments.

Remark: In all cases the value of V_4 is much larger than it would be under the null hypothesis that the function words are truly distributed uniformly and randomly. For example, for Scalia, the *chisq* statistic is equal to $6.059062 \times 63 \times (156 - 1) = 59166.74$; under the null hypothesis this would have the chi-squared distribution with $63(156 - 1) = 9765$ degrees of freedom, for which the value 59166.74 corresponds to a p-value of about $\exp(-15912)$ which is completely negligible. So, the null hypothesis is definitively rejected. However, we still feel that *chisq* (or in particular the related quantity V_4) is the most appropriate measure of writing-style variability in this case, even though it no longer corresponds to an actual chi-squared distribution.

Remark: Of course, while a larger V_i value indicates that one justice has a more variable writing style than another, it does not directly determine whether the justice relies more heavily on law clerks. Alternative explanations include that the justice edits his/her clerks work more carefully, or that some clerks are better than others at copying their justice's writing style, or that some justices naturally have a more variable writing style even when writing entirely on their own. So, we view the V_i measurements as *one* window into the reliance of justices on their clerks, but not a completely definitive one.

3   Bootstrap test of significance

The question remains whether the results from the previous section (e.g., that Kennedy has larger writing variability than Scalia) are the result of mere chance or are actually illustrative of different amounts of writing-style variability. Since we have already rejected the null hypothesis (that the null hypothesis that the function words are truly distributed uniformly and randomly), the quantity V_4 no longer follows a chi-squared distribution, so no simple analytic test of statistical significance is available.

So, to test significance, we perform a *bootstrap* test. Specifically, for each justice we shall select cases $a_1, a_2, \ldots, a_{100}$ uniformly at random (with repetition allowed, though this could also be done without repetition). For each such choice of 100 cases, we shall compute the variability measure V_4 as above. We shall repeat this 1000 times for each justice, thus giving a list of 1000 different possible values of V_4, depending on which list of 100 cases was randomly selected.

If we do this for two different justices, say for Kennedy and for Scalia, then this gives us 1000 x 1000 = 1000000 pairs of V_4 values. We then simply count the fraction of pairs

under which the V_4 for Kennedy is larger than the V_4 for Scalia, to give us an estimate of the *probability* that V_4 for Kennedy is larger than V_4 for Scalia, for a randomly-chosen selection of their judgments. We also use the pairs to estimate the distribution function for the difference of the V_4 variability for Kennedy, minus that for Scalia, and then use this estimated distribution function to compute the 95% confidence interval for the difference of V_4 for Kennedy minus that for Scalia. If this confidence interval is entirely positive, this indicates that Kennedy judgments have a more variable writing style than Scalia judgments, and that this conclusion is robust and statistically significant, rather than merely the result of chance variation.

Note that this bootstrap set-up has the additional advantage that, since the same number of judgments (100) are chosen for each justice at each step, any concerns about comparing different numbers of judgments are avoided.

3.1 Variability bootstrap results

We developed software [14] to compare the V_4 bootstrap values as above, using 1000 bootstrap samples each of size 100. We then ran this software to compare Kennedy and Scalia in this manner, obtaining the following results:

mean(Kennedy)	mean(Scalia)	P(Kennedy>Scalia)	95% C.I. for Kennedy−Scalia
8.181746	6.044971	0.997141	(0.579946, 3.857004)

That is, this bootstrap test determines that the probability that a randomly-selected sample of Kennedy's writings is more variable than a randomly-selected sample of Scalia's writings is over 99.7%, which is a near certainty. Furthermore, the 95% confidence interval (0.579946, 3.857004) for the difference in variabilities is entirely positive. So, we can conclude with confidence that, based on the V_4 chi-squared test, Kennedy's writings have more variable writing style than Scalia's.

Similarly, when comparing Souter to Scalia, we obtain:

mean(Souter)	mean(Scalia)	P(Souter>Scalia)	95% C.I. for Souter−Scalia
8.224238	6.028648	0.999629	(0.896046, 3.685900)

Or, comparing Kennedy's majority opinions to Kennedy's dissents, we obtain:

mean(majority)	mean(dissent)	P(majority>dissent)	95% C.I. for majority−dissent
8.190330	4.739693	0.999990	(1.843168, 5.175277)

Thus, we conclude with confidence that, as expected, both Kennedy's and Souter's majority opinion writing is more variable than that of Scalia, and furthermore Kennedy's majority opinion writing is more variable than his dissent opinion writing.

Similarly, when comparing Stevens to Scalia, we obtain:

mean(Stevens)	mean(Scalia)	P(Stevens>Scalia)	95% C.I. for Souter−Scalia
8.303315	6.008145	0.999903	(0.944279, 3.859911)

while comparing Breyer to Scalia, we obtain:

mean(Breyer)	mean(Scalia)	P(Breyer>Scalia)	95% C.I. for Souter−Scalia
6.886884	6.026134	0.936723	(−0.235533, 1.967974)

Thus, we can conclude that Stevens also has greater writing variability than does Scalia, while Breyer *may* have greater writing variability than does Scalia but that assertion is not definitive. (The conclusion about Stevens may be surprising, since Stevens also has a reputation for doing his own writing, see [3] p. 31. So, this result may indicate either that larger writing variability does not necessarily imply greater reliability on clerks, or that Stevens actually relied on clerks more than is generally believed).

3.2   Within-justice comparisons

It is possible to use this same V_4 bootstrap approach to compare different collections of judgments by the same justice.

For example, as justices age, their writings might get less variable (since they develop a more consistent style), or more variable (if they come to rely more on their law clerks). To test this, we perform V_4 bootstrap tests, as above, except now comparing a justice's majority opinions from the 1990s decade, to the same justice's opinions from the 2000s decade. Our results are as follows:

justice	mean(1990s)	mean(2000s)	P(1990s<2000s)	95% C.I. for 2000s−1990s
Kennedy	7.798363	8.218258	0.660653	(−1.497181, 2.377671)
Scalia	5.835250	6.161246	0.761957	(−0.588609, 1.228179)
Souter	8.796725	7.363916	0.036366	(−3.100154, 0.122444)
Stevens	7.897011	8.675261	0.792364	(−1.006673, 2.612827)
Breyer	8.053281	6.008839	0.001249	(−3.265005, −0.795313)
Ginsburg	7.560492	7.650178	0.565943	(−1.217240, 1.326025)
Thomas	8.172542	7.539790	0.231573	(−2.361786, 1.111958)
Rehnquist	5.934025	5.661661	0.265052	(−1.126469, 0.590608)

Looking at these results, there is no clear pattern. The only statistically significant result is that Breyer apparently had greater variability in the 1990s than in the 2000s. Since this is just one test out of many performed and does not conform to any obvious interpretation or "story", we are inclined to regard this as a mere chance event.

Another way to compare a justice's writing is to look at those judgments which were in the first half of a session (i.e., September through March) versus those judgments in the second half (i.e., April through August). The reason why judgments early in a session may appear different from those later in a session is because law clerks rotate annually; thus, writing variability over the course of a session may increase if a given justice delegates more work to his clerks, or may diminish if clerks better learn the preferences of their justice. That is, increasing variability may indicate a justice's increased trust, and therefore increased delegation or lower oversight to the clerk; conversely, decreasing variability could reflect increased understanding by the clerks of their justice's preferred writing style.

Our results for this comparison are as follows:

justice	mean(first)	mean(second)	P(first<second)	95% C.I. for second−first
Kennedy	7.141271	8.959166	0.951162	$(-0.318166, 3.982379)$
Scalia	5.983945	5.987886	0.505288	$(-0.913589, 0.899150)$
Souter	9.308629	7.473313	0.024944	$(-3.809809, -0.000795)$
Stevens	9.018360	7.694468	0.079028	$(-3.372943, 0.466490)$
Breyer	6.002088	7.567111	0.996030	$(0.406776, 2.743431)$
Ginsburg	7.735728	7.586115	0.401026	$(-1.404674, 1.176858)$
Thomas	7.078534	8.669194	0.966751	$(-0.111958, 3.334510)$
Rehnquist	5.247192	6.612502	0.999459	$(0.505359, 2.226674)$

This time, it appears that Souter is statistically significantly more variable in the *first* half of court sessions, while Breyer and Rehnquist are statistically significantly more variable in the *second* half of court sessions. This is an interesting finding and may warrant further investigation, though its significance and interpretation at this point are not completely clear.

4   Authorship identification

A related question is whether it is possible to identify which justice is the (recorded) author of a judgment, based only on the writing style. Enquiries with constitutional law scholars even those who follow the USSC closely cannot do this effectively (they can perhaps identify authorship based on recognition of the case, or the decision's conclusion or arguments, but

not based purely on writing style). We now consider the extent to which this can be done by appropriate computer algorithms. (This question is thus similar in spirit to the Shakespeare authorship question [15, 2, 18], and also to the Federalist Papers authorship question [10]. Of course, there is one important different here, since recorded authorship of USSC judgments is already completely *known*. However, we still view this as a useful test of the extent to which different USSC justices have identifiably distinct writing styles.)

We shall consider both naive Bayes classifiers and linear classifiers, and shall see that each performs quite well at this task, achieving success rates as high as 90%. (Other possible approaches include neural networks, support vector machines, etc., but for simplicity we do not consider them here.)

In each case, we shall consider a particular pair of justices (say, Justice A and Justice B). We shall consider the collection of all USSC judgments whose recorded author is either A or B, and shall partition this collection into a disjoint training set and testing set. Using only the training set, we develop a model for classifying judgments as being authored by the A or B. We then test to see if our model classifies authorship correctly on the testing set.

4.1   Naive Bayes classifier

We begin with a naive Bayes classifier. More specifically, we assume that conditional on the recorded author being Justice A, the conditional distribution of the fraction f_j of function word j appearing in the judgment is normally distributed. We further assume that the corresponding mean and variance are given by the sample mean and variance of all judgments by Justice A in the training set. In addition, we assume (since we are being "naive") that these different fractions f_j (over different function words j) are all conditionally independent.

Together with the uniform prior distribution on whether the author is Justice A or B, this gives the log-likelihood for a given judgment being authored by Justice A, namely

$$loglike(A) \ = \ C - \sum_{j=1}^{63} \left(\frac{1}{2} \log(v_j) + (f_j - m_j)^2/2v_j \right),$$

for some constant C, where f_j is the fraction of words which are reference word j in the test judgment under consideration, and where m_j and v_j are the sample mean and variance of the fraction of words which are reference word j, over all judgments in the training set authored by A.

Similarly, we can compute $loglike(B)$. The model then classifies the test judgment as being authored by A if $loglike(A) > loglike(B)$, otherwise it classifies it as being authored by B.

4.2   Linear classifier

Another approach is a *linear classifier*. Specifically, let $\mathcal{T}$ be a training set consisting of various judgments by A or B, with $|\mathcal{T}| = n$. We consider the linear regression model

$$\mathbf{Y} = \mathbf{x}\beta + \epsilon,$$

where ϵ is an $n \times 1$ vector of independent zero-mean errors. Here $\mathbf{Y}$ is an $n \times 1$ vector of ± 1, which equals -1 for each judgment in the training set authored by A, or $+1$ for each judgment in the training set authored by B. Also, $\mathbf{x}$ is the $n \times 64$ matrix given by

$$\mathbf{x} = \begin{pmatrix} 1 & f_{1,1} & f_{1,2} & \cdots & f_{1,63} \\ 1 & f_{2,1} & f_{2,2} & \cdots & f_{2,63} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & f_{n,1} & f_{n,2} & \cdots & f_{n,63} \end{pmatrix},$$

where $f_{i,j}$ is the fraction of words in judgment i (in the training set) which are function word j. For this model, the usual estimate for β (which corresponds to the least-squares estimate if the ϵ_i are assumed to be iid normal) is given by

$$\hat{\beta} = \left(\mathbf{x}^T \mathbf{x}\right)^{-1} \mathbf{x}^T \mathbf{Y}.$$

Once we have this estimate $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_n)$, then given a fresh test judgment having function word fractions $g_1, g_2, \ldots, g_{63}$, we can compute the linear fit value

$$\ell = \hat{\beta}_0 + \sum_{j=1}^{63} \hat{\beta}_j \, g_j.$$

Then, if $\ell < 0$ we classify the test judgment as being authored by A, otherwise we classify it as being authored by B.

Below we shall consider both the linear classifier and the naive Bayes classifier. We shall see that, generally speaking, the linear classifier outperforms the naive Bayes classifier, sometimes significantly so.

4.3   Testing accuracy via cross-validation

To test the accuracy of our model, we use *leave-one-out cross-validation*. That is, for each judgment by either A or B, we consider that one judgment to be the test set, with all other judgments by either A or B comprising the training set. We then see whether or not our model classifies the test judgment correctly. Finally, we count the number of correct classifications, separately over all judgments by A, and over all judgments by B.

4.3.1   Results: naive Bayes classifier

We ran software [14] to perform the cross-validation test using the naive Bayes classifier, for various pairs of justices A and B. Our results were as follows:

Justice A	Justice B	success(A)	success(B)
Scalia	Kennedy	$128/156 = 0.820513$	$127/147 = 0.863946$
Scalia	Souter	$125/156 = 0.801282$	$113/143 = 0.790210$
Scalia	Stevens	$117/156 = 0.750000$	$119/148 = 0.804054$
Scalia	Rehnquist	$126/156 = 0.807692$	$102/127 = 0.803150$
Kennedy	Souter	$119/147 = 0.809524$	$106/143 = 0.741259$
Kennedy	Stevens	$113/147 = 0.768707$	$116/148 = 0.783784$
Kennedy	Rehnquist	$115/147 = 0.782313$	$101/127 = 0.795276$
Souter	Stevens	$100/143 = 0.699301$	$121/148 = 0.817568$
Rehnquist	Breyer	$114/127 = 0.897638$	$99/121 = 0.818182$
Rehnquist	Stevens	$80/127 = 0.629921$	$105/148 = 0.709459$
Rehnquist	Thomas	$100/127 = 0.787402$	$66/140 = 0.471429$
Scalia	Scalia dissent	$145/156 = 0.929487$	$90/108 = 0.833333$
Stevens	Stevens dissent	$128/148 = 0.864865$	$138/205 = 0.673171$
Scalia	Stevens dissent	$143/156 = 0.916667$	$145/205 = 0.707317$

We see from these results that our naive Bayes classifier performs fairly well, often achieving a success rate near 80%. (This is fairly consistent across all pairings, not just those shown in the above table; in particular, the success rate for majority opinions is over 70% for all possible pairings except for five: Scalia-Thomas, Souter-Stevens, Rehnquist-Stevens, Stevens-Thomas, and Rehnquist-Thomas.) This appears to be quite a good performance, especially considering the minimal assumptions that have gone into the model. (Presumably a more sophisticated model could achieve even higher success rate.) So, we see this as evidence that USSC judgment authors can indeed be distinguished by their writing style, in fact just by the pattern of fractions of function words used.

We also note that there is some variability in which justices' writing styles are most easily distinguished. For example, Rehnquist and Breyer are apparently relatively easy to distinguish from one another, while Rehnquist and Thomas are rather more difficult.

The algorithm does not perform as well on the dissent opinions, presumably because they tend to be shorter and more variable. In fact, when comparing dissent to majority opinions, the algorithm tends to classify too many judgments as being from the majority collection, and this weakness remains whether the majority and minority collections are from the same justice or from two different justices.

4.3.2 Results: linear classifier

We also ran software [14] to perform the cross-validation test using a linear classifier, again for various pairs of justices A and B. Our results were as follows:

Justice A	Justice B	success(A)	success(B)
Scalia	Kennedy	$135/156 = 0.865385$	$134/147 = 0.911565$
Scalia	Souter	$136/156 = 0.871795$	$128/143 = 0.895105$
Scalia	Stevens	$127/156 = 0.814103$	$134/148 = 0.905405$
Scalia	Rehnquist	$137/156 = 0.878205$	$106/127 = 0.834646$
Kennedy	Souter	$123/147 = 0.904762$	$127/143 = 0.888112$
Kennedy	Stevens	$133/147 = 0.904762$	$132/148 = 0.891892$
Kennedy	Rehnquist	$135/147 = 0.918367$	$108/127 = 0.850394$
Souter	Stevens	$121/143 = 0.846154$	$125/148 = 0.844595$
Rehnquist	Breyer	$118/127 = 0.929134$	$109/121 = 0.900826$
Rehnquist	Stevens	$90/127 = 0.708661$	$124/148 = 0.837838$
Rehnquist	Thomas	$81/127 = 0.637795$	$93/140 = 0.664286$
Scalia	Scalia dissent	$148/156 = 0.948718$	$95/108 = 0.879630$
Stevens	Stevens dissent	$125/148 = 0.844595$	$167/205 = 0.814634$
Scalia	Stevens dissent	$149/156 = 0.955128$	$180/205 = 0.878049$

Comparing these results with those from the previous subsection shows that the linear classifier performs even better than the naive Bayes classifier, with success rates often close to 90%. (This is again fairly consistent across all pairings; in particular, the success rate is above 80% for all possible majority opinion pairings with the exception of Rehnquist-Stevens and those involving Thomas.) This provides further, even stronger evidence that it is indeed possible to distinguish between different USSC justices' judgments solely on the basis of writing style.

Once again, there is some variability in which justices' writing styles are most easily distinguished. For example, success rates for distinguishing Rehnquist from Breyer are over 90%, while those for distinguishing Rehnquist and Thomas are more like 65%.

Regarding the dissenting opinions, generally the linear classifier performs better there too. In particular, it is much less prone to incorrectly classifying almost all judgments as being from the majority opinion collection. Rather, it usually classifies over 80% of the dissenting opinions as indeed being dissenting opinions. So, this again illustrates quite good performance.

5   Summary

In this paper, we have presented methodology and software for two different investigations of USSC judgments, by using statistical properties of function words.

Firstly, we have investigated the variability of writing style over various collections judgments, in particular of majority decisions written by different justices. We have seen that it is possible to uncover statistically significant evidence that one USSC justice (e.g. Kennedy) has greater writing-style variability than another justice (e.g. Scalia), which may indicate (though we cannot directly prove this) that the first justice relies on law clerk assistance to a greater extent than does the second justice.

Secondly, we have investigated the extent to which unknown authorship of USSC judgments can be determined based solely on function word statistics. We have seen that both naive Bayes classifiers and linear classifiers perform fairly well at this task, achieving cross-validation success rates approaching 90%. While authorship is typically known for all USSC opinions, our approach reveals that justices – even with contributions by clerks – have writing styles which are distinguishable from one another. (In a different direction, one could perhaps use function words to identify authorship for the handful of *per curiam* decisions in which the Court does not reveal authorship, though we do not pursue that here.)

Overall, we hope that the methodology and software [14] presented here will provide useful insights into USSC writings, as well as a helpful starting point for other statistical investigations into other bodies of writing in other contexts.

References

[1] S. Argamon and S. Levitan (2005), Measuring the usefulness of function words for authorship attribution. ACH/ALLC 2005.

[2] K. Burns (2006), Bayesian inference in disputed authorship: A case study of cognitive errors and a new system for decision support. Information Sciences **176(11)**, 1570–1589.

[3] W. Domnarski (1996), In the Opinion of the Court.

[4] M.J. Evans and J.S. Rosenthal (2003), Probability and Statistics: The Science of Uncertainty. W.H. Freeman Publishers, New York.

[5] C.C. Fries (1952), The Structure of English. Harcourt, Brace, and Company, New York.

[6] K.H. Jamieson and M. Hennessy (2007), Public understanding of and support for the courts: survey results. Georgetown Law Journal **95**, 899.

[7] A. Kozinski (1999), Conduct Unbecoming (book review). Yale Law Journal **108**, January 1999, 835.

[8] E. Lazarus (1998), Closed Chambers: the Rise, Fall, and Future of the Modern Supreme Court.

[9] G.A. Miller, E.B. Newman, and E.A. Friedman (1958), Length-frequency statistics of written English. Information and Control **1**, 370–389.

[10] F. Mosteller and D.L. Wallace (1964), Applied Bayesian and Classical Inference: The Case of The Federalist Papers. Springer-Verlag, New York.

[11] D.M. O'Brien (1986), Storm Center: The Supreme Court in American Politics.

[12] T.C. Peppers (2006), Courtiers of the Marble Palace: the Rise and Influence of the Supreme Court Law Clerk.

[13] W.H. Rehnquist (2002), The Supreme Court.

[14] J.S. Rosenthal (2009), Software for downloading and analysing USSC texts. To be made freely available at: probability.ca/ussc

[15] O. Seletsky, T. Huang, and W. Henderson-Frost (2007), The Shakespeare authorship question. Unpublished manuscript, Dartmouth College.

[16] Supreme Court collection. Cornell University Law School.
http://www.law.cornell.edu/supct/

[17] A. Ward and D.L. Weiden (2006), Sorcerers' apprentices: 100 years of law clerks at the United States Supreme Court.

[18] Wikipedia, Shakespeare authorship question.
http://en.wikipedia.org/wiki/Shakespeare_authorship_question
(Retrieved Nov. 16, 2009.)

[19] B. Woodward and S. Armstrong (1979), The Brethren: Inside the Supreme Court.

[20] J.L. Wyatt (1980), Can function word distribution indicate authorship? ACM SIGLASH Newsletter, ACM, New York.