Appendices

for

"Diversity Effects or Dissent Aversion? Identification and Estimation in Judicial Panel Voting"

Charles M. Cameron
Center for the Study of Democratic Politics, Princeton University

Lealand Morin
College of Business, University of Central Florida

Harry J. Paarsch
College of Business, University of Central Florida

Friday 4th June, 2021

Appendix A: Simulation Evidence

We conducted a series of simulations to investigate the properties of the econometric model that we described in the body of the paper. In each simulation, we generated data from the econometric model and estimated the parameters in that model by maximizing the logarithm of the likelihood function. The goal of this simulation exercise was twofold: (1) to validate the functions in the `TVN_Probit_Lib.R` library; (2) to verify that the parameters in the model are identified numerically in the samples that we encounter in practice. The parameters in the model include the slope coefficients for the judge-specific covariates $\boldsymbol{\beta}$, the diversity effects on the same covariates from other judges, $\boldsymbol{\gamma}$, and the dissent aversion parameter δ. The functions in the `TVN_Probit_Lib.R` calculate and optimize the logarithm of the likelihood function to estimate these parameters.

The covariates included a constant and two other randomly-drawn binary variables, from a Bernoulli distribution with equal probabilities, for each of five judges. The judges were matched in all permutations of the five judges, comprising a set of sixty distinct judicial panels. Each judicial panel met three times, making a total sample of 180 cases. Three pairs of judge-specific covariates were allocated to each case from the matrix of covariates, to match the judges on the judicial panel. The true values of the parameters were set to $\boldsymbol{\beta} = [0.25, 1, 2]^\top$, for the intercept and the slope coefficients on the judge-specific covariates. The slope coefficients for the peer effects, on the cross-judge covariates, were set to $\boldsymbol{\gamma} = [-0.5, -0.1]^\top$. The dissent aversion parameter was set to $\delta = 0.1$.

We drew 100 realizations of the dataset with the sample of 180 cases. The innovations $\varepsilon_i, i = 1, 2, 3$, for the latent intent equations were generated from a trivariate standard normal distribution, which has no correlation between the three variables. The first simulation was conducted by initializing the optimization at the true values, to reduce the computation time required. In Table 1, we present summary statistics for the realizations of the estimates. It appears as though the model is numerically identified, even in such small samples, and the maximum-likelihood estimator is consistent.

Parameter	β_0	β_1	β_2	γ_1	γ_2	δ
True Value	0.25000	1.0000	2.000	−0.50000	−1.0000	0.10000
Minimum	−0.24589	0.3436	1.536	−0.88316	−1.5815	0.06157
1st Quartile	0.09368	0.8597	1.877	−0.63246	−1.1608	0.05601
Median	0.23648	1.0318	2.037	−0.50860	−1.0027	0.09797
Mean	0.24700	1.0692	2.071	−0.50175	−1.0320	0.09106
3rd Quartile	0.37203	1.2236	2.249	−0.38777	−0.9029	0.12248
Maximum	0.97762	2.3202	2.816	−0.03476	−0.5310	0.23362

Tab. 1: Simulation of Estimates (starting at the true values)

In the above round of simulations, the optimization algorithm started from the true values of parameters. This is not realistic, however, because the true values are typically unknown in practice. We considered the possibility that the performance could be affected by local maxima

or an otherwise poorly-behaved likelihood function. To investigate this possibility, we primed the optimization algorithm with the zero vector as the starting values in the next round of simulations. In Table 2, we present summary statistics for the realizations of the second set of estimates. The

Parameter	β_0	β_1	β_2	γ_1	γ_2	δ
True Value	0.2500	1.0000	2.000	−0.500000	−1.0000	0.10000
Minimum	−0.2932	0.4555	1.455	−1.151891	−1.5465	−0.04055
1st Quartile	0.1221	0.8240	1.867	−0.589969	−1.1761	0.05862
Median	0.2930	1.0150	1.987	−0.495165	−1.0226	0.08485
Mean	0.2799	1.1002	2.009	−0.479760	−1.0319	0.08394
3rd Quartile	0.4188	1.2664	2.153	−0.354121	−0.8934	0.11668
Maximum	0.7542	3.1339	2.673	0.003675	−0.6514	0.19248

Tab. 2: Simulation of Estimates (starting at the zero vector)

optimization appears to work just as well when the true values are unknown. Overall, we conclude that our likelihood function and the optimization functions are coded correctly, the estimates are unbiased, and the parameters in the model are identified.

Appendix B: Data Description

Westlaw Database

Structure of Documents

The primary raw source of data is the Westlaw database, which includes a collection of documents that describe cases heard in the United States Courts of Appeals. As a preliminary test of our ability to organize the data, we collected information for all cases that contain the term "sexual harassment" that Westlaw has classified under the "Labor and Employment" category. We selected a sample of cases heard during the twenty-year period spanning the years 2000 through 2019. The court documents are available in several formats: Microsoft Word 97/2003 `doc` format, rich text `rtf` format, and `pdf` format. We downloaded the files in `doc` format and translated them to `txt` format, using the Python module `win32com`.

Our data-collection strategy is built on the systematic structure of these `txt` files. There are a few variants of documentation for different case types, and some changes in format over the years, but the information is mainly organized with information listed in the same order and written with stable patterns in the text.

We wrote a Python module called `legalbeagle`, which tracks down information from the text in court documents, and fetches these to be stored in a data frame, with one row for each case. The data in each line of text are categorized into one of several fields using functions of the form `is_[field name](line)`, which identify whether the line of text matches the characteristics of one of the fields of interest. These functions are used either within a function of the form

get_[field name](file, last_line) or to trigger a call to such a function. These functions, in turn, either read from the file or continue from the last line read by the previous function to parse a field from the court document.

Some of the fields are described over a number of lines that varies across court cases, so the function reads until another field type is recognized, signaling the conclusion of the data collection for the previous field. For this reason, other types of fields are collected to ensure the reliability of the collection of the contents of subsequent fields, even if some of these fields are not directly used in our statistical analysis.

An example will clarify the data collection process. The most common layout of a file, once converted into txt format, begins as follows:

```
KeyCite Yellow Flag - Negative Treatment
  Distinguished by Wells v. Hi Country Auto Group, D.N.M., November 13, 2013
656 F.3d 1277
United States Court of Appeals,
Tenth Circuit.
Christie HELM, Plaintiff-Appellant,
v.
State of KANSAS, Defendant-Appellee.
No. 10-3092.
|
Sept. 7, 2011.
Synopsis
Background: Administrative assistant brought action against state,
alleging sexual harassment over 10 year period ...

Holdings: The Court of Appeals, Ebel, Circuit Judge, held that:

[1] judge was not alter ego of the state;

[2] judge's sexual harassment of assistant did not culminate in assistant's
termination;

[3] state exercised reasonable care to prevent sexual harassment;

[4] state acted reasonably to correct harassing behavior in response to
assistant's complaint; and

[5] assistant unreasonably failed to take advantage of preventive or corrective
opportunities provided by state to avoid harm.
```

```
Affirmed.

Procedural Posture(s): On Appeal; Motion for Summary Judgment.


West Headnotes (13)


[1]

Civil RightsPractices prohibited or required in general;  elements
CivilRightsHostile environment;  severity, pervasiveness, and frequency


Actionable sexual harassment under Title VII includes not only economic
or tangible discrimination but also discriminatory intimidation, ridicule,...

3 Cases that cite this headnote

...
```

The first lines are blank, corresponding to the upper margin in the original Word document in `doc` format. Although not interesting in itself, this first non-field illustrates a source of irregularity in the data: consecutive fields are occasionally separated by an unknown number of blank spaces. Furthermore, some fields span multiple lines and some span a variable number of lines.

For example, the next line begins with `KeyCite Yellow Flag` and is an addition from Westlaw, which indicates, over two lines, that this case was referenced in a later case. The next line `656 F.3d 1277` is an identifier that—perhaps, incorrectly—we call a `case_code`. It is a sequence of three strings separated by spaces; the first and last are sequences of digits and the middle string is an alphanumeric code. The next line is `United States Court of Appeals,` (comma optional) which is followed by the circuit number, such as `Tenth Circuit`. In most files, these fields are listed in a fixed sequence within the first five non-empty lines.

The next several lines list the parties involved in the case. In the simplest form, as in the example above, the plaintiff-appellant is named, complete with this labelling, followed by a line containing only `v.` On the next line, the name of the defendant-appellee is listed and labelled similarly. Some court records list multiple parties before or after the `v.`, and some court records list multiple parties separated by multiple `v.`s. In some cases with multiple parties, the multiple parties are listed on multiple lines but in many other cases the parties are listed within a single line. Although this may be useful information for each case later on, we collected the list of names of parties but did not separate nor classify them, as we do not yet have a need for this data.

Following the list of parties—in all files—is the case number, perhaps called the "docket number," which is a unique identifier that can be used as a key to join with data from other databases. The case number is commonly written as above, in the form `No. YY-1234`. There exists some variety with which this information is listed. For some files, the case number is written as either `Docket No. YY-1234` or `No. YYYY-1234`, or, without the label `No.`, as simply `YY-1234` or even `YYYY-1234`. In some files, the case number has the suffix `-cv` appended, as in `No. YY-1234-cv`. In other files, multiple case numbers are listed, as in `Nos. YY-1234, YY-4567`, and a few files have many case numbers listed. Later versions of the `legalbeagle` module will include functions for parsing the case number in the form `YY-1234` to join with information contained in other databases, such as those available on the webpage of the Department of Justice.

The text after the case number is often separated by a single line with a pipe symbol, `|`. The next field is a date. In the simplest form, as above, only a single date is listed. If the sequence of events in the case took place over multiple days, these events and dates will also be listed on the following lines, often separated by blank lines or another `|`. For example, one case lists the following:

```
|
Argued: Jan. 13, 2011.
|
Decided and Filed: June 28, 2011.
|
```

Since we have no immediate need for this information, the dates are collected and stored only to continue the flow of the program through the file.

The next line describes the case in sentences. The first line of this description contains only the word `Synopsis`. On the next non-empty line, the case is described in sentences following the header `Background:`, although, in the files from cases heard before 2004, the text of the background is written without the header `Background:`. Still, it is easy to collect this information because it usually appears in one line of text, even though it often spans multiple lines with word-wrapping. We have no plans to parse any information from this field because it would be more difficult, due to the unstructured nature of the field, however, we might find a need for this information later. Note that the `Synopsis` and/or `Background:` lines do not appear in every case: some cases are judged *per curiam* and the case file is abbreviated. For this reason, and perhaps several others, the background information is not verified as being recorded for a material fraction of the cases. Perhaps as much as ten percent of cases do not have the background recorded, although accurately measuring this fraction is problematic because of the unstructured nature of the field, when the header `Background:` is omitted.

After a space, the next set of information is a sequence of double-spaced points describing the holdings. The holdings are preceded by a single line in the form above, as in

```
Holdings: The Court of Appeals, Ebel, Circuit Judge, held that:

[1] ...
```

The next several lines comprise a sequence of statements, enumerated in square brackets, as in `[1]` above. After the last numbered point, the next line is blank and is followed by a statement of the outcome of the case.[1] In many cases, this is simply the word `Affirmed.` In others, the outcome of the case takes on a hybrid form, such as

```
Affirmed in part and reversed in part.
```

A complete listing of the outcomes of the case is listed in Table 3. We should think about how we

Frequency	Outcome
9,999	Examples
1,000	`Affirmed.`
250	`Affirmed in part and reversed in part.`
1,000	Other

Tab. 3: Outcomes of Cases in U.S. Courts of Appeals

use this field to determine which cases to include in our analysis and how to categorize them.

The holdings and outcome is followed by a statement of procedural posture. In the simplest cases, it may take on the form

```
Procedural Posture(s): On Appeal; Motion for Summary Judgment.
```

In other cases, there may be several items listed in this field, such as

```
Procedural Posture(s): On Appeal; Motion for Summary Judgment;
Motion for Judgment as a Matter of Law (JMOL)/Directed Verdict.
```

These are listed in a single line of text, but the second statement is shown above on a separate line to show the added item. I don't understand these terms enough to know what to do with these, but it seems as though these are structured in a such way that it will be easy to parse into separate categories: the items are separated by semicolons and the items take on only so many values.

The next section is often a lengthy listing of quotations from legal documents. It is a numbered list of notes under the heading `West Headnotes (X)`, which indicates the enumerated list of notes from `[1]` to `[X]`. The current version of the `legalbeagle` module skips this section.

A few pages later, the next section begins with the header `Attorneys and Law Firms`. A typical example takes on the following form:

[1] I suspect the technical term for this outcome is "the verdict." Nevertheless, I think we should go through the exercise of identifying the proper terminology for all the fields in the court documents, including the term "court documents." For example, during one of my meetings with the Westlaw representatives, I learned that terminology for seemingly similar features differ between cases in the courts of appeals and trials in trial courts. The distinction between the terms "case" and "trial" is just such an example that I do not yet precisely understand.

Attorneys and Law Firms
*505 ARGUED: Justin S. Gilbert, Gilbert, Russell, McWherter PLC, Jackson, Tennessee, for Appellant. Christopher W. Cardwell, Gullett, Sanford, Robinson & Martin, PLLC, Nashville, Tennessee, for Appellee. ON BRIEF: Justin S. Gilbert, Gilbert, Russell, McWherter PLC, Jackson, Tennessee, Gregory G. Paul, Morgan & Paul, PLLC, Sewickley, Pennsylvania, for Appellant. Christopher W. Cardwell, Mary Taylor Gallagher, Gullett, Sanford, Robinson & Martin, PLLC, Nashville, Tennessee, for Appellee.
Before: MERRITT, ROGERS, and WHITE, Circuit Judges.

MERRITT, J., delivered the opinion of the court. ROGERS (pp. 513-14), and WHITE (pp. 514-20), JJ., delivered separate opinions concurring in part and dissenting in part.

The first line contains a list of attorneys and law firms representing the plaintiff-appellant. The second line usually contains a list of attorneys and law firms representing the defendant-appellee. This passage often spans three lines (without word-wrapping), but sometimes the attorneys are listed in a single line. In any case, the last line in this section is especially important for our research question, since it lists the names of the judges in the judicial panel. This line usually takes the form shown above, as in `Before: MERRITT, ROGERS, and WHITE, Circuit Judges.` It might, however, list two judges as, for example, `Circuit Judges` and a third judge with another title. The structure of this sentence is standardized enough that it should not be too difficult to separate the names of the judges. The judges' names are often—but not always—stated in upper case letters. The names are sometimes listed with first and middle initials and sometimes with first names and middle initials. It is a reasonable possibility that the last names of judges will not be unique. For example, we will have to distinguish between judges.

It is possible, though I have not verified this claim, that judges who share the same surname are listed with initials or first names. In case this does not the case, we should find another strategy for recording the judges' names along with a unique identifier, possibly by scraping a hyperlink from from the documents in another format, such as `doc` or `pdf`. Regardless of how we identify the judges, an important step is to compile a master list of the unique names of judges on these judicial panels and attempt to match it to the information in the database of judges. We have not yet collected information for judges.

A related set of information is the list of opinions. It is typically labelled as the opinion of one of the judges and may be followed by the opinions of some of the other two judges, particularly in the case of a dissenting opinion. The opinions are often written with excerpts from oral arguments or testimony and also excerpts from other legal documents. For now, this information is skipped, since it takes on an irregular format, however, it is worth investigating in order to characterize the outcomes of cases with partial verdicts, such as `Affirmed in part and reversed in part.` For our application, it matters whether the "reversed in part" part is a result of a disagreement

between the judges or a unanimous decision to reverse part of the verdict in the trial case that was appealed.

For instance, the case in file number `088`, heard in 2011, has the outcome:

```
Holdings: The Court of Appeals, Merritt, Circuit Judge, held that:

[1] employee filed "charge" with Equal Employment Opportunity Commission, ...

[2] supervisor's derogatory statements to employee were based on race, ...

[3] other adverse treatment that employee suffered was not race-based; and

[4] supervisor's statements were not sufficiently severe or pervasive ...

Affirmed in part and reversed in part.

Rogers, Circuit Judge, filed an opinion concurring in part and dissenting in part.

Helene N. White, Circuit Judge, filed an opinion concurring in part and
dissenting in part.
```

We will have to consider carefully how we categorize this sort of case. An important next step is to tabulate the frequency of each outcome, to determine whether this outcome is unusual. In some documents, the judges' opinions constitute the bulk of the court document. In others, the opinions are briefly stated, often in a single sentence, stating little more than the verdict.

Database Constructed from Westlaw Documents

After running the scripts with the functions in the `legalbeagle` module on the court documents, the results are compiled into a data frame. I collected files over the twenty-year period from 2000 to 2019. The sample includes files that contain the phrase "sexual harassment" and cases that Westlaw characterized in the "Labor and Employment" category. On average, about 200 cases were heard each year but the cases in this category became less frequent over time: over 300 such cases were heard in 2000 and the number of these cases declined over the sample, with around 150 cases per year in the past decade.

- `file_name`: A string of the form `"001 - Helm v Kansas.txt"`. The number is generated by the Westlaw GUI when the files are downloaded in batches of 200—the limit in Westlaw, which is close to the number of cases per year. Once downloaded, I changed the file names in unix to follow a sequence counting up to the number of cases for a particular year. The rest of the file name is the listing of the main parties involved in the case.

- case_code: A string of the form 123 abcdef 456 representing the... I'm not sure yet, but it is easy to collect. Examples: 656 F.3d 1277, 643 F.3d 502, or 175 Fed.Appx. 207.

- circ_num: A string of the form Nth Circuit., which represents the circuit number from First Circuit to Twelfth Circuit, and also D.C. Circuit.

- pla_appnt_1 to pla_appnt_3: A string listing the name(s) of the plaintiff(s)-appellee(s), such as Pamela D. FYE, Plaintiff-Appellee, but possibly a list of multiple plaintiffs-appellees on one line. Three fields are collected since the parties are sometimes listed on several lines.

- def_appee_1 to def_appee_4: A string listing the name(s) of the defendant(s)-appellant(s), such as Pamela D. FYE, Plaintiff-Appellee, but possibly a list of multiple defendants-appellants on one line. Four fields are collected since the parties are sometimes listed on several lines. I can check the numbers later but it seems as though more cases have a longer list of defendant(s)-appellant(s) than plaintiff(s)-appellee(s). I suppose this is because each person has eight fingers to point at other people.

- case_num: A string of the form No. YY-1234, YY-1234, YYYY-1234, or Docket No. YY-1234, which is probably called the docket number. Sometimes multiple docket numbers are listed in a single string.

- case_date_1 to case_date_4: A string of the form Month DD, YYYY, for the cases in which one date is listed. Sometimes several dates are listed, each with the name of the event that took place, such as Submitted, Filed,Argued, etc.

- background: A string containing a description of the case, in several sentences, often preceded by the header Background:. Follows a line with the heading Synopsis.

- holdings_hdr: A string of the form Holdings: The Court of Appeals, Smith, Circuit Judge, held that:, which is followed by a list of holdings in the case.

- outcome: A string with a single word, such as Affirmed, but other possibilities are more complex, particularly in the case of dissenting opinions, such as, Affirmed in part and reversed in part. Other terms include Vacated, Granted, Denied, and Remanded.

- posture: A string beginning with the header Procedural Posture(s):, indicating the procedural posture, whatever that means. Procedural Posture(s): On Appeal; Motion for Summary Judgment.. Sometimes multiple items are listed.

- judicial_panel: A string of the form Before: MERRITT, ROGERS, and WHITE, Circuit Judges. that lists the names of the judges on the judicial panel. The header Present: is sometimes in the place of the header Before:.

Next Steps

- Create a table of the number of cases per year.
- Create tables of frequencies of outcomes for each field.
- Start with the outcomes of cases, or verdicts, to classify these outcomes or make decisions about cases to exclude, if any.
- Parse the judges' names from the `judicial_panel` variable.
- Compile a master list of judges' names as keys to `LEFT JOIN` with information from a database of judges.
- Compile a list of case numbers (docket numbers?), of the form `YY-1234` to `LEFT JOIN` with information from the databases on the Website of the Department of Justice. Use this joined table to validate against other variables that are common to both tables.
- Read some files related to the cases in the sample in the `dta` datasets to compare against the fields read by the functions in the `legalbeagle` module.

Database of Corresponding Trials in the Westlaw Database

To be added.

Database of Judges' Characteristics from Westlaw Litigation Analytics

To be added.

Database of Cases in the U.S. Courts of Appeals from the Department of Justice

To be added.