Statistical Modeling Of International Migration Flow Tables

Guy Abel
g.j.abel@soton.ac.uk

Division of Social Statistics, University of Southampton, United Kingdom

Sponsor: Economic and Social Research Council

8th July 2009

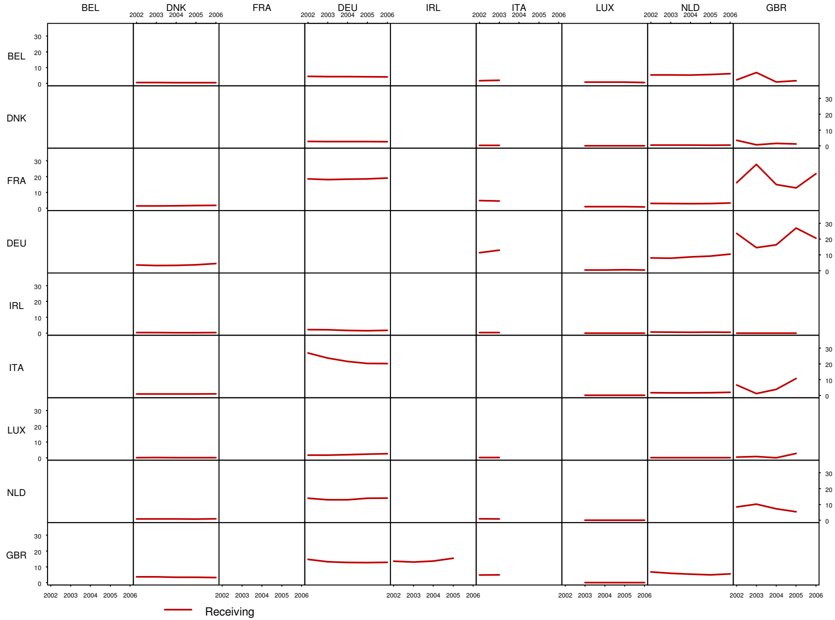
Motivation

- Migration flow data inform policy makers, the media and academic community to the level and direction of population movements
- Comparable migration data can help concerned parties to manage policy and understand people's movements better.
- In comparability from
  1. Differences in data production: collection, definitions, coverage
  2. Differences in data dissemination: completely available, partially available or completely unavailable
- Propose a methodology to:
  1. Scaling for inconsistencies
  2. Impute for incomplete data
- Concentrate on missing data.

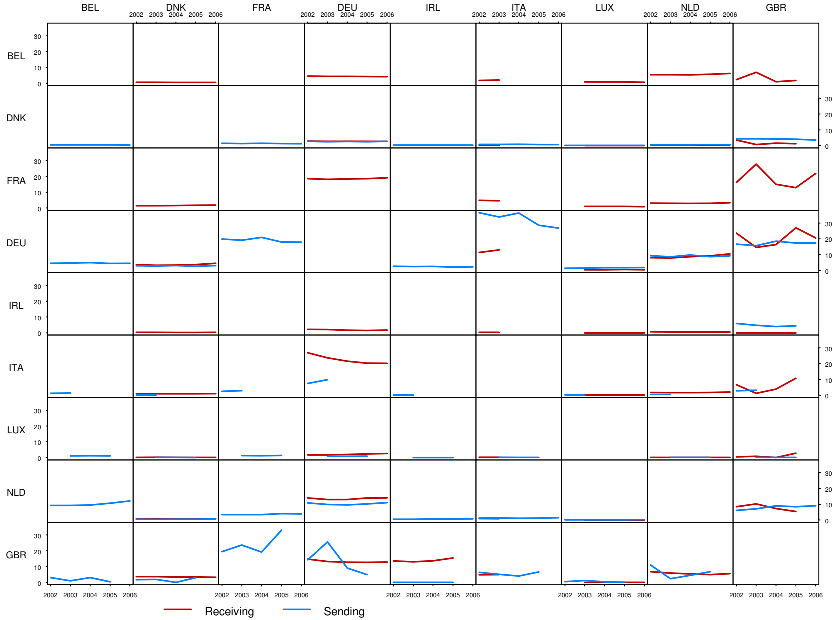
EC9 Data (Eurostat), 2006

	BEL	DNK	FRA	DEU	IRL	ITA	LUX	NLD	GBR
BEL		529		4115			605	6149	
DNK	413		1145	2563 2690	256	675	17 166	488 615	3538
FRA		1755		19095			927	3357	21821
DEU	4540	4471 3115	17790		2330	26807	455 1864	10424 9189	20505 17319
IRL		270		1724			14	471	
ITA		1044		20130			58	1966	
LUX		153		2611				170	0
NLD	12008	939 599	3842	14054 11006	713	1422	29 234		9032
GBR		3235	25962	12903			35	5552	

Receiving Data of EC9, 2002-2006



Sending Data of EC9, 2002-2006



Erf (2007) Ratings of Migration Data for EU9

Country	Receiving			Sending		
	Timing	Completeness	Accuracy	Timing	Completeness	Accuracy
BEL	3	9	9	3	9	9
DNK	2(3)	4(4)	4(4)	3	4	4
FRA	3	2	9			
DEU	2	4	4	2	4	4
IRL	2	2	2	2	2	2
ITA	2(3)	3(3)	3(3)	4	3	3
LUX	2	3	3	2	3	3
NLD	3	4	4	4	4	4
GBR	4	2	2	4	2	2

0:Worst 1:Worse 2:Insufficient 3:Reasonable 4:Good 5:Excellent 9:Unknown

Data Inventory

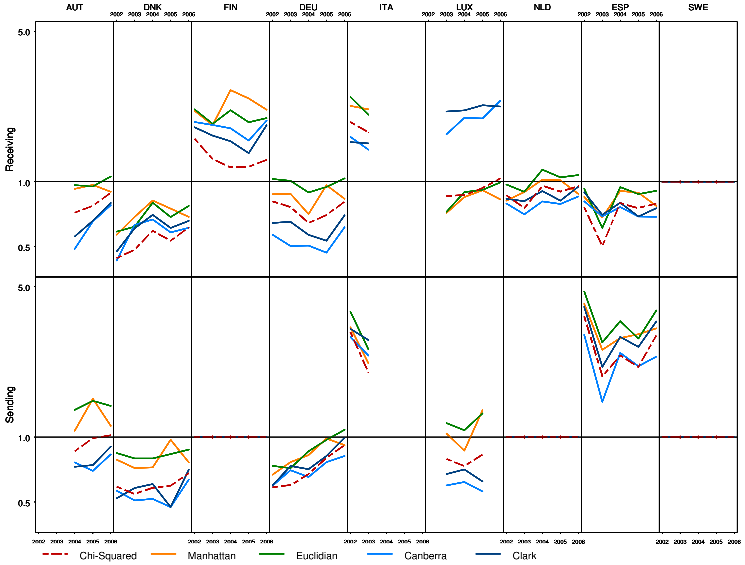
- Detailed literature on these data sources exist
- Can be used to deduce which rows or columns require adjustment good receiving data to 12 month definition of EU9
- Use scaling adjustments estimated from constrained optimization, extending work of Poulain (2007)

$$r_j m_{ij1} = s_i m_{ij2}, \tag{1}$$

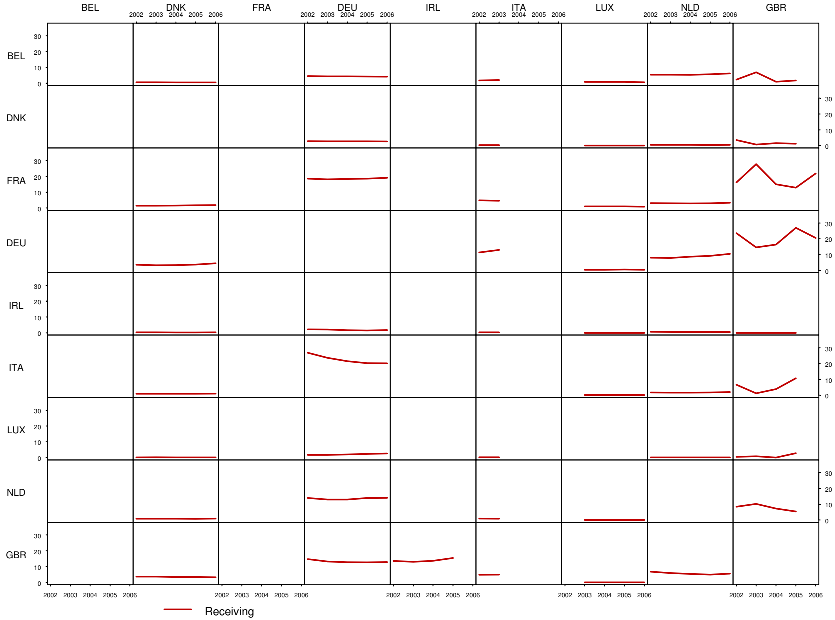
- Different distance measures, analyze variation over time

$$f(r_j, s_i | m_{ijk}) = \sum_{i,j} (r_j m_{ij1} - s_i m_{ij2})^2. \tag{2}$$

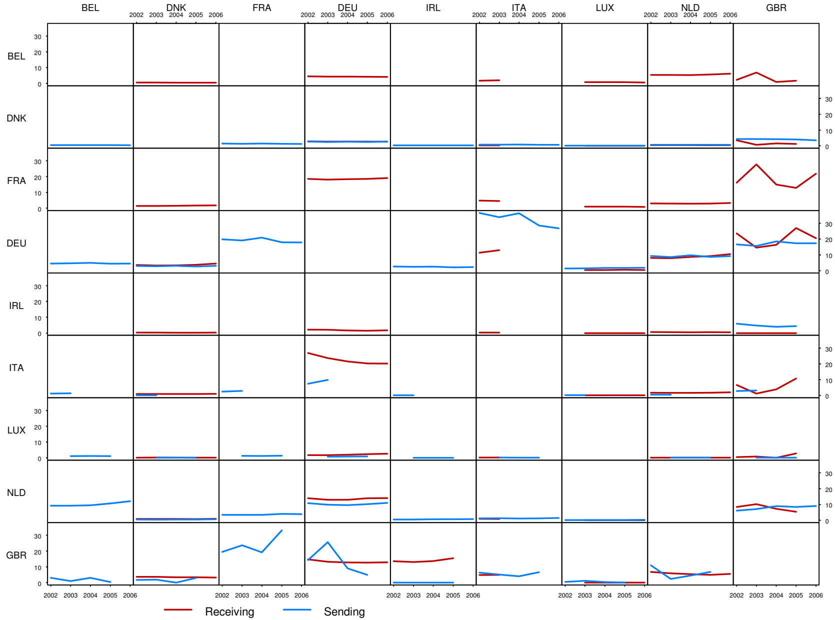
Distance Measures EU15



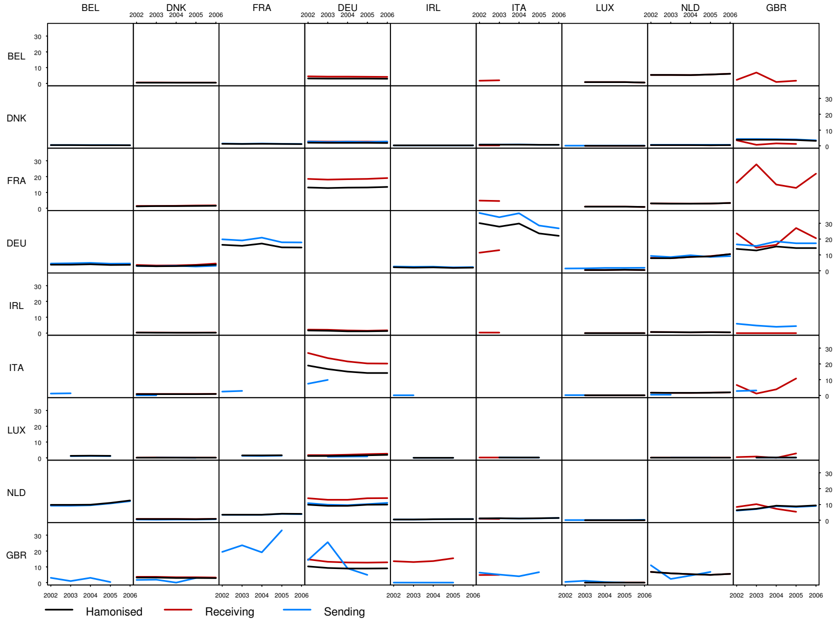
Receiving Data of EC9, 2002-2006



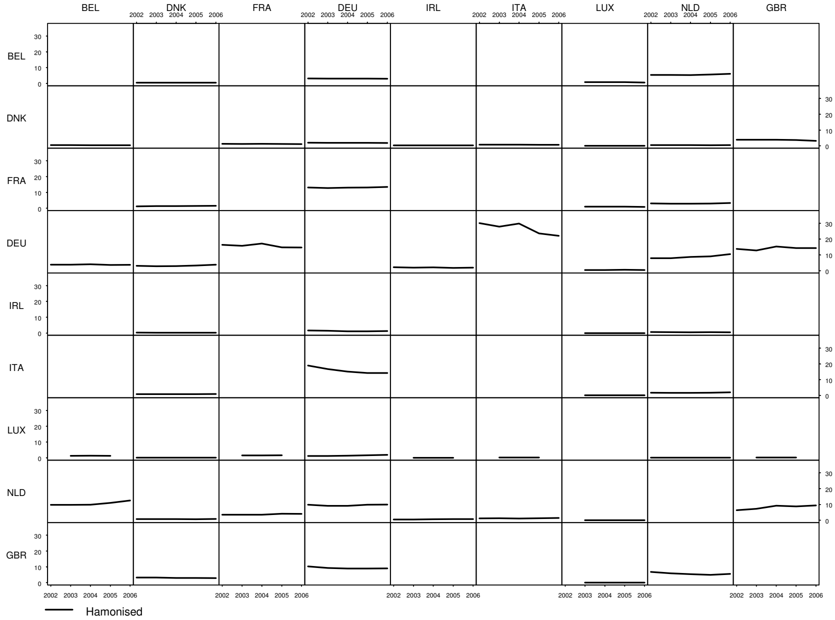
Sending Data of EC9, 2002-2006



Harmonized Flows of EC9, 2002-2006



Harmonized Flows of EC9, 2002-2006



Spatial Interaction Models

- Spatial interaction models have commonly been applied to internal mobility tables
- Willekens (1983) and Flowerdew (1991) showed Poisson regression models with either a row or column dummy covariate are equivalent to origin or destination constrained spatial interaction model

$$\log \mu = \mathbf{X}\beta$$

where $Y \sim Po(\mu)$ and $\beta = (\beta^O, \beta^D \ldots)$

- Poisson regression models fitted to internal migration data still lack a good fit
- Congdon (1991) recommends the use of negative binomial regression models

Additional Covariates

- Expanded basic spatial interaction models to consider factors suggested by literature to influence international migration
- Economic
  1. Difference in Gross Domestic Product per capita
  2. Logarithm of trade volume between each origin and destination
  3. Origin-Destination ratio of unemployment rates
  4. Origin-Destination ratio of rankings in Global Competitiveness Report of World Economic Forum
- Geographical
  1. Logarithm of distance in kilometers between capitol cities
  2. Contiguity
  3. Region covariate
     - European Coal and Steel Community (1957)

Additional Covariates

- Demographic
    1. Logarithm of total populations of origin and destinations
    2. Logarithm of migrant stocks
    3. Language
        - English
        - French
        - German
        - Dutch
- Time treated as continuous to account for correlation of repeated counts over time
- Comparisons of potential models were undertaken using the `stepAIC` function in the MASS library, Venables and Ripley (1999) on observed data

$$AIC = 2k - 2l(\theta|\mathbf{y}), \quad (3)$$

Underlying Distribution

- Assume a spatial interaction model with origin and destination constraints for a response variable, y_{ijt} of n migrations from origin i to destination j at time t, where $i, j = 1, 2, \ldots, r$ for $r = 9$ countries and $t = 5$.

$$E\left(y_{ijt} \left| \theta \right.\right) = \mathbf{X}\beta \qquad i \neq j, \tag{4}$$

where $\theta = \left(\beta_i^O, \beta_j^D, \beta^{GDP}, \ldots\right)$

- Hence $Y \sim NB\left(g\left(\mu\right), a\right)$ where $g\left(\mu\right) = \log \mu = \mathbf{X}\beta$

E and M Steps

- Decide upon some arbitrary initial parameter estimates θ^0

1. The E-Step (expectation step) finds expected fitted values of y_{ijt} given θ^0

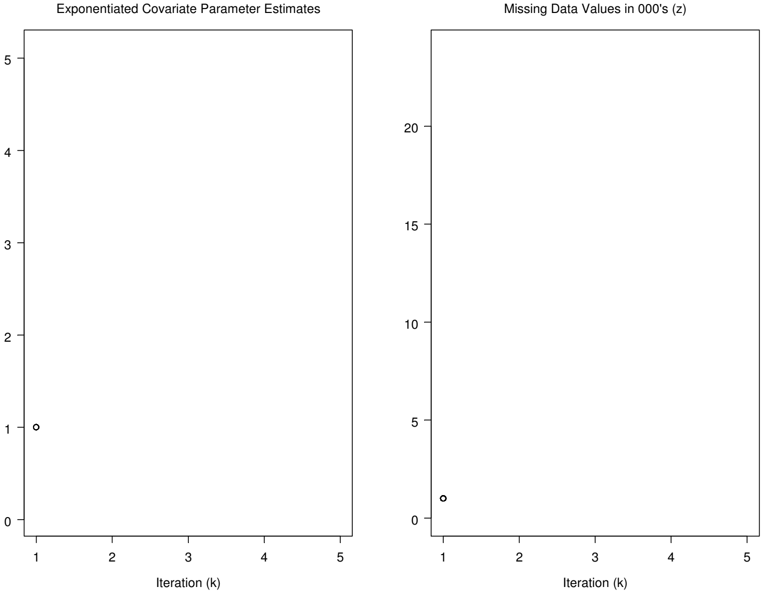
   Use expected fitted value for missing cells, z_{ijt}, to allow the creation of a temporary, but complete explanatory variable.

2. The M-step (maximization step) estimates a new temporary set of model parameters $\theta^1 = (\beta_i^{Ok}, \beta_j^{Dk}, \beta^{GDPk}, \ldots)$

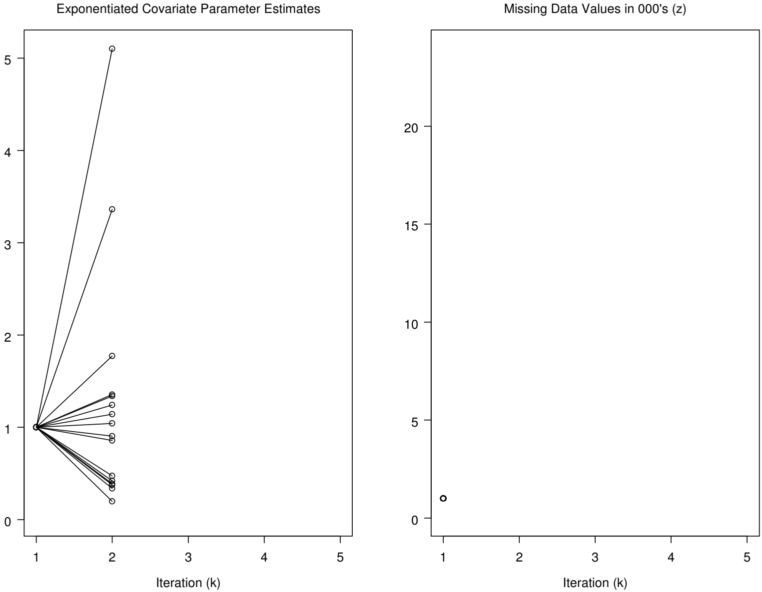
   Can be easily undertaken using the `glm.nb` function in the MASS library of S-Plus/R, Venables and Ripley (1999)

- Re-estimate new expected values of z_{ijt} given θ^1, and so on until convergence in augmented likelihood $\left|\left|Q\left(\theta^{k+1}\left|\theta^k\right.\right) - Q\left(\theta^k\left|\theta^k\right.\right)\right|\right|$

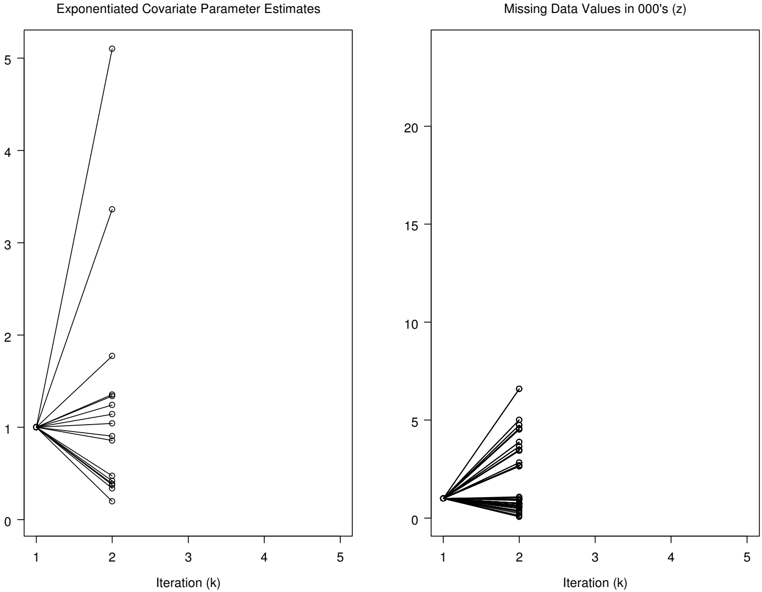
Trace of Model Fit using EM algorithm



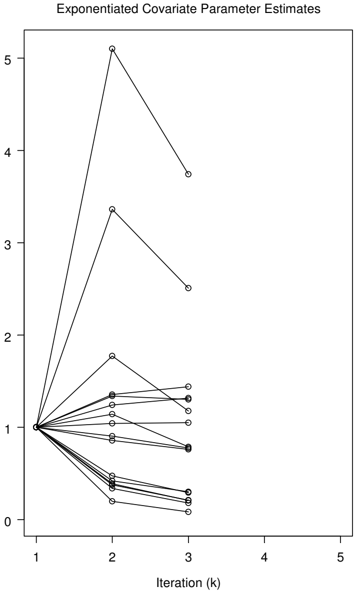
Trace of Model Fit using EM algorithm

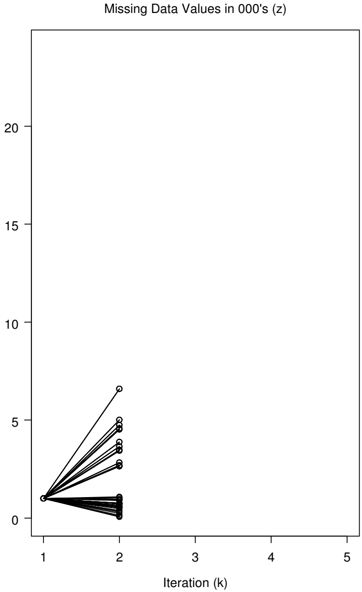


Trace of Model Fit using EM algorithm

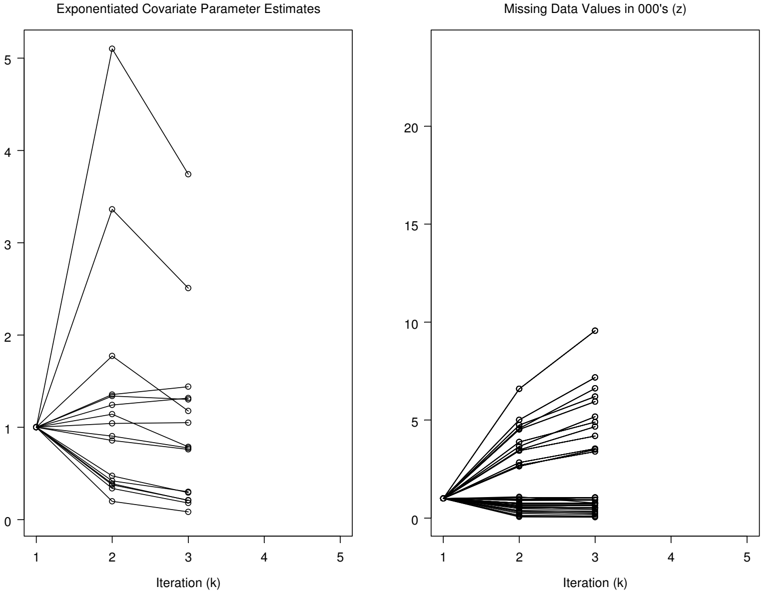


Trace of Model Fit using EM algorithm

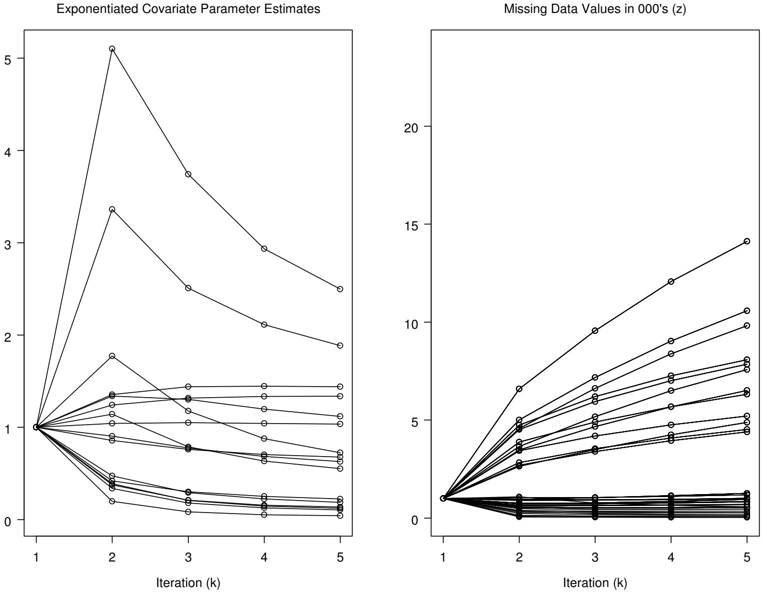




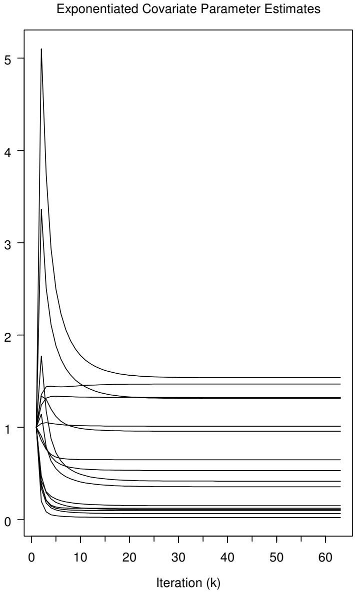
Trace of Model Fit using EM algorithm

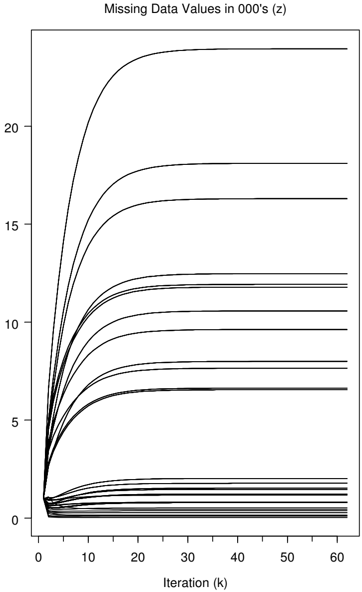


Trace of Model Fit using EM algorithm



Trace of Model Fit using EM algorithm

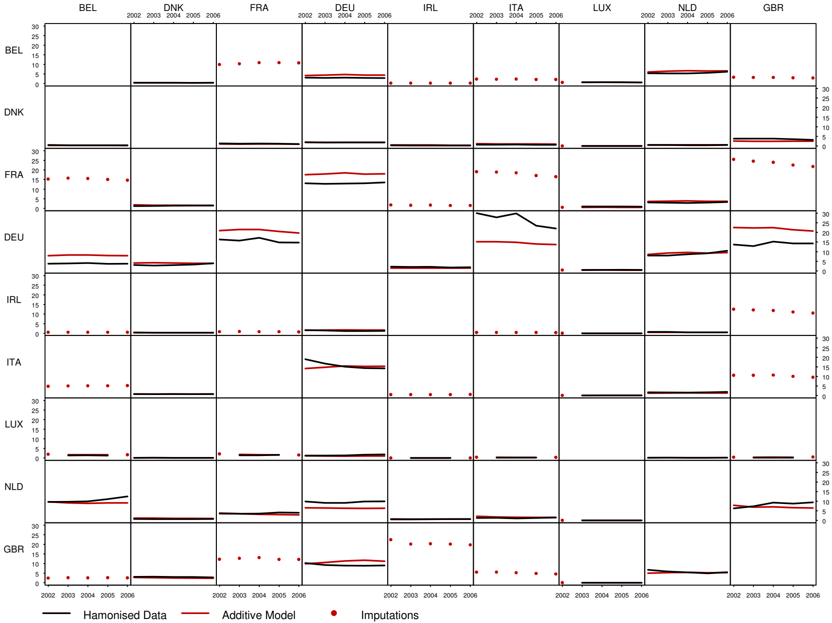




Exponentiated Parameter Estimates of Additive Model

Covariate	β	se(β)
Dispersion	0.0757	1.0189
Time	0.9367	1.0117
log(Trade)	1.6044	1.0385
Unemployment	1.2321	1.0669
Competitiveness	1.1061	1.0153
log(Population)	1.6795	1.0521
log(Stock)	1.3992	1.0309
French	1.8717	1.0904
Dutch	1.9226	1.1236
German	1.7533	1.0694
Contiguity	1.4370	1.0698

Model Fits on Migration Flows ('000s)



Supplemented EM algorithm

- For a single parameter estimate, the EM algorithm describes a mapping $\theta \rightarrow M(\theta)$ from the parameter space of θ, Θ to itself
- For multiple parameters the mappings in the neighborhood of θ^* can be defined as

$$DM = \left(\frac{\partial M_j(\theta)}{\partial \theta_i}\right)\Big|_{\theta=\theta^*} \tag{5}$$

a $d \times d$ Jacobian matrix for $M(\theta) = (M_1(\theta), \ldots, M_d(\theta))$

- The variance covariance matrix can be found by

$$V = I_{OC}^{-1} + \Delta V \tag{6}$$

where $\Delta V = I_{OC}^{-1} DM(I - DM)^{-1}$

Supplemented EM algorithm

- Supplemented EM algorithm of Meng and Rubin (1991)
  1. Run EM algorithm to obtain θ^{t+1}
  2. Calculate $\theta^t(i) = M_j(\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i^{(t)}, \theta_{i+1}^*, \ldots, \theta_d^*,)$ and run one more iterate of EM algorithm
  3. Obtain a single element of DM matrix, r_{ij}

$$r_{ij} = \frac{\partial M_j(\theta)}{\partial \theta_i} \tag{7}$$

$$= \lim_{\theta_t \rightarrow \theta_t^*} \frac{M_j(\theta_1^*, \ldots, \theta_{i-1}^*, \theta_i^{(t)}, \theta_{i+1}^*, \ldots, \theta_d^*,) - M_j(\theta^*)}{\theta_i - \theta_i^*}$$

$$= \lim_{t \rightarrow \infty} \frac{M_j(\theta^t(i)) - \theta_j^*}{\theta_i - \theta_i^*}$$

- Repeat steps 2 and 3 until all of $r_{ij}^{t^*}$, $r_{ij}^{t^*+1}$ for i are stable

Supplemented EM algorithm

- Converged DM matrix used to calculate parameter variance covariance estimates, V

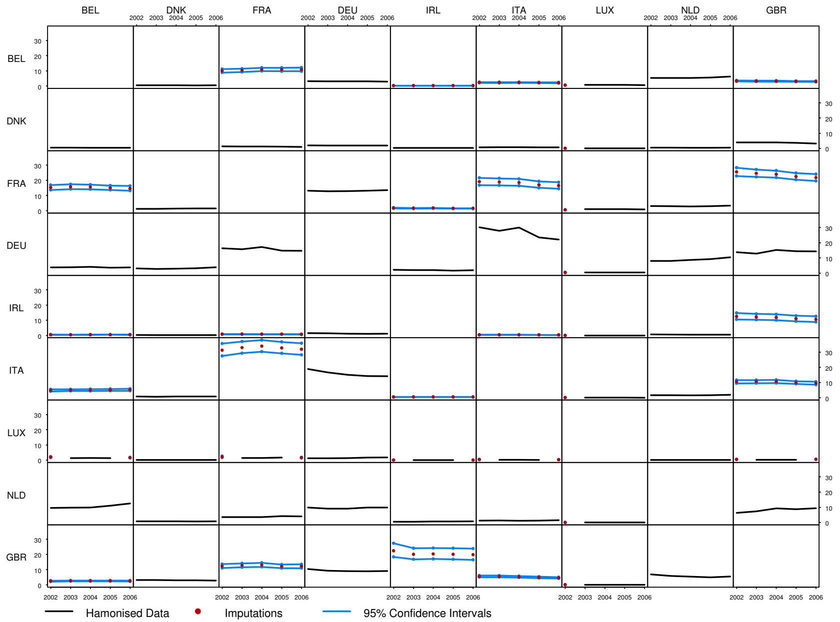
$$V = I_{OC}^{-1} + \Delta V \tag{8}$$

where $\Delta V = I_{OC}^{-1} DM(I - DM)^{-1}$

- Derive confidence intervals for migration flow imputations

$$\log z_{ijt} \pm 1.96 \mathbf{X} V \mathbf{X}^{\mathbf{T}}$$

Standard Errors for Missing Flows ('000s)



Summary

- Formal statistical framework for creating complete flow tables given some harmonized international migration flows
- Imputed missing values where no satisfactory data existed
- EM algorithm successfully fitted models with missing data and provide imputations
- Provide standard errors of missing flow estimates
- $$y_{ijt1}|r_j, \alpha, \boldsymbol{\beta}, \mathbf{x}_i^T \sim NB(r_j \mu_{ijt}, \alpha)$$
  $$y_{ijt2}|s_i, \alpha, \boldsymbol{\beta}, \mathbf{x}_i^T \sim NB(s_i \mu_{ijt}, \alpha)$$
  where $\log \mu_{ijt} = \mathbf{x}_i^T \boldsymbol{\beta}$.
- This allows the joint posterior distribution for all parameters,

$$p(s_i, r_j, \alpha, \boldsymbol{\beta}|y_{ijt1}, y_{ijt2}, \mathbf{x}_i^T) = p(r_j)p(s_i)p(\alpha)p(\boldsymbol{\beta}) \, p(y_{ijt1}|r_j, \alpha, \boldsymbol{\beta}, \mathbf{x}_i^T)p(y_{ijt2}|s_i, \alpha, \boldsymbol{\beta}\mathbf{x}_i^T)$$