Appendix to:
A Dynamic State-Space Model of Coded Political Texts

Martin Elff
Universität Konstanz
Fachbereich Politik- und Verwaltungswissenschaft
Box D84
Universitätsstraße 10
78464 Konstanz
Germany
E-mail: martin.elff@uni-konstanz.de

March 12, 2013

A Textual data and spatial models: The need for a measurement model with dynamics

The reconstruction of political positions from political texts usually starts with the identification of semantic or grammatical tokens and proceeds with classifying these tokens into politically relevant categories and concludes with counting the occurrence of these political categories. These tokens may be words or they may be sentences or quasi-sentences. With regards to the next step, one can further distinguish between approaches that classify these token into more broader categories of political content, and approaches that leave the tokens, in this case usually words, as they are, without further classification. The result of this preparation is a set of counts or percentages, where the counts refer to the number tokens that correspond to the categories of classification or the number of tokens themselves while the percentages correspond to the number of tokens within the respective categories relative to the total number of tokens present in the text.

When such counts or percentages of political categories are obtained one has not yet arrived at positions in a political space. In spatial models of politics, the usual assumption is that actors have positions or ideal points in a uni-dimensional or multidimensional Euclidean space and that the choice of a particular option or the support for a particular policy depends on the distance between the actor's ideal point and the location of the option or policy in the political space. Positions in a Euclidean space are in general represented by coordinate values with respect to a particular coordinate system, where these values can be positive, negative or zero,

depending of the location of the position relative to the origin of the coordinate system. Counts and percentages of coding categories or words, as they result from a quantitative preparation of a political text, are always non-negative, hence there is no simple match or linear relation between these counts or percentages and positions in a political space.

This problem can be illustrated by the example of the general left-right index (also known as the RiLe index) provided by Comparative Manifesto Project (CMP) (Budge et al. 2001; Klingemann et al. 2006; Volkens et al. 2010), which is now widely used to examine the causes and consequences of parties positions that they take in their electoral platforms. This left-right index is constructed based on 26 of the 56 coding categories used by the CMP. Of these 26 categories 13 are classified as "left" while another 13 categories are classified as "right". The left-right index for each electoral platform covered by the CMP is then constructed by summing all percentages referring to the "left" categories and summing all the percentages referring to the "right" categories and then by subtracting the sum of "left" percentages from the sum of "right" percentages. The resulting index thus attains both positive and negative values which are interpreted by the CMP researchers as coordinate values on a left-right political axis. If N is the total number of quasi-sentences of an electoral platform, L is the total number of quasi-sentences that fall into one of the coding categories interpreted as "leftist" and R is the total number of quasi-sentences that fall into one of the coding categories interpreted as "rightist" then the RiLe index is constructed as

$$\text{RiLe} = 100\frac{R - L}{N} \tag{1}$$

where the multiplication by 100 reflects the fact that the RiLe index is a difference in percentages. While this index is simple and straightforward, it circumvents rather than solves the problem of the absence of a direct relation between locations in a Euclidean space and counts/percentages of coding categories and thus has three major limitations.

The first limitation is that this index requires the political space to be uni-dimensional: That is, if a party is liberal or leftist in terms of economic policy it will also be liberal or leftist in terms of social policy. The second limitation is that this index rests on the implicit assumption that there is no real variation among "leftist" categories and among the "rightist" categories in terms of the location on the left-right axis. The third limitation is that it does not make use of the particular structure of the data: Left-right scores for each electoral platform are computed in isolation, without making use of the fact that several of the platforms covered in the data come from the same party and that there is a temporal order in the platforms published by the same party.[1]

It is hard to check empirically whether the first limitation is really a serious one. A principal components analysis or factor analysis is not appropriate for this kind of data, because the

[1]Klingemann et al. (2006) use some a time series technique to discuss the reliability of the scores, but this does not yet mean that the construction of the scores themselves was designed to take into account the pooled time-series structure of the data.

relations among percentages from a given total are typically non-linear. Nevertheless, various authors have attempted to use PCA or factor analysis to check the dimensionality of the CMP data (e.g. Budge et al. 1987; Gabel and Huber 2000). Yet the results of a PCA are barely useful. For example, a replication of Gabel and Huber's principal components analysis leads to no less than 20 components with eigenvalues larger than unity. The first component, on which Gabel and Huber base their "vanilla scores" (Gabel and Huber 2000), captures just 5.07 per cent of the total variance.

An alternative use of the CMP data, which posits the existence of two substantially different left-right dimensions, an economic one and a social one, was proposed by Laver and Garry (2000) (see also Benoit and Laver 2006).[2] They also devise a way to account for variations in the salience of these two dimensions in electoral platforms by forming a relative proportional difference: If L is the frequency of all (quasi-)sentences together that refer to "leftist" categories and R is the frequency of all (quasi-)sentences that refer to "rightist" categories then their salience-corrected economic left-right index is defined as

$$LR_{\text{LG}} = \frac{R - L}{R + L}. \tag{2}$$

Probably as a reaction to the criticism brougth forward against the idea of a single overarching left-right dimension, recent editions of the CMP data set (Klingemann et al. 2006; Volkens et al. 2010) now also contain some domain specific left-right indices, notably a PlanEco and a FreeMarket index which are both variants of an index of a economic left-right dimension. That is, they are also constructed according to equation (1), yet with R and L defined differently. While these indices allow for the possibility of several political dimensions specific for particular policy areas, they do not allow to separate parties' positions within a policy area from the salience of the policy area. One could expect that they are even stronger affected by a variation in the salience of a policy area than the overarching RiLe index, which can be seen as a summary of left-right positions in several policy areas.

Laver and Garry's (2000) approach was further developed by Lowe, Benoit, Mikhaylov, and Laver (2011), who, in order to address the basic problem of non-linearity, form the (stabilized) log-odds of the "leftist" sum L and the "rightist" sum R:

$$LR_{\text{LBML}} = \ln \frac{R + .5}{L + .5} = \ln(R + .5) - \ln(L + .5) \tag{3}$$

Yet both Laver and Garry's (2000) and Lowe et al.'s (2011) approaches do not address the second limitation: When a CMP category is considered for either of the two political dimensions, its location on it is either categorically on the left or on the right. Again, there is no variation in the "left-ness" or "right-ness" of particular coding categories.

[2] Laver and Garry (2000) also propose a different coding scheme that uses more explicitly confrontational coding categories.

Approaches that address the problem of non-linearity and overcome the first two limitations by allowing for a multidimensional political space and for variation in the location of the coding categories have been put forward by van der Brug (2001) and Elff (2009). Van der Brug (2001) uses non-metric multidimensional scaling to reconstruct parties positions from their electoral platforms while Elff (2009) uses metric multidimensional unfolding of the log-transformed counts of the CMP categories. Both approaches are essentially exploratory in nature and thus not well-suited to test hypotheses. Van der Brug's (2001) method of reconstructing parties' position also fails to make sure that coordinate axes obtained from the MDS of electoral platforms from different countries or to different points in time have the same interpretation.

Slapin and Proksch's (2008) Wordfish approach, though based on counts of words rather than on counts of broader coding categories, has a formal structure similar to Elff's (2009) unfolding approach in that it involves a logarithmic relation between parties positions or their distances to the locations of policies and observable counts. In Slapin and Proksch (2008) this logarithmic relation is motivated by an explicit probability model in which the counts are assumed to have a Poisson distribution. This model takes the following form (in the notation consistent with the article)

$$\ln \mu_{ijt} = \lambda_{jt} + \psi_i + \alpha_i b_{jt} \tag{4}$$

where λ_{jt} is a fixed effect for the total length of the political text published by actor j at time t, ψ_i is a fixed effect for the total number of counts in which the semantic token (i.e. word) i occurs in all texts covered, α_i is a "weight" that expresses what direction of the political space is represented by the token i and in what strength, and b_{jt} is the position of the actor j at time t in the political space. Yet their model is limited in so far as it allows only for positions on a single political dimension and, like all the other approaches discussed previously, that it does not make use of the pooled time-series structure of the data. Furthermore, the Wordfish model is not a truly spatial model. The word parameters and position parameters of this model can be interpreted as their "sensitivities" with respect to the axis of the political space, but not as spatial locations (Lowe 2008).

The usefulness of a measurement model linking positions to non-metric data is demonstrated by the Nominate approach for the reconstruction of legislators' ideal points from roll-call data (Poole and Rosenthal 1985). The observed data for the Nominate approach is binary: the Yea or Nay vote of each legislator on a bill proposed in the House of Representatives or Senate. In the original version of the Nominate model it is assumed that (1) either option, Yea or Nay, of the vote for a bill proposed at time t is representable by a location $\boldsymbol{\alpha}_{it}$ in a political space (where $i = 1$ for Yea and $i = 0$ for Nay) (2) that each legislator has a constant ideal point $\boldsymbol{\beta}_j$ in the same political space, and (3) the probability that a legislator votes Yea or Nay is determined by the distance between their ideal point and the locations of the to alternatives. More specifically, the

relation between the vote probabilities, the ideal point and the locations of the alternatives is given by

$$\ln \frac{\Pr(V_{jt} = 1)}{\Pr(V_{jt} = 0)} = \eta\left(||\boldsymbol{\alpha}_{1t} - \boldsymbol{\beta}_j||\right) - \eta\left(||\boldsymbol{\alpha}_{0t} - \boldsymbol{\beta}_j||\right) \tag{5}$$

or equivalently

$$\Pr(V_{jt} = 1) = \frac{\exp\left(\eta\left(||\boldsymbol{\alpha}_{1t} - \boldsymbol{\beta}_j||\right)\right)}{\exp\left(\eta\left(||\boldsymbol{\alpha}_{1t} - \boldsymbol{\beta}_j||\right)\right) + \exp\left(\eta\left(||\boldsymbol{\alpha}_{0t} - \boldsymbol{\beta}_j||\right)\right)} \tag{6}$$

where V_{jt} is the vote cast by legislator j in the role call at time t, $||\boldsymbol{\alpha}_{it} - \boldsymbol{\beta}_j||$ is the Euclidean distance between the location of the alternative i $(i = 0, 1)$ and the ideal point of legislator j, and η is a function that is monotonously decreasing. In the original formulation this function was defined as $\eta(x) = \exp\left(-\frac{1}{2}\omega^2 x^2\right)$ for constant ω (Poole and Rosenthal 1985).

Clinton et al. (2004) use a model that differs from equation (5) in that the right-hand side of their model equation does not contain an exponential function and that the left-hand side specifies a probit-link rather than a logit link, which allows them to arrive at the simpler model equation (in a notation adapted to this article):

$$\Phi^{-1}(\Pr(V_{jt} = 1)) = \boldsymbol{\kappa}_{1t}' \boldsymbol{b}_j - \kappa_{2t}, \tag{7}$$

where

$$\boldsymbol{\kappa}_{1t} = 2(\boldsymbol{\alpha}_{1t} - \boldsymbol{\alpha}_{0t}) \quad \text{and} \quad \kappa_{2t} = (\boldsymbol{\alpha}_{1t}' \boldsymbol{\alpha}_{1t} - \boldsymbol{\alpha}_{0t}' \boldsymbol{\alpha}_{0t}) \tag{8}$$

Clinton et al. (2004) consider cases where both the number of legislators and and the number of roll-call votes are large (i.e. if several sessions of the US Congress are considered). For both theoretical as practical reasons, they use a Bayesian approach at recovering the positions $\boldsymbol{b}_j$ and the parameters $\boldsymbol{\kappa}_{1t}$ and κ_{2t} and hence $\boldsymbol{\alpha}_{0t}$ and $\boldsymbol{\alpha}_{1t}$. For both the positions of the legislators $\boldsymbol{b}_j$ and the transformed location parameters $\boldsymbol{\kappa}_{1t}$ and κ_{2t} they use a multivariate distribution as a prior, a relatively flat prior for $\boldsymbol{\kappa}_{1t}$ and κ_{2t} (with a variance value of 25) and a somewhat informative prior for the positions of the legislators $\boldsymbol{b}_j$ (with a variance equal to unity for most legislators except for those whose positions they need to fix for purposes of model identification). It is important to state here how Clinton et al.'s Bayesian approach differs from the marginal likelihood/empirical Bayes treatment of the positions of political actors proposed in this article. In Clinton et al. (2004) the parameters of the prior distribution of the are fixed in advanced to pre-determined values, while in the context of the approach taken in this article, the parameters of the prior distribution of the positions are estimated from the observed data.

Martin and Quinn (2002) extend Clinton et al.'s (2004) approach by a dynamic component in the prior distribution of the positions of the actors, in the case of the application of their model, to the positions of judges in the US Supreme Court. Consequently, the ideal points $\boldsymbol{b}_{jt}$ gain a time index as well. While Clinton et al. (2004) consider an informative prior for the positions

that nevertheless rests on the assumption of serial independence, the prior distribution used by Martin and Quinn (2002) takes into account that the positions of the same actors at different times may be serially correlated and evolve over time in a manner described by

$$\boldsymbol{b}_{jt} = \boldsymbol{\Gamma}\boldsymbol{b}_{jt-1} + \boldsymbol{u}_{jt} \tag{9}$$

where $\boldsymbol{\Gamma}$ is a matrix of auto-regression coefficients, which is the same as equation (5) in the article, with a prior distribution of the initial distribution as described by equation (4) in the article. Martin and Quinn assume the innovations $\boldsymbol{u}_{jt}$ of the ideal points to have a normal distribution and also treat the locations $\boldsymbol{a}_{it}$ as unobserved data with a normal distribution. However, in contrast to the marginal likelihood/empirical Bayes approach taken in this article, Martin and Quinn have to fix the values of the variance of the initial positions $\boldsymbol{\Sigma}_0$ and the variance of the positional innovations $\boldsymbol{\Sigma}_1$ to specific values in order to make their Bayesian analysis of the development of Justices' political positions feasible.

The models of Poole and Rosenthal (1985) and Martin and Quinn (2002) exhibit some of the features a measurement model should have that allows the reconstruction of political positions from coded political texts: (1) Unlike in the index construction of the RiLe index by the principal investigators of the Comparative Manifesto Project and their variations proposed by Laver and Garry (2000) and Lowe et al. (2011), the policy options are not simply categorized as "left" and "right", but may have varying locations in the political space. (2) The models of Poole and Rosenthal (1985) and Martin and Quinn (2002) specify a plausible link between locations in an Euclidean space and observed data that non-metric. (3) Martin and Quinn (2002) specify a dynamic state-space model for the ideal points. Nevertheless, these models cannot be used for reconstruction of positions of political actors from coded political texts without modifications.

On the one hand, in models from the Nominate lineage the observations there are always only two relevant locations of policy options, one for the Yea alternative and one for the Nay alternative. In political texts there are often numerous categories that are relevant for the reconstruction of political positions, which may make this task in this respect more complicated. On the other hand, while roll-call votes typically deal with a different topic at each point in time, categories used for coding political texts are constant over time, however fine-grained they are. That is, the task of reconstructing political positions becomes simpler in another respect. Finally, since the observations are not just binary choices but counts (or percentages obtained from these counts), each observation is able to provide much more information about political positions than in the case of roll-call votes. This is why the model proposed in the article does not have the restriction to a single dimension as the original Nominate model and why it is also possible to estimate the variance parameters in $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ from empirical data instead of setting them to fixed values as in Martin and Quinn (2002).

B Details on the Estimation Method

Below various aspects of the algorithm to obtain estimates of the dynamic state-space model to obtain predictions of political positions based on an estimated model are presented. These methods are implemented in *R* (R Development Core Team 2011) with some extensions in *C++* with the help of the *Rcpp* and *RcppArmadillo* packages (Eddelbuettel and François 2011; Francois et al. 2011). The software implementing the methods is available for download on the *Political Analysis* dataverse (Elff 2012) and on the author's Web site.

B.1 Marginal Maximum Likelihood and Expectation-Maximization

In the main article, a dynamic state-space model was introduced that considers as estimable parameters only its temporally invariant aspects, such as the locations of political objectives and the parameters of the *distribution* of the actors' positions, whereas the positions themselves were considered as unobserved data. The advantage of this is that the number of parameters stays constant, even if more data becomes available, and that positions of political actors can even be predicted out of the sample of political texts.

This approach means that the estimation of the model parameters becomes a missing data problem (Little and Rubin 2002). To obtain maximum likelihood estimates of the parameters of a model with missing data, one needs to construct the likelihood in a way that it does not depend on the unobserved data. This marginal likelihood is constructed by "integrating out" the unobserved data from the complete data likelihood (that depends on both observed and unobserved data), which in the present case takes the form:

$$\mathcal{L}_{\text{cpl}}(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\psi}) = \prod_{j=0}^{J} \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) = \prod_{j=0}^{J} \mathcal{L}(\boldsymbol{b}_j; \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1) \prod_{t=0}^{T_j} \mathcal{L}(\boldsymbol{y}_{jt} | \boldsymbol{b}_{jt}; \boldsymbol{\alpha}) \tag{10}$$

where the vector $\boldsymbol{y}_j$ is constructed by stacking all the observed data points $\boldsymbol{y}_{jt}$ for party j, $\boldsymbol{b}_j$ is constructed analogously from the unobserved data points $\boldsymbol{b}_{jt}$, and the vector $\boldsymbol{\psi}$ is constructed from all free parameters in the model. The corresponding log-likelihood than is:

$$\ell_{\text{cpl}}(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\psi}) = \sum_{j=0}^{J} \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) = \sum_{j=0}^{J} \ell(\boldsymbol{b}_j; \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_1) + \sum_{t=0}^{T_j} \ell(\boldsymbol{y}_{jt} | \boldsymbol{b}_{jt}; \boldsymbol{\alpha}) \tag{11}$$

where $\ell_{\text{cpl}}(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\psi}) = \ln \mathcal{L}_{\text{cpl}}(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\psi})$, $\ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) = \ln \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})$, etc.

As was explained in the main text, the counts of policy emphases are assumed to follow a multinomial distribution, conditional on the policy positions of the electoral platforms, which

is further specified by the locations of the policy objectives as estimable parameters. This conditional log-likelihood takes the form (where $\Delta_{ijt} = ||\boldsymbol{\alpha}_i - \boldsymbol{b}_{jt}||$):

$$\begin{aligned}
\ell(\boldsymbol{y}|\boldsymbol{b};\boldsymbol{\alpha}) &= \sum_{j=1}^{J} \ell(\boldsymbol{y}_j|\boldsymbol{b}_j;\boldsymbol{\alpha}) \\
&= \sum_{j} c(\boldsymbol{y}_j) + \sum_{j=1}^{J}\sum_{t=0}^{T_j}\sum_{i} m_{ijt} \ln \pi_{ijt} \\
&= \sum_{j} c(\boldsymbol{y}_j) + \sum_{j=1}^{J}\sum_{t=0}^{T_j}\sum_{i} m_{ijt} \frac{\exp\left(-\frac{1}{2}\Delta_{ijt}^2\right)}{\sum_h \exp\left(-\frac{1}{2}\Delta_{hjt}^2\right)} \\
&= \sum_{j} c(\boldsymbol{y}_j) - \sum_{j=1}^{J}\sum_{t=0}^{T_j}\sum_{i} m_{ijt} \frac{1}{2}(\boldsymbol{\alpha}_i - \boldsymbol{b}_{jt})'(\boldsymbol{\alpha}_i - \boldsymbol{b}_{jt}) \\
&\qquad - \sum_{j=1}^{J}\sum_{t=0}^{T_j} \ln\left(\sum_h \exp\left(-\frac{1}{2}(\boldsymbol{\alpha}_i - \boldsymbol{b}_{jt})'(\boldsymbol{\alpha}_i - \boldsymbol{b}_{jt})\right)\right).
\end{aligned} \tag{12}$$

The assumptions about the dynamics positions of a party j at times $t = 0, \ldots, T_j$ lead to a multivariate normal distribution with a density the logarithm of which is:

$$\begin{aligned}
\ell(\boldsymbol{b};\boldsymbol{\beta},\boldsymbol{\Gamma},\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}_1) &= \sum_{j=1}^{J} \ell(\boldsymbol{b}_j;\boldsymbol{\beta},\boldsymbol{\Gamma},\boldsymbol{\Sigma}_0,\boldsymbol{\Sigma}_1) \\
&= -\frac{D}{2}\ln(2\pi)\sum_{j=1}^{J}(T_j+1) - \frac{J}{2}\ln|\boldsymbol{\Sigma}_0| - \sum_{j=1}^{J}\frac{T_j}{2}\ln|\boldsymbol{\Sigma}_1| \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}(\boldsymbol{b}_{j0} - \boldsymbol{\beta})'\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{b}_{j0} - \boldsymbol{\beta}) \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}\sum_{t=1}^{T_j}\boldsymbol{\beta}'(\boldsymbol{I} - \boldsymbol{\Gamma})'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{I} - \boldsymbol{\Gamma})\boldsymbol{\beta} \\
&\quad - \sum_{j=1}^{J}\sum_{t=1}^{T_j}\boldsymbol{\beta}'(\boldsymbol{I} - \boldsymbol{\Gamma})'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{b}_{jt} - \boldsymbol{\Gamma}\boldsymbol{b}_{jt-1}) \\
&\quad - \frac{1}{2}\sum_{j=1}^{J}\sum_{t=1}^{T_j}(\boldsymbol{b}_{jt} - \boldsymbol{\Gamma}\boldsymbol{b}_{jt-1})'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{b}_{jt} - \boldsymbol{\Gamma}\boldsymbol{b}_{jt-1})
\end{aligned} \tag{13}$$

There are a couple of constraints imposed on the parameters of the model. The matrices $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ have to be symmetric positive definite in order to qualify as variance-covariance matrices. Further, the coordinates of the locations of the policy objectives sum to zero along each axis, and some coordinates may be fixed to zero. Thus the elements in $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ can be conceived

as re-parameterized such that they are functions of an (unrestricted) parameter vector $\boldsymbol{\psi}$. As a function of $\boldsymbol{\psi}$, the complete-data log-likelihood takes the form

$$\begin{aligned}\ell_{\text{cpl}}(\boldsymbol{y}, \boldsymbol{b}; \boldsymbol{\psi}) &= \sum_{j=1}^{J} \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) \\ &= \sum_{j=1}^{J} \ell(\boldsymbol{y}_j | \boldsymbol{b}_j; \boldsymbol{\alpha}(\boldsymbol{\psi})) + \sum_{j=1}^{J} \ell(\boldsymbol{b}_j; \boldsymbol{\beta}(\boldsymbol{\psi}), \boldsymbol{\Gamma}(\boldsymbol{\psi}), \boldsymbol{\Sigma}_1(\boldsymbol{\psi}), \boldsymbol{\Sigma}_1(\boldsymbol{\psi})).\end{aligned} \tag{14}$$

Inferences can of course only be drawn based on data actually observed. They have to be based on the marginal likelihood or on the marginal log-likelihood

$$\begin{aligned}\ell_{\text{obs}}(\boldsymbol{y}; \boldsymbol{\psi}) &= \sum_{j=1}^{J} \ell_{\text{obs}}(\boldsymbol{y}_j; \boldsymbol{\psi}) = \sum_{j=1}^{J} \ln \mathcal{L}_{\text{obs}}(\boldsymbol{y}_j; \boldsymbol{\psi}) \\ &= \sum_{j=1}^{J} \ln \int_{\mathbb{R}^{S_j}} \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) \, \mathrm{d}\, \boldsymbol{b}_j \\ &= \sum_{j=1}^{J} \ln \int_{\mathbb{R}^{S_j}} \exp\left(\ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) \, \mathrm{d}\, \boldsymbol{b}_j\right)\end{aligned} \tag{15}$$

where the integral ranges over the $S_j := D(T_j + 1)$-dimensional space $\mathbb{R}^{S_j}$.

The gradient of the marginal or *observed-data* log-likelihood takes the form of a conditional expectation given the observed data:

$$\begin{aligned}\frac{\partial \ell_{\text{obs}}(\boldsymbol{y}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} &= \sum_{j=1}^{J} \frac{\partial \ell_{\text{obs}}(\boldsymbol{y}_j; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} = \sum_{j=1}^{J} \frac{\partial}{\partial \boldsymbol{\psi}} \ln \int_{\mathbb{R}^{S_j}} \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) \, \mathrm{d}\, \boldsymbol{b}_j \\ &= \sum_{j=1}^{J} \int_{\mathbb{R}^{S_j}} \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\mathcal{L}_{\text{obs}}(\boldsymbol{y}_j; \boldsymbol{\psi})} \frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \, \mathrm{d}\, \boldsymbol{b}_j \\ &= \sum_{j=1}^{J} \mathrm{E}_{\boldsymbol{\psi}}\left(\frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \middle| \boldsymbol{y}_j \right)\end{aligned} \tag{16}$$

This gradient can be interpreted as a posterior expectation of a function of $\boldsymbol{b}_j$ with the distribution of the positions $\boldsymbol{b}_j$ as a prior, because

$$\int_{\mathbb{R}^{S_j}} \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\mathcal{L}_{\text{obs}}(\boldsymbol{y}_j; \boldsymbol{\psi})} \frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \, \mathrm{d}\, \boldsymbol{b}_j = \frac{\int_{\mathbb{R}^{S_j}} f(\boldsymbol{y}_j | \boldsymbol{b}_j; \boldsymbol{\psi}) f(\boldsymbol{b}_j; \boldsymbol{\psi}) \frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \, \mathrm{d}\, \boldsymbol{b}_j}{\int_{\mathbb{R}^{S_j}} f(\boldsymbol{y}_j | \boldsymbol{b}_j; \boldsymbol{\psi}) f(\boldsymbol{b}_j; \boldsymbol{\psi}), \mathrm{d}\, \boldsymbol{b}_j} \tag{17}$$

where

$$f(\boldsymbol{y}_j | \boldsymbol{b}_j; \boldsymbol{\psi}) = \exp(\ell(\boldsymbol{y}_j | \boldsymbol{b}_j; \boldsymbol{\alpha}(\boldsymbol{\psi}))) \tag{18}$$

is the multinomial-distribution likelihood conditional on $\boldsymbol{b}_j$ and

$$f(\boldsymbol{b}_j; \boldsymbol{\psi}) = \exp(\ell(\boldsymbol{b}_j; \boldsymbol{\beta}(\boldsymbol{\psi}), \boldsymbol{\Gamma}(\boldsymbol{\psi}), \boldsymbol{\Sigma}_1(\boldsymbol{\psi}), \boldsymbol{\Sigma}_1(\boldsymbol{\psi}))) \tag{19}$$

is the prior distribution of $\boldsymbol{b}_j$.

The *observed information*, the negative of the Hessian matrix of the observed-data log-likelihood takes the form of the difference between two matrices, the conditional expectation of the complete-data information matrix and the conditional covariance matrix of the gradient, where both expectations are formed conditional on the observed data:

$$\begin{aligned}
\mathcal{I}_{\text{obs}}(\boldsymbol{\psi}) &= -\frac{\partial^2 \ell_{\text{obs}}(\boldsymbol{y};\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'} \\
&= -\sum_{j=1}^{J}\int_{\mathbb{R}^{s_j}} \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\mathcal{L}_{\text{obs}}(\boldsymbol{y}_j;\boldsymbol{\psi})}\frac{\partial^2 \ell_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\,\mathrm{d}\,\boldsymbol{b}_j \\
&\quad -\sum_{j=1}^{J}\int_{\mathbb{R}^{s_j}} \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\mathcal{L}_{\text{obs}}(\boldsymbol{y}_j;\boldsymbol{\psi})}\left(\frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right)\left(\frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right)'\mathrm{d}\,\boldsymbol{b}_j \\
&\quad +\sum_{j=1}^{J}\left(\frac{\partial \ell_{\text{obs}}(\boldsymbol{y}_j;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right)\left(\frac{\partial \ell_{\text{obs}}(\boldsymbol{y}_j;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}}\right)' \\
&= \sum_{j=1}^{J}\mathrm{E}_{\boldsymbol{\psi}}\left(-\frac{\partial^2 \ell_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'}\middle|\,\boldsymbol{y}_j\right) \\
&\quad -\sum_{j=1}^{J}\mathrm{Cov}_{\boldsymbol{\psi}}\left(\frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}},\frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\partial\boldsymbol{\psi}'}\middle|\,\boldsymbol{y}_j\right) \\
&= \mathrm{E}_{\boldsymbol{\psi}}(\mathcal{I}_{\text{cpl}}(\boldsymbol{\psi})|\boldsymbol{y}_j) - \mathcal{I}_{\text{mis}}(\boldsymbol{\psi})
\end{aligned} \tag{20}$$

where $\mathcal{I}_{\text{cpl}}(\boldsymbol{\psi})$ usually is referred to as the *complete-data information* and $\mathcal{I}_{\text{mis}}(\boldsymbol{\psi})$ as the *missing information* (Little and Rubin 2002: 177).

The usual method for computing maximum-likelihood estimates of models involving missing data is an EM algorithm (Dempster et al. 1977; Little and Rubin 2002). Such an EM algorithm is iterative and involves steps of the following form: Let $\boldsymbol{\psi}^{(s)}$ be the current approximation of the maximum likelihood estimate at iteration s of the algorithm, then the following expectation is computed based on $\boldsymbol{\psi}^{(s)}$ and conditional the observed data $\boldsymbol{y}_j$ – this is the "E-step":

$$Q(\boldsymbol{\psi};\boldsymbol{\psi}^{(s)}) = \sum_j \mathrm{E}\left(\ell_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})\middle|\,\boldsymbol{y}_j;\boldsymbol{\psi}^{(s)}\right) = \sum_j \frac{\int \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi}^{(s)})\ell_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})\,\mathrm{d}\,\boldsymbol{b}_j}{\int \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi}^{(s)})\,\mathrm{d}\,\boldsymbol{b}_j} \tag{21}$$

Holding $\boldsymbol{\psi}^{(s)}$ fixed, the updated approximation of the maximum likelihood estimate $\boldsymbol{\psi}^{(s+1)}$ is identified as the value of $\boldsymbol{\psi}$ that maximizes $Q(\boldsymbol{\psi};\boldsymbol{\psi}^{(s)})$. This is the "M-step". The EM algorithm will repeat these steps until some criterion of convergence is attained, for example that $||\boldsymbol{\psi}^{(s+1)} - \boldsymbol{\psi}^{(s)}|| < \epsilon$ or $||Q(\boldsymbol{\psi}^{(s+1)};\boldsymbol{\psi}^{(s)}) - Q(\boldsymbol{\psi}^{(s)};\boldsymbol{\psi}^{(s)})|| < \epsilon$ for a small positive number ϵ, e.g., $\epsilon = 10^{-7}$.

This maximization step will itself be conducted iteratively, so that the structure of such an EM algorithm is that of an iterative procedure nested in a iterative procedure. These inner iterations involve the gradient of the Q-function

$$\frac{\partial Q(\boldsymbol{\psi};\boldsymbol{\psi}^{(s)})}{\partial \boldsymbol{\psi}} = \sum_j \mathrm{E}\left(\frac{\partial \ell_{\mathrm{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \middle| \boldsymbol{y}_j; \boldsymbol{\psi}^{(s)} \right) \tag{22}$$

and the Hessian

$$\frac{\partial^2 Q(\boldsymbol{\psi};\boldsymbol{\psi}^{(s)})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} = \sum_j \mathrm{E}\left(\frac{\partial^2 \ell_{\mathrm{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \middle| \boldsymbol{y}_j; \boldsymbol{\psi}^{(s)} \right) \tag{23}$$

It is easy to see that the gradient of the observed-data log-likelihood and the gradient of the Q-function coincide at $\boldsymbol{\psi} = \hat{\boldsymbol{\psi}} = \boldsymbol{\psi}^{(\infty)}$, which is the main motivation of the EM algorithm. However, the Hessian of the observed-data log-likelihood and the Hessian of the Q-function are different, with the former being more complicated and, but more importantly, involving a difference between two matrices. So while one could argue that the availability of a Hessian allows for a direct Newton-Raphson algorithm, such an algorithm can be expected to be unstable, because the Hessian can become positive semi-definite, which will lead to a divergence of the algorithm. In contrast, if appropriately constructed, EM-steps will always lead to an increase in the observed-data log-likelihood function (Dempster et al. 1977; Little and Rubin 2002: 172).

B.2 Monte Carlo Methods

The marginal likelihoods, log-likelihoods and their derivatives involve integrals that are multidimensional and do not have a analytical solution. Therefore, they need to be approximated, using quadrature or Monte Carlo techniques. Whatever technique one uses, the approximation of such an integral takes the form

$$\int_{\mathbb{R}^{s_j}} \mathcal{L}_{\mathrm{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) g(\boldsymbol{b}_j) \,\mathrm{d}\, \boldsymbol{b}_j \approx \sum_{r=1}^{R} w_j^{*(r)} g(\boldsymbol{b}_j^{*(r)}) \tag{24}$$

where $\boldsymbol{b}_j^{*(r)}$ are quadrature points or random vectors and $w_j^{*(r)}$ are weights specific for the numerical integration technique.

If one uses a Monte Carlo approximation with samples from a distribution with a density proportional to $\mathcal{L}_{\mathrm{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})$ then the weights would simply be $w_j^{*(r)} = K_j/R$, where K_j is the normalizing constant $K_j = \int_{\mathbb{R}^{s_j}} \mathcal{L}_{\mathrm{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) \,\mathrm{d}\, \boldsymbol{b}_j$. Since such samples are not readily available, the article uses importance sampling, where the random vectors $\boldsymbol{b}_j^{*(r)}$ are generated from a distribution that is similar in shape to $\mathcal{L}_{\mathrm{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})$ (Robert and Casella 2004: 92).

The first step to find determine an importance sampling strategy is a change of variables of integration: Let $\tilde{\boldsymbol{b}}_j$ be the value of $\boldsymbol{b}_j$ that maximizes $\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})$ and also its logarithm $\ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})$. Let further

$$\tilde{\boldsymbol{K}}_j = - \left. \frac{\partial^2 \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{b}_j \partial \boldsymbol{b}_j'} \right|_{\boldsymbol{b}_j = \tilde{\boldsymbol{b}}_j} \tag{25}$$

the negative Hessian of the complete-data log-likelihood for $\boldsymbol{b}_j$ evaluated at $\boldsymbol{b}_j = \tilde{\boldsymbol{b}}_j$. Now using the substitution $\boldsymbol{b}_j = \tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j \Leftrightarrow \tilde{\boldsymbol{K}}_j^{\frac{1}{2}} (\boldsymbol{b}_j - \tilde{\boldsymbol{b}}_j) = \boldsymbol{v}_j$, where $\tilde{\boldsymbol{K}}_j^{-\frac{1}{2}}$ is the inverse of the Cholesky factor of $\tilde{\boldsymbol{K}}_j$, the first step is completed by the change of variables:

$$\int_{\mathbb{R}^{S_j}} \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) g(\boldsymbol{b}_j) \, \mathrm{d}\, \boldsymbol{b}_j = \int_{\mathbb{R}^{S_j}} \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j; \boldsymbol{\psi}) g(\tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j) |\tilde{\boldsymbol{K}}_j|^{-\frac{1}{2}} \, \mathrm{d}\, \boldsymbol{v}_j \tag{26}$$

The next step is to make use of the "importance sampling fundamental identity" (Robert and Casella 2004: 92) to determine the Monte Carlo approximation:

$$\begin{aligned} \int_{\mathbb{R}^{S_j}} \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) g(\boldsymbol{b}_j) \, \mathrm{d}\, \boldsymbol{b}_j &= \int_{\mathbb{R}^{S_j}} \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j; \boldsymbol{\psi}) g(\tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j)}{|\tilde{\boldsymbol{K}}_j|^{\frac{1}{2}} f_{\text{pro}}(\boldsymbol{v}_j)} f_{\text{pro}}(\boldsymbol{v}_j) \, \mathrm{d}\, \boldsymbol{v}_j \\ &\approx \sum_{r=1}^{R} \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j^{*(r)}; \boldsymbol{\psi})}{R |\tilde{\boldsymbol{K}}_j|^{\frac{1}{2}} f_{\text{pro}}(\boldsymbol{v}_j^{*(r)})} g(\tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j^{*(r)}) \end{aligned} \tag{27}$$

so that in equation (24)

$$\boldsymbol{b}_j^{*(r)} = \tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j^{*(r)} \quad \text{and} \quad w_j^{*(r)} = \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}} \boldsymbol{v}_j^{*(r)}; \boldsymbol{\psi})}{R |\tilde{\boldsymbol{K}}_j|^{\frac{1}{2}} f_{\text{pro}}(\boldsymbol{v}_j^{*(r)})}$$

where $f_{\text{pro}}()$ is the density function of the distribution from which the random vectors $\boldsymbol{v}_j^{*(r)}$ are sampled. This *proposal distribution* should be chosen such that its tails are heavy enough so that the importance weights $w_j^{*(r)}$ are bounded and the Law of Large Numbers applies to the approximation (24). Following Booth and Hobert (1999), the article employs a standard multivariate Student distribution with $F_j = 7 \cdot S_j$ degrees of freedom, where $S_j = D(T_j + 1)$. That is, the resulting random vectors $\boldsymbol{b}_j^{*(r)}$ come from multivariate Student distribution with mean $\tilde{\boldsymbol{b}}_j$ and scale matrix $\tilde{\boldsymbol{K}}_j^{-1}$.

The computation of maximum likelihood estimates and of test statistics requires expectations of some functions $g(\boldsymbol{b}_j)$ of unobserved data (the parties' political positions) conditional on observed data $\boldsymbol{y}_j$ (the emphases of policy objectives in political texts, such as electoral manifestos). Using the methods explained in this section these conditional expectations are approximated as

$$\mathrm{E}(g(\boldsymbol{b}_j) | \boldsymbol{y}_j) = \frac{\int_{\mathbb{R}^{S_j}} \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) g(\boldsymbol{b}_j) \, \mathrm{d}\, \boldsymbol{b}_j}{\int_{\mathbb{R}^{S_j}} \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) \, \mathrm{d}\, \boldsymbol{b}_j} \approx \frac{\sum_{r=1}^{R} w_j^{*(r)} g(\boldsymbol{b}_j^{*(r)})}{\sum_{r=1}^{R} w_j^{*(r)}} \tag{28}$$

While importance sampling may be suitable to approximate marginal likelihoods and conditional expectations, this method is less suitable to obtain quantiles of the posterior distribution that one may want for the construction of prediction intervals. For this, one needs to generate unweighted samples from the posterior distribution of the positions

$$f(\boldsymbol{b}_j|\boldsymbol{y}_j;\boldsymbol{\psi}) = \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})}{\int \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\boldsymbol{\psi})\,\mathrm{d}\,\boldsymbol{b}_j}. \tag{29}$$

For this an accept-reject sampling method (Robert and Casella 2004: 41) is better suited, because then importance weights are not needed.

This accept-reject sampling method rests, firstly, on the fact that the posterior distribution is proportional to $\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\hat{\boldsymbol{\psi}})$ and, secondly, on the simple identity

$$\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\hat{\boldsymbol{\psi}}) = \mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\boldsymbol{b}_j;\hat{\boldsymbol{\psi}})\frac{f_{\text{pro}}(\tilde{\boldsymbol{K}}_j^{\frac{1}{2}}(\boldsymbol{b}_j-\tilde{\boldsymbol{b}}_j))}{f_{\text{pro}}(\tilde{\boldsymbol{K}}_j^{\frac{1}{2}}(\boldsymbol{b}_j-\tilde{\boldsymbol{b}}_j))}. \tag{30}$$

Thus to obtain random vectors $\boldsymbol{b}_j^{**}$ from the posterior distribution one generates random vectors $\boldsymbol{v}_j^*$ from a proposal distribution with density $f_{\text{pro}}(\boldsymbol{v}_j)$, and accepts each candidate random vector $\boldsymbol{b}_j^* = \tilde{\boldsymbol{b}}_j + \tilde{\boldsymbol{K}}_j^{-\frac{1}{2}}\boldsymbol{v}_j^*$ to the sample with probability

$$p_j^* = \frac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\tilde{\boldsymbol{b}}_j+\tilde{\boldsymbol{K}}_j^{-\frac{1}{2}}\boldsymbol{v}_j^*;\hat{\boldsymbol{\psi}})}{C_j f_{\text{pro}}(\boldsymbol{v}_j^*)} \tag{31}$$

and rejects it with probability $1-p_j^*$, where C_j is a number chosen such that $p_j^* \leq 1$ is assured. This accept/reject step is achieved by generating a random number u_j^* from the uniform distribution to accompany each candidate vector. The candidate vector is then accepted if $u_j^* \leq p_j^*$ and rejected if $u_j^* > p_j^*$ (Robert and Casella 2004: 40).

The literature on accept-reject sampling does not provide a simple method to choose the normalizing constant C_j. So in the implementation of this algorithm for the article it is chosen adaptively, starting with 1.5 times the maximum value of $\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j,\tilde{\boldsymbol{b}}_j+\tilde{\boldsymbol{K}}_j^{-\frac{1}{2}}\boldsymbol{v}_j^*;\hat{\boldsymbol{\psi}})/f_{\text{pro}}(\boldsymbol{v}_j^*)$ from an initial sample. Whenever one of the values of p_j^* (call it $p_j^\dagger$) happens to be larger than one, the accept-reject sampling process is restarted with a new C_j equal to $1.5p_j^\dagger$ times the old C_j.

B.3 The MCEM Algorithm

Like the marginal log-likelihood and its derivatives the Q-function and its derivatives involve multidimensional analytically intractable integrals. Fortunately, the Monte Carlo approximation techniques discussed in the previous section of this appendix are applicable to the Q-function

and its derivatives as well. The Monte Carlo approximation $Q^*(\boldsymbol{\psi}; \boldsymbol{\psi}^{(s)})$ of the Q-function $Q(\boldsymbol{\psi}; \boldsymbol{\psi}^{(s)})$ takes the form

$$Q^*(\boldsymbol{\psi}; \boldsymbol{\psi}^{(s)}) = \frac{\sum_{r=1}^{R} w_j^{*(r;s)} \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j^{*(r;s)}; \boldsymbol{\psi})}{\sum_{r=1}^{R} w_j^{*(r;s)}} = \sum_{r=1}^{R} \tilde{w}_j^{(r;s)} \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j^{*(r;s)}; \boldsymbol{\psi}). \tag{32}$$

with $\tilde{w}_j^{(r;s)} = w_j^{*(r;s)} / \sum_{r=1}^{R} w_j^{*(r;s)}$. The gradient and Hessian also take form of weighted sums:

$$\begin{aligned} \frac{\partial Q^*(\boldsymbol{\psi}; \boldsymbol{\psi}^{(s)})}{\partial \boldsymbol{\psi}} &= \sum_j \sum_{r=1}^{R} \tilde{w}_j^{(r;s)} \frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j^{*(r;s)}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \\ \frac{\partial^2 Q^*(\boldsymbol{\psi}; \boldsymbol{\psi}^{(s)})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} &= \sum_j \sum_{r=1}^{R} \tilde{w}_j^{(r;s)} \frac{\partial^2 \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j^{*(r;s)}; \boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'} \end{aligned} \tag{33}$$

Consequently, the M-step of the MCEM algorithm consists in maximizing a complete data log-likelihood, where the missing data are "filled in" and weighed based on importance sampling. The complete data log-likelihood and its derivatives have a relatively simple structure, so that the M-step can be conducted using a more or less straightforward Newton-Raphson (e.g. Agresti 2002: 143) or Fisher-Scoring (e.g. Agresti 2002: 145) algorithm.[3]

Crucial for the unbiasedness of estimates obtained from a MCEM algorithm is that the Monte Carlo error is not too large, that is, that the Monte Carlo Sample size is not too small. Earlier steps of the MCEM algorithm are relatively large so that a relatively small Monte Carlo sample size is sufficient in the earlier iterations. The more the algorithm advances, the smaller the step sizes become, and the larger is the danger that improvements are "swamped" by Monte Carlo error. It is therefore important to assess the Monte Carlo error at each step of the MCEM algorithm and to increase the sample size if necessary. Fortunately, the elements $\boldsymbol{b}_j^{*(r;s)}$ of the Monte Carlo samples are independent and identically distributed, so that the assessment of the Monte Carlo error is relatively straightforward. The method to assess the Monte Carlo error used in the article follows Caffo et al. (2005).[4]

Following Caffo et al. (2005) the MCEM algorithm employed in the article starts with a small Monte Carlo sample size, which is automatically increased if the improvement of approximate the Q-function achieved in step s of the EM algorithm, denoted by $\Lambda^{(s)} = Q^*(\boldsymbol{\psi}^{(s+1)}; \boldsymbol{\psi}^{(s)}) - Q^*(\boldsymbol{\psi}^{(s)}; \boldsymbol{\psi}^{(s)})$, is "statistically insignificant". That is, if the ratio $z^{(s)} = \Lambda^{(s)} / SE(\Lambda^{(s)})$ is smaller than the $1 - \alpha/2$ quantile of the standard normal distribution (where the conventional significance level $\alpha = 0.05$ is applied here) then the new sample size is $R_{\text{new}} = (3/2)R_{\text{old}}$.

[3] The Hessian for the policy objective location parameter vector $\boldsymbol{\alpha}$ involves a difference between matrices again. Therefore a simplified version of the Hessian is used (its expectation under the current estimates of $\boldsymbol{\psi}$), which leads to a Fisher-Scoring algorithm. Had the model been linear in $\boldsymbol{\alpha}$ as in generalized linear models with canonical link, Newton-Raphson and Fisher Scoring algorithms would have been identical (McCullagh and Nelder 1989: 42)

[4] For another approach at gauging the Monte Carlo error in MCEM algorithms, see Booth and Hobert (1999)

The MCEM algorithm proceeds as follows:

1. Starting values for $\boldsymbol{\psi}$ are obtained and the iteration counter is set $s \leftarrow 1$.

2. **While** the algorithm has not converged, the following steps iterated:

2.a For the current parameter vector $\boldsymbol{\psi}^{(s)}$, for each unit j R random vectors $\boldsymbol{b}_j^{*(s,1)}$, …, $\boldsymbol{b}_j^{*(s,R)}$ are generated together with corresponding weights $w_j^{*(s,1)}, \ldots, w_j^{*(s,R)}$, needed to compute approximations $Q^*(\boldsymbol{\psi}; \boldsymbol{\psi}^{(s)})$ of the Q-function and approximations of its derivatives. (This is the "E-step" of the MCEM algorithm)

2.b Based on the random vectors and weights computed in the previous step, the (approximate) Q-function is maximized for $\boldsymbol{\psi}$ and the maximizing value is accepted as updated estimate $\boldsymbol{\psi}^{(s+1)}$.

2.c $\Lambda^{(s)} = Q^*(\boldsymbol{\psi}^{(s+1)}; \boldsymbol{\psi}^{(s)}) - Q^*(\boldsymbol{\psi}^{(s)}; \boldsymbol{\psi}^{(s)})$ and its standard error $SE(\Lambda^{(s)})$ are computed.

If $\Lambda^{(s)}/SE(\Lambda^{(s)}) < z_{1-\alpha/2}$

(where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution), it is concluded that the improvement $\Lambda^{(s)}$ is not statistical significant. The Monte Carlo sample size adjusted $R \leftarrow R + R/2$ and the algorithm goes back to step *a*.

Else if $\Lambda^{(s)} < \epsilon$

(where ϵ is a small positive number, e.g. 10^{-7}) the MCEM algorithm is stopped and $\boldsymbol{\psi}^{(s+1)}$ is declared as the maximum likelihood estimate $\hat{\boldsymbol{\psi}}$.

2.d The algorithm has not yet converged and the next iteration is started with $s \leftarrow s+1$.

3. The MCEM agorithm has converged, the random vectors $\boldsymbol{b}_j^{*(s,1)}, \ldots, \boldsymbol{b}_j^{*(s,R)}$ with corresponding weights $w_j^{*(s,1)}, \ldots, w_j^{*(s,R)}$ are used to compute the observed-data Hessian $\frac{\partial^2 \ell_{\text{obs}}(\boldsymbol{y};\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}'}$, and the standard errors for the estimate $\hat{\boldsymbol{\psi}}$.

The generation of the random vectors and the computation importance weights is not completely trivial and will be therefore described in the following. Suppose the MCEM algorithm is in iteration s, then the generation of the random vectors $\boldsymbol{b}_j^{*(r;s)}$ and importance weights to approximate the Q-function at this iteration takes the following steps:

For each unit (i.e. party) j do:

1. For the current value $\boldsymbol{\psi}^{(s)}$ of the parameter vector, compute the value $\tilde{\boldsymbol{b}}_j^{(s)}$ of the latent data vector that maximizes the complete data log-likelihood and compute the curvature (matrix of second derivatives) $\tilde{\boldsymbol{K}}_{j,s}$ of the complete-data log-likelihood at this maximum.

2. Generate R $(T_j + 1)D$-dimensional random vectors $\boldsymbol{v}_j^{*(1)}, \ldots, \boldsymbol{v}_j^{*(R)}$ from a standard multivariate t-distribution with F_j degrees of freedom.

3. Compute R random vectors $\boldsymbol{b}_j^{*(r;s)} = \tilde{\boldsymbol{K}}_{j,s}^{-\frac{1}{2}} \boldsymbol{v}_j^{*(r)} + \tilde{\boldsymbol{b}}_j^{(s)}$.

4. Compute R importance weights $w_{j,s,r} = \dfrac{\mathcal{L}_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j^{*(r;s)}; \boldsymbol{\psi}^{(s)})}{|\tilde{\boldsymbol{K}}_{j,s}| f_{\mathcal{T}}(\boldsymbol{v}_j^{*(r)})}$.

B.4 Details of the M-step

Note that the gradient of the complete-data log-likelihood with respect to $\boldsymbol{\alpha}$ can be written as

$$\begin{aligned}\frac{\partial \ell_{\text{cpl}}}{\partial \boldsymbol{\alpha}} &= \sum_j \sum_t \sum_i \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}} \sum_h \frac{\partial}{\partial \eta_{ijt}} m_{hjt} \ln \pi_{hjt} \\ &= \sum_j \sum_t \sum_i \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}} (m_{ijt} - n_{jt}\pi_{ijt}) \\ &= \sum_j \boldsymbol{X}'_{\boldsymbol{\alpha},j} \boldsymbol{r}_j = \boldsymbol{X}'_{\boldsymbol{\alpha}} \boldsymbol{r}\end{aligned} \tag{34}$$

where

$$\pi_{ijt} = \frac{\exp(\eta_{ijt})}{\sum_h \exp(\eta_{hjt})} \quad \text{and} \quad \eta_{ijt} = -\frac{1}{2}(\boldsymbol{\alpha}_i - \boldsymbol{b}_{jt})'(\boldsymbol{\alpha}_i - \boldsymbol{b}_{jt}), \tag{35}$$

$$\frac{\partial \ln \pi_{hjt}}{\partial \eta_{ijt}} = \frac{\partial}{\partial \eta_{ijt}} \left(\eta_{hjt} - \ln \sum_k \exp(\eta_{kjt}) \right) = \begin{cases} 1 - \pi_{ijt} & \text{if } h = i \\ -\pi_{ijt} & \text{if } h \neq i, \end{cases} \tag{36}$$

$$\frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}} = \begin{bmatrix} \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}_1} \\ \vdots \\ \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}_I} \end{bmatrix} \quad \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}_h} = \begin{cases} \boldsymbol{b}_{jt} - \boldsymbol{\alpha}_i & \text{if } h = i \\ \boldsymbol{0} & \text{if } h \neq i, \end{cases} \tag{37}$$

$$\boldsymbol{X}_{\boldsymbol{\alpha}} = \begin{bmatrix} \boldsymbol{X}_{\boldsymbol{\alpha},1} \\ \vdots \\ \boldsymbol{X}_{\boldsymbol{\alpha},J} \end{bmatrix} \quad \boldsymbol{X}_{\boldsymbol{\alpha},j} = \begin{bmatrix} \boldsymbol{X}_{\boldsymbol{\alpha},j0} \\ \vdots \\ \boldsymbol{X}_{\boldsymbol{\alpha},jT_j} \end{bmatrix} \quad \boldsymbol{X}_{\boldsymbol{\alpha},jt} = \begin{bmatrix} \frac{\partial \eta_{1jt}}{\partial \boldsymbol{\alpha}'} \\ \vdots \\ \frac{\partial \eta_{Ijt}}{\partial \boldsymbol{\alpha}'} \end{bmatrix} \tag{38}$$

$$\boldsymbol{r} = \begin{bmatrix} \boldsymbol{r}_1 \\ \vdots \\ \boldsymbol{r}_J \end{bmatrix} \quad \boldsymbol{r}_j = \begin{bmatrix} \boldsymbol{r}_{j0} \\ \vdots \\ \boldsymbol{r}_{jT_j} \end{bmatrix} \quad \boldsymbol{r}_{jt} = \begin{bmatrix} m_{1jt} - n_{jt}\pi_{1jt} \\ \vdots \\ m_{Ijt} - n_{jt}\pi_{Ijt} \end{bmatrix} \tag{39}$$

Further

$$\begin{aligned}\frac{\partial^2 \ell_{\text{cpl}}}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'} &= \frac{\partial}{\partial \boldsymbol{\alpha}'} \sum_j \sum_i \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}} (m_{ijt} - n_{jt}\pi_{ijt}) \\ &= \sum_j \sum_t \sum_i \frac{\partial^2 \eta_{ijt}}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}} (m_{ijt} - n_{jt}\pi_{ijt}) - \sum_j \sum_t n_{jt} \sum_i \sum_h \frac{\partial \pi_{hjt}}{\partial \eta_{ijt}} \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}} \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}}\end{aligned} \tag{40}$$

$$
\begin{aligned}
-\mathrm{E}\left(\frac{\partial^2 \ell_{\mathrm{cpl}}}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'}\right) &= \sum_j \sum_t n_{jt} \sum_i \sum_h \frac{\partial \pi_{hjt}}{\partial \eta_{ijt}} \frac{\partial \eta_{hjt}}{\partial \boldsymbol{\alpha}} \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}'} \\
&= \sum_j \sum_t n_{jt} \left(\sum_i \pi_{ijt} \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}} \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}} - \sum_i \sum_h \pi_{hjt} \pi_{ijt} \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}} \frac{\partial \eta_{ijt}}{\partial \boldsymbol{\alpha}'} \right) \\
&= \sum_j \sum_t \boldsymbol{X}'_{\boldsymbol{\alpha},jt} \boldsymbol{W}_{jt} \boldsymbol{X}_{\boldsymbol{\alpha},jt} = \sum_j \boldsymbol{X}'_{\boldsymbol{\alpha},j} \boldsymbol{W}_j \boldsymbol{X}_{\boldsymbol{\alpha},j} = \boldsymbol{X}'_{\boldsymbol{\alpha}} \boldsymbol{W} \boldsymbol{X}_{\boldsymbol{\alpha}}
\end{aligned} \tag{41}
$$

$$
\boldsymbol{W} = \begin{bmatrix} \boldsymbol{W}_1 & & \\ & \ddots & \\ & & \boldsymbol{W}_J \end{bmatrix} \quad \boldsymbol{W}_j = \begin{bmatrix} \boldsymbol{W}_{j0} & & \\ & \ddots & \\ & & \boldsymbol{W}_{jT_j} \end{bmatrix} \tag{42}
$$

$$
\boldsymbol{W}_{jt} = n_{jt} \begin{bmatrix} \pi_{1jt} & & \\ & \ddots & \\ & & \pi_{Ijt} \end{bmatrix} - n_{jt} \begin{bmatrix} \pi_{1jt}\pi_{1jt} & \dots & \pi_{1jt}\pi_{Ijt} \\ \vdots & & \vdots \\ \pi_{Ijt}\pi_{1jt} & \dots & \pi_{Ijt}\pi_{Ijt} \end{bmatrix} \tag{43}
$$

Earlier it was stated that maximizing the likelihood for $\boldsymbol{\alpha}$ subject to linear restrictions of the form $\boldsymbol{T}\boldsymbol{\alpha} = \boldsymbol{t}$ can be achieved by maximizing the likelihood for $\boldsymbol{\phi}$ with $\boldsymbol{\alpha} = \boldsymbol{r} + \boldsymbol{Q}\boldsymbol{\phi}$, where $\boldsymbol{r}$ and $\boldsymbol{Q}$ are constructed from $\boldsymbol{t}$ and $\boldsymbol{T}$. Note that

$$
\frac{\partial \boldsymbol{\alpha}'}{\partial \boldsymbol{\phi}} = \boldsymbol{Q}' \tag{44}
$$

and consequently

$$
\frac{\partial \ell_{\mathrm{cpl}}}{\partial \boldsymbol{\phi}} = \frac{\partial \boldsymbol{\alpha}'}{\partial \boldsymbol{\phi}} \frac{\partial \ell_{\mathrm{cpl}}}{\partial \boldsymbol{\alpha}} = \boldsymbol{Q}' \boldsymbol{X}'_{\boldsymbol{\alpha}} \boldsymbol{r} \tag{45}
$$

and

$$
-\mathrm{E}\left(\frac{\partial^2 \ell_{\mathrm{cpl}}}{\partial \boldsymbol{\phi} \partial \boldsymbol{\phi}'}\right) = -\frac{\partial \boldsymbol{\alpha}'}{\partial \boldsymbol{\phi}} \mathrm{E}\left(\frac{\partial^2 \ell_{\mathrm{cpl}}}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}'}\right) \frac{\partial \boldsymbol{\alpha}}{\partial \boldsymbol{\phi}'} = \boldsymbol{Q}' \boldsymbol{X}'_{\boldsymbol{\alpha}} \boldsymbol{W} \boldsymbol{X}_{\boldsymbol{\alpha}} \boldsymbol{Q} \tag{46}
$$

With $\boldsymbol{X}_{\boldsymbol{\phi}} = \boldsymbol{X}_{\boldsymbol{\alpha}} \boldsymbol{Q}$ the Fisher scoring iteration (with complete data) take the form:

$$
\boldsymbol{\phi}^{(s+1)} = \boldsymbol{\phi}^{(s)} + \left(\boldsymbol{X}'_{\boldsymbol{\phi}} \boldsymbol{W} \boldsymbol{X}_{\boldsymbol{\phi}}\right)^{-1} \boldsymbol{X}'_{\boldsymbol{\phi}} \boldsymbol{r} \tag{47}
$$

where $\boldsymbol{X}_{\boldsymbol{\phi}}$, $\boldsymbol{W}$, and $\boldsymbol{r}$ are evaluated at $\boldsymbol{\phi} = \boldsymbol{\phi}^{(s)}$.

The complete-data Fisher scoring algorithm can now easily adapted to the M-step of the EM-algorithm. Let $\boldsymbol{X}^{*(r;s)}_{\alpha,j}$ denote the matrix of derivatives $\boldsymbol{X}_{\alpha,j}$ evaluated at $\boldsymbol{b}_j = \boldsymbol{b}^{*(r;s)}_j$, where $\boldsymbol{b}^{*(r;s)}_j$ is the the r-th sampled value of the position vector in the s-th iteration of the EM algorithm, that is,

$$
\boldsymbol{X}^{*(r;s)}_{\boldsymbol{\alpha},j} = \boldsymbol{X}_{\boldsymbol{\alpha},j}\big|_{\boldsymbol{b}_j = \boldsymbol{b}^{*(r;s)}_j} \quad \text{and} \quad \boldsymbol{X}^{*(r;s)}_{\boldsymbol{\phi},j} = \boldsymbol{X}^{*(r;s)}_{\boldsymbol{\alpha},j} \boldsymbol{Q} \tag{48}
$$

Further, let

$$
\boldsymbol{r}^{*(r;s)}_j = \boldsymbol{r}_j\big|_{\boldsymbol{b}_j = \boldsymbol{b}^{*(r;s)}_j} \quad \text{and} \quad \tilde{\boldsymbol{r}}^{(r;s)}_j = \tilde{w}^{(r;s)}_j \boldsymbol{r}^{*(r;s)}_j \tag{49}
$$

where $\tilde{w}_j^{(r;s)}$ are the normalized importance weights introduced above, and let

$$\boldsymbol{W}_j^{*(r;s)} = \boldsymbol{W}_j\big|_{\boldsymbol{b}_j=\boldsymbol{b}_j^{*(r;s)}} \quad \text{and} \quad \tilde{\boldsymbol{W}}_j^{(r;s)} = \tilde{w}_j^{(r;s)}\boldsymbol{W}_j^{*(r;s)}. \tag{50}$$

Now with

$$\boldsymbol{X}_{\boldsymbol{\phi}}^{*(r;s)} = \begin{bmatrix} \boldsymbol{X}_{\boldsymbol{\phi},1}^{*(r;s)} \\ \vdots \\ \boldsymbol{X}_{\boldsymbol{\phi},J}^{*(r;s)} \end{bmatrix} \quad \tilde{\boldsymbol{W}}^{(r;s)} = \begin{bmatrix} \tilde{\boldsymbol{W}}_1^{(r;s)} & & \\ & \ddots & \\ & & \tilde{\boldsymbol{W}}_J^{(r;s)} \end{bmatrix} \quad \tilde{\boldsymbol{r}}^{(r;s)} = \begin{bmatrix} \tilde{\boldsymbol{r}}_1^{(r;s)} \\ \vdots \\ \tilde{\boldsymbol{r}}_J^{(r;s)} \end{bmatrix} \tag{51}$$

the gradient and the expected information matrix with respect to $\boldsymbol{\phi}$ of the Monte Carlo approximation to the Q-function take the form

$$\frac{\partial Q^*(\boldsymbol{\psi};\boldsymbol{\psi}^{(s)})}{\partial\boldsymbol{\phi}} = \sum_r \boldsymbol{X}_{\boldsymbol{\phi}}^{*(r;s)\prime}\tilde{\boldsymbol{r}}^{(r;s)} \tag{52}$$

and

$$-\operatorname{E}\left(\frac{\partial^2 Q^*(\boldsymbol{\psi};\boldsymbol{\psi}^{(s)})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'}\right) = \sum_r \boldsymbol{X}_{\boldsymbol{\phi}}^{*(r;s)\prime}\tilde{\boldsymbol{W}}^{(r;s)}\boldsymbol{X}_{\boldsymbol{\phi}}^{*(r;s)}, \tag{53}$$

so that the q-th inner iteration of the s-th M-step of the Monte Carlo EM algorithm takes the form

$$\boldsymbol{\phi}^{(q+1;s)} = \boldsymbol{\phi}^{(q;s)} + \left(\sum_r \boldsymbol{X}_{\boldsymbol{\phi}}^{*(q,r;s)\prime}\tilde{\boldsymbol{W}}^{(q,r;s)}\boldsymbol{X}_{\boldsymbol{\phi}}^{*(q,r;s)}\right)^{-1} \sum_r \boldsymbol{X}_{\boldsymbol{\phi}}^{*(q,r;s)\prime}\tilde{\boldsymbol{r}}^{(q,r;s)} \tag{54}$$

where $\boldsymbol{X}_{\boldsymbol{\phi}}^{*(q,r;s)}$ denotes $\boldsymbol{X}_{\boldsymbol{\phi}}^{*(r;s)}$ evaluated with $\boldsymbol{\phi} = \boldsymbol{\phi}^{(q;s)}$, etc.

Note that the complete-data log-likelihood of the positions (see equation (13)) can be written as

$$\begin{aligned}
\ell(\boldsymbol{b};\boldsymbol{\psi}) = &-\frac{D}{2}\ln(2\pi)\sum_j(T_j+1) + \frac{J}{2}\ln|\boldsymbol{\Theta}_0| + \sum_j\frac{T_j}{2}\ln|\boldsymbol{\Theta}_1| \\
&- \frac{J}{2}\boldsymbol{\beta}'\boldsymbol{\Theta}_0\boldsymbol{\beta} - \sum_j\frac{T_j}{2}\boldsymbol{\beta}'(\boldsymbol{I}-\boldsymbol{\Gamma})'\boldsymbol{\Theta}_1(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{\beta} \\
&+ \sum_j\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{\Theta}_0\boldsymbol{b}_{j0} + \sum_j\frac{1}{2}\boldsymbol{\beta}'(\boldsymbol{I}-\boldsymbol{\Gamma})'\boldsymbol{\Theta}_1(\boldsymbol{s}_{j2}-\boldsymbol{\Gamma}\boldsymbol{s}_{j1}) \\
&+ \sum_j\frac{1}{2}\boldsymbol{b}_{j0}'\boldsymbol{\Theta}_0\boldsymbol{\beta} + \sum_j\frac{1}{2}(\boldsymbol{s}_{j2}-\boldsymbol{\Gamma}\boldsymbol{s}_{j1})'\boldsymbol{\Theta}_1(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{\beta} \\
&- \sum_j\frac{1}{2}\operatorname{tr}(\boldsymbol{\Theta}_0\boldsymbol{S}_{j0}) - \sum_j\frac{1}{2}\operatorname{tr}(\boldsymbol{\Theta}_1(\boldsymbol{S}_{j22}-\boldsymbol{\Gamma}\boldsymbol{S}_{j12}-\boldsymbol{S}_{j21}\boldsymbol{\Gamma}'+\boldsymbol{\Gamma}\boldsymbol{S}_{j11}\boldsymbol{\Gamma}'))
\end{aligned} \tag{55}$$

where $\boldsymbol{\Theta}_0 = \boldsymbol{\Sigma}_0^{-1}$, $\boldsymbol{\Theta}_0 = \boldsymbol{\Sigma}_0^{-1}$ $\boldsymbol{s}_{j1} = \sum_t \boldsymbol{b}_{jt-1}$, $\boldsymbol{s}_{j2} = \sum_t \boldsymbol{b}_{jt}$, $\boldsymbol{S}_{j22} = \sum_t \boldsymbol{b}_{jt}\boldsymbol{b}_{jt}'$, $\boldsymbol{S}_{j12} = \sum_t \boldsymbol{b}_{jt-1}\boldsymbol{b}_{jt}'$, and $\boldsymbol{S}_{j11} = \sum_t \boldsymbol{b}_{jt-1}\boldsymbol{b}_{jt-1}'$. Taking derivatives for $\boldsymbol{\theta}_0 = \operatorname{vec}\boldsymbol{\Theta}_0$ and $\boldsymbol{\theta}_1 = \operatorname{vec}\boldsymbol{\Theta}_1$ leads to

$$\frac{\partial\ell(\boldsymbol{b};\boldsymbol{\psi})}{\partial\boldsymbol{\theta}_0} = \frac{J}{2}\operatorname{vec}\boldsymbol{\Sigma}_0 - \frac{1}{2}\operatorname{vec}\left(\sum_j\boldsymbol{S}_{j0} - \sum_j\boldsymbol{\beta}\boldsymbol{b}_{j0}' - \sum_j\boldsymbol{b}_{j0}\boldsymbol{\beta}' + J\boldsymbol{\beta}\boldsymbol{\beta}'\right) \tag{56}$$

and

$$\begin{aligned}\frac{\partial \ell(\boldsymbol{b};\boldsymbol{\psi})}{\partial \boldsymbol{\theta}_1} &= \sum_j \frac{T_j}{2} \operatorname{vec} \boldsymbol{\Sigma}_1 - \sum_j \frac{1}{2} \operatorname{vec}\left(\boldsymbol{S}_{j22} - \boldsymbol{\Gamma}\boldsymbol{S}_{j12} - \boldsymbol{S}_{j21}\boldsymbol{\Gamma}' + \boldsymbol{\Gamma}\boldsymbol{S}_{j11}\boldsymbol{\Gamma}'\right) \\ &\quad + \sum_j \frac{1}{2} \operatorname{vec}\left((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{\beta}(\boldsymbol{s}_{j2} - \boldsymbol{\Gamma}\boldsymbol{s}_{j1})' + (\boldsymbol{s}_{j2} - \boldsymbol{\Gamma}\boldsymbol{s}_{j1})\boldsymbol{\beta}'(\boldsymbol{I}-\boldsymbol{\Gamma})'\right) \\ &\quad - \sum_j \frac{T_j}{2} \operatorname{vec}\left((\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{\beta}\boldsymbol{\beta}'(\boldsymbol{I}-\boldsymbol{\Gamma})'\right)\end{aligned} \tag{57}$$

It is therefore possible to find, for any given values of $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$, values of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$ that maximize $Q^*(\boldsymbol{\psi};\boldsymbol{\psi}^{(s)})$ by

$$\boldsymbol{\Sigma}_0^{(s)}(\boldsymbol{\beta}) = \frac{1}{J}\left(\bar{\boldsymbol{S}}_0^{(s)} - \boldsymbol{\beta}\bar{\boldsymbol{s}}_0^{(s)\prime} - \bar{\boldsymbol{s}}_0^{(s)}\boldsymbol{\beta}'\right) + \boldsymbol{\beta}\boldsymbol{\beta}' \tag{58}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_1^{(s)}(\boldsymbol{\beta},\boldsymbol{\Gamma}) &= \frac{1}{\sum_j T_j}\left(\bar{\boldsymbol{V}}^{(s)} - (\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{\beta}(\bar{\boldsymbol{s}}_2^{(s)} - \boldsymbol{\Gamma}\bar{\boldsymbol{s}}_1^{(s)})' + (\bar{\boldsymbol{s}}_2^{(s)} - \boldsymbol{\Gamma}\bar{\boldsymbol{s}}_1^{(s)})\boldsymbol{\beta}'(\boldsymbol{I}-\boldsymbol{\Gamma})'\right) \\ &\quad + (\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{\beta}\boldsymbol{\beta}'(\boldsymbol{I}-\boldsymbol{\Gamma})'\end{aligned} \tag{59}$$

where $\bar{\boldsymbol{S}}_0^{(s)} = \sum_j \sum_r \tilde{w}_j^{(r;s)} \boldsymbol{S}_{j0}^{*(r;s)}$, $\bar{\boldsymbol{s}}_0^{(s)} = \sum_j \sum_r \tilde{w}_j^{(r;s)} \boldsymbol{b}_{j0}^{*(r;s)}$, $\bar{\boldsymbol{s}}_1^{(s)} = \sum_j \sum_r \tilde{w}_j^{(r;s)} \boldsymbol{s}_{j1}^{*(r;s)}$, $\bar{\boldsymbol{s}}_2^{(s)} = \sum_j \sum_r \tilde{w}_j^{(r;s)} \boldsymbol{s}_{j2}^{*(r;s)}$, and $\bar{\boldsymbol{V}}^{(s)} = \sum_j \sum_r \tilde{w}_j^{(r;s)} (\boldsymbol{S}_{j22}^{*(r;s)} - \boldsymbol{\Gamma}\boldsymbol{S}_{j12}^{*(r;s)} - \boldsymbol{S}_{j21}^{*(r;s)}\boldsymbol{\Gamma}' + \boldsymbol{\Gamma}\boldsymbol{S}_{j11}^{*(r;s)}\boldsymbol{\Gamma}')$, where $\boldsymbol{S}_{j0}^{*(r;s)}$ is $\boldsymbol{S}_{j0}$ evaluated with $\boldsymbol{b}_j = \boldsymbol{b}_j^{*(r;s)}$, etc.

The derivative of the complete-data log-likelihood for $\boldsymbol{\beta}$ is

$$\frac{\partial \ell(\boldsymbol{b};\boldsymbol{\psi})}{\partial \boldsymbol{\beta}} = \sum_j \boldsymbol{\Theta}_0(\boldsymbol{b}_{j0} - \boldsymbol{\beta}) - \sum_j T_j(\boldsymbol{I}-\boldsymbol{\Gamma})'\boldsymbol{\Theta}_1(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{\beta} + (\boldsymbol{I}-\boldsymbol{\Gamma})'\boldsymbol{\Theta}_1 \sum_j (\boldsymbol{s}_{j2} - \boldsymbol{\Gamma}\boldsymbol{s}_{j1}) \tag{60}$$

and its derivative for $\boldsymbol{\gamma} = \operatorname{vec}\boldsymbol{\Gamma}$ is

$$\frac{\partial \ell(\boldsymbol{b};\boldsymbol{\psi})}{\partial \boldsymbol{\gamma}} = \boldsymbol{I} \otimes \boldsymbol{\Theta}_1 \sum_j \operatorname{vec}\left(\boldsymbol{U}_{j21} - \boldsymbol{\Gamma}\boldsymbol{U}_{j11}\right) \tag{61}$$

where $\boldsymbol{U}_{j11} = \boldsymbol{S}_{j11} - \boldsymbol{\beta}\boldsymbol{s}_{j1}' - \boldsymbol{s}_{j1}\boldsymbol{\beta}' + T_j\boldsymbol{\beta}\boldsymbol{\beta}'$ and $\boldsymbol{U}_{j21} = \boldsymbol{S}_{j21} - \boldsymbol{\beta}\boldsymbol{s}_{j1}' - \boldsymbol{s}_{j2}\boldsymbol{\beta}' + T_j\boldsymbol{\beta}\boldsymbol{\beta}'$. It is obvious that there is no closed-form solution for the roots of these gradients. Therefore, as part of the M-step of the EM algorithm the following objective function

$$\begin{aligned}Q^*(\boldsymbol{\beta},\boldsymbol{\Gamma};\boldsymbol{\psi}^{(s)}) &= -\frac{J}{2}\ln\left|\boldsymbol{\Sigma}_0^{(s)}(\boldsymbol{\beta})\right| - \frac{\sum_j T_j}{2}\ln\left|\boldsymbol{\Sigma}_1^{(s)}(\boldsymbol{\beta},\boldsymbol{\Gamma})\right| \\ &\quad - \frac{J}{2}\boldsymbol{\beta}'\left(\boldsymbol{\Sigma}_0^{(s)}(\boldsymbol{\beta})\right)^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}'\left(\boldsymbol{\Sigma}_0^{(s)}(\boldsymbol{\beta})\right)^{-1}\bar{\boldsymbol{s}}_0^{(s)} \\ &\quad - \frac{\sum_j T_j}{2}\boldsymbol{\beta}'(\boldsymbol{I}-\boldsymbol{\Gamma})'\left(\boldsymbol{\Sigma}_1^{(s)}(\boldsymbol{\beta},\boldsymbol{\Gamma})\right)^{-1}(\boldsymbol{I}-\boldsymbol{\Gamma})\boldsymbol{\beta} \\ &\quad + \boldsymbol{\beta}'(\boldsymbol{I}-\boldsymbol{\Gamma})'\left(\boldsymbol{\Sigma}_1^{(s)}(\boldsymbol{\beta},\boldsymbol{\Gamma})\right)^{-1}(\bar{\boldsymbol{s}}_2^{(s)} - \boldsymbol{\Gamma}\bar{\boldsymbol{s}}_1^{(s)}) \\ &\quad - \frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{\Sigma}_0^{(s)}(\boldsymbol{\beta})\right)^{-1}\bar{\boldsymbol{S}}_0^{(s)}\right) - \frac{1}{2}\operatorname{tr}\left(\left(\boldsymbol{\Sigma}_1^{(s)}(\boldsymbol{\beta},\boldsymbol{\Gamma})\right)^{-1}\bar{\boldsymbol{V}}^{(s)}\right)\end{aligned} \tag{62}$$

is maximized numerically for $\boldsymbol{\beta}$ and $\boldsymbol{\Gamma}$, which is the part of the Q^*-function that depends on these parameters.

B.5 Constructing the Proposal Distribution

Following Booth and Hobert (1999) and Caffo et al. (2005), both the importance sampler and the accept-reject sampler employed for Monte Carlo approximations rest on a Laplace approximation of the posterior distribution of the unobserved data. This Laplace approximation can be constructed as follows: Note that the contribution to the complete data log-likelihood dependent on $\boldsymbol{b}_j$ can be expressed as:

$$\ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi}) = \ell(\boldsymbol{y}_j|\boldsymbol{b}_j; \boldsymbol{\alpha}(\boldsymbol{\psi})) + \frac{1}{2}\ln|\boldsymbol{\Omega}_j| - \frac{1}{2}(\boldsymbol{b}_j - \boldsymbol{\beta}_j)'\boldsymbol{\Omega}_j(\boldsymbol{b}_j - \boldsymbol{\beta}_j) \tag{63}$$

where $\boldsymbol{\beta}_j$ is constructed by stacking the parameter vector $\boldsymbol{\beta}$ such that it is conformable with $\boldsymbol{b}_j$ and

$$\boldsymbol{\Omega}_j = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\Gamma} & & \\ & \boldsymbol{I} & -\boldsymbol{\Gamma} & \\ & & \ddots & \ddots \\ & & & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_0^{-1} & & & \\ & \boldsymbol{\Sigma}_1^{-1} & & \\ & & \ddots & \\ & & & \boldsymbol{\Sigma}_1^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & & & \\ -\boldsymbol{\Gamma} & \boldsymbol{I} & & \\ & \ddots & \ddots & \\ & & -\boldsymbol{\Gamma} & \boldsymbol{I} \end{bmatrix} \tag{64}$$

because the determinant of $\boldsymbol{\Omega}_j$ equals the determinant of the middle factor constructed of $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\Sigma}_1$. To wit, the outer factors constructed of an identity matrix and $\boldsymbol{\Gamma}$ are triangular and have a determinant equal to unity.

The first derivative of this contribution to the complete-data log-likelihood is

$$\frac{\partial \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{b}_j} = \boldsymbol{Z}_j'\boldsymbol{r}_j - \boldsymbol{\Omega}_j(\boldsymbol{b}_j - \boldsymbol{\beta}_j) \tag{65}$$

while the negative Hessian is

$$-\frac{\partial^2 \ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})}{\partial \boldsymbol{b}_j \partial \boldsymbol{b}_j'} = \boldsymbol{Z}_j'\boldsymbol{W}_j\boldsymbol{Z}_j + \boldsymbol{\Omega}_j \tag{66}$$

where

$$\frac{\partial \eta_{ijt}}{\partial \boldsymbol{b}_j} = \boldsymbol{\alpha}_i - \boldsymbol{b}_{jt} \tag{67}$$

$$\boldsymbol{Z}_j = \begin{bmatrix} \boldsymbol{Z}_{j0} & & \\ & \ddots & \\ & & \boldsymbol{Z}_{jT_j} \end{bmatrix} \quad \boldsymbol{Z}_{jt} = \begin{bmatrix} \frac{\partial \eta_{1jt}}{\partial \boldsymbol{b}_{jt}'} \\ \vdots \\ \frac{\partial \eta_{Ijt}}{\partial \boldsymbol{b}_{jt}'} \end{bmatrix}. \tag{68}$$

Note that the negative Hessian is equal to the expected information given the parameter values and $\boldsymbol{b}_j$ because

$$\sum_i \frac{\partial^2 \eta_{ijt}}{\partial \boldsymbol{b}_{jt} \partial \boldsymbol{b}'_{jt}} (m_{ijt} - n_{jt}\pi_{ijt}) = 0, \tag{69}$$

since

$$\frac{\partial^2 \eta_{1jt}}{\partial \boldsymbol{b}_{jt} \partial \boldsymbol{b}'_{jt}} = \cdots = \frac{\partial^2 \eta_{Ijt}}{\partial \boldsymbol{b}_{jt} \partial \boldsymbol{b}'_{jt}} = \boldsymbol{I}_D \tag{70}$$

and

$$\sum_i (m_{ijt} - n_{jt}\pi_{ijt}) = 0. \tag{71}$$

This leads to the following Newton-Raphson iterations to find $\tilde{\boldsymbol{b}}_j$ to maximize $\ell_{\text{cpl}}(\boldsymbol{y}_j, \boldsymbol{b}_j; \boldsymbol{\psi})$ for $\boldsymbol{b}_j$:

$$\boldsymbol{b}_j^{(s+1)} = \boldsymbol{b}_j^{(s)} + \left(\boldsymbol{Z}_j' \boldsymbol{W}_j \boldsymbol{Z}_j + \boldsymbol{\Omega}_j\right)^{-1} \left(\boldsymbol{Z}_j' \boldsymbol{r}_j - \boldsymbol{\Omega}_j (\boldsymbol{b}_j^{(s)} - \boldsymbol{\beta}_j)\right). \tag{72}$$

In much of the literature on state-space models (for an overview, see Harvey 1991) Kalman filtering and smoothing steps have been used instead of performing the above Newton-Raphson steps. Kalman filtering and smoothing in combination are an iterative procedure to compute the right-hand side of equation (72), which makes use of the known structure of $\boldsymbol{Z}_j' \boldsymbol{W}_j \boldsymbol{Z}_j + \boldsymbol{\Omega}_j$ and thus has the computational advantage of avoiding the storage of the potentially large matrices $\boldsymbol{Z}_j$ and $\boldsymbol{\Omega}_j$. Nowadays there are software libraries available, such as the *Matrix* package of R (Bates and Maechler 2010), that handle computation with sparse matrices such $\boldsymbol{Z}_j$ and $\boldsymbol{\Omega}_j$ in a way that is efficient both computational terms and in terms of memory requirement.[5] Thus it is now possible to compute the matrix expression on the right-hand side of equation (72) directly.

Once the value $\tilde{\boldsymbol{b}}_j$ of $\boldsymbol{b}_j$ maximizing the complete-data log-likelihood is found, the scaling matrix $\tilde{\boldsymbol{K}}_j$ needed for the proposal distribution can be obtained as

$$\tilde{\boldsymbol{K}}_j = \boldsymbol{Z}_j' \boldsymbol{W}_j \boldsymbol{Z}_j \big|_{\boldsymbol{b}_j = \tilde{\boldsymbol{b}}_j} + \boldsymbol{\Omega}_j. \tag{73}$$

B.6 Standard Errors

After convergence of the MCEM algorithm, standard errors for the parameters of the model can be computed from the square-root of the diagonal of

$$(\mathcal{I}_{\text{obs}}(\boldsymbol{\psi}))^{-1} = \left(\text{E}_{\boldsymbol{\psi}}(\mathcal{I}_{\text{cpl}}(\boldsymbol{\psi}) | \boldsymbol{y}_j) - \mathcal{I}_{\text{mis}}(\boldsymbol{\psi})\right)^{-1}. \tag{74}$$

[5] Sparse matrices are matrices most elements of which are zero. Linear algebra packages handle such matrices efficiently, for example, by storing only their non-zero elements. A specialized package such as Bates and Maechler (2010) can also efficiently compute the inverse of sparse symmetric matrix with the help of a sparse version of the Cholesky decomposition.

A Monte Carlo approximation of $\mathcal{I}_{\text{mis}}(\boldsymbol{\psi})$ can be obtained from the gradients described in the previous section. What remains to be shown is how to obtain a Monte Carlo approximation of $\mathrm{E}_{\boldsymbol{\psi}}(\mathcal{I}_{\text{cpl}}(\boldsymbol{\psi})|\boldsymbol{y}_j)$. This approximation is

$$-\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\psi}\partial\boldsymbol{\psi}'} = \begin{bmatrix} -\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} & & & & \\ & -\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} & & & \\ & & -\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} & & \\ & & & -\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\sigma}_0\partial\boldsymbol{\sigma}_0'} & \\ & & & & -\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\sigma}_1\partial\boldsymbol{\sigma}_1'} \end{bmatrix} \tag{75}$$

with

$$-\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}'} = \sum_r \boldsymbol{X}_{\boldsymbol{\phi}}^{*(q,r;s)\prime}\tilde{\boldsymbol{W}}^{(q,r;s)}\boldsymbol{X}_{\boldsymbol{\phi}}^{*(q,r;s)}, \tag{76}$$

$$-\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}'} = \sum_j \boldsymbol{\Sigma}_0^{-1} + \sum_j T_j(\boldsymbol{I}-\boldsymbol{\Gamma})'\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{I}-\boldsymbol{\Gamma}) \tag{77}$$

$$-\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} = \bar{\boldsymbol{U}}_{11}\otimes\boldsymbol{\Sigma}_1^{-1}, \tag{78}$$

$$-\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\sigma}_0\partial\boldsymbol{\sigma}_0'} = J\boldsymbol{\Sigma}_0^{-1}\otimes\boldsymbol{\Sigma}_0^{-1}, \tag{79}$$

and

$$-\mathrm{E}\,\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\sigma}_1\partial\boldsymbol{\sigma}_1'} = \frac{\sum_j T_j}{2}\boldsymbol{\Sigma}_1^{-1}\otimes\boldsymbol{\Sigma}_1^{-1}, \tag{80}$$

where further $\bar{\boldsymbol{U}}_{11} = \sum_j\sum_r \tilde{w}_j^{(r)}(\boldsymbol{S}_{j11}^{*(r)} - \boldsymbol{\beta}\boldsymbol{s}_{j1}^{*(r)\prime} - \boldsymbol{s}_{j1}^{*(r)}\boldsymbol{\beta}' + T_j\boldsymbol{\beta}\boldsymbol{\beta}')$. It should be noted here that the matrix in equation (75) has its block-diagonal structure because all "cross-derivatives", like $\frac{\partial^2 Q^*(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\psi}})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\gamma}'}$ etc., have a vanishing expectation.

B.7 Identifying Constraints on the Location Parameters

In the main text, two issues in the identification of spatial models were discussed. Unless further constraints are imposed on the location parameters, these models are not identified with respect to translation, rotation, and reflection, that is, to any linear transformation that preserves the distance between pairs of points. It is stated that linear constraints on the location parameters make such models identified with respect to rotation and translation.

In the following it is shown how linear constraints can be implemented in an iterative algorithm without augmenting the objective function with Lagrange multipliers, but by constructing a unique mapping between the space of linearly constrained parameters and a space of unconstrained parameters with a reduced dimension.

The linear restrictions take the form

$$\boldsymbol{T}\boldsymbol{\alpha} = \boldsymbol{t} \tag{81}$$

where $\boldsymbol{T}$ is a $(p \times q)$ matrix with $p \leq I \cdot D$ rows and $q = I \cdot D$ columns and $\boldsymbol{t}$ is a q-dimensional vector. If the first D rows of $\boldsymbol{T}$ are equal to $\mathbf{1}_I' \otimes \boldsymbol{I}_{(D\times D)}$, and the first D elements of $\boldsymbol{t}$ are equal to zero, where $\mathbf{1}_I$ is an I-dimensional vector with all elements equal to unity and $\boldsymbol{I}_{(D\times D)}$ is the $(D \times D)$ identity matrix, then this corresponds to the restriction that $\sum_i \boldsymbol{\alpha}_i = \mathbf{0}$. If the remaining rows of $\boldsymbol{T}$ are equal to some of the rows of a $(p \times p)$ identity matrix and the remaining elements of $\boldsymbol{t}$ also are zero, then this means that some of the elements of $\boldsymbol{\alpha}$ are restricted to equal zero.

If $p = q$ and $\boldsymbol{T}$ is of full rank, then the linear restrictions are an fully determined linear system and equation (81) has exactly one solution, namely $\boldsymbol{T}^{-1}\boldsymbol{t}$. In that case, the spatial model is identified, but the restriction are of course too strong to make the model useful for empirical analysis. If $p = q$ and $\boldsymbol{T}$ is not of full row rank or, equivalently, if it is of full row rank, but with number of rows $p < q$, then the set of all solutions to equation (81) is a r-dimensional subset of the parameter space (where $r = q - p$) the elements of which are of the form

$$\boldsymbol{\alpha} = \boldsymbol{T}^{-}\boldsymbol{t} - (\boldsymbol{I}_{(q\times q)} - \boldsymbol{T}^{-}\boldsymbol{T})\boldsymbol{\phi}^{*} \tag{82}$$

where $\boldsymbol{T}^{-}$ is any generalized inverse of $\boldsymbol{T}$ (Harville 1997: 141, eq. 2.1). Now because $\boldsymbol{I}_{(q\times q)} - \boldsymbol{T}^{-}\boldsymbol{T}$ is singular, the vector $\boldsymbol{\phi}^{*}$ is not uniquely determined by $\boldsymbol{\alpha}$. The problem now is to construct a mapping between the set of q-dimensional vectors $\boldsymbol{\alpha}$ that satisfy the under-determined linear system (81) and a space of r-dimensional vectors $\boldsymbol{\phi}$ of the form

$$\boldsymbol{\alpha} = \boldsymbol{r} + \boldsymbol{Q}\boldsymbol{\phi}. \tag{83}$$

This construction rests on the (reduced-rank) QR decomposition

$$\boldsymbol{I}_{(q\times q)} - \boldsymbol{T}^{-}\boldsymbol{T} = \boldsymbol{Q}\boldsymbol{R} \tag{84}$$

where $\boldsymbol{Q}$ is a $(q \times r)$ matrix with orthogonal columns (that is $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_{(r\times r)}$) and $\boldsymbol{R}$ is a $(r \times q)$ matrix formed of the first r rows of a $(q \times q)$ upper triangular matrix (Harville 1997: 66, 277). Thus equation (82) can be rewritten as

$$\boldsymbol{\alpha} = \boldsymbol{T}^{-}\boldsymbol{t} - \boldsymbol{Q}\boldsymbol{R}\boldsymbol{\phi}^{*} \tag{85}$$

If one sets

$$\boldsymbol{R}\boldsymbol{\phi}^{*} = \boldsymbol{\phi} \tag{86}$$

and

$$\boldsymbol{r} = \boldsymbol{T}^{-}\boldsymbol{t} \tag{87}$$

then equation (85) leads to equation (83).

There are several ways to construct a generalized inverse of $\boldsymbol{T}$ (unless it is of full rank) and equation (82) holds independently from the choice of the generalized inverse. If one chooses the generalized inverse

$$\boldsymbol{T}^{-} = \boldsymbol{T}'(\boldsymbol{T}\boldsymbol{T}')^{-1} \tag{88}$$

(note that if $\boldsymbol{T}$ has full row rank, then $\boldsymbol{T}\boldsymbol{T}'$ is non-singular) and if one further notes that for each $\boldsymbol{\phi}$ one can find a $\boldsymbol{\phi}^*$ that satisfies (86) by setting

$$\boldsymbol{\phi}^* = \boldsymbol{R}'(\boldsymbol{R}\boldsymbol{R}')^{-1}\boldsymbol{\phi} \tag{89}$$

it is easy to verify that $\boldsymbol{\alpha}$ given by equation (83) satisfies equation (81):

$$\begin{aligned} \boldsymbol{T}\boldsymbol{\alpha} &= \boldsymbol{T}(\boldsymbol{r} + \boldsymbol{Q}\boldsymbol{\phi}) = \boldsymbol{t} + \boldsymbol{T}\boldsymbol{Q}\boldsymbol{R}\boldsymbol{\phi}^* \\ &= \boldsymbol{t} + \boldsymbol{T}(\boldsymbol{I}_{(q\times q)} - \boldsymbol{T}^{-}\boldsymbol{T})\boldsymbol{\phi}^* = \boldsymbol{t} + (\boldsymbol{T} - \boldsymbol{T})\boldsymbol{\phi}^* \\ &= \boldsymbol{t} \end{aligned} \tag{90}$$

It is also worth noting that, since $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}_{(q\times Q)}$, equation (83) leads to

$$\boldsymbol{\phi} = \boldsymbol{Q}'(\boldsymbol{\alpha} - \boldsymbol{r}). \tag{91}$$

so that for every q-dimensional vector $\boldsymbol{\alpha}$ that satisfies equation (81) one can find exactly one corresponding r-dimensional vector $\boldsymbol{\phi}$. Maximizing either the observed-data or complete-data log-likelihood $\ell(\boldsymbol{\alpha})$ for $\boldsymbol{\alpha}$ subject to the constraints (81) is therefore equivalent to maximizing $\ell(\boldsymbol{r} + \boldsymbol{Q}\boldsymbol{\phi})$ for $\boldsymbol{\phi}$ without any constraints.

In the application of the dynamic state-space model to the space of economic policy, the locations of the seven policy objectives in the one-dimensional space of economic policy are elements of a location vector $\boldsymbol{\alpha}$ with seven elements. If the restriction is posed that these seven elements sum to zero then this leads to an unrestricted vector $\boldsymbol{\phi}$ with six elements. This corresponds to the fact that under the sum-to-zero constraints any of the seven elements of $\boldsymbol{\alpha}$ can be expressed as the negative of the sum of the remaining six elements.

With appropriate constraints (81) in place, a spatial model can be made identified with respect to rotation and translation, but may remain unidentified with respect to reflection. However, then it is only globally unidentified, not locally. That is, for any $\boldsymbol{\alpha}$ and any reflection matrix $\boldsymbol{M}$ (a diagonal matrix with elements at the diagonal either equal to 1 or -1) there is an ϵ ($0 < \epsilon < ||\boldsymbol{\alpha}||$) such that $\boldsymbol{M}\boldsymbol{\alpha}$ is outside the ϵ-neighborhood of $\boldsymbol{\alpha}$ (i.e. $||\boldsymbol{M}\boldsymbol{\alpha} - \boldsymbol{\alpha}|| > \epsilon$). Consequently, if for all updating steps of an iterative procedure to find a maximum likelihood estimate $\hat{\boldsymbol{\alpha}}$

$$||\boldsymbol{\alpha}^{(s)} - \boldsymbol{\alpha}^{(s-1)}|| < ||\boldsymbol{\alpha}||, \tag{92}$$

where $\boldsymbol{\alpha}^{(s)}$ is the approximate value of $\hat{\boldsymbol{\alpha}}$ at the s-th iteration, the procedure can converge. That is, if starting values are given that are close enough to the maximum likelihood value so that the

sign of the elements of $\boldsymbol{\alpha}^{(s)}$ do not change as $s \to \infty$, the signs of the elements of the starting value $\boldsymbol{\alpha}^{(0)}$ determine the signs of the elements of $\boldsymbol{\alpha}^{(\infty)} = \hat{\boldsymbol{\alpha}}$ and thus make the model identified with respect to reflection.

In contrast, if a spatial model is unidentified with respect to rotation or translation, then it is not only globally but also locally unidentified. To illustrate that for the case of non-identification with respect to translation: For any $\boldsymbol{\alpha}$ and any $\epsilon > 0$ there exists a $\boldsymbol{c}$ such that $\|(\boldsymbol{\alpha} + \boldsymbol{c}) - \boldsymbol{\alpha}\| < \epsilon$ (it suffices that $\|\boldsymbol{c}\| < \epsilon$). Local non-identification of $\boldsymbol{\alpha}$ means that an iterative procedure to find a maximum likelihood value $\hat{\boldsymbol{\alpha}}$ cannot lead to convergence, hence the necessity to explicitly implement restrictions such as those of equation (81).

B.8 Starting Values

Iterative procedures to compute ML estimates usually need starting values to proceed to their first iteration. Unless the likelihood surface is globally concave, there is no guarantee that iterations will converge to a global maximum. Rather, if "bad" starting values are provided, the algorithm may get stuck in a local, non-global maximum. Also, if starting values are far off those value that maximize the likelihood, numerical instabilities may lead the iterative algorithm into divergence. Therefore, a method to automatically obtain starting values that are already near the maximum of the likelihood function is desirable. An example for this is the method proposed by Nelder and Wedderburn (1972) to obtain starting values for the Fisher-scoring procedure that they discuss fitting generalized linear models (such as logit and probit models).

The method for generating starting values for the MCEM algorithm to find the maximum marginal likelihood estimates for the model introduced in the article is the same method used in Elff (2009) to obtain approximate estimates of the positions of political parties based on CMP data. It rests on the idea that the counts that are used to represent the emphases of policy objectives have expected values that are related to the distances between the locations of policy objectives and the political positions expressed in political texts. Therefore one may obtain approximate values of the distances Δ_{ijt} directly from the count data. From these approximate distances one can obtain approximate locations of the policy objectives and approximate political positions by multidimensional metric unfolding. If the number of approximate distances is large relative to the number of locations to be reconstructed, Schönemann's algorithm can be used, which only employs a finite number of linear algebra operations (Schönemann 1970).[6]

C Checking the Proposed Estimator with Simulated Data

It is well known from theoretical statistics that maximum likelihood estimators are, under suitable regularity conditions, consistent, asymptotically efficient, and asymptotically normal dis-

[6] This algorithm is implemented in the *R* package `munfold` (Elff 2010), available at the Comprehensive R Archive Network.

Table 1: Parameter values of the model generating simulated data

Locations of objectives ($\boldsymbol{\alpha}$)	
Objective 1	−1.0
Objective 2	−0.5
Objective 3	−0.1
Objective 4	0.1
Objective 5	0.5
Objective 6	1.0
Distribution of positions	
Mean ($\boldsymbol{\beta}$)	0.1
Auto-regression slope ($\boldsymbol{\Gamma}$)	0.9
Variance	
Between actors ($\boldsymbol{\Sigma}_0$)	1.0
Between time points ($\boldsymbol{\Sigma}_1$)	0.1

tributed. But these are properties that these estimators have only as the number of observations approaches infinity. Therefore, there is in principle no guarantee that estimates obtained from finite samples are in any way unbiased or "close enough" to the true parameter values. Therefore this section reports results from a small simulation study. This simulation study can also be perceived as a test case for the software implemented to compute the estimates. In this simulation study, data were generated from a model with parameter values as shown in Table 1.

The simulation of data from this model involved two steps: In the first step time series of political positions one hundred actors based on the settings of the values for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}_1$ are generated. The lengths of the time series in the first step are also generated as random values with a Poisson distribution with mean parameter 10, so that the average length of the time series is 10. In the second step emphasis counts are simulated from multinomial distributions with cell probabilities determined by the policy objective locations and the actor positions, as specified in equation (3) of the main text. The denominators of the multinomial distributions are themselves random numbers, generated from a Poisson distribution. There are two runs of the simulation study. In the first run the mean parameter of the Poisson distribution is set to $\lambda = 20$ so that the mean text length is 20. In the second run, the Poisson distribution parameter is set to $\lambda = 2000$.

Table 2 shows maximum likelihood estimates for the parameters of the model from which the simulated data are generated. The estimates are all quite close to the true values (shown in Table 1). In the simulation with $\lambda = 2000$ all the estimates do not depart from their "true" values by more than twice the standard error, with the exception of the auto-regression slope. The result for the simulation with $\lambda = 20$ is almost the same, in addition only one of the estimates of the location of the objectives departs from its "true" value by more than twice its standard error (but not much more than that).

Table 2: Parameter estimates from the two runs of the simulation study — maximum likelihood estimates with standard errors in parentheses.

Mean text size:	$\lambda = 20$		$\lambda = 2000$	
Locations of objectives:	Estimate	Std.Err.	Estimate	Std.Err.
Objective 1	−0.981	(0.024)	−0.996	(0.002)
Objective 2	−0.506	(0.023)	−0.499	(0.002)
Objective 3	−0.117	(0.021)	−0.104	(0.002)
Objective 4	0.050	(0.021)	0.100	(0.002)
Objective 5	0.493	(0.022)	0.496	(0.002)
Objective 6	1.061	(0.027)	1.002	(0.003)
Distributions of positions:	Estimate	Std.Err.	Estimate	Std.Err.
Means	0.116	(0.074)	0.105	(0.073)
Auto-regression slope	0.882	(0.000)	0.885	(0.000)
Variances				
Between actors	1.360	(0.210)	1.341	(0.190)
Between time points	0.104	(0.012)	0.103	(0.005)
Summary statistics:				
Likelihood:	−7744.5		−21031.6	
Deviance:	4374.1		7480.1	
Totals:				
Number of actors:	100		100	
Number of observations:	974		974	
Sum of counts:	19293		1947627	

As discussed previously, the parameters of the model only describe how the positions of political actors evolve, they do not describe the positions themselves. To reconstruct the positions of the actors one needs to obtain predictions based on the model and the observed data of counts recorded in the individual texts. This is illustrated by Figure 1, which compares the (simulated) actors' positions that were generated in the first stage of the simulation study with posterior mean predictions from the model estimated under the condition of $\lambda = 20$. The black lines in the figure represent the development of the actors "true" positions, while the dark gray lines represent the posterior mean predictions from the estimated model. The lines representing the point predictions are enclosed in 99 per cent prediction bands, which represent the uncertainty about the predicted positions.

As an inspection of Figure 1 reveals, the point predictions about the actors' position mainly obviously follow their general dynamics. If the "true" positions of an actor move to the "center" of the latent political space, so do the predicted positions and if they move away from the center, so do the predicted positions. There are of course also departures of the predictions from the true positions as one would expect from the limited information about the positions, present in the observed counts that follow multinomial distributions with a denominator of about 20. Yet in general, the "true" positions stay within the prediction bands most of the time.

If there is an abundance of information about actors positions in political texts the match between "true" and predicted positions is much closer, as Figure 2 makes clear. Here the predictions are obtained from the model fit to simulated data with a mean text length of 2000 sentences. Figure 2 is constructed to contain, like the preceding one, lines representing "true" positions, predicted positions, and 99 per cent prediction bands. However, the prediction bands are so narrow that they cannot be distinguished from the lines presenting the true and predicted positions. Furthermore, the lines of the true and the predicted positions also coincide in general.

D Extended and Supplementary Results

D.1 A Wald Test of Hypotheses on the Locations of CMP Economic Policy Categories in the Policy Space

Many of the more traditional CMP-based indices of political positions group policy categories into broader "leftist" and "rightist" categories, without allowing for a variation in their "left-ness" or "right-ness". Based on the method proposed in the article one now has maximum likelihood estimates of the locations of economic policy categories on an economic left-right dimension, so one can use likelihood-based inferential techniques to test whether such variations are statistically significant. Table 3 shows the results of three Wald tests, one for the difference between the mean locations of the interventionist and the laissez-faire objectives, a second one for the equality of all interventionist objectives, and a third one for the the equality of all laissez-faire objectives. All three Wald tests clearly lead to a rejection of the null hypothesis, that is, while

Figure 1: "True" positions and predictions from estimated model with mean text length 20 sentences, with 99 per cent prediction bands.

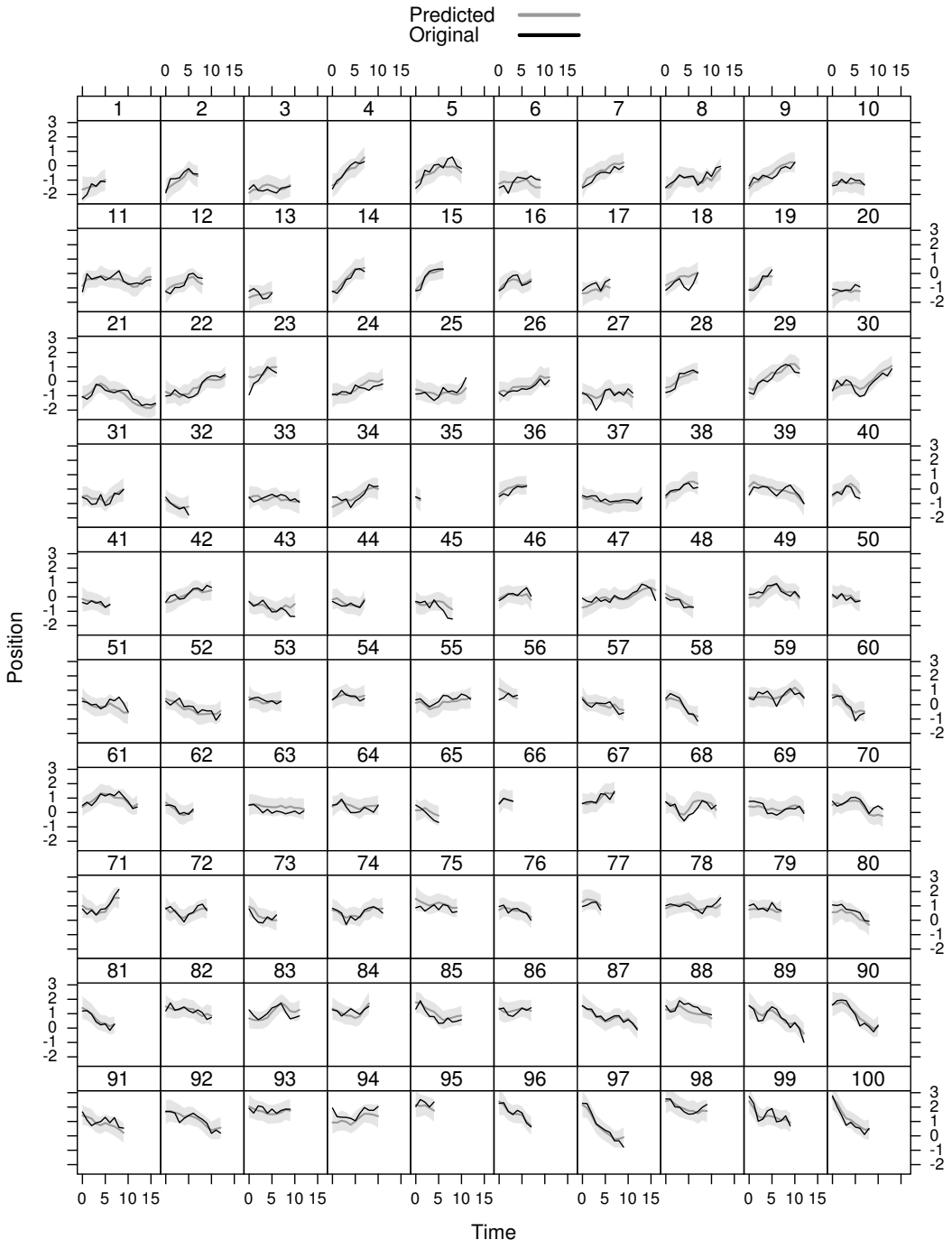


Figure 2: “True” positions and predictions from estimated model with mean text length 2000 sentences, with 99 per cent prediction bands.

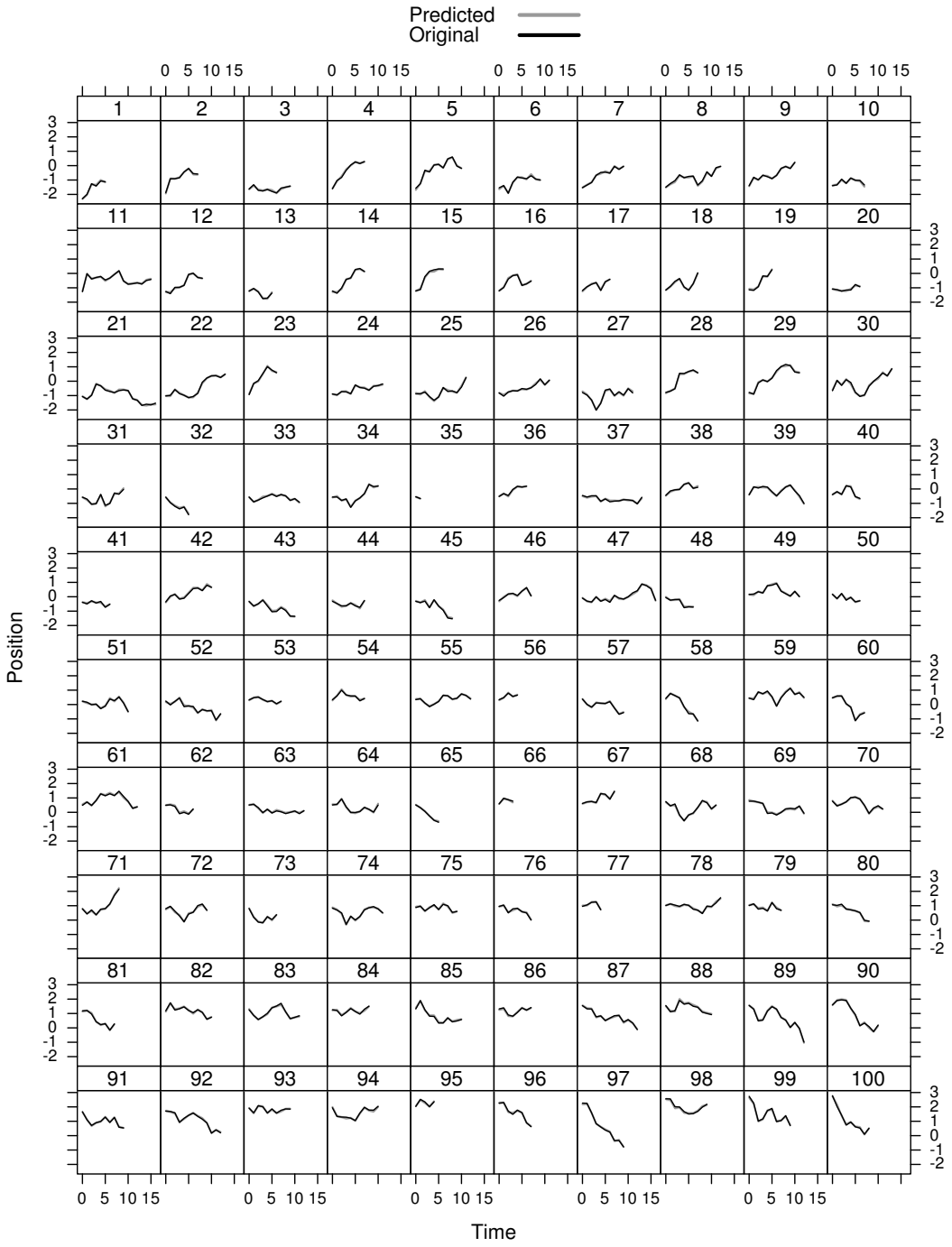


Table 3: Wald tests of hypotheses about the locations of policy objectives in the economic policy space.

	W	df	$\Pr(\chi^2 > W)$
Mean difference between interventionist and laissez-faire objectives	99393.6	1	0.000
Variation among interventionist objectives	27309.6	3	0.000
Variation among laissez-faire objectives	9460.5	2	0.000

Data source: The Manifesto Project
(Budge et al. 2001; Klingemann et al. 2006; Volkens et al. 2010).

the difference between the mean locations of the interventionist and the laissez-faire objectives is statistically significant, so are the variations among the locations of the interventionist objectives and among the locations of the laissez-faire objectives. That is, while one can distinguish between objectives that have more interventionist locations and objectives that have more laissez-faire locations, it is clear that these two groups of objectives do not form homogeneous blocks. Thus the indices that rest on a simple dichotomy between economically "leftist" objectives and economically "rightist" objectives, as they are often used in the literature, do not find support by a suitable measurement model for political texts.

D.2 The Evolution of the Positions of West European Conservative and Social Democratic Parties from the Perspective of Various Approaches

Figures 3 through 6 track the evolution of major conservative and social democratic parties from six countries from Western Europe: France, Germany, the Netherlands, Spain, Sweden, and the United Kingdom. They allow to get an impression of how the method proposed in the article performs in comparison to some other common approaches at the reconstruction of parties' political positions based on data from the Comparative Manifestos Project.

Figure 3 shows the development of the parties as it appears based on the still widely used general left-right index (RiLe) provided with the CMP data set. Figure 4 shows the development of the parties based on the PlanEco index which is included in newer editions of the CMP data set (Klingemann et al. 2006; Volkens et al. 2010). It is constructed from the difference between the CMP variables 'Economic orthodoxy" (Per414) and "Free enterprise" (Per401), which are categorized as "rightist", and the CMP variables "Controlled economy" (Per412), "Economic planning" (Per404), and "Market regulation" (Per403), which are categorized as "leftist". Figure 5 uses Lowe et al.'s (2011) LogPlanEco index, which employs the same variables as the PlanEco index, but is based on equation (3) instead of (1). Figure 6 (which is identical to Figure 3 in the main article and is only included in this appendix to ease the comparison) finally tracks the

development of the variables using posterior means predictions generated from the dynamic state-space model introduced in the article.

Figures 3 through 5 are enhanced by 95% condidence intervals based on the bootstrap method proposed by Benoit et al. (2009), to give an impression of the associated uncertainty. Figure 6 shows prediction bands obtained from the simulated 2.5% and 97.5% quantiles of the posterior distribution of the parties' positions.

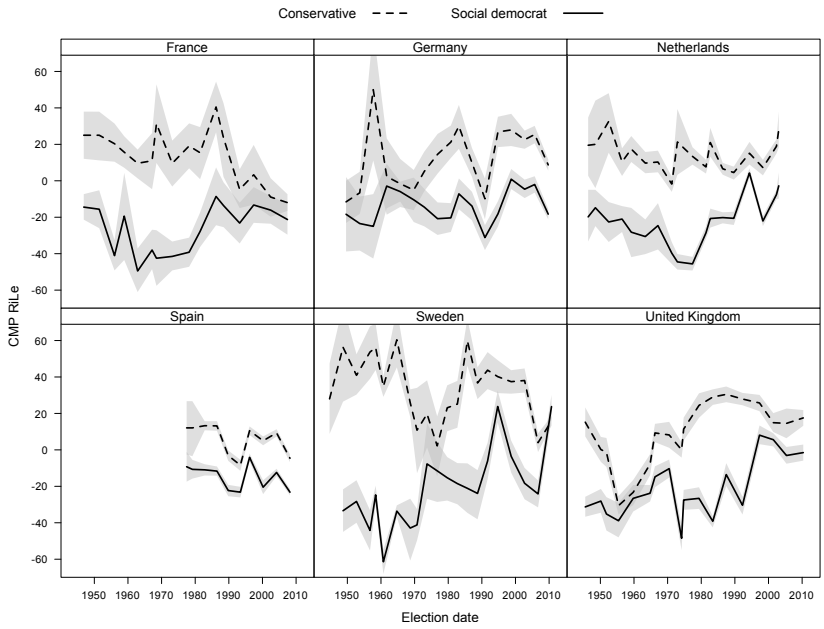
As Figure 3 shows, conservative and social democratic parties appear to converge in terms of their left-right positions at least in four of the six countries, yet in German and Spain no clear pattern of convergence can be discerned. The German parties even shows a slight tendency of increasing polarisation. A notable feature of the dynamics of the parties' positions according to the RiLe index is its occasionally strong volatility. For example, the the CDU/CSU takes a one-off tour into the radical right in the late 1950s, which seems to dwarf the moderation of the SPD position in the early 1950s.

A look at Figure 4 leads to different conclusions about the evolution of some parties' positions: Now one does not find a convergence between the two French parties, yet a more clear pattern of convergence between the two Dutch parties. The German CDU/CSU now shows fluctuations in its position that are larger than its typical distance towards the SPD. The Swedish parties now show an even clearer pattern of convergence yet also a highly volatile parallel movement during the later decades. Also notable are the extreme rightist but also highly volatile positions of the Swedish conservatives in the 1950s and 1960s. In general, the positions appear much more volatile when measured using the PlanEco index, except for the Spanish and British parties.

If one now considers the development of the parties from the perspective of the LogPlanEco index one comes to different impressions and conclusions again. The movement of the social democratic parties in Figure 5 appears more volatile than in Figure 3 and 4. On the other hand, the positions of the Swedish conservatives look much less volatile until 2000, yet the leftward movement of both Swedish parties after 2000 now looks much more pronounced. Finally, while it does not seem to be very consequential for the apparent volatility of the two British parties whether one uses the CMP Rile index or the CMP PlanEco index, the positions of the British parties appear more volatile if the LogPlanEco index is used.

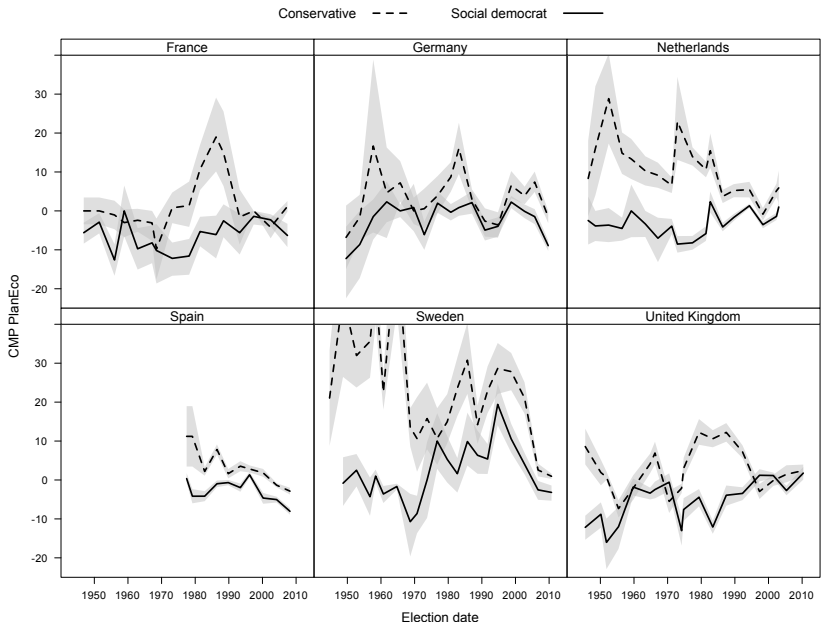
The posterior means predictions obtained from the dynamic state-space model introduced by the article, shown in Figure 6, look considerably less volatile than positions reconstructed from the LogPlanEco index. This applies especially to the positions of the French and Dutch socialist parties, both German parties and both British parties. In contrast to Figures 3 through 5 one now also finds that the Spanish parties converge. On the other hand, the prediction bands look, relative to the movements of the parties, wider in Figure 6 than in 5. Consequently, inferences about policy shifts based on the method proposed in the article will be more conservative than inferences based on the LogPlanEco index.

Figure 3: A comparison of the evolution of the political positions of major conservative and social democratic parties of Western Europe based on the CMP RiLe Index



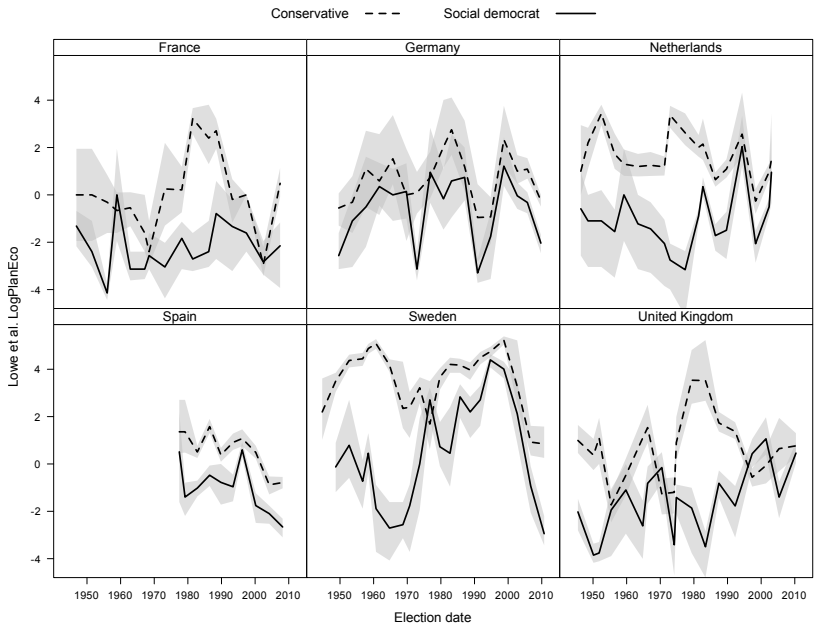
Note: The conservative parties considered here are the *Gaullistes* of France (in their various guises as RPR, UMP, etc.), the CDU/CSU of Germany, the VVD of the Netherlands, the Popular Alliance/ Popular Party of Spain, the Moderates of Sweden, and the Conservatives of the United Kingdom. The social democratic parties considered are the *Parti Socialiste* of France, the SPD of Germany, the PvdA of the Netherlands, PSOE of Spain, the Socialist Workers' Party of Sweden, the Labour Party of the United Kingdom.
The dashed and solid lines connect the index values for the positions of the respectives parties on occasion of the respective elections. The gray areas are 95% confidence bands based on the bootstrap method propsed by Benoit et al. (2009).
Data source: The Manifesto Project (Budge et al. 2001; Klingemann et al. 2006; Volkens et al. 2010).

Figure 4: A comparison of the evolution of the political positions of major conservative and social democratic parties of Western Europe based on the CMP PlanEco Index



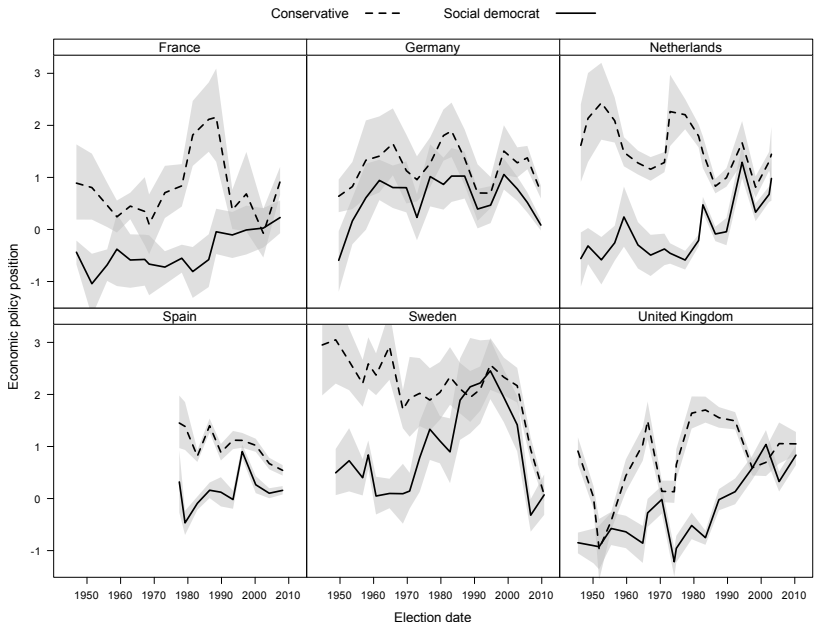
Note: The conservative parties considered here are the *Gaullistes* of France (in their various guises as RPR, UMP, etc.), the CDU/CSU of Germany, the VVD of the Netherlands, the Popular Alliance/ Popular Party of Spain, the Moderates of Sweden, and the Conservatives of the United Kingdom. The social democratic parties considered are the *Parti Socialiste* of France, the SPD of Germany, the PvdA of the Netherlands, PSOE of Spain, the Socialist Workers' Party of Sweden, the Labour Party of the United Kingdom.
The dashed and solid lines connect the index values for the positions of the respectives parties on occasion of the respective elections. The gray areas are 95% confidence bands based on the bootstrap method propsed by Benoit et al. (2009).
Data source: The Manifesto Project (Budge et al. 2001; Klingemann et al. 2006; Volkens et al. 2010).

Figure 5: A comparison of the evolution of the political positions of major conservative and social democratic parties of Western Europe based on Lowe et al.'s LogPlanEco Index



Note: The conservative parties considered here are the *Gaullistes* of France (in their various guises as RPR, UMP, etc.), the CDU/CSU of Germany, the VVD of the Netherlands, the Popular Alliance/ Popular Party of Spain, the Moderates of Sweden, and the Conservatives of the United Kingdom. The social democratic parties considered are the *Parti Socialiste* of France, the SPD of Germany, the PvdA of the Netherlands, PSOE of Spain, the Socialist Workers' Party of Sweden, the Labour Party of the United Kingdom.
The dashed and solid lines connect the index values for the positions of the respectives parties on occasion of the respective elections. The gray areas are 95% confidence bands based on the bootstrap method propsed by Benoit et al. (2009).
Data source: The Manifesto Project (Budge et al. 2001; Klingemann et al. 2006; Volkens et al. 2010).

Figure 6: A comparison of the evolution of the political positions of major conservative and social democratic parties of Western Europe based on the method proposed in the article.



Note: The conservative parties considered here are the *Gaullistes* of France (in their various guises as RPR, UMP, etc.), the CDU/CSU of Germany, the VVD of the Netherlands, the Popular Alliance/ Popular Party of Spain, the Moderates of Sweden, and the Conservatives of the United Kingdom. The social democratic parties considered are the *Parti Socialiste* of France, the SPD of Germany, the PvdA of the Netherlands, PSOE of Spain, the Socialist Workers' Party of Sweden, the Labour Party of the United Kingdom.
The dashed and solid lines connect the simuliated posterior expectations of the parties' positions on occasion of the respective elections. The gray areas are 95% prediction bands based on the simulated 2.5% and 97.5% quantiles of the posterior distribution of the parties' positions.
Data source: The Manifesto Project (Budge et al. 2001; Klingemann et al. 2006; Volkens et al. 2010).

References

Agresti, A. (2002). *Categorical Data Analysis* (Second ed.). New York: Wiley.

Bates, D. and M. Maechler (2010). Matrix: Sparse and dense matrix classes and methods. R package version 0.999375-46. `http://CRAN.R-project.org/package=Matrix`.

Benoit, K. and M. Laver (2006). *Party Policy in Modern Democracies*. London and New York: Routledge.

Benoit, K., S. Mikhaylov, and M. Laver (2009). Treating words as data with error: Uncertainty in text statements of policy positions. *American Journal of Political Science 53*(2), 495–513.

Booth, J. G. and J. P. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61*(1), 265–285.

Budge, I., H.-D. Klingemann, A. Volkens, J. Bara, E. Tanenbaum, et al. (2001). *Mapping Policy Preferences: Estimates for Parties, Electors, and Governments 1945-98*. Oxford: Oxford University Press.

Budge, I., D. Robertson, and D. Hearl (1987). *Ideology, Strategy, and Party Change: Spatial Analysis of Post-War Election Programs in 19 Democracies*. Cambridge: Cambridge Univerity Press.

Caffo, B. S., W. Jank, and G. L. Jones (2005). Ascent-based Monte Carlo expectation-maximization. *Journal of the Royal Statistical Society. Series B (Methodological) 67*(2), 235–251.

Clinton, J., S. Jackman, and D. Rivers (2004). The statistical analysis of roll call data. *American Political Science Review 98*(2), 355–370.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological) 39*(1), 1–38.

Eddelbuettel, D. and R. François (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software 40*(8), 1–18.

Elff, M. (2009). Social divisions, party positions, and electoral behaviour. *Electoral Studies 28*(2), 297–308.

Elff, M. (2010). munfold: Metric unfolding. R package version 0.2. `http://CRAN.R-project.org/package=munfold`.

Elff, M. (2012). Replication data for: A dynamic state-space model of coded political texts. http://hdl.handle.net/1902.1/19198 IQSS Dataverse Network [Distributor] V1 [Version].

Francois, R., D. Eddelbuettel, and D. Bates (2011). Rcpparmadillo: Rcpp integration for armadillo templated linear algebra library. R package version 0.2.18. http://CRAN.R-project.org/package=RcppArmadillo.

Gabel, M. and J. Huber (2000). Putting parties in their place: Inferring party left-right ideological positions from party manifestos data. *American Journal of Political Science 44*(1), 94–103.

Harvey, A. C. (1991). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge, UK: Cambridge University Press.

Harville, D. A. (1997). *Matrix Algebra From a Statistician's Perspective*. New York: Springer.

Klingemann, H.-D., A. Volkens, J. Bara, I. Budge, and M. McDonald (2006). *Mapping Policy Preferences II. Estimates for Parties, Electors, and Governments in Eastern Europe, the European Union and the OECD, 1990-2003*. Oxford: Oxford University Press.

Laver, M. J. and J. Garry (2000). Estimating policy positions from political texts. *American Journal of Political Science 44*(3), 619–634.

Little, R. J. and D. B. Rubin (2002). *Statistical Analysis with Missing Data* (Second ed.). Hoboken, NJ: Wiley.

Lowe, W. (2008). Understanding wordscores. *Political Analysis 16*(4), 356–371.

Lowe, W., K. Benoit, S. Mikhaylov, and M. Laver (2011). Scaling policy preferences from coded political texts. *Legislative Studies Quarterly 34*(1), 123–155.

Martin, A. D. and K. M. Quinn (2002). Dynamic ideal point estimation via markov chain monte carlo for the u.s. supreme court, 1953–1999. *Political Analysis 10*(2), 134–153.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models. Second Edition*. London, New York: Chapman and Hall.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General) 135*(3), 370–384.

Poole, K. T. and H. Rosenthal (1985). A spatial model for legislative roll call analysis. *American Journal of Political Science 29*(2), 357–384.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Robert, C. P. and G. Casella (2004). *Monte Carlo Statistical Methods* (Second ed.). New York: Springer.

Schönemann, P. H. (1970). On metric multidimensional unfolding. *Psychometrika 35*(3), 349–366.

Slapin, J. B. and S.-O. Proksch (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science 52*(3), 705–722.

van der Brug, W. (2001). Analysing party dynamics by taking partially overlapping snapshots. In M. Laver (Ed.), *Estimating the Policy Positions of Political Actors*, pp. 115–132. London and New York: Routledge.

Volkens, A., O. Lacewell, S. Regel, H. Schultze, and A. Werner (2010). *The Manifesto Data Collection*. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). `https://manifestoproject.wzb.eu/`.