RUNNING HEAD: How many countries?

How many countries for multilevel modeling?
A comparison of Frequentist and Bayesian approaches.

Forthcoming *American Journal of Political Science*
http://onlinelibrary.wiley.com/doi/10.1111/ajps.12001/abstract

Daniel Stegmueller
Nuffield College, University of Oxford
1 New Road
Oxford OX1 1NF, United Kingdom
mail@daniel-stegmueller.com

First page footnote / acknowledgements:

I am indebted to my reviewers and editor, Rick Wilson, whose criticisms and suggestions improved the paper considerably. Equal thanks to Jeff Gill, Thomas Gschwend, Adam Ramey, Tobias Heinrich, Ray Duch, Thomas Pluemper, Vera Troeger, Tom Scotto, and Michel Becher for their excellent comments and advice. Earlier versions of this paper were presented at the annual meeting of the Midwest Political Science Association in Chicago, 2011, and the first conference of the European Political Science Association in Dublin, 2011. I thank participants and discussants for helpful comments and suggestions. Replication data and supplemental material is available at the author's website.

Abstract

Researchers in comparative research increasingly use multilevel models to test effects of country level factors on individual behavior and preferences. However, the asymptotic justification of widely employed estimation strategies presumes large samples and applications in comparative politics routinely involve only a small number of countries. Thus researchers and reviewers often wonder if these models are applicable at all. In other words, how many countries do we need for multilevel modeling? I present results from a large scale Monte Carlo experiment comparing the performance of multilevel models when few countries are available. I find that maximum likelihood estimates and confidence intervals can be severely biased, especially in models including cross-level interactions. In contrast, the Bayesian approach proves to be far more robust and yields considerably more conservative tests.

Keywords: comparative research, multilevel models, maximum likelihood, Bayesian

Integration of micro- and macro-data is now seen as state of the art in many subfields of political science. This trend is most marked in comparative political research, where researchers link macro-level differences across countries to individual preferences and behavior (Anderson and Singer 2008). Indeed, some claim that "all comparative politics is multilevel" (Kedar and Shively 2005: 2). It is thus not surprising that multilevel models (Steenbergen and Jones 2002; Gelman and Hill 2007) are increasingly popular.

The majority of studies employ pooled individual level survey data with matched country level information to estimate micro and macro effects. Using this multilevel setup, a diverse range of topics have been studied: policy diffusion (Gilardi 2010), attitudes towards immigration (O'Rourke and Sinnott 2006), ethnic and social tolerance (Weldon 2006; Andersen and Fetner 2008), right-wing voting (Arzheimer 2009), social and political trust (Hooghe et al. 2009), satisfaction with democracy (Anderson and Singer 2008), political participation (van der Meer, van Deth, and Scheepers 2009), the political economy of the gender vote gap (Iversen and Rosenbluth 2006) and support for European integration (Hooghe and Marks 2004). Multilevel analysis is not restricted to comparative politics. Research in American politics using states or neighborhoods (e.g. Lax and Phillips 2009), studies of judicial decision making (Voeten 2008), and research on legislative politics (Franchino and Hoyland 2009) or the politics of economic reforms (e.g. Denisova et al. 2009) similarly do (or would) benefit from multilevel models.[1]

Multilevel models and software have mainly been developed in the context of educational research (Aitkin and Longford 1986; Goldstein et al. 1993), where practitioners enjoy rather generous sample sizes. The conditions for those models to work might not be met in comparative politics applications. For example, in a simulation study of the

[1] Note that a seemingly simple way to avoid multilevel modeling, namely the use of "cluster-robust" standard errors is highly dangerous, as they can be severely biased if the number of countries is small (e.g. Angrist and Pischke 2008: ch. 8.2).

effects of small sample sizes, Maas and Hox's (2004*b*) condition for "small" sample size is 30 groups, whereas in most comparative survey data sets the number of countries is substantially lower. The fact that comparative applications usually involve a smaller number of countries matters, because the basis for the widely used maximum likelihood inference is asymptotic and assumes large sample sizes. Thus standard errors are biased downwards and researchers who rely on "levels of significance" greatly overstate the level of their tests – leading to spurious significant effects (Maas and Hox 2004*b*).

Thus many researchers and reviewers wonder if multilevel models are applicable in such cases. In other words, how many countries does one need, in order to properly test hypotheses in a multilevel framework? The literature gives varying rules of thumb ranging from just 8 or 10 to 30, 50 or even 100 groups (Rabe-Hesketh and Skrondal 2008; Kreft and de Leeuw 1998; Afshartous 1995)![2]

My goal in this paper is to provide some evidence on the necessary number of countries in typical 'multilevel data sets'. Furthermore, I emphasize key differences between the widely employed frequentist and the alternative Bayesian approach. I conduct a large simulation study, which analyzes the behavior of maximum likelihood and Bayesian estimation strategies when a small number (5 to 30) of countries is used. My focus lies on how strongly estimates of theoretically relevant parameters are biased, and to what extent hypothesis tests will be misleading.

[2] Most simulation studies focus on simulations that mimic data often found in educational research. Comprehensive examples are the studies conducted by Maas and Hox (2004*b*, 2005). For an overview of earlier studies, which use approximate estimation techniques, see Rodríguez and Goldman (1995) and Kreft (1996). A comprehensive study by Moineddin, Matheson, and Glazier (2007), oriented towards health care applications, takes an approach 'inverse' to mine, by using a large number of level two units with only few level one units. A study by Normand and Zou (also a medical application), includes 10 groups as a study condition, however, their maximum number of individuals is 125 (Normand and Zou 2002). The study by Browne and Draper (2006) includes six level two units among its conditions, but focuses only on estimation of variance components.

Table 1: Six commonly used multilevel setups

	Random effects	Macro variables	Macro-micro interactions
Linear model	I	III	V
Non-linear model	II	IV	VI

Hierarchical models

I start by describing the basic types of multilevel models used in applied research and outlining key differences between the dominant frequentist paradigm and Bayesian approaches. Accessible introductions to multilevel modeling are given by Rabe-Hesketh and Skrondal (2008) and Gelman and Hill (2007). More thorough introductions are given by Snijders and Bosker (1999) and Goldstein (2010) in a frequentist framework, and by Gill (2008: ch.10), Jackman (2009: ch.7) and Draper (2008) in a Bayesian framework. In-depth treatments covering a broader variety of estimation strategies are available from McCulloch and Searle (2001) and Jiang (2007).

Table 1 lists multilevel specifications for continuous and binary dependent variables that are considered in this paper.[3] The most basic model (type I and II) tries to capture systematic differences between countries by including country random effects. This model is often termed 'random intercept model', since it can be understood as providing country specific intercepts, while all other effects are constant across countries (for a graphical illustration, see Gelman and Hill 2007: 238). These unexplained differences between countries often are central objects of study in applied comparative research. For example, a researcher might be interested in testing whether individual preferences for

[3] Linear and binary models constitute the vast majority of all applications. Extensions to ordered dependent variables are straightforward (e.g. Agresti et al. 2000; Agresti and Natarajan 2001) and results obtained here should hold for them as well. Allowing for multinomial outcomes is slightly more involved and requires a more specific simulation study.

income redistribution differ between countries as a function of welfare policies. Then, the random intercept model is extended by including country-level variables, such as measures of institutional features or income inequality, in order to explain variation in the dependent variable that is not captured by characteristics of individuals (model types III and IV).

Somewhat more formally, one models the response of individual i $(i = 1, \dots, n_j)$ living in country j $(j = 1, \dots, J)$ as function of individual and macro level variables, x_{ij} and z_j, respectively:

$$y_{ij}^{(*)} \sim N(\alpha_j + \beta x_{ij}, \sigma_y^2) \tag{1}$$

$$\alpha_j \sim N(\gamma_0 + \gamma_1 z_j, \sigma_\alpha^2) \tag{2}$$

Here β captures the individual level effect of covariate x_{ij} and α_j are country specific intercepts, which are assumed to follow a normal distribution with freely estimated variance σ_α^2. Since the goal here is to explain variation in country specific intercepts, they are modeled by a regression equation including country characteristics. An overall intercept γ_0 represent the 'country averaged level' or grand mean of the dependent variable, and the systematic effect of country characteristic z_j is captured by γ_1.

To reduce notational complexity, equation (1) refers to both continuous and binary dependent variables. With continuous outcomes this yields a standard linear model with freely estimated individual level variance σ_y^2. For binary dependent variables this is a probit model, where y^* is a latent variable which generates observed categorical responses, such that one observes $y = 1$ if $y^* > 0$ and 0 otherwise. Here the individual level variance has to be fixed at some value, usually $\sigma_y^2 = 1$.[4]

[4] For readers unfamiliar with the latent variable interpretation of probit models, King (1998: ch. 5.3) provides a quick introduction. Albert and Chib (1993) provide a Bayesian perspective.

More sophisticated comparative theories often include hypotheses which specify that the effect of an individual level variable varies as a function of country level characteristics (type V and VI). For example, a researcher might want to test if the relationship between income and left-right self-placement is stronger in countries characterized by high levels of income inequality. In this setup, often termed "cross level interaction" (Snijders and Bosker 2012: 81), intercept and covariate effects vary over countries, and one tries to explain (some of) the covariate's effect variation by an explanatory country level variable:

$$y_{ij}^{(*)} \sim N(\alpha_j + \beta_j x_{ij}, \sigma_y^2) \tag{3}$$

$$\begin{bmatrix} \alpha_j \\ \beta_j \end{bmatrix} \sim N\left(\begin{bmatrix} \gamma_0 + \gamma_1 z_j \\ \delta_0 + \delta_1 z_j \end{bmatrix}, \mathbf{\Sigma} \right) \tag{4}$$

The key addition in equation (3) is the country subscript of β_j signifying that covariate x_{ij} has a different slope in different countries. These different slopes are now, too, modeled by a regression equation. The country-average effect is captured by δ_0, and systematic differences in slopes between countries are predicted by country characteristic z_j and its associated effect coefficient δ_1. As before, systematic variation in the level of the dependent variable is explained by a macro level covariate z_j with coefficient γ_1. The fact that intercept and slope vary over countries leads to a more complex variance structure: one generally assumes that intercept and slope come from a common multivariate normal distribution, with variance covariance matrix $\mathbf{\Sigma}$.[5] Thus, with one intercept and one random slope, there are now three variance parameters to estimate: variances of intercept

[5] The intercept-slope covariance or correlation should always be included in the model, since setting it to zero a priori is a rather strong assumption (see Snijders and Bosker 2012: 76).

(σ_α^2) and slope (σ_β^2), and the covariance between intercept and slope ($\sigma_\alpha\sigma_\beta$):

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_\alpha^2 & \sigma_\alpha\sigma_\beta \\ \sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{bmatrix}.$$

Frequentist vs. Bayesian multilevel models

However, are multilevel models even applicable to political science problems? What does it mean to test the 'significance' of country level effects? Many authors take a strong sampling-based perspective (e.g. Snijders and Bosker 2012), where multilevel models are applicable when groups are sampled from a larger population.[6] This works well in fields such as educational research, where researchers design studies sampling schools and pupils. Comparative (political) research is markedly different. Most data sets do not contain a random sample of countries, but have a rather strong regional focus (e.g. *Euro*barometer, *Latino*barometer). Furthermore, a random sample from the population of all countries is often not even desirable, since middle-range theories are limited in their applicability to, say, advanced industrialized countries.

It is well known that, using the maximum likelihood estimate $\hat{\theta}$ and its associated standard error s.e.($\hat{\theta}$), the frequentist confidence interval is constructed by $\hat{\theta} \pm q \times$ s.e.($\hat{\theta}$), where q is the appropriate quantile from the normal sampling distribution. In contrast, Bayesian confidence intervals (called 'credible intervals') can be constructed without reference to a hypothetical sampling distribution (Jaynes 1976). As the full posterior probability distribution of a parameter is available, credible intervals are simply the

[6] This interpretation follows straightforwardly from the logic of classical inference: significance tests on group level coefficients have to refer to a larger population for the frequentist interpretation of probability to make sense. For a short discussion of key differences between classical and Bayesian inference see the introductory chapter of Jackman (2009).

corresponding quantiles of that distribution (Gill 2008: 45).[7] Thus, a Bayesian credible interval simply gives the posterior (i.e. after looking at the data) probability that the coefficient lies in that interval – without any reference to a 'population' of countries.[8]

When employing a Bayesian approach, prior distributions are needed for all parameters, which, multiplied with the data likelihood, yield the full posterior distribution. For a graphical illustration of the role of priors in Bayesian analysis see Jackman (2009: 15-17).[9] Many applied researchers prefer those priors to be 'non-informative', in other words they should exert as little influence on the resulting posterior distribution as possible.[10] Therefore, I consider commonly employed, non-informative or vague prior distributions in my simulations.

- Residuals (level 1 variances) in linear model specifications have (conjugate) inverse gamma priors, $\sigma^2_y, \sim \Gamma^{-1}(\epsilon, \epsilon)$ with ϵ set to 0.001 (initial Monte Carlo experiments show that the choice of ϵ is not consequential, since the data clearly dominate the prior.)

- Diffuse priors for random effects (level 2 variances) are either distributed (conjugate) inverse gamma $\sigma^2_y, \sim \Gamma^{-1}(\epsilon, \epsilon)$, with ϵ being a small constant, e.g. 0.001, or distributed uniform on the standard deviation, $\sqrt{\sigma^2_y} \sim c$, as suggested by Gelman (2006).

- In models containing a random coefficient, the variance covariance matrix $\mathbf{\Sigma}$ has

[7] For example, a 95% credible interval is constructed by taking the 2.5th and 97.5th quantile of a posterior distribution. Another widely used interval estimate is the highest posterior density region (Gill 2008: 49).

[8] In this sense, only the Bayesian approach provides a straightforward interpretation of confidence/credible intervals (for a lucid discussion see Jackman 2009, Chapter 1. An in depth discussion is given in Robert 2007, Chapter 5.)

[9] In this simple discussion I ignore the constant of proportionality as well as the discussion about subjective, diffuse and 'objective' priors. A detailed discussion of priors is given in O'Hagan and Forster 2009 and Jaynes 2003.

[10] But see Jackman and Western (1994) on the benefits of using informative priors.

the inverse of the Wishart distribution as its prior, $\boldsymbol{\Sigma} \sim W^{-1}(\boldsymbol{S}, d)$, with d degrees of freedom, set to the dimension of the variance-covariance matrix (2) plus one, and diagonal scale matrix $\boldsymbol{S} = \boldsymbol{I}_2$. This produces a marginal prior for the correlation between intercept and slope, which is uniform on $[-1, 1]$, and distributed $\Gamma^{-1}(1, 1/2)$ for the two variances. An inverse Wishart prior which posits twice the size for the (diagonal) variances serves as an alternative specification.

More sophisticated priors have been proposed (Natarajan and Kass 2000; Gelman 2006); however, I employ specifications which are commonly used by researchers (e.g. Spiegelhalter et al. 1997; Gelman and Hill 2007).

Monte Carlo study setup

For this Monte Carlo study I focus on quantities that are usually at the center of comparative researchers' interest: effect estimates and uncertainties of individual and country level variables. I use a setup which mimics data structures commonly found in comparative (survey) research, and which differs starkly from those found in educational research: It includes large numbers (usually thousands) of individuals nested within a small number of countries (often less than twenty). While systematic differences exist between individuals from different countries, individuals vary on a number of (often unobserved) factors, so that the ratio of between to overall country heterogeneity, the intraclass correlation, is rather low.[11]

[11] Intraclass correlation (ICC) is defined as variation between countries divided by total (i.e. individual and country level) variance. It thus indicates the proportion of variance that is accounted for by the country level (cf. Snijders and Bosker 2012: 17f).

Experimental design

For each of the six multilevel model types, I use a full factorial $6 \times 3 \times 7$ design. The following factors are varied:

- The main factor of interest, the number of countries: I use a commonly available set of countries ranging from 5 to 30 in increments of 5 (6 conditions). Each country contains 500 individuals.
- Intraclass correlation: I use three values typical for comparative (survey) research, namely 0.05, 0.10 and 0.15 (3 conditions).
- Estimators: Models are estimated via Maximum likelihood (1 condition) and Gibbs sampling with two different prior specifications outlined above, each time summarized using three posterior point summaries: expectation, median and mode ($2 \times 3 = 6$ conditions).

For each of those 756 conditions (126 experimental conditions for six types of multilevel models), I generated 1000 data sets and calculated estimates and confidence intervals using maximum likelihood and fully Bayesian estimation. More specifically, likelihood estimation is carried out using a standard EM algorithm (McLachlan and Krishnan 2008), and integration needed in the probit model is done numerically via adaptive Gauss-Hermite quadrature using 15 points (Pinheiro and Bates 1995; Rabe-Hesketh, Skrondal, and Pickles 2002).[12] I use a Gibbs sampler for the Bayesian model implementations (Gelfand and Smith 1990; Gelman et al. 2004: ch.11), running 2 chains for 4000 iterations,

[12] I also estimated ML models using restricted maximum likelihood. This leads to somewhat better coverage properties of the ML estimates in the simple linear model case (typically by 2-5 percentage points). However, in more complex random slope and non-linear models quite drastic non-coverage problems became apparent. Therefore, I present full maximum likelihood results here, which did not show these problems.)

and compute posterior mean, median and mode as ‘point estimates’ summarizing the posterior distribution.[13]

Reported quantities

I concentrate on two central quantities summarizing the Monte Carlo simulations, which are of primary importance for applied research: bias of estimates and non-coverage of confidence intervals. First, in order to assess the bias of point estimates, I calculate the percent relative bias, which is simply the difference between estimated and true value expressed as a proportion of the true value:

$$\frac{\hat{\theta}-\theta}{\theta}\times 100.$$

Second, the quality of interval estimates (i.e. the 95% confidence or credible intervals), is of special interest, since researchers will use them to accept or reject theories. At each Monte Carlo run, I create an indicator variable, which registers if the calculated 95% confidence interval contains the true population parameter. Averaging these values yields the coverage of the confidence interval; subtracting the nominal 95% interval coverage level (950 out of 1000) and multiplying by 100 yields its level of non-coverage in percentage points, which I report below.[14]

[13] To learn about the rate of convergence of the sampler, I carried out initial runs for each model. The usual tests (see Cowles and Carlin 1996; Gelman and Rubin 1992) suggested that the chains reached their equilibrium distribution at less than 1000 iterations.

[14] Accordingly, the estimated coverage of the true 95% interval will have a simulation accuracy of 1.35% ($1.96\sqrt{0.05*0.95/1000}=0.0135$).

Results

To reduce the number of graphs and the complexity of the presentation, I do not display results for simple linear and probit random effect models not containing a country level explanatory variable (type I and II). Results for those model types are similar to those from models including a macro predictor (types III/IV).[15] Below, I focus on models with an intra-class correlation of 0.10 and use the mean as posterior summary for Bayesian models[16], since results show that both choice of ICC and posterior summary measure do not substantially influence the central conclusions of this study. More detailed versions of these plots including posterior mean, mode, and median are available in the online appendix. The following subsection documents bias in individual and country level covariate effects for different estimation strategies in linear and probit random effects models (type III and IV). Models with random coefficients (or random slopes, type V and VI) are discussed next. Finally, I discuss the role of intraclass correlation and different variance prior choices in the last two subsections.

Hierarchical models with macro variables

Before examining estimates of substantive variables, I turn to results for the residuals, σ_y^2, and random effects, σ_α^2. In all model types, individual level residuals are estimated well by both the Frequentist and Bayesian approach, which is to be expected given the large number of individuals. The situation looks less rosy for the estimated random effect variance, where both maximum likelihood and Bayesian estimates exhibit strong bias (the Bayesian posterior mean estimate bias is more than 100% when only five countries

[15] Furthermore, in absence of country variables of interest, the question of model specification reduces to 'fixed versus random effects', which is not the topic of this paper.

[16] All Bayesian results presented below are estimated using inverse gamma priors. More on the role of priors for variances in subsection "Priors" below.

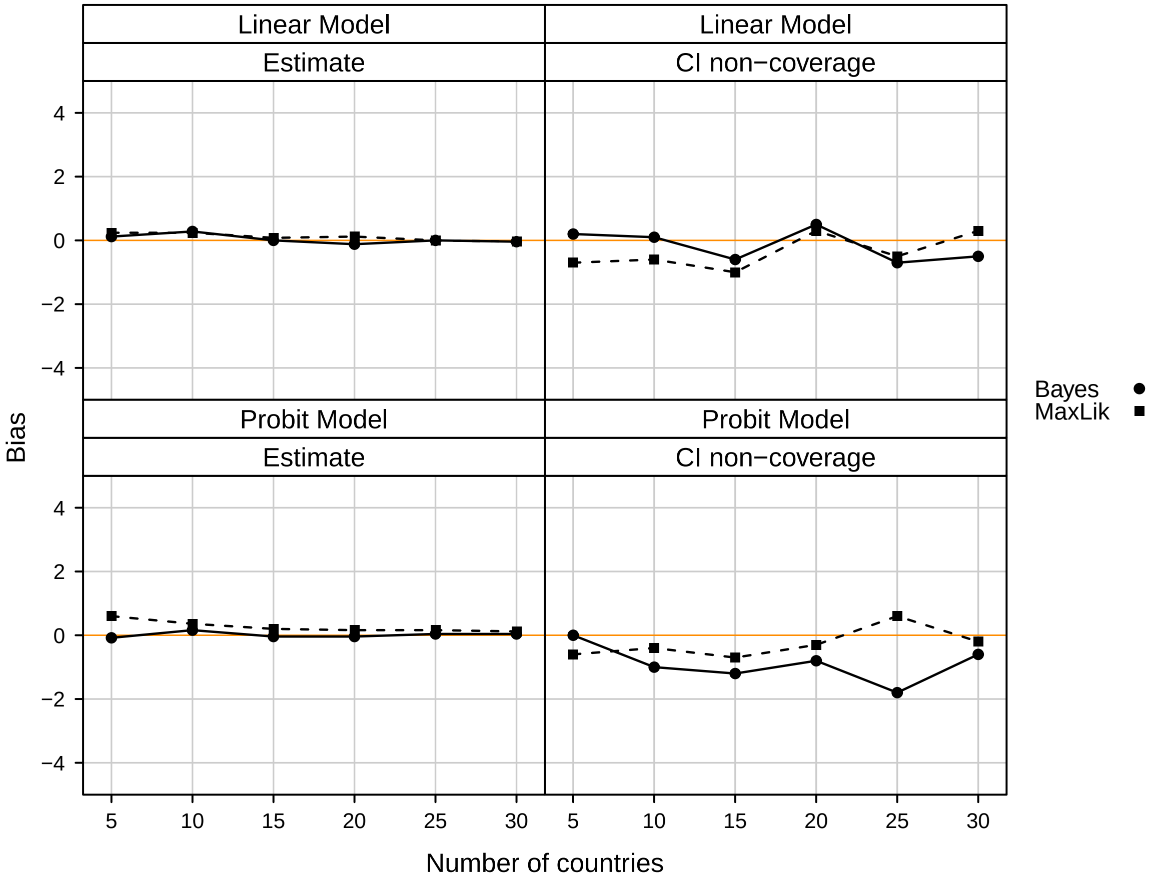

Figure 1: Performance of point and interval estimates of individual level covariate effect β in hierarchical linear and probit models (type III/IV). Displayed are relative bias (in %) of estimate and 95% confidence interval non-coverage as a function of the number of countries used, and ML and Bayes estimation.

are available). Contrarily, the 95% interval coverage of the Bayesian estimator is excellent independent of the number of countries used. Maximum likelihood results show greater non-coverage even when 30 countries are available (similar results are presented by Browne and Draper 2000; Maas and Hox 2004*a*; Kreft and de Leeuw 1998). For a much more extensive simulation study of variance component estimation see Browne and Draper (2006).[17]

While unbiased estimation of variance components is certainly important, most applied researchers are arguably most interested in the quality of estimates of their theoretical variables. Therefore, I now examine the extent of bias in estimates of individual and country level covariates. Figure 1 on the previous page shows how well the effect β of an individual level covariate x_{ij} is estimated, both in linear (upper panel) and probit (lower panel) hierarchical models. For linear and binary dependent variables, both estimation methods produce quite reliable coefficient estimates, even when using only five countries. Similarly the coverage level of the 95% intervals is very close (within 2%) to the nominal 95% level. In practice this means that for the models considered here, which contain large numbers of individuals, individual level estimates are robust to small country level sample sizes (cf. Maas and Hox 2005).

However, the main interest in multilevel models with macro variables lies in testing effects of country level characteristics on individual level outcomes. Therefore, Figure 2 on the following page displays the relative bias of effect estimates γ_1 of a macro variable z_j; and here we see that estimation strategy matters quite a bit for substantial results. At any number of countries used, Bayesian estimates are within ± 5 percent of the true population value. In contrast, maximum likelihood estimates are sharply biased upwards

[17] The term variance component refers to the fact that one tries to decompose the observed variance of a dependent variable y into several components of variation. For example, in the random intercept linear model, observed total variance is modeled as the sum of an individual level (σ_y^2) and a country level (σ_α^2) variance component.

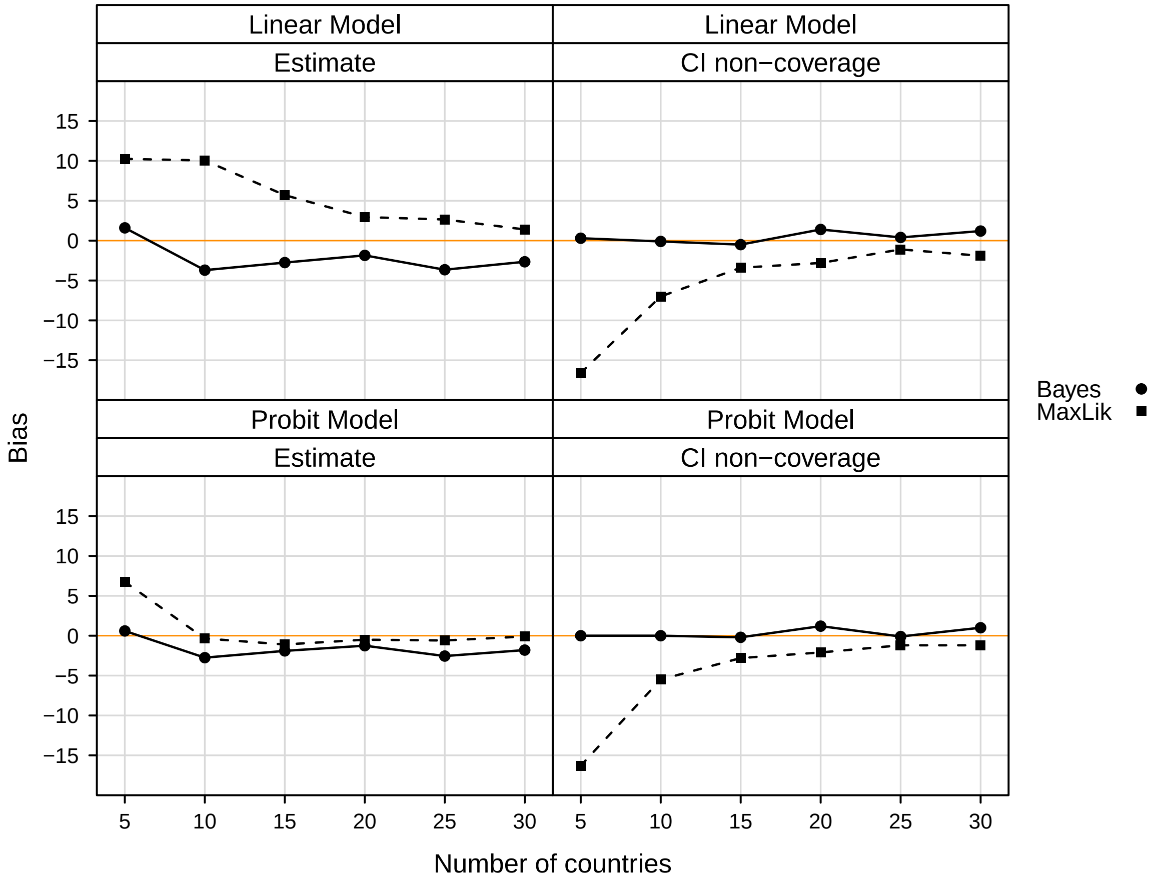

Figure 2: Performance of point and interval estimates of country level covariate effect γ_1 in hierarchical linear and probit models (type III/IV). Displayed are relative bias (in %) of estimate and 95% interval non-coverage as a function of the number of countries used, and ML and Bayes estimation.

when the number of countries is less than 20 in a hierarchical linear model. Matters get worse when examining actual versus nominal coverage of 95% intervals: when less than 15 countries are used, maximum likelihood confidence intervals are 5 to 15 percentage points too narrow, while Bayesian credible intervals are virtually congruent with their nominal level.

Models including random coefficients

In this subsection I consider random coefficient or slope models, which specify the effect of a covariate as varying between countries. As discussed above, there are three central parameters of interest: the effect of a country-level variable on the outcome, denoted γ_1; the average effect δ_0 of an individual-level covariate that varies over countries; and the effect of country-level variables on that varying variable, δ_1.

Estimates of γ_1 are displayed in Figure 3 on the next page. Both maximum likelihood and Bayesian point estimates are biased when very small samples are used, but display good properties as sample size grows. The coverage of estimated 95% intervals looks much more damaging. In contrast to simpler models (cf. Figure 2 on the preceding page), Bayesian credible intervals for γ_1 differ from their nominal level, even when larger numbers of countries are available. Compared to Bayesian credible intervals, the extent of non-coverage is larger for maximum likelihood confidence intervals. However, the most striking difference between both approaches lies in the 'direction' of non-coverage. Bayesian intervals are too wide and consequently provide overly conservative tests of hypotheses, whereas maximum likelihood intervals underestimate uncertainty of the effect of γ_1 and will provide hypothesis tests that are much more lenient than indicated by their nominal level.

Figure 4 on page 18 shows bias and non-coverage results for δ_0, which represents the country-average effect of an individual level covariate.[18] Results show that this coefficient is estimated reasonably well by Bayesian as well as maximum likelihood procedures – with ML estimates exhibiting somewhat more bias when country level sample sizes are small. This picture looks less favorable for the classical approach when considering actual coverage of confidence bounds. Both, Bayesian credible intervals and ML based

[18] This estimate corresponds to the individual level covariate effect β in models III and IV.

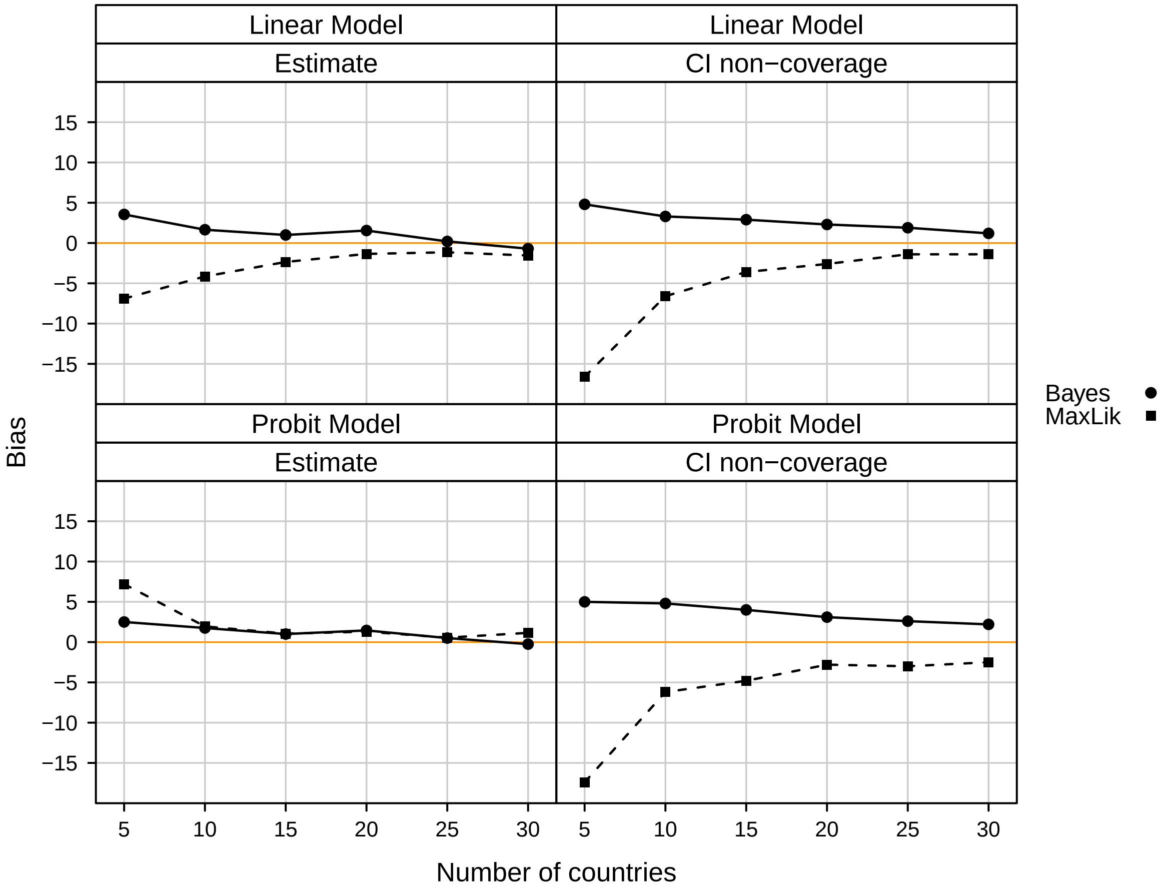

Figure 3: Performance of point and interval estimates of country level covariate effect γ_1 in hierarchical linear and probit models (type V/VI). Displayed are relative bias (in %) of estimates and 95% interval non-coverage as a function of the number of countries used, and ML and Bayes estimation.

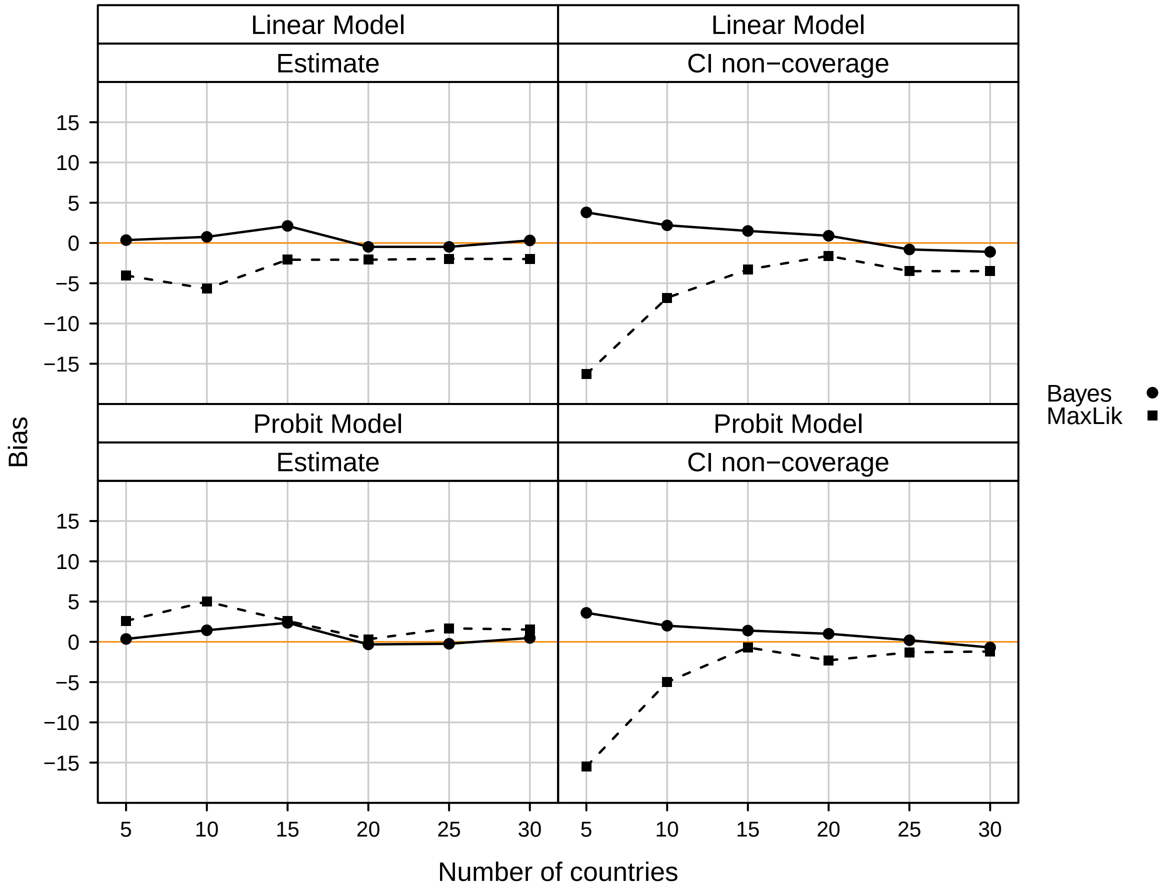

Figure 4: Performance of point and interval estimates of individual level covariate random coefficient δ_0 in hierarchical linear and probit models (type V/VI). Displayed are relative bias (in %) and 95% interval non-coverage as a function of the number of countries used, and ML and Bayes estimation.

confidence intervals get close to their nominal level as sample size gets larger, but when less than 15 countries are used, coverage of maximum likelihood confidence intervals is again strongly anti-conservative.

Finally, I examine parameter δ_1, which is of primary interest in random coefficient models: it estimates how a country level variable z_j influences the strength of the relationship between changes in x_{ij} and the dependent variable, often denoted “cross-level interaction effect”. Values of this coefficient are influenced by estimates of variance components σ^2_β and again differ sharply between maximum likelihood and Bayesian approaches. Figure 5 on the following page shows that the relative bias in estimates is most substantial when maximum likelihood estimation is used with a small number of countries (less than 15). In hierarchical linear models this leads to an underestimation of the true effect by 10 to 15 percent, while the converse is true for hierarchical probit models. Bayesian point estimates produces less bias for small numbers of countries and are virtually identical with ML estimates when 25 or 30 countries are available. Considering the actual coverage of nominal 95% intervals the now familiar picture emerges: Bayesian credible intervals are too wide, i.e. they provide more conservative tests of hypotheses, while ML confidence intervals are too short providing test that are potentially very misleading, even at medium sample sizes.

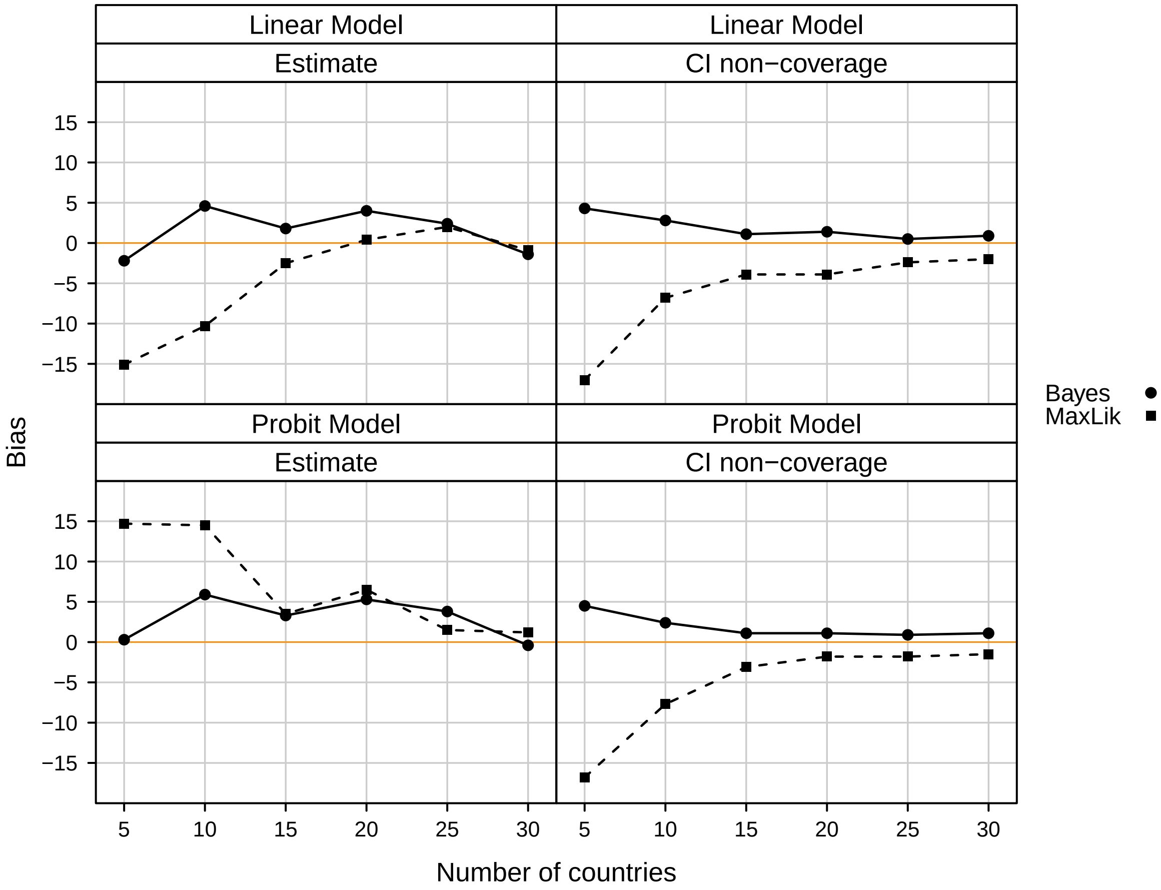

Figure 5: Performance of point and interval estimates of country*individual level interaction effect δ_1 in hierarchical linear and probit models (type V/VI). Displayed are relative bias (in %) and 95% interval non-coverage as a function of the number of countries used, and ML and Bayes estimation.

Intra-class correlation

Previous results were based on an 'average' intraclass correlation value of 0.10, since my results show that it is of minor importance for the kind of model setups considered here. The pattern among the different estimators, established above, is similar at different ICC levels. If anything, higher ICC levels yield biases that are slightly more pronounced – thus making the gap between Bayesian and ML estimators even wider. An instance where an ICC effect is discernible is displayed in Figure 6 on the next page, which shows relative bias in estimates at different ICC levels in a hierarchical probit model of type VI. When high correlations between unobserved characteristics of individuals from the same country exist, Bayesian as well as maximum likelihood procedures produce slightly larger bias when the number of countries is very small. With increasing number of countries, this effect vanishes, with the Bayesian posterior mean getting closer to the true population value at a slightly faster rate than maximum likelihood estimates.

Priors

The previous discussion ignored the role of priors for the variance components in Bayesian analyses. Results presented used inverse gamma priors with small values for shape and scale, which showed to yield reasonable results in my simulations, despite their technical shortcomings (Gelman 2006). The difference between prior choices is illustrated in Figure 7 on the following page, which shows interval coverage for country level covariate effects γ_1 in a hierarchical linear model under different prior specifications. The slight differences between inverse gamma and uniform on the standard deviation priors vanish as more information becomes available. This pattern is similar for other coefficient estimates in both linear and probit models. When inverse Wishart priors are used as priors for variance-covariance matrices, I find similar results: differences in prior values

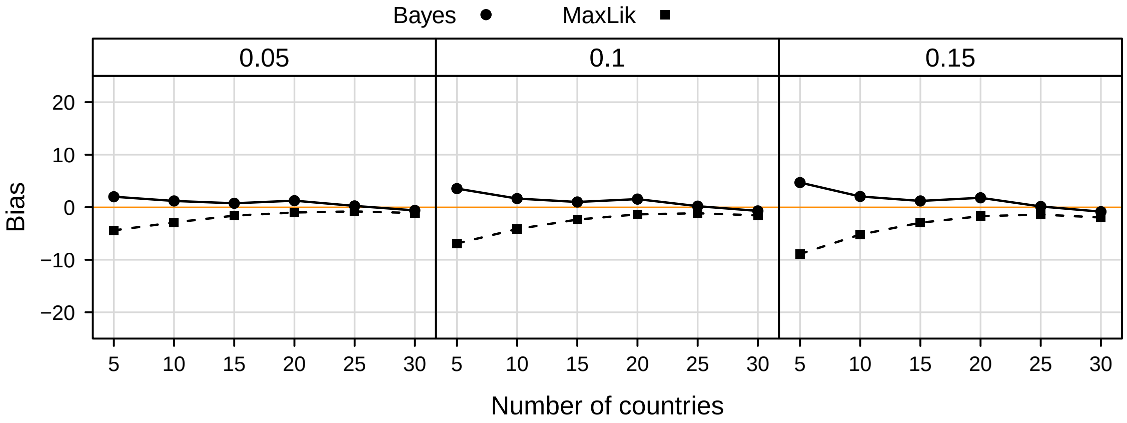

Figure 6: Effect of intraclass correlation on performance of point estimates of country level effects γ_1 in a hierarchical probit model (type VI). Displayed is relative bias (in %) of the estimate as a function of three levels of intraclass correlations, the number of countries used, and ML and Bayes estimation.

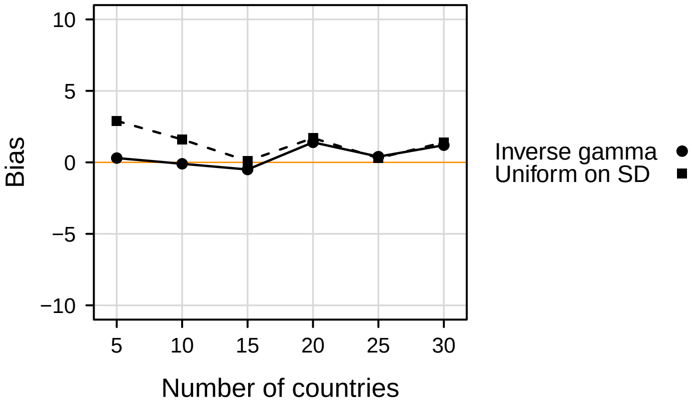

Figure 7: Effect of different variance component priors on performance of point and interval estimates of γ_1 in a hierarchical linear model (type III). Displayed is 95% interval non-coverage as a function of the number of countries used, and inverse gamma and SD uniform priors.

are only relevant when little country level information is available. This suggest that applied researchers who have to work with small samples, should test the robustness of their findings by estimating several models using varying prior values (cf. Gill 2008: 204f.).[19]

An illustration: Support for the European Union

To illustrate the behavior of different estimators in a practical setting, I continue the example of Steenbergen and Jones' (2002) introductory paper. They model citizen support for the European Union (see also Hooghe and Marks 2004) as a function of individual, party, and country characteristics. To that end, they employ a three-level model with a continuous dependent variable (our type III) where individuals are nested in parties and countries. To stay in line with the setup of this paper, and to keep this example simple, I will focus on individual and country level factors. Using the Eurobarometer 46.0 survey, fielded in September and October 1996, I find that – as expected from my Monte Carlo study results – maximum likelihood and Bayesian estimates agree with each other as far as individual level effects are concerned.[20] What I will focus on here is the effect of country level variables.

Suppose a researcher is interested in testing the claim that, as dependence between EU

[19] One of the strengths of the Bayesian approach is the possibility of utilizing prior information when little data is available (see Jackman and Western (1994) for an illustration of using informative priors). In this study I relied on priors which are generally seen as 'uninformative' in the sense that they try to provide no or little *a priori* information. In applications it will often be helpful to use more informative priors, if only as a robustness check.

[20] For estimates see the online appendix. Steenbergen and Jones discard all individuals for which they have no matching party information. Since I do not use party information for this example, my sample size is slightly higher, comprising 10,777 individuals. Eurobarometer 46.0 was fielded in fifteen countries. Data on trade balance by Eurostat was not available for Luxembourg, which leaves fourteen countries for my analysis: Austria, Belgium, Denmark, Finland, France, Greece, Ireland, Italy, the Netherlands, Portugal, Spain, Sweden, United Kingdom, and West Germany.

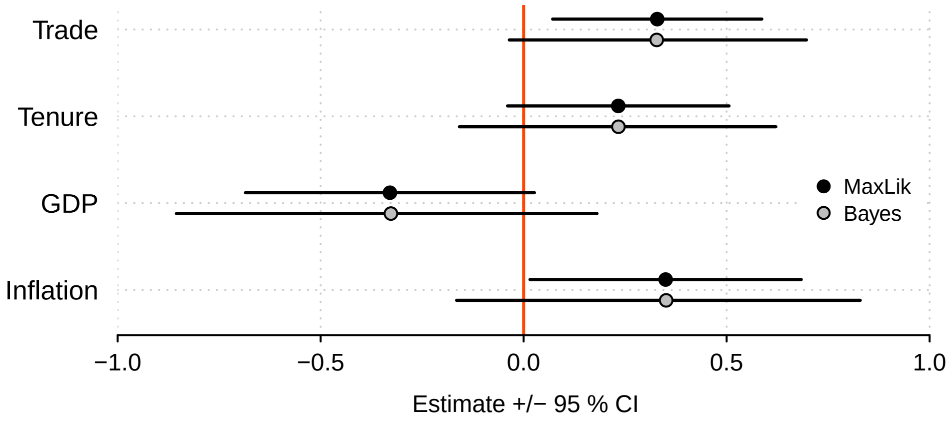

Figure 8: Country level determinants of support for the European Union. Maximum likelihood estimates and normal based 95% confidence intervals. Bayesian estimates (posterior means) and 95% credible intervals. Sample size: 10,777 individuals, 14 countries.

countries increases, citizens realize the importance of the common market, which leads to higher popular support for the European Union (Eichenberg and Dalton 1993). Following Steenbergen and Jones, this dependence is measured using a country's trade balance – the ratio of a country's intra-EU trade to its total trade. I also include a country's tenure, i.e. the number of years a country has been a member of the union, in order to capture the effect that public opinion in new member states is often negative (Steenbergen and Jones 2002: 228). Furthermore, since countries differ widely in their level of economic development and economic performance, I include their gross domestic product per capita and rate of inflation (see Mahler, Taylor, and Wozniak 2000 for more detailed analyses). All macro variables are standardized, i.e. centered around the overall mean and divided by their standard deviation.[21]

Figure 8 plots estimates and confidence intervals for our macro variables of interest.[22]

[21] GDP per capita and inflation data are from Eurostat's online database. Information on trade balance is calculated from Eurostat 2004.

[22] Bayesian estimates are calculated from two chains run for 30,000 iterations, of which with the first 10,000 were discarded as burn-in. Point summaries are computed from the remaining 20,000 iterations thinned by a factor of 5.

We see the now familiar 'overconfident' confidence intervals produced by maximum likelihood estimation when the number of countries is small. While parameter estimates of macro effects produced by both estimation strategies are virtually identical, maximum likelihood's 95% confidence intervals are considerably shorter than those obtained when following a fully Bayesian approach. With respect to our main variable of interest, trade balance, two researchers could arrive at two different theoretical conclusions simply by using different estimation strategies. A researcher relying on maximum likelihood estimation might confidently conclude to have demonstrated the link between trade dependence and public opinion. Contrarily, a researcher relying on Bayesian estimation would have to conclude that such an effect cannot be unambiguously demonstrated with the data at hand. Clearly, even if one is only interested in substantive questions when applying multilevel models to comparative data, attention should be paid to the implications of maximum likelihood versus Bayesian inference.

Conclusion

In this Monte Carlo study I examined the effects of using a small number of countries in multilevel models. I used an extensive design, covering both linear and non-linear models estimated via maximum likelihood in a classical framework and via MCMC sampling in a Bayesian framework, different posterior summaries, different prior choices and different levels of intraclass correlation.

Results are rather sobering from a classical maximum likelihood perspective. The simulations confirm, once more, the literature on problems with maximum likelihood inference for multilevel models when the number of groups is small – a problem that arises when the log-likelihood is not close to being a quadratic function of the parameters. Simple linear or probit models containing only a random intercept are the best case scenario. Here ML estimates and confidence interval coverage of estimated macro effects are only biased to a limited extent, as long more than 15 or 20 countries are available. But even in this optimal setting, using fewer countries quickly leads to confidence intervals that are far from their declared level. While confidence interval non-coverage is undesirable in itself, it is the direction of this bias which is cause for concern. Without exception, ML produces confidence intervals that are too short, so that hypothesis tests are anti-conservative.

Many comparative theories build on interactions between effects of individual level variables and country level characteristics, the so called 'cross-level interactions'. Here, the problems of ML estimation are most apparent. Even in their most simple specification (as employed in this simulation study), those models include three variance parameters: intercept and slope variance and the covariance between them. Their estimation is difficult even with 20 or more countries. ML estimates of cross-level interactions tend to be biased upwards in probit models, whereas the opposite occurs for linear models. More problematic is, again, the fact that actual and nominal confidence interval level do not match: ML confidence intervals are invariably too short. Furthermore, in these more complicated models testing the effect of a country level covariate on the dependent variable is problematic as well. Estimated with 15 or 20 available countries, ML confidence intervals are almost 5% too short – in other words, researchers are more likely to obtain 90% confidence intervals, rather than the 95% intervals announced by their software package.[23]

In contrast, estimates obtained using a Bayesian approach show far better properties, especially with respect to confidence interval coverage. When using only a small number of countries (<10), Bayesian point estimates are biased as well. But the magnitude of

[23] One should keep in mind that Monte Carlo studies represent optimal conditions. Thus in real-life data analyses bias and non-coverage problems are likely to be worse.

this bias is much smaller as in the case of ML estimates: under conditions considered in this study, Bayesian point estimates were biased at most 5%, whereas ML estimates reached 10 or 15%. The clearest advantage of employing a Bayesian approach to multilevel modeling lies in its excellent confidence interval coverage. Bayesian credible intervals are closer to their nominal level than their ML counterparts. What is more, when they are biased, they usually are too long. Thus, one could claim that researchers using Bayesian multilevel models put their hypotheses to more rigid tests than their colleagues relying on ML estimates! However, there is no magic bullet. A small numbers of countries combined with complex models can present problems for the Bayesian approach as well. This is evident in complex cross-level interaction models, where the credible interval for this interaction effect is consistently too large, even with 20 or more countries.

This point is worth repeating. As already discussed above, estimation of models with just one cross-level interaction can already prove to be difficult. For practical applications this means that researchers should be cautious when fitting complex models with a large number of macro-micro interactions. Researchers might be tempted to include many such cross-level interactions to specify flexible models where 'everything depends on context'. A seemingly simple strategy then starts from a specification with many interactions and removes those with "non-significant" variances one-by-one. My results strongly suggest that this is doomed to fail: confidence intervals of cross-level interactions will likely be severely biased and theoretical conclusion drawn from such a procedure are not particularly trustworthy. This suggests that analyses using many cross-level interactions with a limited set of countries should be met with a healthy dose of skepticism.

Despite these reservations, my results show that the integration of micro- and macro-data is a worthwhile enterprise – provided one's tools are chosen wisely. Thus, for researchers in comparative politics (and adjacent 'comparative' fields), who are interested in multilevel analyses, turning to a Bayesian approach might be a fruitful choice.

References

Achen, Christopher H. 2005. “Two-Step Hierarchical Estimation: Beyond Regression Analysis.” *Political Analysis* 13(4): 447–456.

Angrist, Joshua D. and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist’s Companion.* Princeton: Princeton University Press.

Afshartous, David. 1995. “Determination of Sample Size for Multilevel Model Design.” Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Agresti, Alan, and Ranjini Natarajan. 2001. “Modeling Clustered Ordered Categorical Data: A Survey.” *International Statistical Review* 69(3): 345–371.

Agresti, Alan, James G. Booth, James P. Hobert, and Brian Caffo. 2000. “Random-Effects Modeling of Categorical Response Data.” *Sociological Methodology* 30(1): 27–80.

Aitkin, M., and N. Longford. 1986. “Statistical Modelling Issues in School Effectiveness Studies.” *Journal of the Royal Statistical Society A* 149(1): 1–43.

Albert, J. H., and S. Chib. 1993. “Bayesian Analysis of Binary and Polychotomous Response Data.” *Journal of the American Statistical Association* 88(422): 669–679.

Andersen, Robert, and Tina Fetner. 2008. “Economic Inequality and Intolerance: Attitudes toward Homosexuality in 35 Democracies.” *American Journal of Political Science* 52(4): 942–958.

Anderson, Christopher J., and Matthew M. Singer. 2008. “The Sensitive Left and the Impervious Right: Multilevel Models and the Politics of Inequality, Ideology, and Legitimacy in Europe.” *Comparative Political Studies* 41(4/5): 564–599.

Arzheimer, Kai. 2009. “Contextual Factors and the Extreme Right Vote in Western Europe, 1980–2002.” *American Journal of Political Science* 53(2): 259–275.

Browne, William C., and David Draper. 2006. “A comparison of Bayesian and likelihood-based methods for fitting multilevel models.” *Bayesian Analysis* 1(3): 473–514.

Browne, William J., and David Draper. 2000. “Implementation and Performance Issues in the Bayesian and Likelihood Fitting of Multilevel Models.” *Computational Statistics* 15(3): 391–420.

Cowles, Mary Kathryn, and Bradley P. Carlin. 1996. “Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review.” *Journal of the American Statistical Association* 91(434): 883–904.

Denisova, Irina, Markus Eller, Timothy Frye, and Ekaterina Zhuravskaya. 2009. “Who Wants To Revise Privatization? The Complementarity of Market Skills and Institutions.” *American Political Science Review* 103(2): 284–304.

Draper, David. 2008. “Baysian Multilevel Analysis and MCMC.” In *Handbook of Multilevel Analysis*, ed. Jan de Leeuw, and Erik Meijer. New York: Springer pp. 77–139.

Eichenberg, Richard, and Russell Dalton. 1993. “Europeans and the European Community: The Dynamics of Public Support for European Integration.” *International Organization* 47(4): 507–534.

Eurostat. 2004. *External and intra-European Union trade – Statistical Yearbook.* Luxembourg: Office for Official Publications of the European Communities.

Fahrmeir, Ludwig, and Gerhard Tutz. 1997. *Multivariate Statistical Modelling Based on Generalized Linear Models. Second edition.* New York: Springer.

Franchino, Fabio and Bjorn Hoyland. 2009. "Legislative Involvement in Parliamentary Systems: Opportunities, Conflict, and Institutional Constraints " *American Political Science Review* 103(4): 607–621.

Franzese, Robert J. Jr. 2005. "Empirical Strategies for Various Manifestations of Multilevel Data." *Political Analysis* 13(4): 430–446.

Gelfand, Alan E., and Adrian F. M. Smith. 1990. "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85(410): 398–409.

Gelman, Andrew. 2006. "Prior distributions for variance parameters in hierarchical models." *Bayesian Analysis* 1(3): 515–534.

Gelman, Andrew, and Donald Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7(4): 457–511.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel / Hierarchical Models.* Cambridge University Press.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis.* Boca Raton: Chapman & Hall.

Gilardi, Fabrizio. 2010. "Who Learns from What in Policy Diffusion Processes?" *American Journal of Political Science* 54(3): 650–666.

Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3): 647–674.

Gill, Jeff. 2008. *Bayesian Methods. A Social and Behavioral Sciences Approach.* Boca Raton: Chapman & Hall.

Goldstein, Harvey. 2010. *Multilevel Statistical Models..* Chicester: Wiley.

Goldstein, Harvey, Jon Rasbash, Min Yang, Geoffrey Woodhouse, Huiqi Pan, Desmond Nuttall, and Sally Thomas. 1993. "A Multilevel Analysis of School Examination Results." *Oxford Review Of Education* 19(4): 425–433.

Hooghe, Liesbet, and Gary Marks. 2004. "Does Identity or Economic Rationality Drive Public Opinion on European Integration?" *PS: Political Science & Politics* 37(3): 415–420.

Hooghe, Marc, Tim Reeskens, Dietlind Stolle, and Ann Trappers. 2009. "Ethnic Diversity and Generalized Trust in Europe. A Cross-National Multilevel Study." *Comparative Political Studies* 42(2): 198–223.

Iversen, Torben, and Frances Rosenbluth. 2006. "The Political Economy of Gender: Explaining Cross-National Variation in the Gender Division of Labor and the Gender Voting Gap." *American Journal of Political Science* 50(1): 1–19.

Iversen, Torben, and Frances Rosenbluth. 2006. “Bayesian Inference for Comparative Research.” *American Political Science Review* 88(2): 412–423.

Jackman, Simon D. 2009. *Bayesian Analysis for the Social Sciences.* New York: Wiley.

Jaynes, E. T. 1976. “Confidence intervals vs. Bayesian Intervals.” In *Foundations of Probability theory, statistical inference, and statistical theories of science.* Dordrecht: D. Reidel pp. 175–257.

Jaynes, E. T. 2003. *Probability Theory. The Logic of Science.* Cambridge: Cambridge University Press.

Jiang, Jiming. 2007. *Linear and Generalized Linear Models Mixed Models and Their Applications.* New York: Springer.

Jusko, Karen Long, and W. Philips Shively. 2005. “Applying a Two-Step Strategy to the Analysis of Cross-National Public Opinion Data.” *Political Analysis* 13(4): 327–344.

Kedar, Orit, and W. Philips Shively. 2005. “Introduction to the Special Issue” *Political Analysis* 13(4): 1–4.

King, Gary. 1998. *Unifying Political Methodology. The Likelihood Theory of Statistical inference.* Ann Arbor: The University of Michigan Press.

Kreft, Ita. 1996. “Are multilevel techniques necessary? An overview including simulation studies.” California State University, Los Angeles.

Kreft, Ita G. G., and Jan de Leeuw. 1998. *Introducing Multilevel Modeling.* Thousand Oaks: Sage.

Lax, Jeffrey R., and Justin H. Phillips. 2009. “Gay Rights in the States: Public Opinion and Policy Responsiveness.” *American Political Science Review* 103(3): 367–386.

Maas, C. J. M., and J. J. Hox. 2004*a*. “The influence of violations of assumptions on multilevel parameter estimates and their standard errors.” *Computational Statistics and Data Analysis* 46(3): 427–440.

Maas, Cora J. M., and Joop J. Hox. 2004*b*. “Robustness issues in multilevel regression analysis.” *Statistica Neerlandica* 58(2): 127–137.

Maas, Cora J. M., and Joop J. Hox. 2005. “Sufficient Sample Sizes for Multilevel Modeling.” *Methodology* 1(3): 85–91.

Mahler, Vincent A., Bruce J. Taylor, and Jennifer R. Wozniak. 2000. “Economics and Public Support for the European Union: An Analysis at the National, Regional, and Individual Levels.” *Polity* XXXII(3): 429–453.

McCulloch, Charles E., and Shayle R. Searle. 2001. *Generalized, Linear, and Mixed Models.* New York: Wiley.

McLachlan, Geoffrey J., and Thriyambakam Krishnan. 2008. *The EM Algorithm and Extensions.* New York: Wiley.

Moineddin, Rahim, Flora I. Matheson, and Richard H. Glazier. 2007. “A simulation study of sample size for multilevel logistic regression models.” *BMC Medical Research Methodology* 7: 34.

Natarajan, Ranjini, and Robert E. Kass. 2000. “Reference Bayesian methods for generalized linear mixed models." *Journal of the American Statistical Association* 95(449): 227–237.

Normand, Sharon-Lise T., and Kelly H. Zou. 2002. "Sample size considerations in observational health care quality studies." *Statistics in Medicine* 21(3): 331–345.

O'Hagan, Anthony, and Jonathan Forster. 2009. *Kendalls Advanced Theory of Statistic: Bayesian Inference.* 2nd ed. Chichester: Wiley.

O'Rourke, Kevin H., and Richard Sinnott. 2006. "The determinants of individual attitudes towards immigration." *European Journal of Political Economy* 22(4): 838–861.

Pinheiro, José C., and Douglas M. Bates. 1995. "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model." *Journal of Computational and Graphical Statistics* 4(1): 12–35.

Rabe-Hesketh, Sophia, and Anders Skrondal. 2008. *Multilevel and Longitudinal Modeling Using Stata.* College Station: Stata Press.

Rabe-Hesketh, Sophia, Anders Skrondal, and Andrew Pickles. 2002. "Reliable estimation of generalized linear mixed models using adaptive quadrature." *The Stata Journal* 2(1): 1–21.

Robert, Christan P. 2007. *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation.* New York: Springer.

Rodríguez, Germán, and Noreen Goldman. 1995. "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses." *Journal of the Royal Statistical Society A* 158(1): 73–89.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models.* Boca Raton: Chapman & Hall.

Snijders, Tom, and Roel Bosker. 1999. *Multilevel Analysis. An introduction to basic and advanced multilevel modeling.* Thousand Oaks: Sage.

Snijders, Tom, and Roel Bosker. 2012. *Multilevel Analysis. An introduction to basic and advanced multilevel modeling. 2nd edition..* Thousand Oaks: Sage.

Spiegelhalter, D. J., A. Thomas, N. Best, and W. R. Gilks. 1997. *BUGS: Bayesian Inference Using Gibbs Sampling. Manual.* Cambridge: Medical Research Council Biostatistics Unit.

Steenbergen, Marco R., and Bradford S. Jones. 2002. "Modeling Multilevel Data Structures." *American Journal of Political Science* 46(1): 218–237.

van der Meer, Tom W. G., Jan W. van Deth, and Peer L. H. Scheepers. 2009. "The Politicized Participant: Ideology and Political Action in 20 Democracies." *Comparative Political Studies* 42(11): 1426–1457.

Verbeek, Marno. 2004. *A Guide to Modern Econometrics.* Chicester: Wiley.

Voeten, Erik. 2008. "The Impartiality of International Judges: Evidence from the European Court of Human Rights." *American Political Science Review* 102(4): 417–433.

Weldon, Steven A. 2006. "The Institutional Context of Tolerance for Ethnic Minorities: A Comparative, Multilevel Analysis of Western Europe." *American Journal of Political Science* 50(2): 331–349.