Working Paper 2013:5

Department of Economics
School of Economics and Management

Using Lasso-Type Penalties to Model Time-Varying Covariate Effects in Panel Data Regressions -

A Novel Approach Illustrated by the 'Death of Distance' in International Trade

Wolfgang Hess
Maria Persson
Stephanie Rubenbauer
Jan Gertheiss

February 2013

LUND UNIVERSITY

Using Lasso-Type Penalties to Model Time-Varying Covariate Effects in Panel Data Regressions – A Novel Approach Illustrated by the 'Death of Distance' in International Trade*

Wolfgang Hess† Maria Persson‡ Stephanie Rubenbauer§ Jan Gertheiss§

February 18, 2013

Abstract

When analyzing panel data using regression models, it is often reasonable to allow for time-varying covariate effects. We propose a novel approach to modelling time-varying coefficients in panel data regressions, which is based on penalized regression techniques. To illustrate the usefulness of this approach, we revisit the well-known empirical puzzle of the 'death of distance' in international trade. We find significant differences between results obtained with the proposed estimator and those obtained with 'traditional' methods. The proposed method can also be used for model selection, and to allow covariate effects to vary over other dimensions than time.

Keywords: Penalized Regression; Lasso-type Penalties; Varying Coefficient Models; Gravity; Death of Distance; Missing Globalization Puzzle.

JEL Classification: C23; C52; F10.

* An earlier version of this paper was presented at the European Trade Study Group (ETSG) Thirteenth Annual Conference, Copenhagen, Denmark, September 2011. Financial support from the Jan Wallander and Tom Hedelius Foundation under research grant numbers W2009-0352:1 (Persson) and W2010-0305:1 (Hess) is gratefully acknowledged.

†Department of Economics and Centre for Economic Demography, Lund University, Sweden

‡Department of Economics, Lund University, and Research Institute of Industrial Economics (IFN), Sweden

§Department of Statistics, Ludwig-Maximilians-Universität Munich, Germany

1 Introduction

Over the past decades, the increasing availability of panel data sets has triggered a rapid development of econometric tools to efficiently exploit the information contained therein (see e.g. Arellano, 2003, Hsiao, 2003, and Baltagi, 2008, for extensive overviews of these methods). When analyzing panel data using regression models, the question arises whether the data can be pooled or not. Typically, it is asked whether the data can be pooled across different cross-sectional units such as persons, firms, or countries (see e.g. Baltagi *et al.*, 2008, and references therein). Less attention has been paid to poolability over time. In many empirical studies it is simply assumed that regression coefficients do not vary over time. However, the panels available often cover rather long time periods, introducing the question of whether it is reasonable to expect that the effects of explanatory variables remain constant over time. Therefore, we propose a novel approach to modelling time-varying covariate effects in panel data regressions. In doing so, we envision scenarios where the cross-sectional dimension is larger than the temporal dimension.

An obvious solution to allowing variables' effects to vary flexibly over time is to simply incorporate interactions of covariates and time dummies into the regression model. This simple solution leads to certain problems, though, the most important one being over-fitting of the model, i.e. overly wiggly and hard-to-interpret covariate effects. We therefore propose using penalized regressions in this context. Specifically, the basic idea of our approach is to incorporate flexible interactions of covariates and time into the regression model, and then penalize the differences between adjacent coefficients.

This approach has several virtues. First, the flexible interactions of explanatory variables and time allow for covariate effects that vary freely over time without being restricted by parametric assumptions. Yet, the penalization of differences between adjacent coefficients avoids the problem of over-fitting. Second, our proposed method is rather flexible with respect to the type of penalization used. We will focus on two types of penalties: the group lasso (least absolute shrinkage and selection operator) and the fused lasso. The former predominantly produces covariate effects that vary rather smoothly over time, and the latter allows for piecewise constant covariate effects that may exhibit distinct 'jumps' at particular points in time. Which of these ways for coefficients to change over time that is most empirically reasonable depends on the particular application. Third, penalties may not only be imposed on differences between adjacent (time-varying) coefficients but also on other coefficients. Since lasso-type penalties can shrink coefficients to be exactly zero, our proposed approach can therefore also be used for model selection (in fact, as the term 'selection operator' indicates, this is the original purpose of lasso-type penalties). Fourth, our approach can be applied to the broad class of generalized linear models (GLMs), which constitutes the most widely used framework in applied econometrics. Fifth and last, our approach can be implemented using standard software, and the only thing required for model estimation is an adequate preparation of the data set. This makes the proposed methodology particularly useful for applied researchers who look for a flexible,

yet practically manageable way to estimate time-varying effects in panel data models.

To illustrate the usefulness of the proposed methodology, we revisit a well-known empirical puzzle in international trade: the so-called 'death of distance'. Researchers in international economics have for quite some time discussed that due to the falling costs of transportation, distance – which is one of the key variables for explaining variations in the size of bilateral trade flows – should become less important as a trade barrier over time (see e.g. Cairncross, 1997). Interestingly, however, conventional wisdom among empirical researchers investigating this phenomenon is that distance, if anything, becomes *more* important over time (see e.g. Disdier and Head, 2008, and references therein). We therefore apply our penalized regression framework to estimate the well-known gravity model of international trade, allowing the effects of distance to vary over time. Unlike the majority of previous studies, we do not find a temporal trend in the distance effect.

The major goal of this paper is to introduce a novel approach to modelling time-varying covariate effects in panel regressions. To the best of our knowledge, penalized regression techniques have up to now not been used to model time-varying coefficients in a panel data context. Moreover, since penalized regression in general has only been discussed rather sparingly in the economics literature, we aim to make this method more accessible to the empirical economic researcher. To assist empirical researchers who wish to employ this methodology, we have included step-by-step instructions on how to prepare the data set and perform the regressions in the Appendix. Our empirical analysis is interesting in its own right, and we intend to also make a contribution to the field of international trade. As mentioned above, the 'death of distance' has been extensively discussed in the trade literature. It is therefore noteworthy that our empirical findings differ from the majority of existing empirical results based on standard gravity models.

Despite the focus on the panel gravity model, the method we propose can be used in a broad range of economic applications where the effects of explanatory variables can be expected to vary over time. Moreover, our proposed method can also be used to allow covariate effects to vary over other dimensions than calendar time. The only requirement is that these dimensions have a natural ordering. A short list of examples includes duration time in (discrete-time) hazard models, an individual's age or income in micro panel studies, and the size of geographical units (countries, states, etc.) in macro panel studies.

The remainder of the paper is organized as follows. Section 2 introduces our proposed method. It shows how penalized estimation can be used to estimate time-varying covariate effects in panel data models, and provides a detailed discussion of the lasso-type penalties that we propose. Section 3 contains the empirical application of our proposed method to the 'death of distance' in international trade. It gives a brief overview of the existing literature in that field, and provides a thorough discussion of our empirical findings. Section 4 contains concluding remarks.

2 Modelling Time-Varying Coefficients in Panel Regressions

Let y_{it} denote the outcome of a specific cross-sectional unit i ($i = 1, \ldots, N$) in period t ($t = 1, \ldots, T$), and let $u_{it} = (u_{1,it}, \ldots, u_{m,it})^\top$ be a vector of realizations of explanatory variables that may vary over time.[1] In GLMs it is assumed that the density of y_{it} belongs to the linear exponential family (LEF) of densities, i.e. $f(y_{it}) = \exp\{a(\mu_{it}) + b(y_{it}) + c(\mu_{it})y_{it}\}$, with conditional mean $\mu_{it} = E(y_{it}|u_{it})$. Standard GLMs specify μ_{it} to be of the single-index form

$$\mu_{it} = g(\alpha + u_{it}^\top \varphi), \tag{1}$$

with intercept α, parameter vector $\varphi = (\varphi_1, \ldots, \varphi_m)^\top$, and response function $g(\cdot)$ that varies across models, depending for example on restrictions on the range of y_{it}. Intercept α and parameter vector φ are typically estimated using (quasi-)maximum likelihood techniques. The log-likelihood to be maximized is then

$$l(\alpha, \varphi) = \ln \mathcal{L}(\alpha, \varphi) = \sum_{i=1}^{N} \sum_{t=1}^{T} a(g(\alpha + u_{it}^\top \varphi)) + b(y_{it}) + c(g(\alpha + u_{it}^\top \varphi))y_{it}. \tag{2}$$

The quasi-maximum likelihood estimator (MLE) maximizes this log-likelihood, but it is no longer assumed that the LEF density is correctly specified (see e.g. Cameron and Trivedi, 2005, Ch. 5.7). However, even with a misspecified LEF density, the quasi-MLE is consistent provided that $E(y_{it}|u_{it}) = g(\alpha + u_{it}^\top \varphi)$ (see Gourieroux *et al.*, 1984, for a proof). The class of GLMs depicted above is widely applicable, since it includes many popular models such as Poisson, logit, probit, exponential, or linear regression models as special cases.

In model (1) it is assumed that regression coefficients $\varphi_1, \ldots, \varphi_m$ do not vary over time. Since this is a very strong assumption, we may relax it by writing

$$\mu_{it} = g(\eta_{it}), \tag{3}$$

where

$$\eta_{it} = \alpha_t + \sum_{j=1}^{p} x_{j,it} \cdot \beta_j + \sum_{l=1}^{q} z_{l,it} \cdot \gamma_{l,t},$$

with $p + q = m$. This implies that intercept α_t and coefficients in $\gamma_t = (\gamma_{1,t}, \ldots, \gamma_{q,t})^\top$ are allowed to vary with time t, whereas $\beta = (\beta_1, \ldots, \beta_p)^\top$ is time-invariant. In other words, $x_1, \ldots, x_p$ are the covariates (from $\{u_1, \ldots, u_m\}$) that are restricted to have constant effects across time, and $z_1, \ldots, z_q$ denote the covariates (from $\{u_1, \ldots, u_m\}$) that are allowed to exhibit time-varying effects. Model (3), however, has a lot of parameters, and estimating them using conventional (quasi-)maximum likelihood techniques may lead to instable results. The resulting wiggly coefficient vectors may then be hard to interpret in an economically meaningful manner. Furthermore, some coefficients in γ_t may actually be time-invariant. So on the one hand, the aim of the analysis should be to determine

[1] For notational convenience, we focus on the balanced panel data case. However, as discussed below, unbalanced data can be analyzed accordingly.

which γ-coefficients are time-varying, and which are not. On the other hand, time-varying coefficients should be estimated in an adequate fashion. In order to accomplish this, we propose the use of penalized estimation techniques.

2.1 Penalized Estimation of Time-Varying Coefficients

In empirical applications it is often reasonable to assume that covariate effects do not vary erratically but rather smoothly over time. This implies that adjacent coefficients $\gamma_{l,t}$ and $\gamma_{l,t-1}$ can be expected to be similar or, equivalently, that differences $\delta_{l,t} = \gamma_{l,t} - \gamma_{l,t-1}$ should be small. Therefore, we propose to not maximize the (quasi-)log-likelihood

$$l(\alpha, \beta, \gamma) = \ln \mathcal{L}(\alpha, \beta, \gamma) = \sum_{i=1}^{N} \sum_{t=1}^{T} a(g(\eta_{it})) + b(y_{it}) + c(g(\eta_{it}))y_{it},$$

but its penalized version

$$l_p(\alpha, \beta, \gamma) = l(\alpha, \beta, \gamma) - \lambda J(\gamma), \tag{4}$$

where penalty $J(\gamma)$ penalizes differences between adjacent γ-parameters, $\gamma = (\gamma_1^\top, \ldots, \gamma_T^\top)$. The strength of penalization (and hence the smoothness) is controlled by tuning parameter $\lambda \geq 0$.[2] A particular virtue of this approach is that a variety of penalties $J(\gamma)$ with differing properties can be employed when maximizing (4). A concrete penalty that effects smoothness is, for example, given by

$$J(\gamma) = \sum_{l=1}^{q} \sum_{t=2}^{T} (\gamma_{l,t} - \gamma_{l,t-1})^2 = \sum_{l=1}^{q} \sum_{t=2}^{T} \delta_{l,t}^2. \tag{5}$$

From (4) and (5), the intuition behind the penalization approach becomes obvious. Using a λ-value strictly greater than zero and squared differences of adjacent γ-parameters, large parameter differences have a negative impact on the target function $l_p(\alpha, \beta, \gamma)$ that is to be maximized. Thus, estimated parameter differences will be smaller than they would have been in standard models without penalization.[3] However, for any fixed $\lambda < \infty$, the asymptotic properties of the estimator are not affected by the penalty. As the penalized estimates $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$ maximize $l_p(\alpha, \beta, \gamma)$ at (4), they also maximize

$$\frac{1}{N} l_p(\alpha, \beta, \gamma) = \frac{1}{N} l(\alpha, \beta, \gamma) - \frac{\lambda}{N} J(\gamma).$$

Now, assuming that T is fixed as $N \to \infty$, the penalty term $\lambda J(\gamma)/N$ vanishes, but the ordinary log-likelihood term $l(\alpha, \beta, \gamma)/N$ does not (with probability one). Hence, the penalized estimates tend (almost surely) towards the non-penalized estimates obtained if the usual log-likelihood $l(\alpha, \beta, \gamma)$ is maximized. Consequently, for any given λ, the penalized

[2] With $\lambda = 0$ there is no penalization, and $l_p(\alpha, \beta, \gamma)$ is, of course, equivalent to $l(\alpha, \beta, \gamma)$.

[3] From (4) and (5), it should also be clear that covariates should be standardized to have equal variance in order to avoid that coefficient values, and thus the strength of penalization, are scale dependent. However, as discussed below in Section 2.1.3, there may be instances where varying strengths of penalization are desired. In this case, covariates can be scaled to have different variances.

estimator has the same asymptotic properties as the conventional MLE. In particular, if the latter is consistent, the penalized estimator is consistent, too.

By penalizing squared differences of adjacent γ-parameters, as in (5), large shifts in parameter values are avoided (see e.g. Gertheiss and Tutz, 2009). However, by using (5), it is not possible to distinguish between γ-coefficients that are actually varying across time and those that are not. To see this, recall that a time-constant γ_l implies that $\gamma_{l,1} = \gamma_{l,2} = \ldots = \gamma_{l,T}$; in other words, $\delta_{l,t} = 0$ for all $t = 2, \ldots, T$. When using (5), estimated γ-coefficients are only set equal for the limit case $\lambda \to \infty$, and in this case, γ-coefficients are fit as time-constant for all $l = 1, \ldots, q$. To be able to discriminate between time-varying and time-invariant coefficients, a penalty is needed so that for some $l \in \{1, \ldots, q\}$ the entire group of coefficients $\{\gamma_{l,1}, \ldots, \gamma_{l,T}\}$ is set equal, whereas coefficients for the remaining l are left time-varying. Of course, this group-wise selection should be done in a data-driven way, and an adequate penalty for that purpose is the so-called group lasso (Yuan and Lin, 2006).

2.1.1 Group Lasso

The group lasso

$$J(\gamma) = \sum_{l=1}^{q} \sqrt{\sum_{t=2}^{T} (\gamma_{l,t} - \gamma_{l,t-1})^2} = \sum_{l=1}^{q} \sqrt{\sum_{t=2}^{T} \delta_{l,t}^2} \tag{6}$$

is a modification of the original lasso (Tibshirani, 1996) that allows for group-wise selection of covariates. In order to maximize (4), the group lasso will shrink parameter differences $\delta_{l,t}$, thereby generating smooth time variations in the effects of z-covariates. For a large enough value of λ, the group lasso will (simultaneously) force the whole group of parameter differences $\{\delta_{l,2}, \ldots, \delta_{l,T}\}$ to be zero (see e.g. Yuan and Lin, 2006, or Gertheiss *et al.*, 2011), implying that the effect of z_l is constant over time. For λ-values larger than a distinct value $\lambda_{\max}$, differences $\delta_{l,t}$ will be set to zero for all l (and t), implying that no covariate has time-varying effects. If also β-coefficients are to be penalized, the group lasso penalty term in (6) can be extended to include β-parameters. The result of this is that x-covariates can be excluded from the model in a data-driven way, which is a further virtue of this approach. Of course, this model selection property of the group lasso can also be applied to the time-varying z-covariates. For that purpose, the coefficients $\gamma_{l,1} = \delta_{l,1}$ $(l = 1, \ldots, q)$ have to be included in the penalty term in (6) as separate groups of size one. This way, it can be determined whether a particular covariate effect is time-varying, time-constant, or irrelevant.[4] Note that using this penalization approach allows the researcher to not only assess the statistical significance of covariates and their time interactions but also their (economic) importance. In other words, if covariates or their time interactions contribute relatively little to the maximization of the (quasi-)likelihood, they may be removed from the model even if they are statistically significant by common standards. When working

[4]Note that this may lead to the somewhat hard-to-interpret scenario where a covariate is excluded from the model but its time interactions are not. In practice, however, this is very unlikely to happen.

with very large panel data sets, where estimates tend to be significant at all common significance levels, assessing the relative importance of explanatory variables may be very fruitful. It may help in finding parsimonious model specifications without jeopardizing the model's explanatory power.

For computing estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$, algorithms proposed for the group lasso in GLMs (Meier *et al.*, 2008), and implemented in the R package `grplasso` (Meier, 2009), can be used. Thus, our proposed method can be readily applied to a broad class of models by using freely available standard software. The only thing required for this software to be applicable is a re-parametrization of model (3) using parameters $\delta_{l,t}$ instead of $\gamma_{l,t}$. With $\delta_{l,1} = \gamma_{l,1}$ and $\delta_{l,t} = \gamma_{l,t} - \gamma_{l,t-1}$ (for $t \geq 2$), we have $\gamma_{l,t} = \delta_{l,1} + \ldots + \delta_{l,t}$, and hence

$$\begin{aligned} \eta_{it} &= \alpha_t + \sum_{j=1}^{p} x_{j,it} \cdot \beta_j + \sum_{l=1}^{q} z_{l,it} \cdot \sum_{k=1}^{t} \delta_{l,k} \\ &= \alpha_t + \sum_{j=1}^{p} x_{j,it} \cdot \beta_j + \sum_{l=1}^{q} \sum_{k=1}^{T} \tilde{z}_{l,ik} \cdot \delta_{l,k}, \end{aligned}$$

with

$$\tilde{z}_{l,ik} = \begin{cases} z_{l,it} & \text{if } k \leq t, \\ 0 & \text{otherwise.} \end{cases}$$

If also the time-varying intercept α_t is to be smoothed, it can be re-parameterized analogously by using $\nu_1 = \alpha_1$ and $\nu_t = \alpha_t - \alpha_{t-1}$ (for $t \geq 2$). Thus, it can be easily included in the group lasso penalty by using

$$J(\alpha, \gamma) = \breve{J}(\nu, \delta) = \sqrt{\sum_{t=2}^{T} \nu_t^2} + \sum_{l=1}^{q} \sqrt{\sum_{t=2}^{T} \delta_{l,t}^2}.$$

Unlike $\delta_{l,1}$, however, ν_1 must not be included in the penalty, i.e. the global intercept must not be penalized. The Appendix provides a detailed description of how the above re-parametrization and the required scaling of covariates can be performed in practice.

2.1.2 Fused Lasso

In some instances it may be reasonable to assume that coefficients $\gamma_{l,t}$ are piecewise constant over time. In that case, the influence of covariate z_l does not vary across most of the time periods $t = 1, \ldots, T$, but for some distinct time points $t_1^* < t_2^* < \ldots$ the effect changes, for example due to some external events. When estimating the regression coefficients, these breakpoints, or 'jumps', in the coefficient function should be identified. For simultaneous estimation of coefficients and identification of jumps, fused lasso-type penalties (Tibshirani *et al.*, 2005) can be used. The penalty term in (4) is then specified as

$$J(\gamma) = \sum_{l=1}^{q} \sum_{t=2}^{T} |\gamma_{l,t} - \gamma_{l,t-1}| = \sum_{l=1}^{q} \sum_{t=2}^{T} |\delta_{l,t}|. \tag{7}$$

Using this L_1-penalty, separate differences $\delta_{l,t}$ of adjacent coefficients $\gamma_{l,t}$ and $\gamma_{l,t-1}$ can be fitted as exactly zero, thus yielding piecewise constant coefficient profiles and selecting relevant jumps. Just like the group lasso, the fused lasso penalty can be extended to include parameters β_j, $\gamma_{l,1} = \delta_{l,1}$, and ν_t (for $t \geq 2$). Since, for groups of size one (i.e. separate differences $\delta_{l,t}$), the L_1-norm in (7) is equivalent to the L_2-norm in (6), existing software for computing group lasso solutions (see Section 2.1.1) can also be employed to estimate models with fused lasso-type penalties.[5]

2.1.3 Adaptive Penalties

For improving the estimation and selection performance when the number of observations becomes large, so-called adaptive penalties have been proposed (see e.g. Zou, 2006, Zhang and Lu, 2007, Wang and Leng, 2008, Meier *et al.*, 2009, or Gertheiss and Tutz, 2012). The decisive modification is to weight the penalty terms by the inverse of the respective unpenalized parameter estimates. For the group lasso penalty (6), for example, we obtain the adaptive version

$$J(\gamma) = \sum_{l=1}^{q} w_l \|\delta_l\|, \tag{8}$$

with weights $w_l = \|\tilde{\delta}_l\|^{-1}$ based on the unpenalized estimates $\tilde{\delta}_l = (\tilde{\delta}_{l,2}, ..., \tilde{\delta}_{l,T})^\top$, and the L_2-norm $\|\delta_l\| = \sqrt{\sum_{t=2}^{T} \delta_{l,t}^2}$. The adaptive fused lasso penalty is simply

$$J(\gamma) = \sum_{l=1}^{q} \sum_{t=2}^{T} w_{l,t} |\delta_{l,t}|, \tag{9}$$

with $w_{l,t} = |\tilde{\delta}_{l,t}|^{-1}$.

The intuition behind this weighting procedure is rather straightforward. With very large data sets, unpenalized point estimates can be expected to be rather accurate. Thus, if the unpenalized estimates of parameter differences $\tilde{\delta}_l = (\tilde{\delta}_{l,2}, ..., \tilde{\delta}_{l,T})^\top$ are large, the time variations in the respective covariate effects can be expected to be of significance. Consequently, the corresponding penalization should be small. And this is exactly what is achieved by using $\|\tilde{\delta}_l\|^{-1}$ or $|\tilde{\delta}_{l,t}|^{-1}$, respectively, as weights in the penalty term.

2.2 Unbalanced Panels

So far, for notational convenience, we have focused on balanced panel data. However, it is often the case in practice that not every unit i is observed for all $t \in \{1, \ldots, T\}$. In fact, data on trade volumes, as considered in the empirical part of this paper, are almost always unbalanced. Unbalanced data, however, do not constitute a problem, and regression coefficients can still be estimated as long as the number of missing combinations of i and t is not too large. In fact, if penalty (6) is used, the model remains uniquely

[5] However, if the penalty consists exclusively of fused lasso-type terms, more efficient algorithms (such as the path algorithm by Park and Hastie, 2007) exist.

identified even if there are no observations at all for some $t \in \{1, \ldots, T\}$. In that case, the coefficient estimates $\hat{\delta}_{l,t}$ for time periods without observations would be obtained through linear interpolation of the estimated adjacent coefficients.

2.3 Penalized (Quasi-)Poisson Models

Since the empirical part of this paper is concerned with the estimation of gravity models for trade, and since (quasi-)Poisson regression is the recommended tool for estimating such models (see e.g. Santos Silva and Tenreyro, 2006, or Westerlund and Wilhelmsson, 2009), this section provides a detailed presentation of penalized (quasi-)Poisson models as a special case of the penalized GLMs discussed above.

Let the volume of trade y_{it} between any pair of countries i at time t be modelled as

$$y_{it} = \exp(\eta_{it}) + \upsilon_{it}, \tag{10}$$

where, as before,

$$\eta_{it} = \alpha_t + \sum_{j=1}^{p} x_{j,it} \cdot \beta_j + \sum_{l=1}^{q} z_{l,it} \cdot \gamma_{l,t},$$

and it is merely assumed that $E(\upsilon_{it}|x_{it}, z_{it}) = 0$. Hence, we have

$$\mu_{it} = E(y_{it}|x_{it}, z_{it}) = \exp(\eta_{it}). \tag{11}$$

For estimating regression coefficients in model (11), we follow McCullagh and Nelder (1989). We specify a variance function $Var(y_{it}|x_{it}, z_{it}) = v(\mu_{it})$, and solve the so-called generalized estimation equations. With the specification $v(\mu_{it}) = \mu_{it} = \exp(\eta_{it})$, this leads to solving

$$\begin{aligned} \sum_{i,t} [y_{it} - \exp(\eta_{it})] x_{it} &= 0, \\ \sum_{i} [y_{i1} - \exp(\eta_{i1})] z_{i1} &= 0, \\ &\vdots \\ \sum_{i} [y_{iT} - \exp(\eta_{iT})] z_{iT} &= 0, \end{aligned} \tag{12}$$

with $x_{it} = (x_{1,it}, \ldots, x_{p,it})^\top$ and $z_{it} = (1, z_{1,it}, \ldots, z_{p,it})^\top$ (the 1 accounts for intercept α_t). A desirable feature of this estimator is that it is consistent even if the variance function $v(\mu_{it})$ is misspecified. Only the conditional mean $\mu_{it} = \exp(\eta_{it})$ has to be specified correctly (see e.g. Gourieroux *et al.*, 1984, or McCullagh and Nelder, 1989). Furthermore, the estimator defined by (12) is equivalent to the (quasi-)MLE based on the Poisson log-likelihood. In other words, the estimator defined by (12) is also obtained by maximizing the Poisson (log-)likelihood over α-, β-, and γ-coefficients from model (11). Therefore, a penalized version of this estimator can be obtained by maximizing the penalized log-likelihood

$$l_p(\alpha, \beta, \gamma) = \ln \mathcal{L}(\alpha, \beta, \gamma) - \lambda J(\gamma), \tag{13}$$

where for $\mathcal{L}$ the Poisson likelihood $\mathcal{L} = \sum_{i,t} y_{it} \ln(\mu_{it}) - \mu_{it} + \text{const}$ is used. For $J(\gamma)$, any penalty discussed in Section 2.1 can be used.

2.4 Tuning Parameter Selection

A remaining task in practice is the selection of the tuning parameter λ, which determines the strength of penalization. A common approach for selecting tuning parameters is K-fold cross-validation. For this, the data are (randomly) split into K (roughly) equal-sized parts. Then, for each part $k = 1, \ldots, K$, the model is fit to the remaining $K - 1$ parts of the data, and the prediction error of the fitted model is calculated when predicting the outcome variables y_{it} of the kth sub-sample. Lastly, the K estimates of prediction error are combined (see Hastie *et al.*, 2009), and the resulting measure of prediction error is minimized as a function of the tuning parameter(s) of interest.

One way to evaluate the predictive performance of a model is to use the cumulative (cross-validated) deviance. The deviance for observation i at time t is defined as

$$D_{it} = -2 \left(\ln \mathcal{L}(\hat{\mu}_{it}) - \ln \mathcal{L}(y_{it}) \right), \tag{14}$$

with $\mathcal{L}(\hat{\mu}_{it})$ denoting the likelihood at the estimated (conditional) mean of response y for observation i at time t, and $\mathcal{L}(y_{it})$ being the likelihood if the observed value y_{it} is plugged in instead of $\hat{\mu}_{it}$. In general, the likelihood is maximized if the observed values y_{it} are plugged in, and it should be as large as possible for fitted values $\hat{\mu}_{it}$. Hence we have $D_{it} \geq 0$ for all i, t, and minimizing the cumulative (cross-validated) deviance leads to an adequate λ-parameter.

Another way to determine λ is the rule that the same criterion should be used as has been used to estimate model parameters α, β and γ. When fitting the (quasi-)Poisson model, we use the Poisson likelihood to compute (14). This leads to the so-called quasi-deviance for the general model (11)

$$Q_{it} = -2(y_{it} \ln(\hat{\mu}_{it}) - \hat{\mu}_{it} - (y_{it} \ln(y_{it}) - y_{it})). \tag{15}$$

However, since the term $(y_{it} \ln(y_{it}) - y_{it})$ in (15) does not depend on λ, it can be neglected when minimizing Q_{it}. Consequently, we will use the adjusted quasi-deviance

$$Q^*_{it} = -2(y_{it} \ln(\hat{\mu}_{it}) - \hat{\mu}_{it}) \tag{16}$$

to determine λ in the penalized (quasi-)Poisson model.

3 Empirical Application: The 'Death of Distance' in International Trade

Within the field of international economics, one of the most stable empirical relationships is captured by the gravity model. In this model, bilateral trade between two countries is to a large extent explained by the size of the two countries' economies and the distance

between them. Since the latter variable is typically thought to capture transport costs for shipping goods from the exporter to the importer, it has been a popular prediction that falling transport and communication costs would lead to the 'death of distance' (see e.g. Cairncross, 1997). In other words, the importance of distance as an impediment to trade is expected to decrease over time. At the same time, conventional wisdom among researchers applying gravity models has, to the contrary of this belief, been that distance, if anything, becomes *more* important over time. For instance, Carrère and Schiff (2005) summarize the gravity literature by stating that most gravity model estimations "find that the negative impact of distance on bilateral trade increases over time". In a similar way, Brun *et al.* (2005), note that "when the model is estimated separately for several years, the absolute value of the coefficient almost always increases over time".

The discrepancy between theoretical predictions and empirical findings regarding the historical evolution of the distance effects in international trade makes the issue an important research puzzle with a strong policy relevance. Furthermore, from a methodological perspective, the issue is a suitable example of a research question where our proposed methodology offers clear advantages compared to traditional methods. Empirical research on international trade is typically carried out using very large data sets, implying that most variables will become statistically significant even though they may not necessarily be economically relevant. Our approach is therefore particularly useful as a tool to determine whether changes in the effects of distance over time are truly economically important, rather than merely statistically significant. We will therefore apply our proposed methodology to estimate a standard gravity model where the effect of distance on bilateral trade is allowed to change yearly. It is important to stress, however, that the question of *why* the effect of distance does or does not change over time is beyond the scope of this paper. Instead, we focus on offering a methodologically well-grounded answer to the question of *how* the effect of distance actually evolves over time in a standard gravity model.

We will begin by presenting a brief overview of the previous research in the literature. Thereafter, we will outline our own empirical strategy, and then illustrate how the results differ when we compare our approach with a more traditional methodology.

3.1 Previous Research

The gravity model is one of the most commonly used tools to assess effects of trade policy and economic integration, and there are therefore hundreds of studies available, using a broad range of samples. While most studies do not allow the effect of distance to vary over time, investigating how the estimated distance effects vary in studies investigating different time periods is an indirect way to assess how the distance effect evolves over time. This approach has been used in an ambitious meta analysis of estimated distance effects performed by Disdier and Head (2008). Using a large number of estimated distance coefficients from a wide range of gravity studies, these authors find that there is a significant increase (in absolute terms) in the estimated distance effects after 1970. For example, according to their meta-regression results, distance impedes trade by 37% more after 1990

than it did during the time period 1870 to 1969.

There are also studies that estimate gravity models where the distance effect is allowed to vary over time. For instance, when estimating a standard gravity specification where the distance variable is interacted with a linear time trend and time squared, Brun *et al.* (2005) find that the impact of distance on trade increases over time.[6] In another study, Coe *et al.* (2007) capture changes in the effects of distance over time in two ways: by repeated cross-sectional regressions and by estimating a pooled model where the distance parameter is allowed to shift through the interaction with decade-specific dummies. When using nonlinear models, they find a decreasing distance effect, though not when using log-linear models. Carrère *et al.* (2010) capture changes in the distance coefficient by first conducting repeated cross-sectional regressions on five-year averages and then using a panel framework where the distance variable is interacted with a linear time trend and time squared. When analyzing a broad sample of trading countries, they draw the conclusion that the elasticity of trade with respect to distance becomes larger over time. Looking at trade for disaggregated industries, Berthelon and Freund (2008) draw the same conclusion for many of the industries by comparing the estimated elasticities for two time periods (1985-1990 and 2001-2004).

3.2 Empirical Strategy

As outlined in Section 3.1, there are gravity studies where the effect of distance has been allowed to vary over time. However, this has typically been done by either performing repeated cross-sectional regressions, which is an inefficient way of using the information available in trade data, or by exploiting the panel structure of the data, but then placing strong parametric restrictions on the allowed evolution of the distance effect. In this study, we estimate panel gravity models where the distance effect is allowed to vary arbitrarily over time without being restricted by parametric assumptions. In order to illustrate the usefulness of the methodology we propose, we contrast our preferred penalized approach with a flexible 'traditional' model that could potentially constitute a good approach to capturing the temporal evolution of the distance effect. The flexible 'traditional' model that we use as the benchmark contains interactions of the distance variable with year dummies. This way, separate coefficients for every year can be estimated, and the distance effect is allowed to vary freely over time. In our proposed penalized approach, we use the same flexible model with separate coefficients for every year, but we then additionally penalize the differences between adjacent coefficients (see Section 2). In doing so, we can simultaneously assess how the effect of distance changes over time and whether these changes are economically important. Moreover, by utilizing the model selection capacity of our approach, we can evaluate the relative importance of various explanatory variables

[6] Our main interest lies in the methodological aspects, and therefore we focus on estimations from standard gravity models. As shown by Brun *et al.* (2005) and some of the other studies reviewed, adding important control variables, or focusing on sub-samples of countries with specific characteristics, can reverse the finding of an increasing distance effect.

for explaining the volumes of bilateral trade.

We use data on aggregated bilateral trade between 185 importing countries and 195 exporting countries – at all levels of development – for the years 1962 to 2006 (for a full list of countries included in the sample, see Table A1 in the Appendix). The very long time period covered implies that if the effect of distance on trade changes over time, we should have a good chance of identifying this change. The set of explanatory variables (for a full list of variable definitions and data sources, see Table A2 in the Appendix) is rather standard and resembles what one would typically find in the gravity literature. Against that background we would again like to stress that we do not have the ambition to offer explanations as to *why* the effect of distance does or does not change over time. Our goal is instead to establish whether or not it actually does change by applying the penalized (quasi-)Poisson model described in Section 2.3.

Before presenting our empirical results, we want to point out a methodological detail that has been discussed in the gravity literature. A well-known problem in the trade literature is that the sources for data on trade volumes – in our case COMTRADE – do not typically differentiate between observations of zero trade flows and missing observations. Missing observations could therefore be truly missing observations of zero or positive trade flows, or an actual observation of a trade flow with the value zero. As has been extensively discussed in the literature about estimating gravity models (see e.g. Santos Silva and Tenreyro, 2006), the zeroes contain a lot of information, and failure to take them into account can lead to biased estimates. To separate zero-trade volumes from missing values, we therefore apply the following baseline rule: if a country has no recorded imports from *any* country in a particular year, import volumes between that country and all other countries are considered to be missing for that year; if a country has recorded imports from some (but not all) countries in a particular year, import from the remaining countries is considered to be zero. We also evaluate the ramifications of this proceeding by estimating a model where all non-positive observations are excluded.

3.3 General Results

Table 1 shows the results obtained from a (penalized) Poisson model using various values of the penalization parameter λ.[7] Due to the large number of observations in the analyzed data set, we have used adaptive penalties (see Section 2.1.3). In particular, the results in Table 1 originate from Poisson models penalized by the adaptive group lasso penalty given in (8). Results obtained when using the adaptive fused lasso penalty given in (9) are discussed in Section 3.4 below.

The first column of Table 1 shows the results from a conventional Poisson model where the coefficients were not penalized (i.e. $\lambda = 0$). The results confirm what is typically found in gravity studies. Both the importer's GDP and the exporter's GDP significantly

[7]Note that (for $\lambda > 0$) all the coefficients reported in Table 1, except the main effect of distance, were penalized. As discussed below, including the main effect of distance into the penalty renders the results virtually unaffected.

increase the volume of trade. Both the importer's population and the exporter's population decrease the volume of trade. As compared to GDP, however, the effects are smaller in absolute values, and the effect of the importer's population is not significant on the 1%-level. If the two trading partners share a common border, a common official language, or a common currency, the volume of trade is increased. The same holds if the trading partners have a common colonial history, and if they used to have a common colonizer. If both the importer and the exporter are members of the GATT (or later in the period WTO) or the same RTA (Regional Trade Agreement), trade is increased. The same holds if the importer is an ACP country and the exporter a member of the EU or *vice versa*.[8] As expected, the distance between the two trading partners has a strongly negative effect on their trade volumes. The coefficient may be interpreted as the elasticity of trade with respect to distance, so a 1% increase in bilateral distance is associated with a decrease in trade of about 0.6%. This is well in line with many previous findings (see e.g. Leamer and Levinsohn, 1995, and Disdier and Head, 2008).

In addition to the above mentioned covariates, we have also included four sets of dummy variables into the model. Besides flexible distance-time interactions, that allow the distance effect to vary freely over time, we have included year dummies, importer dummies, and exporter dummies to account for unobserved temporal and country-specific heterogeneity. According to compound Wald-type tests, all the four sets of dummy variables are significant on the 1‰-level, indicating that variations in the distance effect as well as unobserved year- and country-specific effects have a non-negligible impact on trade. Using penalized estimation techniques, we can evaluate the relative importance of these sets of dummy variables (as compared to the other covariates in the model). Since we are primarily interested in (the temporal evolution of) the distance effect, we will put particular focus on the distance-time interactions.

Introducing a small penalty with $\lambda = 1$ slightly changes the estimation results (see the second column of Table 1) and provides a first indication of the relative importance of the included covariates.[9] While most coefficients are hardly affected by this small penalization,

[8] It is noteworthy that both these variables are significant. Broadly speaking, they capture the trade preferences offered by the European Union to African, Caribbean and Pacific countries through the Yaoundé, Lomé and Cotonou agreements. Since these preferences are non-reciprocal, i.e. the ACP countries do not offer better than most favored nation (MFN) market access to EU countries, while they themselves export under preferential conditions, one should theoretically expect an effect only in one direction of trade. It is possible that stringent rules of origin combined with the permission of cumulation of origin when importing intermediate goods from EU countries could be an explanation for the positive effect on trade from the EU to ACP countries.

[9] With the introduction of a penalty term, P-values are no longer reported. Since lasso-type penalties allow for 'corner solutions' when maximizing (4), analytical derivatives of the maximized target function (and thus asymptotic standard errors) may not be available. Of course, standard errors of penalized estimates could be obtained using e.g. bootstrap methods. However, we have not pursued this, since significance is not an issue here. Due to the very large data set that we analyze, all but three coefficients in the unpenalized model are significant even on the 1‰-level. Moreover, the remaining three coefficients are the first to be shrunk to zero as the strength of penalization is increased, and they are no longer contained in the model that we prefer.

Table 1: Estimation Results

	Poisson ($\lambda = 0$)	Penalized Poisson					
		$\lambda = 1$	$\lambda = 10^6$	$\lambda = 10^8$	$\lambda = 10^{8.4}$	$\lambda = 10^{8.8}$	$\lambda = 10^{10}$
Log GDP (importer)	0.7089 (0.000)	0.7289	0.7283	0.7260	0.7281	0.7354	0.7577
Log GDP (exporter)	0.7163 (0.000)	0.6986	0.6982	0.6960	0.6967	0.6988	0.6955
Log population (importer)	−0.1180 (0.027)	−0.0036	−0.0029	0	0	0	0
Log population (exporter)	−0.5162 (0.000)	−0.0101	−0.0096	−0.0020	0	0	0
Common border	0.4887 (0.000)	0.4924	0.4911	0.4792	0.4564	0.4059	0
Common language	0.2786 (0.000)	0.2757	0.2750	0.2541	0.2227	0.1201	0
Common currency	0.0755 (0.012)	0.1130	0.1108	0	0	0	0
Colonial history	0.3427 (0.000)	0.3445	0.3428	0.2439	0.0522	0	0
Common colonizer	0.1049 (0.020)	0.0806	0.0694	0	0	0	0
GATT/WTO	0.3559 (0.000)	0.3171	0.3150	0.2373	0.1572	0	0
RTA	0.6755 (0.000)	0.6619	0.6599	0.6202	0.5738	0.4608	0
ACP to EU	0.4074 (0.000)	0.3517	0.3426	0	0	0	0
EU to ACP	0.4666 (0.000)	0.4210	0.4066	0	0	0	0
Log distance	−0.5990 (0.000)	−0.5389	−0.5374	−0.5496	−0.5747	−0.6239	−0.7056
Distance-time interactions	yes (0.000)	yes	yes	yes	no	no	no
Year dummies	yes (0.000)	yes	yes	yes	yes	yes	no
Importer dummies	yes (0.000)	yes	yes	yes	yes	yes	no
Exporter dummies	yes (0.000)	yes	yes	yes	yes	yes	yes
Observations	774,708	774,708	774,708	774,708	774,708	774,708	774,708
Deviance ($\times 10^{12}$)	−3.423	−3.423	−3.423	−3.422	−3.421	−3.418	−3.374

Note: λ denotes the tuning parameter used in the penalized regression models. All penalized estimates were obtained using the adaptive penalty given in (8). Deviance refers to the (cumulative) adjusted quasi-deviance given in (16). For the unpenalized Poisson model, P-values based on robust Huber-White standard errors (White, 1980) are provided in parentheses. The P-values for groups of dummy variables are based on robust Wald-type tests of the compound hypothesis that all coefficients are equal to zero.

the coefficients on the importer's and the exporter's population are strongly shrunk in absolute values. This suggests that neither the importer's population nor the exporter's population contribute much to explaining trade volumes. The third column of Table 1 shows that increasing the strength of penalization from $\lambda = 1$ to $\lambda = 10^6$ renders the results virtually unaffected. Columns 4 to 7 of Table 1 show that increasing the strength of penalization up to and beyond $\lambda = 10^8$ affects the estimation results quite substantially and causes some coefficient values to be shrunk to zero (i.e. the corresponding covariates are effectively removed from the regression). When imposing a very strong penalization with $\lambda = 10^{10}$, only four explanatory variables remain in the model: the importer's and the exporter's GDP, distance, and the set of exporter dummies. This suggests that these covariates are the most important ones for explaining the volume of trade.

Table 2 provides a more detailed picture of the relative importance of the covariates included in the regression. The table shows the order in which the covariates are removed from the regression and the corresponding λ-values. The first covariate to be dropped is *common colonizer* at a λ-value of $10^{7.2}$. The first set of dummy variables that is removed from the regression is the set of distance-time interactions at a λ-value of $10^{8.4}$. In other words, for a strength of penalization corresponding to $\lambda = 10^{8.4}$ or higher, the effect of distance is constant across time. The covariates that are dropped last are the importer's and the exporter's GDP, suggesting that these covariates are the most important ones for explaining the volume of trade.

Table 2: Importance of Covariates

λ	Covariates excluded
$10^{7.2}$	common colonizer
$10^{7.6}$	log population (importer)
$10^{8.0}$	common currency, ACP to EU, EU to ACP
$10^{8.4}$	distance-time interactions, log population (exporter)
$10^{8.8}$	colonial history, GATT/WTO
$10^{9.2}$	common language
$10^{9.6}$	year dummies, common border, RTA
$10^{10.0}$	importer dummies
$10^{10.4}$	exporter dummies
$10^{10.8}$	
$10^{11.2}$	
$10^{11.6}$	log GDP (importer), log GDP (exporter)

As Table 2 shows, the main effect of distance is never shrunk to zero. This is due to the fact that the corresponding coefficient was not penalized in order to rule out the somewhat hard-to-interpret scenario, where the main effect of distance is excluded from the model but the distance-time interactions are not. However, we also estimated a model

with a penalized distance coefficient, and the scenario mentioned above did not occur. In fact, the distance coefficient was the third last to be removed from the model (at a λ-value of $10^{10.8}$), suggesting that the distance between the trading partners, together with the two GDP-variables, constitute the three most important covariates. Of course, this is a very reassuring result, because it is precisely these variables that constitute the very core of the theoretical gravity model. That they are the last variables to disappear when we progressively increase the strength of penalization is therefore well in line with what we would expect to see. Thus, this example offers a neat illustration of how penalized regression can be used to discriminate between profoundly influential covariates and those that are merely statistically significant. In applications such as ours, where researchers increasingly work with very large data sets and therefore tend to find statistical significance for most covariates, the latter distinction becomes very important to make.

The above results raise the question of how large the strength of penalization should be. To determine the strength of penalization, we consider the (cumulative) adjusted quasi-deviance obtained from five-fold cross validation (see Section 2.4). Figure 1 shows the cross-validated quasi-deviance as a function of the tuning parameter λ. For λ ranging from 1 to $10^{8.4}$ the deviance function is roughly constant. For larger λ-values the deviance function increases markedly. This suggests that the value of λ chosen should not exceed $10^{8.4}$. Since the deviance function does not exhibit a distinct minimum, the exact value of λ that should be chosen is not straightforward to determine. In this case, we propose to choose a relatively heavy penalization corresponding to $\lambda = 10^{8.4}$, as this leads to a parsimonious and readily interpretable model which still has good explanatory power in terms of cross-validated deviance. However, since the deviance function is rather flat for $\lambda \leq 10^{8.4}$, it is important to make sure that the estimation results are not markedly affected by small changes in λ.

As Table 1 shows, increasing the value of λ from 10^8 to $10^{8.4}$ induces only two noteworthy changes in the estimation results. First, the exporter's population is removed from the set of explanatory variables. However, the effect of this covariate is negligible even for lower values of λ. Second, the distance-time interactions are removed from the regression, implying that the effect of distance is constant for $\lambda \geq 10^{8.4}$. Since we are mainly interested in how the effect of distance changes over time, we scrutinize the distance effect and its dependence on λ in the following section.

3.4 The Distance Effect

In what follows, we focus on the evolution of the distance effect over time. Figure 2 shows the distance effects obtained from the penalized Poisson model with four different values for the tuning parameter λ.

Without penalization, the distance effect varies rather erratically over time. As the dash-dotted line shows, the distance effect reaches a maximum (in absolute terms) of about -0.67 in 1969, and a minimum of about -0.47 in 1980. In other words, without penalization, the distance effect is estimated to be about 40% larger in 1969 than in

Figure 1: Cross-validated (Quasi-)Deviance as a Function of the Tuning Parameter λ

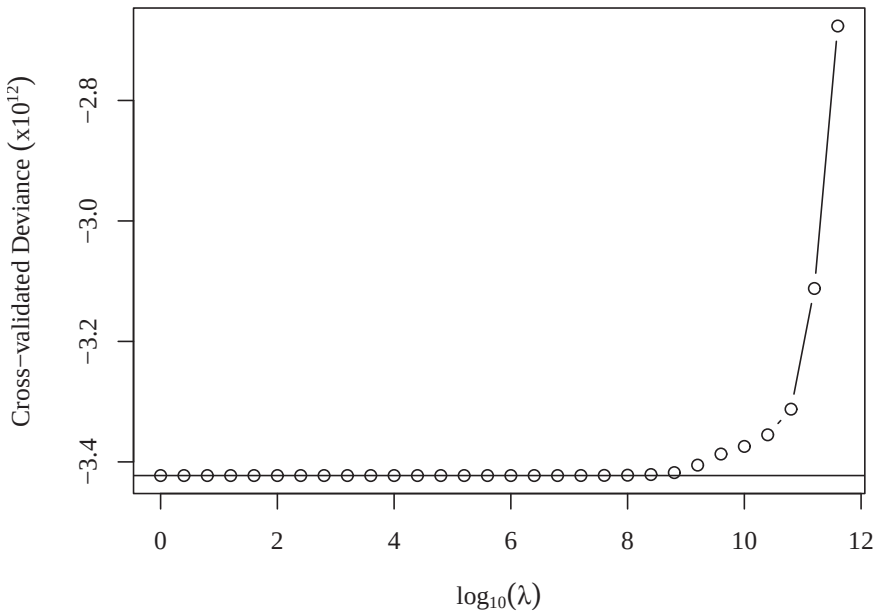
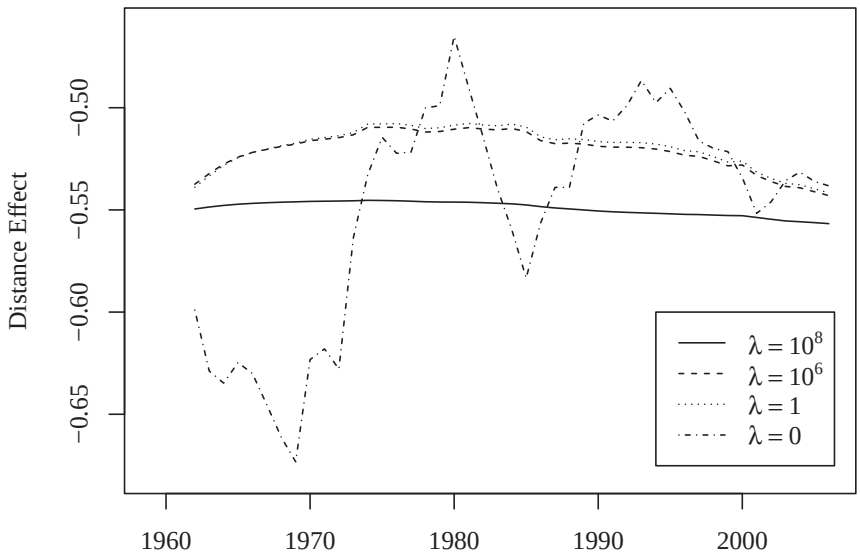


Figure 2: The Varying Effect of Distance over Time



1980. In general, the dash-dotted line obtained from the unpenalized model provides the impression that the distance effect exhibits a decreasing trend (in absolute terms) over time. However, introducing just a small penalty with $\lambda = 1$ substantially changes the estimated distance effect function. As the dotted line shows, the yearly changes in the distance effect are heavily shrunk, making the new, penalized function rather smooth. Moreover, after penalization, the distance effect no longer exhibits a trend in any direction. The penalized distance effect appears to be rather constant in the long run, exhibiting only a slightly inversely u-shaped form. As was the case with the parameter estimates given in Table 1, increasing the tuning parameter λ from 1 to 10^6 hardly affects the estimation results. The dashed line is almost indistinguishable from the dotted line. As the solid line shows, further increasing the tuning parameter to $\lambda = 10^8$ removes the slightly inversely u-shaped form and leads to a virtually constant distance effect. As discussed above, for λ-values of $\lambda = 10^{8.4}$ and larger, the distance effect is exactly constant.

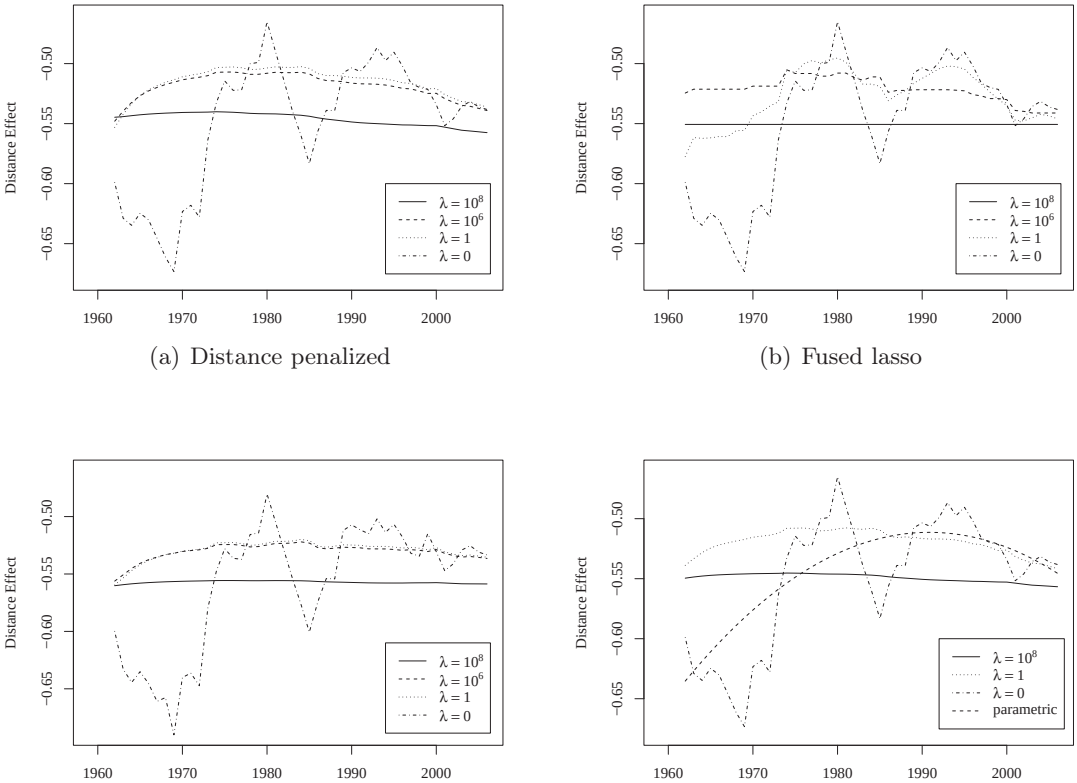
To test the robustness of the above results, we estimated three additional variants of the penalized Poisson model: 1) a model, where the main effect of distance was penalized; 2) a model, where the distance-time interactions were not treated as a group but as individual covariates (fused lasso); and 3) a model where all the 'zeros' were excluded. Moreover, since previous studies have frequently allowed the distance effect to vary over time in a linear and quadratic fashion, we compared our group lasso approach with such a parametric specification. Figure 3 shows the estimated distance effects obtained from these four models.

Figure 3(a) shows the estimated distance effects when the main effect of distance is penalized. The plotted distance functions are almost indistinguishable from their counterparts depicted in Figure 2. Thus, whether the main effect of distance is penalized or not, is not crucial for the estimation results and the conclusions that can be drawn from them. As above, the penalized effect of distance does not appear to exhibit any long-run trend, and, with $\lambda = 10^8$, the distance effect is virtually constant over time.

Figure 3(b) shows the estimated distance effects when differences in adjacent distance coefficients are penalized individually rather than group-wise. Such an individual penalization allows the distance effect to vary rather freely from year to year but also to be constant during longer subperiods. It may be a useful approach when the distance effect exhibits one or more 'jumps' during the long period of time that we study but is constant during the periods in between these jumps. As compared to the group-wise penalization, the distance effect varies more erratically when the strength of penalization is small, i.e. when $\lambda = 1$. However, when λ is increased to a value of 10^6, the distance effect is rather similar to the group-wise penalized effect, and with $\lambda = 10^8$ the effect is constant.

Figure 3(c) shows the estimated distance effects when all zero-trade volumes are excluded from the regression. Comparing Figure 3(c) with Figure 2 shows that the estimates are virtually unaffected by this modification. In other words, our *ad hoc* procedure to separate the 'zeros' from missing values does not seem to be crucial for the resulting distance effect estimates.

Figure 3: The Varying Effect of Distance over Time: Alternative Specifications



(a) Distance penalized

(b) Fused lasso

(c) No 'zeros'

(d) Group lasso vs. parametric trend

Lastly, Figure 3(d) shows the estimated distance effect obtained from a parametric specification with a linear and a quadratic time trend. For comparison, the figure also depicts group lasso estimates for $\lambda = 0$, $\lambda = 1$, and $\lambda = 10^8$, as already shown in Figure 2. Unlike the penalized estimates, the parametric estimates suggest that the effect of distance varies quite substantially over time.[10] In fact, the parametric distance effect function seems to approximate the flexible (unpenalized) distance effect function. Both of these functions suggest that the distance effect is markedly larger (in absolute terms) during the 1960s and early 1970s than during the later years of the observation period. The penalized estimates do not indicate such a shift in the distance effect, not even when the strength of penalization is very small. The reason for this discrepancy between the penalized and the unpenalized estimates might be that the number of observations is comparably small during the 1960s and early 1970s. As a result, the distance effect estimates are relatively shaky for these early years, and, consequently, differences in adjacent distance coefficients are shrunk rather strongly when penalization is introduced.

[10] Both the linear and the quadratic component of the trend are significant on the 1‰-level. Detailed results are available upon request.

To summarize, the above results suggest that our initial finding that the effect of distance on trade is rather constant across time is robust to changes in the strength and form of penalization as well as to changes in the data analyzed. It is noteworthy that this finding of a rather constant distance effect differs from the result obtained when estimating a parametric model specification where the effect of distance is allowed to vary over time in a linear and a quadratic fashion. This suggests that our proposed penalized estimation approach may lead to insights that differ from those obtained in previous studies, where parametric time trends have typically been estimated.

4 Summary and Conclusions

In this paper, we have argued that estimating panel data models with time-varying covariate effects is a new area where penalized regression techniques could be very useful. In particular, we have proposed the use of lasso-type models where differences between adjacent time-varying coefficients are penalized. This approach produces predominantly smooth temporal variations in covariate effects without imposing restrictive parametric assumptions. It also helps us determine whether the temporal variations in covariate effects are (economically) relevant rather than merely statistically significant. If the temporal variations make relevant contributions to the explanatory power of the model, the degree of smoothness will be small, and relevant shifts in covariate effects can still be identified.

In order to illustrate the usefulness of the methodology we propose, we have revisited the well-known empirical puzzle of the 'death of distance' in international trade. By estimating a standard gravity model on a very large panel of trade between 185 importers and 195 exporters over the period 1962-2006, we have compared two ways to investigate whether (and if so how) the effect of distance on bilateral trade changes over time. As a benchmark, we have used a very flexible version of a traditional panel specification, where separate coefficients are estimated for each year by interacting the distance variable with year dummies. For a researcher who does not wish to use our proposed penalized regressions, this would be a simple way to allow the effect of distance to vary in a non-restrictive fashion. We have then compared this benchmark to our proposed method where we additionally have penalized the differences between adjacent coefficients.

By comparing the benchmark regression with the penalized regressions, we have been able to draw some interesting conclusions. Our results suggest that the choice between penalized and unpenalized regressions can influence the empirical conclusions. When we estimate the flexible model without penalties, the estimated distance effects vary quite substantially, suggesting that the effect of distance on trade would jump up and down even over rather limited time periods. For instance, our results suggest that the effect in 1969 would be about 40% larger than in 1980. However, when introducing smoothing penalties – thereby ensuring that only changes that contribute in a relevant way to explaining variations in trade flows are considered – the differences become much more limited, even if the strength of penalization is rather small. This suggests that the potential problem of

over-fitting when not using penalties should be taken seriously. Indeed, at reasonable levels of penalization (according to cross-validation performance), our estimation results suggest that the effect of distance on trade is virtually unchanged over time. In other words, the approach of using lasso-type penalties helps us to avoid drawing misleading conclusions about how the effect of distance on trade evolves over time.

In a further analysis, we have also compared our penalized estimation results with the results obtained from a model where the distance effect is allowed to vary parametrically over time. Since previous studies have frequently allowed the distance effect to vary over time in a linear and quadratic fashion, we have contrasted our penalized estimation approach with such a parametric specification. Again, we have found substantial differences in the results. While the penalized estimates indicate that the distance effect is rather constant over time, the parametric estimates exhibit significant changes over time. This suggests that our proposed penalized estimation approach may lead to insights that differ from those obtained in previous studies, where parametric time trends have typically been estimated.

Our empirical application also highlights other advantages of the lasso-type penalties that we propose. Specifically, in the same way as the penalization assists us in choosing an appropriate structure of temporal evolution for the distance coefficients, it can also help us to find a parsimonious model specification. By penalizing all coefficients (except the global intercept), and progressively increasing the strength of penalization, we can see which variables that are first dropped from the model, and which variables that make relevant contributions to the explanatory power of the model at practically any levels of penalization. Reassuringly, we see that variables where the theoretical expectations of effects are lower disappear long before the core gravity variables, namely the geographical distance between the exporter and the importer and their respective GDPs. In the context of international trade – and any other field where very large data sets are available – this way to discriminate between factors that are statistically significant due to the large number of observations, and those that are truly 'economically' relevant, can be very beneficial for applied researchers who take the assessment of policy relevance seriously.

Appendix A: Auxiliary Tables

Table A1: List of Countries (Exporters and Importers) Included in the Analysis

Afghanistan	Colombia	Haiti	Mauritius	Slovak Rep.
Albania	Comoros	Honduras	Mexico	Slovenia
Algeria	Congo, Dem. Rep.	Hong Kong	Micronesia, Fed. Sts.	Solomon Islands
Andorra	Congo, Rep.	Hungary	Moldova	Somalia
Angola	Costa Rica	Iceland	Mongolia	South Africa
Antigua & Barbuda	Cote d'Ivoire	India	Morocco	Spain
Argentina	Croatia	Indonesia	Mozambique	Sri Lanka
Armenia	Cuba	Iran, Islamic Rep.	Namibia	St. Kitts & Nevis
Aruba	Cyprus	Iraq	Nepal	St. Lucia
Australia	Czech Rep.	Ireland	Netherlands	St. Vincent & the Grenadines
Austria	Denmark	Israel	New Caledonia	Sudan
Azerbaijan	Djibouti	Italy	New Zealand	Suriname
Bahamas, The	Dominica	Jamaica	Nicaragua	Swaziland
Bahrain	Dominican Rep.	Japan	Niger	Sweden
Bangladesh	East Timor	Jordan	Nigeria	Switzerland
Barbados	Ecuador	Kazakhstan	Norway	Syrian Arab Rep.
Belarus	Egypt, Arab Rep.	Kenya	Oman	Tajikistan
Belgium	El Salvador	Kiribati	Pakistan	Tanzania
Belize	Equatorial Guinea	Korea, Rep.	Palau	Thailand
Benin	Eritrea	Kuwait	Panama	Togo
Bermuda	Estonia	Kyrgyz Rep.	Papua New Guinea	Tonga
Bhutan	Ethiopia	Lao PDR	Paraguay	Trinidad & Tobago
Bolivia	Faeroe Islands	Latvia	Peru	Tunisia
Bosnia & Herzegovina	Fiji	Lebanon	Philippines	Turkey
Botswana	Finland	Lesotho	Poland	Turkmenistan
Brazil	France	Liberia	Portugal	Tuvalu
Brunei	French Polynesia	Libya	Puerto Rico	Uganda
Bulgaria	Gabon	Lithuania	Qatar	Ukraine
Burkina Faso	Gambia, The	Luxembourg	Romania	United Arab Emirates
Burundi	Georgia	Macao	Russian Federation	United Kingdom
Cambodia	Germany	Macedonia, FYR	Rwanda	United States
Cameroon	Ghana	Madagascar	Samoa	Uruguay
Canada	Greece	Malawi	San Marino	Uzbekistan
Cape Verde	Greenland	Malaysia	Sao Tome & Principe	Vanuatu
Cayman Islands	Grenada	Maldives	Saudi Arabia	Venezuela
Central African Rep.	Guatemala	Mali	Senegal	Vietnam
Chad	Guinea	Malta	Seychelles	Yemen
Chile	Guinea-Bissau	Marshall Islands	Sierra Leone	Zambia
China	Guyana	Mauritania	Singapore	Zimbabwe

Table A2: Overview of Variables and Data Sources

Variable	Definition & Data Source
Trade	Bilateral imports from the United Nations' Comtrade.
Log GDP	Log of importer's or exporter's GDP. Data from the World Bank's World Development Indicators (WDI) online.
Log population	Log of importer's or exporter's population. Data from the World Bank's World Development Indicators (WDI) online.
Common border	Takes the value one if the trading countries share a border Data from CEPII, `http://www.cepii.fr`.
Common language	Takes the value one if the trading countries share the same official language. Data from CEPII, `http://www.cepii.fr`.
Common currency	Takes the value one if the trading countries share the same currency. Data from Head *et al.* (2010), available via CEPII, `http://www.cepii.fr`.
Colonial history	Takes the value one if the trading countries share a common colonial history, i.e. the importer (exporter) has been a colony to the exporter (importer). Data from CEPII, `http://www.cepii.fr`.
Common colonizer	Takes the value one if the trading countries have had the same colonizer after 1945. Data from CEPII, `http://www.cepii.fr`.
GATT/WTO	Takes the value one if both countries are members of the General Agreement on Tariffs and Trade (GATT) or the World Trade Organization (WTO). Data from Head *et al.* (2010), available via CEPII, `http://www.cepii.fr`.
RTA	Takes the value one if both countries are members of same Regional Trade Agreement. Data from Head *et al.* (2010), available via CEPII, `http://www.cepii.fr`.
ACP to EU	Takes the value one if the exporter is an ACP (African, Caribbean and Pacific) country and the importer is an EU country. Put differently, if the exporter is granted non-reciprocal trade preferences by the European Union through the Yaoundé, Lomé and Cotonou agreements. Data from Head *et al.* (2010), available via CEPII, `http://www.cepii.fr`.
EU to ACP	Same as the 'ACP to EU' variable, but covering trade in the other direction. Data from Head *et al.* (2010), available via CEPII, `http://www.cepii.fr`.
Log distance	Log of distance in km between the trading countries' capitals. Data from (CEPII), `http://www.cepii.fr`.

Appendix B: Practical Guideline

As discussed in Section 2, estimating the models proposed in this paper using the R software package `grplasso` (Meier, 2009) requires a particular preparation of the data set. This Appendix provides a detailed description of how model re-parametrization can be performed in practice, and a step-by-step instruction for model estimation.

To start with, we consider a single explanatory variable z that is allowed to exhibit time-varying effects. Again, for notational convenience, we consider the balanced data case. Using the required interaction of z with a set of time dummies, the corresponding linear predictor is given by

$$\underbrace{\begin{pmatrix} z_{11} & 0 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ z_{N1} & 0 & \dots & \dots & 0 \\ 0 & z_{12} & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & z_{N2} & 0 & \dots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \vdots & \vdots & \vdots & 0 & z_{1T} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & z_{NT} \end{pmatrix}}_{Z} \cdot \underbrace{\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \vdots \\ \gamma_T \end{pmatrix}}_{\gamma}.$$

Now, using the re-parametrization $\delta_1 = \gamma_1$ and $\delta_t = \gamma_t - \gamma_{t-1}$, we have

$$\begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_T \end{pmatrix} = \begin{pmatrix} \delta_1 \\ \delta_2 + \delta_1 \\ \vdots \\ \delta_T + \ldots + \delta_1 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & 1 & \cdots & 1 \end{pmatrix}}_{L} \cdot \underbrace{\begin{pmatrix} \delta_1 \\ \delta_2 \\ \vdots \\ \delta_T \end{pmatrix}}_{\delta}.$$

Thus, the linear predictor $Z\gamma$ can be expressed as $ZL\delta$, where

$$ZL = \begin{pmatrix} z_{11} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & 0 & \dots & 0 \\ \\ z_{12} & z_{12} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ z_{N2} & z_{N2} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ z_{1T} & z_{1T} & \dots & z_{1T} \\ \vdots & \vdots & \ddots & \vdots \\ z_{NT} & z_{NT} & \dots & z_{NT} \end{pmatrix}.$$

If intercept α_t and q time-varying coefficients $\gamma_{l,t}$ are to be smoothed, the corresponding design matrix is given by

$$D = \begin{pmatrix} 1 & 0 & \cdots & 0 & z_{1,11} & 0 & \cdots & 0 & \cdots & z_{q,11} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 & z_{1,N1} & 0 & \cdots & 0 & \cdots & z_{q,N1} & 0 & \cdots & 0 \\ \\ 1 & 1 & \vdots & \vdots & z_{1,12} & z_{1,12} & \vdots & \vdots & \cdots & z_{q,12} & z_{q,12} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \vdots & \vdots & z_{1,N2} & z_{1,N2} & \vdots & \vdots & \cdots & z_{q,N2} & z_{q,N2} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & z_{1,1T} & z_{1,1T} & \cdots & z_{1,1T} & \cdots & z_{q,1T} & z_{q,1T} & \cdots & z_{q,1T} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 & z_{1,NT} & z_{1,NT} & \cdots & z_{1,NT} & \cdots & z_{q,NT} & z_{q,NT} & \cdots & z_{q,NT} \end{pmatrix}.$$

Now, in order to estimate model (4) using the R software package `grplasso`, the following procedure can be applied:

1. Arrange the data as in design matrix D above.
2. Decide whether conventional penalization or adaptive penalization is to be used; if conventional penalization is chosen, go to step 3, otherwise go to step 4.
3. Divide every column in D except the first one (global intercept) by the standard deviation of the respective column entries, σ_c, and continue with step 5.

4. Divide every column in D except the first one by $w_l \cdot \sigma_c$ (group lasso) or $w_{l,t} \cdot \sigma_c$ (fused lasso), where σ_c denotes the standard deviation of the respective column entries, and weights w_l and $w_{l,t}$ are defined as in Section 2.1.3.

5. Determine $\lambda_{\max}$ using the command `lambdamax()`.

6. Select an adequate λ-value (between 0 and $\lambda_{\max}$) through cross-validation (see Section 2.4).

7. Estimate the desired model using the command `grplasso()`.

8. For better interpretability, the estimated coefficients may be back-transformed through division by σ_c, and using the relations $\delta_1 = \gamma_1$ and $\delta_t = \gamma_t - \gamma_{t-1}$.

References

Arellano, M. (2003), *Panel Data Econometrics*, Oxford: Oxford University Press.

Baltagi, B. H. (2008), *Econometric Analysis of Panel Data*, Chichester: John Wiley & Sons.

Baltagi, B. H., Bresson, G. and Pirotte, A. (2008), "To Pool or Not to Pool?", in L. Mátyás and P. Sevestre (eds.), *The Econometrics of Panel Data: Fundamentals and Recent Developments in Theory and Practice*, pp. 517–546, Berlin: Springer.

Berthelon, M. and Freund, C. (2008), "On the Conservation of Distance in International Trade", *Journal of International Economics*, vol. 75, pp. 310–320.

Brun, J.-F., Carrère, C., Guillaumont, P. and de Melo, J. (2005), "Has Distance Died? Evidence from a Panel Gravity Model", *World Bank Economic Review*, vol. 19, pp. 99–120.

Cairncross, F. (1997), *The Death of Distance. How the Communications Revolution is Changing Our Lives*, Boston: Harvard Business School Press.

Cameron, A. C. and Trivedi, P. K. (2005), *Microeconometrics: Methods and Applications*, New York: Cambridge University Press.

Carrère, C., de Melo, J. and Wilson, J. (2010), "Distance and Regionalization of Trade for Low-Income Countries", World Bank Policy Research Working Paper No. 5214.

Carrère, C. and Schiff, M. (2005), "On the Geography of Trade. Distance is Alive and Well", *Revue Economique*, vol. 56, pp. 1249–1274.

Coe, D. T., Subramanian, A. and Tamirisa, N. T. (2007), "The Missing Globalization Puzzle: Evidence of the Declining Importance of Distance", *IMF Staff Papers*, vol. 54, pp. 34–58.

Disdier, A.-C. and Head, K. (2008), "The Puzzling Persistence of the Distance Effect on Bilateral Trade", *Review of Economics and Statistics*, vol. 90, pp. 37–48.

Gertheiss, J., Hogger, S., Oberhauser, C. and Tutz, G. (2011), "Selection of Ordinally Scaled Independent Variables with Applications to International Classification of Functioning Core Sets", *Applied Statistics*, vol. 60, pp. 377–395.

Gertheiss, J. and Tutz, G. (2009), "Penalized Regression with Ordinal Predictors", *International Statistical Review*, vol. 77, pp. 345–365.

Gertheiss, J. and Tutz, G. (2012), "Regularization and Model Selection with Categorial Effect Modifiers", *Statistica Sinica*, vol. 22, pp. 957–982.

Gourieroux, C., Monfort, A. and Trogon, A. (1984), "Pseudo Maximum Likelihood Methods: Applications to Poisson Models", *Econometrica*, vol. 52, pp. 701–720.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009), *The Elements of Statistical Learning*, New York: Springer.

Head, K., Mayer, T. and Ries, J. (2010), "The Erosion of Colonial Trade Linkages after Independence", *Journal of International Economics*, vol. 81, pp. 1–14.

Hsiao, C. (2003), *Analysis of Panel Data*, Cambridge: Cambridge University Press.

Leamer, E. E. and Levinsohn, J. (1995), "International Trade Theory: The Evidence", in G. M. Grossman and K. Rogoff (eds.), *Handbook of International Economics*, vol. 3, pp. 1339–1394, Amsterdam, North-Holland: Elsevier.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman & Hall, 2nd ed.

Meier, L. (2009), *grplasso: Fitting User Specified Models with Group Lasso Penalty*, R package version 0.4-2.

Meier, L., van de Geer, S. and Bühlmann, P. (2008), "The Group Lasso for Logistic Regression", *Journal of the Royal Statistical Society, Series B*, vol. 70, pp. 53–71.

Meier, L., Van de Geer, S. and Bühlmann, P. (2009), "High-Dimensional Additive Modeling", *The Annals of Statistics*, vol. 37, pp. 3779–3821.

Park, M. Y. and Hastie, T. (2007), "L_1-Regularization Path Algorithm for Generalized Linear Models", *Journal of the Royal Statistical Society, Series B*, vol. 69, p. 659–677.

Santos Silva, J. M. C. and Tenreyro, S. (2006), "The Log of Gravity", *Review of Economics and Statistics*, vol. 88, pp. 641–658.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso", *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Kneight, K. (2005), "Sparsity and Smoothness via the Fused Lasso", *Journal of the Royal Statistical Society, Series B*, vol. 67, pp. 91–108.

Wang, H. and Leng, C. (2008), "A Note on Adaptive Group Lasso", *Computational Statistics & Data Analysis*, vol. 52, pp. 5277–5286.

Westerlund, J. and Wilhelmsson, F. (2009), "Estimating the Gravity Model without Gravity Using Panel Data", *Applied Economics*, vol. 41, pp. 1–9.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica*, vol. 48, pp. 817–838.

Yuan, M. and Lin, Y. (2006), "Model Selection and Estimation in Regression with Grouped Variables", *Journal of the Royal Statistical Society, Series B*, vol. 68, pp. 49–67.

Zhang, H. and Lu, W. (2007), "Adaptive Lasso for Cox's Proportional Hazards Model", *Biometrika*, vol. 94, pp. 691–703.

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties", *Journal of the American Statistical Association*, vol. 101, pp. 1418–1429.