WORKING PAPER 7/2024 (STATISTICS)

Yet another case of Nordic exceptionalism?: A quantitative approach to an intra-Nordic and an international comparison of supreme courts' constitutional reasoning

Nicklas Pettersson and Katalin Kelemen

ISSN 1403-0586

Örebro University School of Business
SE-701 82 Örebro, Sweden

Yet another case of Nordic exceptionalism?

A quantitative approach to an intra-Nordic and an international comparison of supreme courts' constitutional reasoning

Nicklas Pettersson, Katalin Kelemen

Abstract

We present a systematic quantitative approach how to analyze the reasons that judges in Nordic countries publicly adduce for their decisions in constitutional matters, as implemented in the Nordic CONREASON Project. Based on encodings of forty (per court) purposively selected landmark cases, common traits and patterns of constitutional argumentative practices in each of the Nordic supreme courts were identified and an international comparison were made to courts from related studies. Our results provided strong support that, regarding specific aspects (on a univariate level), one or more courts typically tended to deviate from the other Nordic courts. Also, in a multivariate worldwide comparison there were variation between the Nordic supreme courts. However, although not detached from other supreme courts, the Nordic supreme courts seemed to occupy an area of their own on the international map of constitutional reasoning.

Keywords: constitutional reasoning, quantitative comparative law,
empirical legal studies, Nordic exceptionalism, realist decision-making

JEL Codes: C10, C38, C53, K10, K40, N40

Contents

1 Background and purpose

The aim of the Nordic CONREASON Project was a twofold comparative analysis of argumentative practices of supreme courts in constitutional cases: (1) among the Nordic countries, and (2) between Nordic countries and other countries worldwide. The focus in (1) was to discover common traits and trends among the Nordic countries, and in (2) to compare the Nordic region to other countries worldwide. The aim was intended to be accomplished by systematically documenting the argumentative practice of the Nordic supreme courts in their leading constitutional cases, and through incorporating the methodology of and the data collected within previous similar projects.

The primary research question may be phrased as: In what aspects are the constitutional reasoning of Nordic supreme courts exceptional? In other words, to what extent does the style of reasoning of Nordic supreme courts resemble that of other supreme courts around the world? To address the question, a combination of qualitative and quantitative methods are employed. Through triangulation, the results from the quantitative and the qualitative analysis control and complement each other, increasing the comparability across the jurisdictions and contributing to the identification of various argumentation patterns.

The following paper documents the quantitative methodology used in the project, and provide some of the main results and scopes for further research.

2 Design and data collection

The Nordic CONREASON project collected data on the reasoning-practice of the supreme courts of the five Nordic countries: the Swedish Högsta domstolen, the Norwegian Høgsterett, the Icelandic Hæstiréttur, the Danish Højesteret, and the Finnish Korkein oikeus. Two of the Nordic countries also have a supreme administrative court (the Swedish Högsta förvaltningsdomstolen and the Finnish Korkein hallinto-oikeus), which were also included in the analysis. Their inclusion were motivated by the fact that administrative courts also have the power to review the constitutionality of the laws they apply, thus engaging in constitutional reasoning. The inclusion of administrative supreme courts in the study also allows us to compare argumentative practices of these courts with those of the ordinary supreme courts.

In order to compare the reason-giving practices of the Nordic countries with supreme and constitutional courts in other parts of the world, the project used the methodology of the CONREASON-project (2011-2016), adapting it to the Nordic context. The CONREASON-project, led by András Jakab, developed a groundbreaking conceptual and methodological framework to enable a comprehensive and systematic analysis of constitutional reasoning in 18 supreme jurisdictions (Jakab, Dyevre and Itzcovich 2017), without, however, extending its geographical reach to Northern Europe.

The first regional follow-up of the CONREASON project, the CORE Latam Project, was launched in Latin America in 2018, headed by Johanna Fröhlich at the Pontifical Catholic University of Chile. This project examined the reason-giving practice of 15 Latin American jurisdictions, including the Inter-American Court of Human Rights and the Caribbean Commonwealth. The CORE Latam project also relied on the methodology of the first CONREASON project, while adapting it to the Latin American constitutional context.

In line with the two related projects, our analysis of the reasoning practices of a court was based on landmark (or leading) cases, i.e. cases where the rulings were deemed the most important in the legal community, with no time limit as to the year of the delivery of the judgments. Such cases tend to set the tone of a court's jurisprudence, as they often provide the lens through which court watchers recognize the defining traits of a court's approach to constitutional argumentation.

2.1 Expert consensus and inter-coder reliability

The selection of forty leading cases within each Nordic high court was based on the expert opinion of the project participant responsible for the given jurisdiction, as accompanied by five constitutional law scholars of their own jurisdiction to review the choice of cases.

We employed Gwet's AC1 agreement coefficient among multiple raters (Gwet, 2008) to evaluate consensus and inter-coder reliability. An important property of this estimator is that the correction for chance agreement is adjusted consistently to the prevalence of the underlying phenomenon (as measured by the observed ratings). In the first situation, we evaluated the agreement among experts in the proposal of cases, and found it on average to be very high (0.94), smallest (but still high) in Iceland (0.86) and in Finland the agreement was perfect (1.00). Secondly, we compared variable ratings of 89 variables among coders during their training in the pilot phase. Initially, the outcome was fairly decent, where 64 variables were found to be almost perfect (0.8-1.0), 17 substantial (0.6-0.8), 7 moderate (0.4-0.6), 1 fair(0.2-0.4) and none slight (0-0.2) or poor (<0). The results were then used as a basis for additional training in order to improve the agreement of the final codings, focusing mostly on improving the consensus of coding the low agreement variables.

2.2 Variables collected in the Nordic CONREASON Project

The court and country indicators were coded as unique variables. A complete description of the variable coded from cases may be found in the Nordic CONREASON Project Codebook (Kelemen, 2024). Variables labels were noted by a capital letter 'N' followed by the variable number, sometimes with a subcategory _# added. For example, variable N3_6 refers to the sixth subcategory of variable N3.

The variables coded from cases were sorted, based on their topic, into the following four groups:

- Variables related to the characteristics of the case (N1-N11)
- Variables related to the arguments used in the case (N12-N30)
- Variables related to the conceptualisation (N31-N51)
- Conclusive measurement variables (N52-N53).

Two variables only had characters (N1, N6), four were multinomial (N4, N7, N17, N53), three were continuous (N2, N8, N9, and in addition the ratio $N9\_8 = \frac{N9}{N8}$), one included voting counts (N10), and two were ordinal (N12, N52). All other variables were coded as binary (Yes/No, except for N10: unanimous/non-unanimous decision making panel, and N11: ordinary/strengthened panel) including those with multiple separated (N3, N5, N13, N14, N16, N18, N23, N24, N25, N26, N31, N34, N38, N41, N45, N50) and/or aggregated (N3, N25, N26) subcategories.

Two variables had the same value recorded on all 280 cases: the fourth subcategory of N25, N25_4 (only '0=NO' answers); and N52 (only 'M=majority of the arguments mentioned in the opinion were ratio decidendi or obiter dictum arguments' answers).

2.3 Set of variables after merging with other projects

Since both the Core Latam and the Nordic CONREASON projects adapted the metodology and variable definitions from the CONREASON project in combination with mutual exchange and collaboration between the principal investigators of all three projects, it was reasonable to assume a high comparability among them. However, some adaptations to regional conditions were still made in both of the latter projects, such as the addition of extra variables or subcategories of existent variables.

One example of the latter were the categorizations of law system. In our international comparison, the groupings of the legal systems were adapted from the CONREASON project. A first classification stated whether the legal system were civil, common, mix of civil and common, regional (international), or Nordic. The second classification were whether the model of judicial review was centralized, diffused, a hybrid of centralized and diffused, or regional. An extra variable with subcategories of diffused(Nordic) and diffused(Other) were also added, enabling comparison between Nordic and other countries with a diffused judicial review model. In all variables, the category regional was used for the European Court of Human Rights (ECHR), the European Court of Justice (ECJ) and the Inter-American Court of Human Rights (IAC). In the Nordic CONREASON project, we did not include any regional court.

The case variable labels in the merged dataset kept the same numbering as in Nordic CONREASON, but replaced the initial 'N' with the letter 'R'. The merged dataset included all the 37 original variables from CONREASON, with extra subcategories suppressed via merging them back into the same form used in the CONREASON project. Variables that were included in only Core Latam and Nordic CONREASON were also included (R28, R32, R47-R52) and thus had missing values on the courts in CONREASON. More details on the specific variable codings in the merged file are given in table A1 in Appendix.

3 Inference and methods

The output of our quantitative approach were sample descriptions and inferential statistics in relation to our project aims. In order to make generalizable statements, we tried and translate the research objectives into tractable statistical models. However, the use of (by experts) purposively selected samples did put certain limits on the type of inference that was feasible.

We studied univariate distributions (typically proportions) and bivariate correlations both within or in comparison between Nordic courts (or countries). The analysis was often employed on a within or between subgroups (typically administrative and non-administrative cases) level, or was undertaken with regards to the subgroups (e.g. when controlling for country we could better comprehend how the administrative differ from the non-administrative cases in the Nordic region).

Comparisons were also made between the Nordic courts/countries/region and other international courts or legal systems and several of the analyses were based on various types of regression models. Part of the analysis were undertaken on an aggregate court data level (e.g. the average proportions of argumentation vs conceptualization between international courts) and potential underlying structures and patterns in the data were explored via dimensional reduction, clustering and other multivariate techniques, including multivariate hypotheses tests.

Trends over time were also analysed, and when found to be relevant, accounted for in the analysis. The most recent Nordic cases were from the year 2022, while the oldest (Danish) case dated back to 1920. However, among Nordic cases the distribution over time, as observed also in the CONREASON and CORE Latam projects, was highly skewed, with only eight cases from before 1970. Therefore, in our statistical analysis, we usually approximated the growth of the number of leading cases as exponential over time and employed a log-transformed scale. However, for ease of interpretation, afterwards we always back-transformed the results to the original (yearly) scale, see Figure 1.

3.1 Descriptives

Tables, graphs and summary statistics were used for description and exploration. When described, most variables were presented as proportions with accompanying confidence intervals estimated using Wilson scores (Wilson, 1927), which does not suffer from the problems of ordinary normal approximation intervals. Since we have fairly few observations and proportions often are high or low, we also employed a Yate's continuity correction (Yate, 1934).

For nominal variables with more than two categories (N4, N7, N17, N53) we employed simultaneous confidence intervals for multinomial probabilities (Sison and Glaz, 1995). Estimated means (with confidence intervals based on normal approximations) and medians (with non-parametric confidence intervals) were used with the continuous variables (N2, N8, N9, N8_9).

3.2 Inference

The inferential analysis relied on the assumption that there exists a data generating process (governed by a set of parameters) from which our set of observations constitutes one, out of an infinite number of, possible outcomes. Alternatively, one may view our statistical statements as referring to a context of imagined (long-run repeatedly sampled) alternative realizations. The data generating process was thus

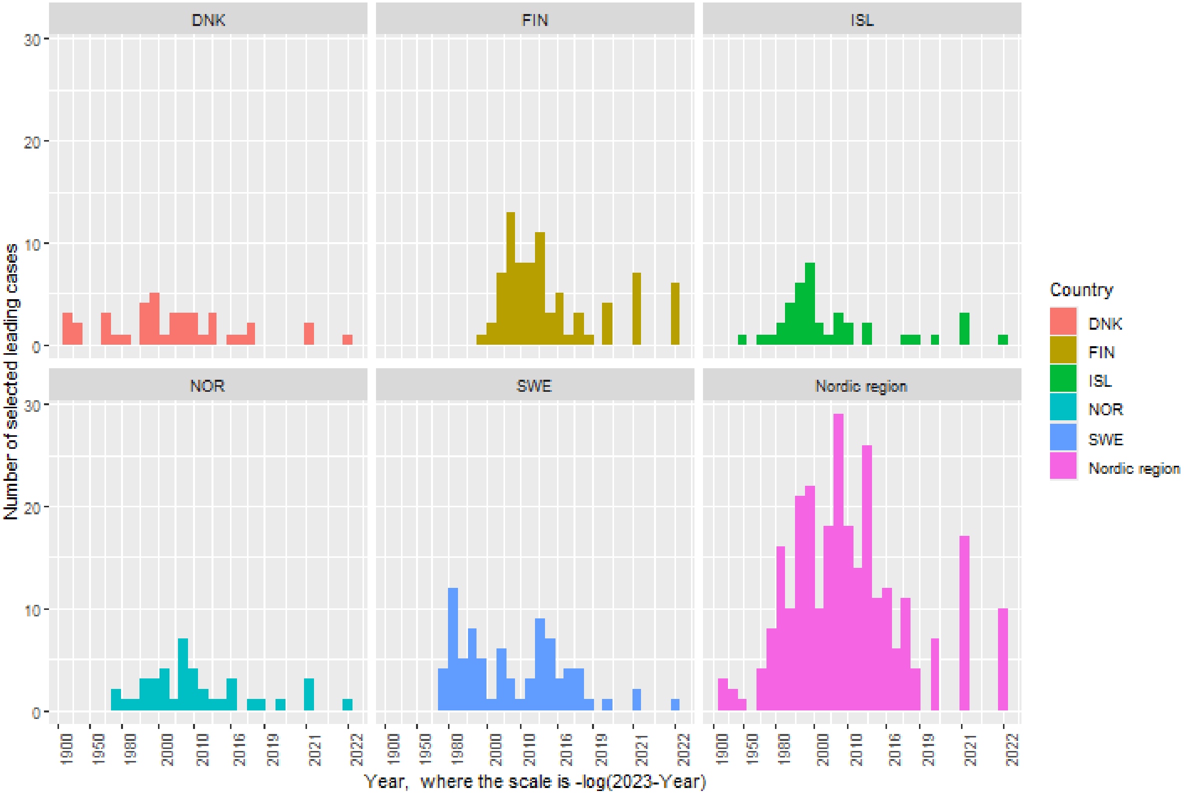


Figure 1: Histogram of leading cases by year in the Nordic countries and region

an abstraction of how the supreme courts' leading constitutional cases came about, and our inference thus refers to this abstraction, or more specifically, the parameters by which it was governed.

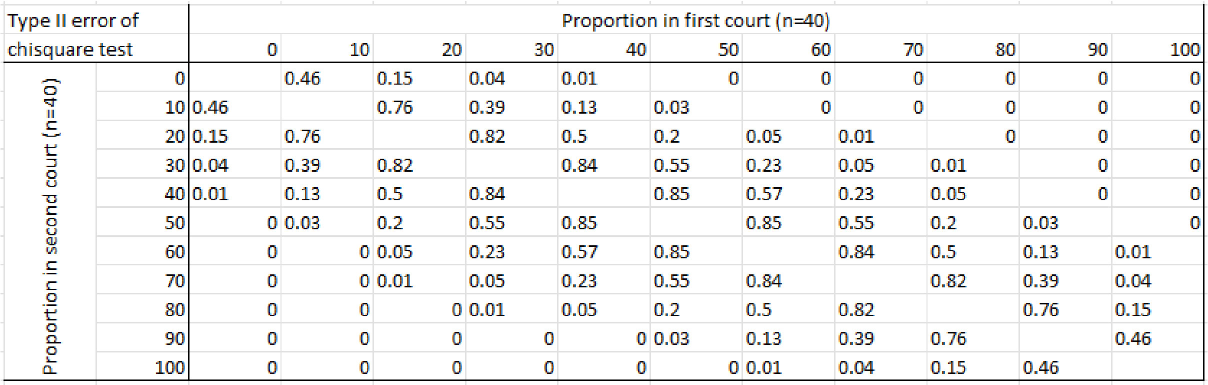
Inference may take the form of point estimation where e.g. a parameter might be the proportion of cases with 'arguments from silence' (*argumentum ex silentio*) in a court, as measured by variable (N21); in interval estimation it could be a confidence interval for the same proportion; and in hypothesis testing it could be whether there was support against, or not, that the proportion parameters of the data generating process were the same among different (Nordic) courts.

In null hypothesis testing, we consistently used a significance level (i.e. the type I error) of $\alpha = 0.05$. Whenever possible, we opted for two-tailed alternative hypotheses. Similarly, when utilizing confidence intervals, we always employed double-sided intervals with a confidence level of 95%. Viewed from a repeated sampling perspective, given that the null hypothesis would be true, we expect that 1 in 20 (i.e. 5%) of the realizations would result in the false decision to reject the null hypothesis.

Similarly, given that the null hypothesis was false the type II error (one minus the power of the test) of β% (conditional on the specific test, effect, and sample size) represents the amount of (long-run) realizations where the conclusion of the test incorrectly would be that the null hypothesis should be rejected. As an example, in a test situation with a two-sample t-test of difference in proportion (or correspondingly a chi-square test of independence given a two-by-two contingency table) where the null hypothesis was that the proportions of a binary variable in two courts (each with 40 observed cases) were the same, the type II error of the test would depend on the two true underlying proportions, see Table 1:

Aside from the risk of drawing the wrong conclusion in null hypothesis testing, although a rejection of the null hypothesis would be correct (when the p-value was smaller than the chosen significance level), a practical significance was not automatically implied and the estimated effect sizes was considered. A substantial reduction of an observed p-value (below the chosen significance level) does neither automatically imply a significant change in the belief about the underlying parameter(s) of the data generating

Table 1: The type II error of a chi-square test of independence between two binary variables, one depicting the question of interest and the other the two courts (each with n=40 cases)



Type II error of		Proportion in first court (n=40)										
chisquare test		0	10	20	30	40	50	60	70	80	90	100
Proportion in second court (n=40)	0		0.46	0.15	0.04	0.01	0	0	0	0	0	0
	10	0.46		0.76	0.39	0.13	0.03	0	0	0	0	0
	20	0.15	0.76		0.82	0.5	0.2	0.05	0.01	0	0	0
	30	0.04	0.39	0.82		0.84	0.55	0.23	0.05	0.01	0	0
	40	0.01	0.13	0.5	0.84		0.85	0.57	0.23	0.05	0	0
	50	0	0.03	0.2	0.55	0.85		0.85	0.55	0.2	0.03	0
	60	0	0	0.05	0.23	0.57	0.85		0.84	0.5	0.13	0.01
	70	0	0	0.01	0.05	0.23	0.55	0.84		0.82	0.39	0.04
	80	0	0	0	0.01	0.05	0.2	0.5	0.82		0.76	0.15
	90	0	0	0	0	0	0.03	0.13	0.39	0.76		0.46
	100	0	0	0	0	0	0	0.01	0.04	0.15	0.46	

process.

Although our explicit inferential claims were limited to the specific setting of leading constitutional cases in high courts, given a reasonable modelling one might increase the external validity of any results to parameters of similar models regarding higher court' or constitutional cases in general, but less likely to lower courts' or court cases in general. However, to make any such claims one would specifically need to consider such situations in detail.

The availability of related population data was fairly limited, but some comparison were carried out using a database containing a partial population (with civil and criminal cases) from the Swedish Supreme Court (Lindholm et al, 2023). In our sample, a majority of cases (26 out of 40) from the Swedish Högsta domstolen belonged to the (2482) cases in this population. The sample deviated in various other aspects such as being on average more recent (year 2005) compared to the population (year 1999), had a higher proportion of criminal cases (73%) compared to the population (45%), and with cases classified as covering more types of legal areas, on average 1.61 vs 1.23 (out of 24 types categorized under traditional domestic law).

While inference is related to the question of *why* statistical analysis is undertaken, there is a preceding need of statistical methods (based on underlying algorithmic techniques) for *how* to do the analysis. The choices of inferential methods are presented and motivated in sections 4-7 of this report, including relevant examples from the main results. An overview of the data and analysis are also given in Figure 2 (sections 4-6) and Figure 3 (section 7).

4 Analysis within each Nordic country or court

The purpose of the analysis in this section was to understand the relationship between variables within a country. In Finland and Sweden, where there were two courts, we also studied the relationships within each of the two courts. In addition, we also compared whether there were differences between groups of cases (e.g. administrative vs non-administrative) within countries/courts.

4.1 Degree of association and tests of relationships between variables

The degree of linear association was measured by Pearson's correlation coefficient r, or in case of two binary variables the equivalent Phi correlation coefficient, see Table 1. Asymptotic confidence interval were calculated by use of a Fisher transformation.

Variable N12 was the only ordered variable with more than one type of value recorded, and was also the only one with three categories. Thus for N12 we used Tau-c to measure the degree of ordinal association, since this measure is suitable when the other variable is measured by a different scale (with only two, or more than three, unique values). Confidence intervals were found by Monte Carlo simulation.

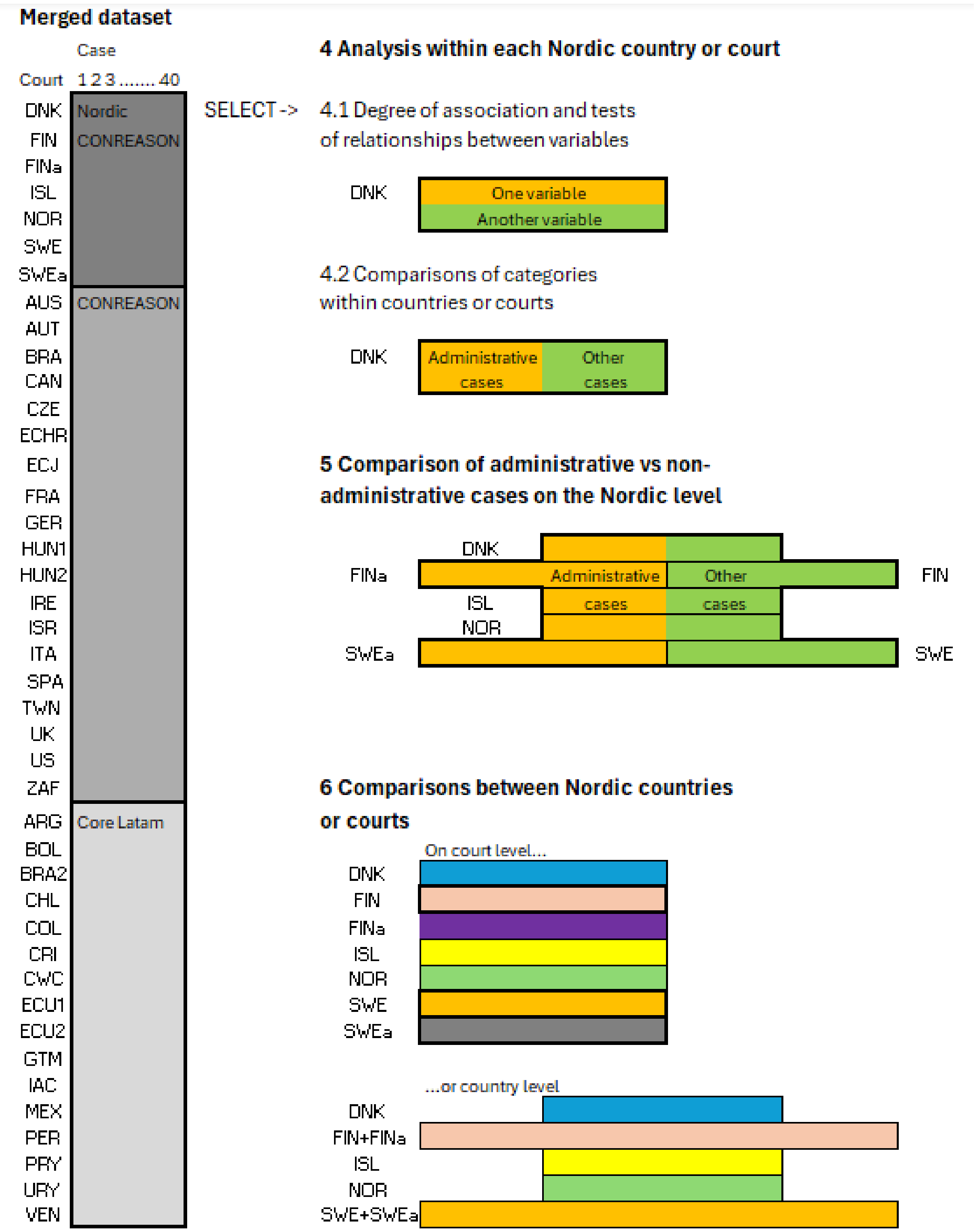


Figure 2: Overview of data and analysis in sections 4-6

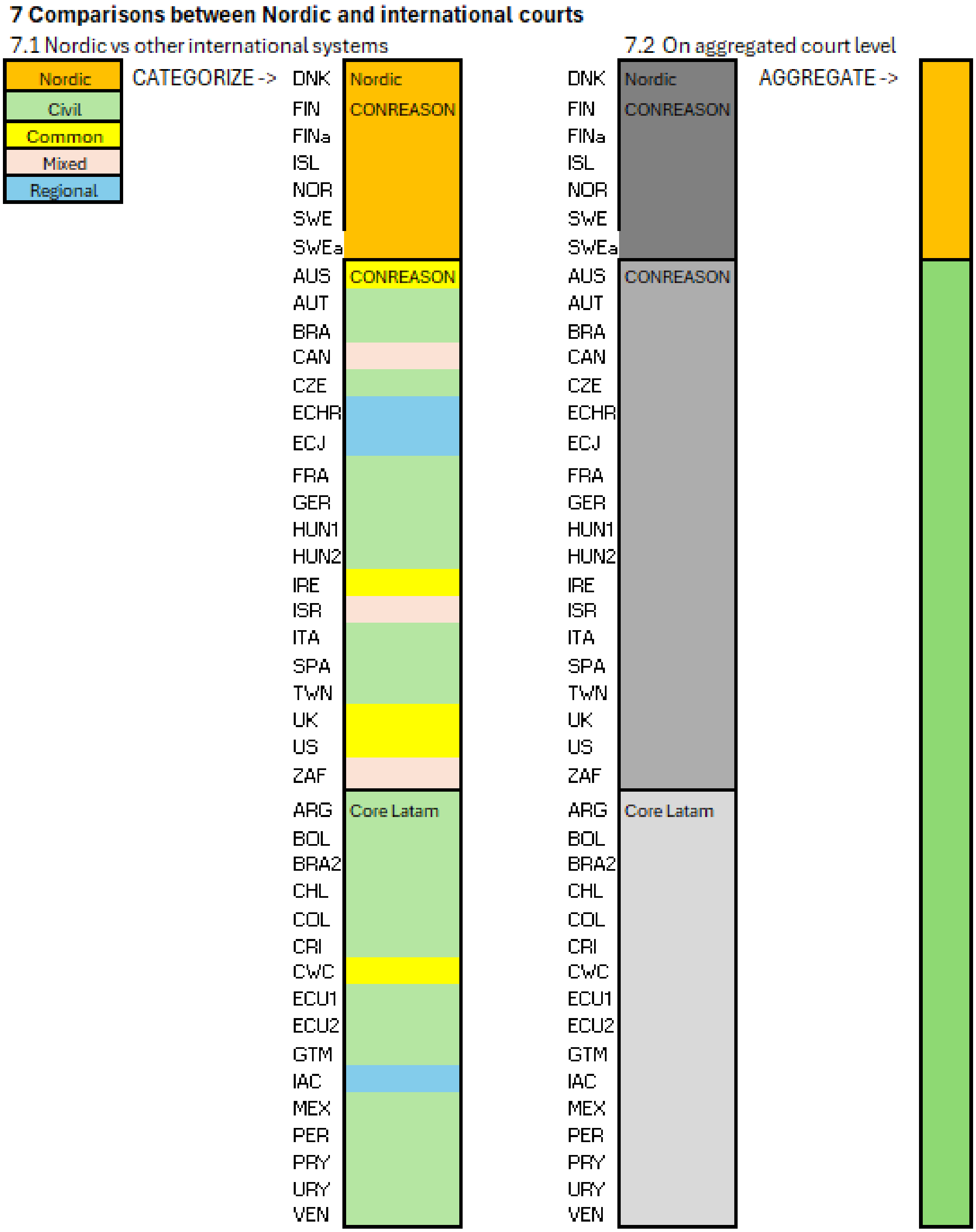


Figure 3: Overview of data and analysis in section 7

Within each Nordic country, we also estimated partial (conditional) correlations between variables (Kunihiro, 2004), controlling for the influence of administrative vs non-administrative cases. The partial

Table 2: Correlation measure by scale of variables

	1st variable is		
2nd variable is	binary	ordinal	continuous
binary	Phi	Tau-c	Pearson's r
continouos	Pearson's r	Tau-c	Pearson's r

correlations tries to equalize the Nordic countries in the sense that if e.g. the correlation between two variables among all the countries would have been stronger among administrative cases, then we expect the estimated correlation to be stronger within a country if it relatively had more administrative cases. However, this was controlled for in the partial correlation.

A partial correlation was found by first regressing each of the two variables of interest on an administrative indicator variable, and then calculating the correlation between the two resulting sets of residuals. The indicator variables for administrative vs non-administrative cases where deducted directly in Sweden and Finland from the fact that these two types of cases are divided between separate high courts. In the remaining Nordic countries, the indicator variables where deducted from variable N7.

When at least one nominal variable (with more than two categories) is involved the concept of correlation is not applicable. However, we did undertake chi-square tests of independence between nominal and other (nominal, binary or ordinal) variables. The p-values were then simulated (Hope, 1968) from 10000 Monte Carlo replicates as to always avoid the unreliable chi-square approximation when there were small expected counts in pairwise outcomes of two variables.

By regressing a continuous variable on a nominal variable, we were also able to undertake a null hypothesis test of no significant difference in total (using F-tests) or of no significant difference of individual categories (by t-tests).

4.2 Comparisons of categories within countries or courts

We tested whether there was a difference between pairs of groups of cases based on types of disputes (N7) within Nordic countries, and case disposition (N4) and general topic (N5) in Nordic high court. The groups were thus

- N7: administrative vs non-administrative cases
- N4: whether the court found (at least partially) against the law/decision/act (of the government) challenged (all three pairwise sets: Yes vs No; Yes vs Not applicable; No vs Not applicable);
- N5: the general topic (all six pairwise sets based on the four topics; Fundamental rights; State organisation; Procedural; Other). There are no Procedural cases in the Finnish administrative court, and no State organization cases in either of the Swedish courts so those categories are not considered there.

Each of the four general topics (N5) may be applicable to the same case so they were not mutually exclusive. In order to test if the impact of a variable was equal among two such topics, we first regressed our variable of interest on indicator variables of the four general topics. Then we used a general linear hypothesis test (Bretz et all, 2010) of the equivalence of the beta parameters of the two topic indicators.

For the two other variables (N4 and N7), the equality of means of continuous variables between two categories were tested by Welch's (unequal variances) t-tests. To test the equal location null hypothesis between two groups for the ordinal variable (N12) we employed a Wilcoxon-Mann-Whitney test. The equality of binary variable proportions between two groups were tested using a chi-square test (with monte-carlo based p-values). A chi-square test was also employed to test the equality between two groups of proportions for nominal variables (that has more than two categories). Fairly few observations were involved in these comparisons, and we seldom found any significant and substantial differences.

5 Comparison of administrative vs non-administrative cases on the Nordic level

In order to test whether location (parameters) or proportions were equal between administrative and non-administrative cases on a Nordic level, we used all the Nordic cases, and then regressed the variable of interest on an administrative indicator variable, and added a set of indicator variables representing the Nordic countries as controls. For

- continuous variables we used linear regression.
- binary variables we used logistic regression.
- the ordinal variable (N12) we used ordinal regression.
- nominal variables we used multinomial regression.

We then tested the significance of the administrative indicator parameter in linear regression (t-test), logistic regression (Wald test), and ordinal or multinomial regression (by profile likelihoood based confidence intervals).

We found there to be reasonable substantial and significant differences in several situations, such as fewer opinions invoking freedom of expression rights (N45) among the administrative cases.

6 Comparisons between Nordic countries or courts

In a comparison of the Nordic countries, we started with an unconditional view (without accounting for differences in either time or administrative cases) and tested null hypotheses that the Nordic countries had equal:

- means among continuous variables, via F-tests in an ANOVA.
- medians among ordinal variables and the continuous variables, using Kruskall-Wallis-test.
- proportions among binary or ordinal variables, using chi-square tests.

The reason to also apply the Kruskall-Wallis-test of (median) location to continuous variables was the robustness to non-normal underlying data generation processes, since some distributions were somewhat skew.

The null hypothesis of equality was rejected almost in every case. The only exceptions were the case disposition (N4) and some of the conceptual variables (N38, N39, N42, N45, N50). Typically, one or two countries (most often Denmark and/or Norway) differed from the others. This was also seen in tests of equality (pairwisely) of proportions (using chi-square test) between the Nordic courts.

In order to test if trends over time differed between courts, we regressed the binary variables on; a random court intercept (as control), and a court indicator, a time variable, and their interaction. In this way, we simultaneously tested whether a court had a different level or trend over time compared to other courts. Each court were found to have a few (four variables on average) significantly different levels or trends compared to the others courts. Deviations were slightly more common for Norway, and also among the argumentative variables.

7 Comparisons between Nordic and international courts

The purpose of employing the methods in this section was to explore whether the Nordic higher courts, in terms of constitutional reasoning, distinguished themselves from other constitutional and international courts. We thus made use of the merged data from the Nordic CONREASON, the CONREASON, and the Core Latam projects, as described in section 2.3.

7.1 The Nordic system vs other international systems

In order to make comparisons between the Nordic and other constitutional systems, we employed the previously described legal system categorizations:

- Nordic; Civil; Common; Mixed; Regional
- Diffused(Nordic); Diffused(Other); Centralised; Hybrid; Regional

Depending on which projects a variable was measured in, see Table A.1 in Appendix, we had at most 42 courts, each with 40 cases each. These courts may not be considered as a random sample of all potential courts, not at least since both the Core Latam and the Nordic Conreason project included high courts in specific geographical regions. In order to model this, we conceptualized a level effect of each court to be a random draw from a much larger set of potential levels emanating from a common underlying process. At the same time we treated the potential differences between the systems of our interest as structural.

We then choose to regress the variable we investigate on (indicators of) the court systems of interest (with Nordic as the reference category) but also to include a random intercept for each of the courts in our model. In this way we tried to control for the within-court case dependence.

When the variables of interest was continuous we estimate a generalized linear mixed models. We choose to transform the nominal variables to sets of binomial variables, so that when the variable was not continuous, we estimated binomial generalized linear mixed models. We then choose to use the Nordic category as reference, and then we tested whether the (fixed) effects of each of all the other systems, i.e. either {Civil; Common; Mixed; Regional} or {Diffused(Other); Centralised; Hybrid; Regional} differed significantly from the Nordic.

In order to visually study the development over time, given each variable of interest, we first fitted LOESS (Cleveland et al, 1992) curves (including confidence intervals) on an accumulated 'global' level, and then did the same for each constitutional systems (but without accounting for court). In this way we could visually get an idea of any potential trends, although without performing any formal tests. As a way of trying to control for any overall trends, we added natural cubic splines (Hastie, 1992) to the previously fitted regression models.

We also tested whether the Nordic countries differed in level and/or in the linear trend over time compared to other countries, through a binomial generalized linear mixed model, in a similar way as described in section 6 among Nordic courts. Thus we regressed the variable of interest on an indicator of the Nordic courts, a time variable, and their interaction. As previously, we included a random intercept for each of the courts. In addition to these formal tests, we visualized the predicted model outcomes with confidence intervals. We found about the same number of significant differences as among the Nordic courts, a third of them exactly the same.

7.2 Nordic vs international courts on aggregated court level

In this section of the analysis we aggregated data on court level. The analyses focused on the argumentative (R12-R26, R28-R29) and conceptual (N31-R50) variables. All variables were thus aggregated into binary proportions on court level. Some of the variables were missing from all the courts in the CONREASON project (see Table A.1 in Appendix), and also from one court (ECU1) in the Core Latam project.

The available variables, conditional on the projects included, are presented in Table 2. Since variable R12 was ordinal with three mutually exclusive categories, it was represented by two category variables. Unless anything else is stated, we choose to proceed with all of the projects (i.e. 41 court observations) and thus in total with 32 variables observed.

We then compared proportions of all of the courts for each variable at a time. We also calculate the average proportion among argumentative and conceptual variables and compared these two averages among all the courts.

Table 3: Available variables conditional on courts/projects included

Projects included			
Nordic Conreason	x	x	x
Core Latam	x	x	
Conreason	x		
Total number of courts	41	22	7
Argumentative variables	R12-R26, R29	R12-R26, R28-R29	N12-N30
Number of variables	17	18	20
Conceptual variables	R31, R33-R46	R31-R50	N31-N51
Number of variables	15	20	21

The Nordic courts did not stick out particularly on either of the scales compared to other courts, irrespective if we compared to all courts with fewer variables, or only Core Latam courts including more variables.

7.2.1 Pairwise and canonical correlations

In order to better understand the relations both within the two sets of (argumentative and conceptual) variables we studied the pairwise (Pearson's r) correlations between all pairs of variables, and considered them by each set. Correlations were on average slightly stronger among conceptual (mean=0.24, median=0.22, sd=0.18) than argumentative (mean=0.18, median=0.22, sd=0.21) and on average (although only) somewhat weaker between them (mean=0.16, median=0.15, sd=0.19).

Due to the sparseness with very few observations (41) relative to the number of variables (32), it was sometimes hard to directly extract reliable multivariate (i.e. based on several variables simultaneously) results. Such a small dataset may principally be framed as a 'curse of dimensionality' case. In other words, if dimensions (i.e. variables) are added (linearily) the observations gets more sparse (since the volume of the space spanned up by the variables grows exponentially) and each observation thus seems to become more distant from the other observations. When fitting a model with in such a situation, with very few observations, the risk of overfitting to the data is a potential problem. (On the other hand, given that the variables to a high degree are relevant to the phenomena studied, that would help out in e.g. discrimination of groups).

To investigate the overall correlation between argumentative and conceptual, we applied canonical correlation analysis (where one construct canonical variates, that is, linear combinations in each of the set of variables, such that the correlations between the canonical variate pairs are maximized under the constraint that the canonical variate within a set is orthogonal to other variates of the same set). However, the results strongly suggests overfitting to the data.

We therefore choose a more robust approach and redid the analysis repeatedly, using only a random draw of three variables in each of the two sets. On average, we had one (out of three) canonical variates being significantly different from zero, thus lending some support for a correlational structure between the two sets of variables.

However, considering both the pairwise and the canononical correlations, it was suggested that the structural distinction between argumentative and conceptual variables were not necessarily that strong. Although the matter could be further investigated, the availability of data was still a major limitation. In the further analysis, initially we therefore did not distinguish between the two sets of variables, but still looked for any anomalous differences between them.

7.2.2 Dimensionality reduction and multiple factor analysis

One way of dealing with the sparsity was to reduce the dimensionality of the data while trying to extract the relevant information, i.e. to try and 'increase the signal and reduce the noise' in the data.

In practice this typically means that we projected (all, or many of) our variables onto a (much) smaller set of variables in such a way that they still captured a lot of the information held within the original variables. Calculating average number of occurrences, as we did for the argumentative and conceptual variables, reduced the dimension to two (potentially correlated) variables, and might thus not be an efficient way of preserving the information content.

Instead, first we tried out non-linear dimensionality reduction techniques. The similarity between observation were modelled either (in an Euclidean space) using t-distributed stochastic neighbor embedding (t-SNE) ()van der Maaten & Hinton (2008), or (in a Riemann space) with Uniform manifold approximation and projection (UMAP) (McInnes et al, 2018). The results were then visualized in two-dimensions.

While UMAP focuses more on preserving global structures, the t-SNE leans more towards the local, while the results are more variable (and therefore more repetitions are needed to ensure that the results are reliable). However, the outcome of both methods are still highly dependent on the choice of hyperparameters and exploration of these are therefore also needed. Although, given a set of hyperparameters, it might have seemed that different groups of observations would be clustered, the distances between groups were not directly interpretable and specific outcomes should mainly be use as suggestions. However, after a thorough exploration, both of the methods seemed to suggest that all the Nordic courts, perhaps together with the IAC court, might form a group of observations on their own.

In contrast, linear dimensional reduction methods (Gower, 1966) such as principal component analysis (PCA) do have interpretable distances (Mardia, 1979). The method is based on finding linear combinations of the original variables, such that the linear combinations explains as much of the variance in the original variables as possible. At the same time, the linear combinations are themselves orthogonal to each other and are thus not correlated. The direction of the first principal component will therefore fall along that in which the observations vary the most, and so on for the following principal components.

The first principal component explained 24% of the variation, the second 11%, the third 8% and so on. When the scores of the six first components (comprising 62% of the variation) were compared, we notice that, except for component four and five, most of the Nordic courts tend to either have only positive (or negative) scores, indicating that they, as a group, tended to deviate from the overall mean of all the courts, see Figure 4.

When plotted among the first two components simultaneously, see Figure 5, to a large extent we confirmed that the Nordic supreme courts did tend to deviate and we got results similar to those with the non-linear dimensional reductions techniques.

The Nordic courts seemed to take a fringe position, although SWEa was a bit further away from the others (primarily along the first dimension). Now IAC lied on its own, but with the Nordic (and ECHR) as its nearest neighbours. AUS also lied on its own, but on the opposite side of the second dimension. As was indicated in Figure 4, the third component also contributed in discriminating between the Nordic and other courts.

None of the (conceptual or argumentative) variables seemed to have a dominant impact on the principal components, and there were no clear pattern among the variable loadings on the principal components with regard to the argumentative vs conceptual. On average there the absolute loadings on the first (mean=0.17, median=0.17, sd=0.06), the second (mean=0.15, median=0.15, sd=0.09) and the third (mean=0.15, median=0.12, sd=0.10) component were also fairly similar. It was therefore hard to interpret the principal components, beyond their ability to extract the signal in the data.

In order to further investigate the influence of the two sets (of argumentative and conceptual) variables we undertook a multiple factor analysis, similar to a PCA, accounting (via weights) for that the two sets of variables had different number of variables. The argumentative and conceptual sets of variables essentially had the same impact (50%) each on the first dimension, and very similar on the second (54% and 46%) and third (53% and 47%), although slightly less even on the fourth (41% and 59%). As with the similarity among loadings, this provides further support that there was not so much of an argumentative vs conceptual structure observable in the data.

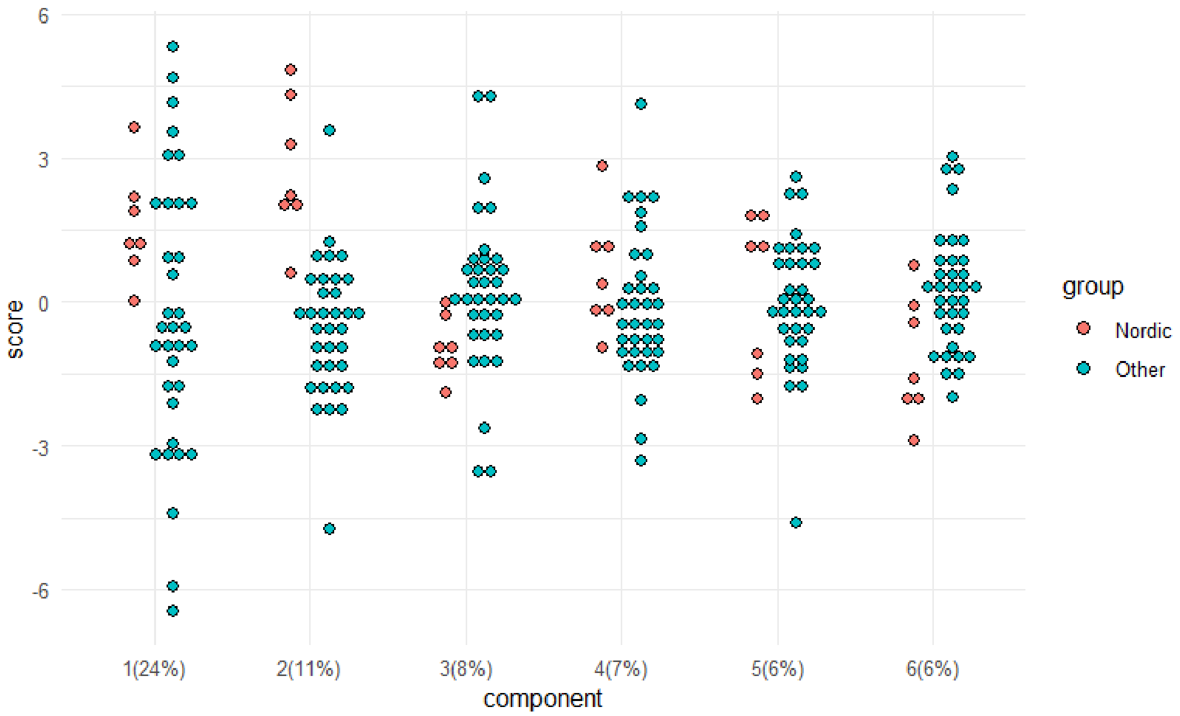


Figure 4: Dotplot of court scores for each of the six first principal components extracted from argumentative and conceptual variables

7.2.3 Cluster analysis

Although the Nordic courts, with regard to the projection unto a few dimensions, seemed to lie along the boundary of the data, it was not clear whether any separation into groups/clusters seemed reasonable. In order to investigate whether one could find such a separation, we employed clustering techniques on (at least two) of the principal components.

Clustering is to a large extent an exploratory methodology, so we choose to search through a broad spectrum of clustering techniques including:

- k-means clustering
- PAM (partitoning around mediods) - a more robust alternative to k-means clustering
- hierarchical clustering (using various linkages and similarity measures)
- SOTA (self-organizing tree algorithm) clustering
- DBSCAN (density-based spatial clustering of applications with noise)
- FANNY (fuzzy analysis clustering) allowing observations to belong to more than one cluster
- model-based expectation-maximization (EM) clustering

However, none of these methods gave any strong additional insights and there were never any clear separation into clusters. We also transposed the data and undertook principal component analysis using variables as observations and courts as variables, followed by similar clustering attempts, but without any new substantial insights.

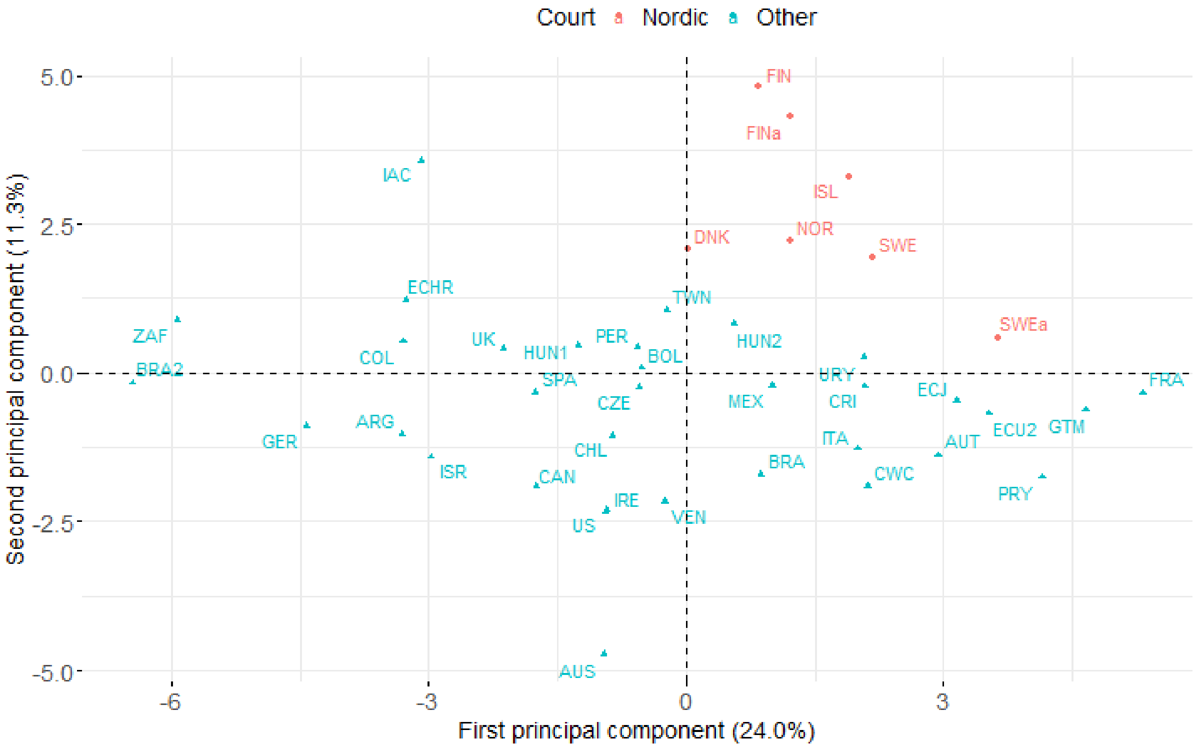


Figure 5: Court scores on the two first principal components extracted from argumentative and conceptual variables

7.2.4 Discriminatory power

An alternative approach to investigate whether there was anything distinct to the Nordic vs other courts was to apply supervised (statistical) learning and test the discriminatory power of such a model. As in clustering, there are a plethora of methods available.

For a small dataset with fairly many variables in relation to the number of observations, a partial-least-squares discriminatory-analysis (PLS-DA) would likely perform reasonably well. PLS resembles PCA, but do not only focus on explaining the variance of the other variables but also accounts for the variable of interest in the chosen projection. (So, while PCA is unsupervised, PLS involves supervised dimensional reduction technique). In PLS-DA, the variable of interest was qualitative, that is, the Nordic court indicator variable.

The modelling was performed in the following way:

1. We randomly selected 35 (i.e. 85% of all) courts, conditional on that six of them were Nordic and 29 were not (i.e. a stratified train-test split). In this way, we ensured that the proportion of Nordic cases was the same (17%) both among the selected and the non-selected courts. (One could allow for more variation in the number of Nordic vs non-Nordic cases, but the chosen balance seemed to be pragmatic here).

2. Based on the 35 selected courts, utilising both the variables and the Nordic indicator variable, we estimated a model which calibrated to have good discriminatory power. The model depended on the selection of a hyperparameter (namely the number of components to be used, analogous to the number of principal components selected in PCA) automatically chosen through a leave-one-out cross-validation (LOOCV) procedure.

3. Based on the estimated model and the variable values of the six excluded (including one Nordic) court observations, we then predicted the class (Nordic or not) among each of them.

4. In order to evaluate the discriminatory power, we then compared the predictions to the true values (Nordic or not) which had been held out during the prediction phase.

The above procedure was repeated many (10000) times such that each of the Nordic courts on average would be predicted $\frac{10000}{7} \approx 1429$ times each (and the other courts $\frac{(6-1)*10000}{34} \approx 1471$ times each). We then evaluated the average discriminatory power, i.e. how often the classifications were correct. All of the Nordic countries where always classified correctly, except for NOR (which was correctly classified 85.7% of the times). Among the other courts, HUN2 were falsely classified as Nordic 12.7% of the time, and IAC 0.6% of the times, while all the other courts were always classified correctly. With regards to some of the previous (non-linear) results this did not seem conspicuous since HUN2 lied fairly adjacent to the Nordic courts, and NOR sometimes lied fairly close to IAC. Thus, given the imbalance of categories in the data, overall the discriminatory ability were very strong with a misclassification rate of $\frac{9+187+210}{60000} = \frac{406}{60000} = 0.07\%$.

Although one could use many other setups, and for example balance the dataset through oversampling, the results turned out to be fairly robust to alternative specifications. This lends even more support to that there was something measured in the data which made the Nordic courts distinct from the other courts. Still, it was hard to interpret exactly why we observed the discrimination, and there is of course scope for examining how the individual variables enters the model.

Regarding the hyperparameter, three components were selected 80%, two components 19.7%, and four components 0.3% of the times. This seemed to be consistent with the previous results. A discriminatory model such as the ones used are likely to rather select a (too) large then a (too) small model. Notably, the PCA also seemed to indicate two (or perhaps three) components.

7.2.5 Multivariate hypothesis testing of differences in means

Another approach to compare the Nordic to other courts was to formally test whether there was a significant difference in means between them. Such tests may of course be undertaken directly on variables, and would to a large extent align with the results in 9.1 where the Nordic was compared to other international systems. In eight of those cases (involving variables R19, R21, R22, R26, R31, R34, R35, and R42) it was generally supported that the Nordic system differed significantly from the other legal systems.

In order to take on a more comprehensive approach, we applied a multivariate test. Due to the data sparsity, and the nature of the principal components (which after visual inspection seemed to not unlikely have been draws from bell shaped distributions), we choose to use the Hotelling's t-squared distribution as a reference distribution. Then we stated a null hypothesis that the multivariate mean of the difference in principal components (extracted in section 9.2.2) between Nordic and other courts were zero, before we undertook the test (Hotelling, 1931).

It turned out that the null hypothesis was rejected irrespective of the number of components included. However, tests of normal skewness and kurtosis where both rejected with more than four components. (When components were tested individually, the null hypothesis of no difference were rejected for the first three and the sixth and seventh principal components). Considering the small number of observations and previous results, this still lends huge support in favour of the Nordic courts (as a group) being different from the other courts.

8 Concluding remarks

The availability of the data from the Nordic CONREASON project, including the merging with data from previous similar projects worldwide, opened up entirely new possibilities for studying and generalizing about constitutional reasoning in supreme courts, both on a Nordic and on an international level.

Since the field was largely unexplored, we applied a systematic approach where we carefully analysed the data using different selections, conditionings and levels of analysis. In addition, through applying a wide flora of relevant methods, hopefully we did not miss out essential structures or failed to explain variation based on the factors that were available.

However, it cannot be ruled out that, in particular, part of the observed differences between the Nordic and other supreme courts may be due to systematic differences in the measurement process of the Nordic project. To investigate that was beyond the scope of this analysis, but rather focus on more qualitative validations, in addition to those that had already been undertaken.

9 References

Baba, K., Shibata, R., Sibuya, M. (2004). Partial correlation and conditional correlation as measures of conditional independence. Australian and New Zealand Journal of Statistics. 46 (4): 657–664.

Bretz, F., Hothorn, T., Westfall, P. (2010). Multiple Comparisons Using R, CRC Press, Boca Raton.

Kelemen, K. (2024). Nordic CONREASON kollektion av data om konstitutionellt resonemang i de nordiska högsta domstolarna (Version 1) [Dataset]. Örebro universitet. Available at: https://doi.org/10.60689/zmk7-0z14

Cleveland, W. S., Grosse, E., Shyu, W. M. (1992). Local regression models. Chapter 8 of Statistical Models in S eds Chambers J.M., Hastie, T.J. Wadsworth & Brooks/Cole.

Fisher, R. (1936). The use of multiple measurements in taxonomic problems. Annals Of Eugenics, 7(2), 179-188.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53, 325–328.

Gwet, K.L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. British Journal of Mathematical and Statistical Psychology, 61, 29-48.

Hastie, T.J. (1992). Generalized additive models. Chapter 7 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

Hotelling, H. (1931). The generalization of Student's ratio. Annals of Mathematical Statistics. 2 (3): 360–378.

Hope, A.C.A. (1968). A simplified Monte Carlo significance test procedure. Journal of the Royal Statistical Society Series B, 30, 582–598.

Lindholm, J., Derlén, M., Naurin, D. (2023). Swedish High Court Database (version 0.9.1, 1 May 2023). Available at: https://github.com/jojolindholm/sehc.

Mardia, K.V., Kent, J.T., Bibby, J.M. (1979). Multivariate Analysis, London: Academic Press.

McInnes, L., Healy, J., Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Sison, C.P., Glaz, J. (1995). Simultaneous confidence intervals and sample size determination for multinomial proportions. Journal of the American Statistical Association, 90:366-369.

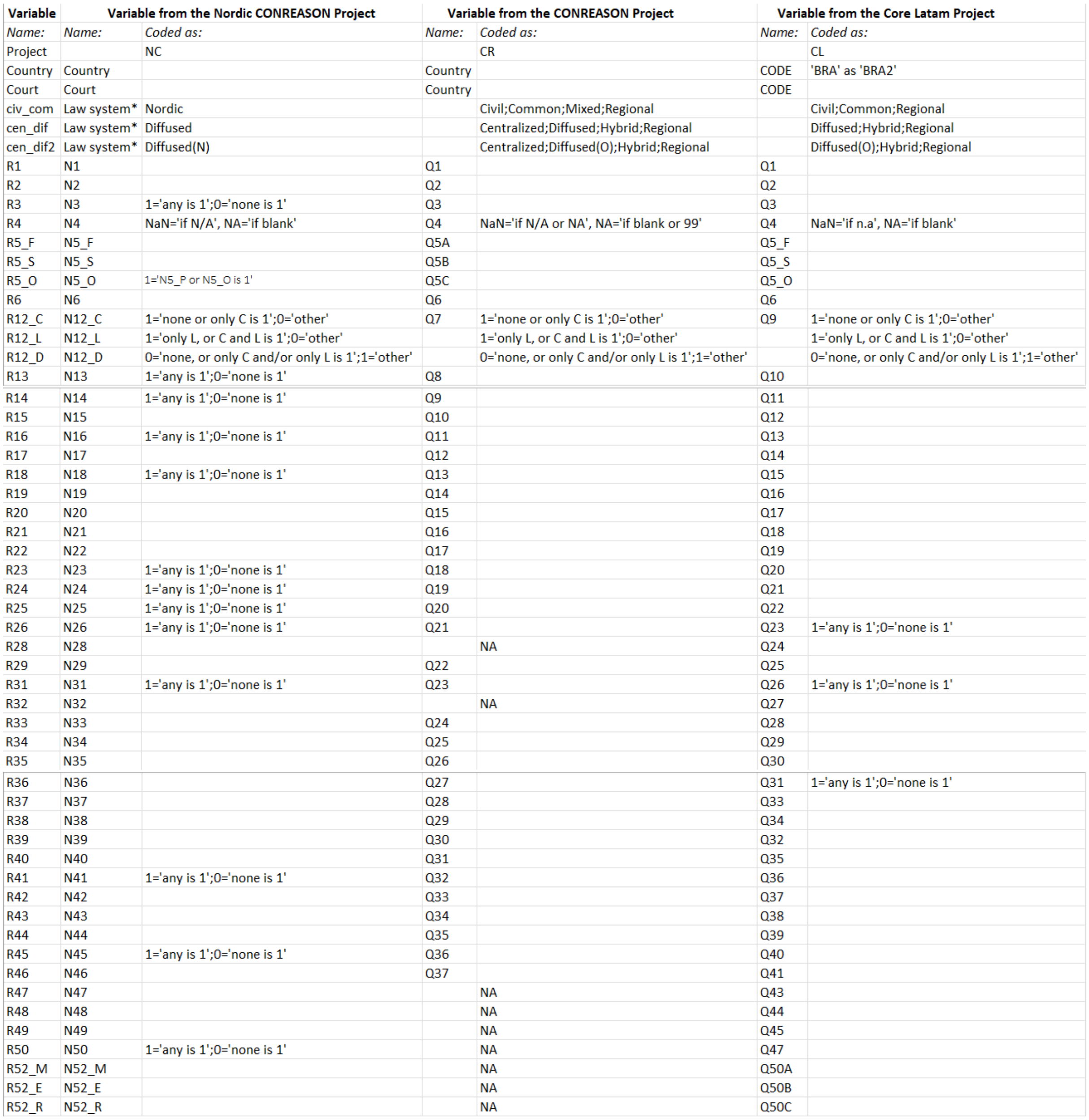
van der Maaten, L.J.P., Hinton, G.E. (2008). Visualizing High-Dimensional Data Using t-SNE. Journal of Machine Learning Research, 9, pp.2579-2605.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. Journal of the American Statistical Association. 22 (158): 209–212.

Yates, F. (1934). Contingency table involving small numbers and the χ^2 test. Supplement to the Journal of the Royal Statistical Society 1(2): 217–235.

A Variable coding when merging data from the projects

The variable codings are presented in Table A1. An empty cell in the columns 'Coded as' means that the variable was kept as it was in the original dataset. And an 'NA' in a cell of 'Coded as' columns means that the observations on this variable was not available in that project and was thus missing in the merged dataset.



Variable	Variable from the Nordic CONREASON Project		Variable from the CONREASON Project		Variable from the Core Latam Project	
Name:	*Name:*	*Coded as:*	*Name:*	*Coded as:*	*Name:*	*Coded as:*
Project		NC		CR		CL
Country	Country		Country		CODE	'BRA' as 'BRA2'
Court	Court		Country		CODE	
civ_com	Law system*	Nordic		Civil;Common;Mixed;Regional		Civil;Common;Regional
cen_dif	Law system*	Diffused		Centralized;Diffused;Hybrid;Regional		Diffused;Hybrid;Regional
cen_dif2	Law system*	Diffused(N)		Centralized;Diffused(O);Hybrid;Regional		Diffused(O);Hybrid;Regional
R1	N1		Q1		Q1	
R2	N2		Q2		Q2	
R3	N3	1='any is 1';0='none is 1'	Q3		Q3	
R4	N4	NaN='if N/A', NA='if blank'	Q4	NaN='if N/A or NA', NA='if blank or 99'	Q4	NaN='if n.a', NA='if blank'
R5_F	N5_F		Q5A		Q5_F	
R5_S	N5_S		Q5B		Q5_S	
R5_O	N5_O	1='N5_P or N5_O is 1'	Q5C		Q5_O	
R6	N6		Q6		Q6	
R12_C	N12_C	1='none or only C is 1';0='other'	Q7	1='none or only C is 1';0='other'	Q9	1='none or only C is 1';0='other'
R12_L	N12_L	1='only L, or C and L is 1';0='other'		1='only L, or C and L is 1';0='other'		1='only L, or C and L is 1';0='other'
R12_D	N12_D	0='none, or only C and/or only L is 1';1='other'		0='none, or only C and/or only L is 1';1='other'		0='none, or only C and/or only L is 1';1='other'
R13	N13	1='any is 1';0='none is 1'	Q8		Q10	
R14	N14	1='any is 1';0='none is 1'	Q9		Q11	
R15	N15		Q10		Q12	
R16	N16	1='any is 1';0='none is 1'	Q11		Q13	
R17	N17		Q12		Q14	
R18	N18	1='any is 1';0='none is 1'	Q13		Q15	
R19	N19		Q14		Q16	
R20	N20		Q15		Q17	
R21	N21		Q16		Q18	
R22	N22		Q17		Q19	
R23	N23	1='any is 1';0='none is 1'	Q18		Q20	
R24	N24	1='any is 1';0='none is 1'	Q19		Q21	
R25	N25	1='any is 1';0='none is 1'	Q20		Q22	
R26	N26	1='any is 1';0='none is 1'	Q21		Q23	1='any is 1';0='none is 1'
R28	N28			NA	Q24	
R29	N29		Q22		Q25	
R31	N31	1='any is 1';0='none is 1'	Q23		Q26	1='any is 1';0='none is 1'
R32	N32			NA	Q27	
R33	N33		Q24		Q28	
R34	N34		Q25		Q29	
R35	N35		Q26		Q30	
R36	N36		Q27		Q31	1='any is 1';0='none is 1'
R37	N37		Q28		Q33	
R38	N38		Q29		Q34	
R39	N39		Q30		Q32	
R40	N40		Q31		Q35	
R41	N41	1='any is 1';0='none is 1'	Q32		Q36	
R42	N42		Q33		Q37	
R43	N43		Q34		Q38	
R44	N44		Q35		Q39	
R45	N45	1='any is 1';0='none is 1'	Q36		Q40	
R46	N46		Q37		Q41	
R47	N47			NA	Q43	
R48	N48			NA	Q44	
R49	N49			NA	Q45	
R50	N50	1='any is 1';0='none is 1'		NA	Q47	
R52_M	N52_M			NA	Q50A	
R52_E	N52_E			NA	Q50B	
R52_R	N52_R			NA	Q50C	

Table A1: Variable coding when merging the three CONREASON projects into one dataset. *For a full explanation, see the CONREASON Project documentation.