Improved estimation in threshold regression with applications to price transmission modeling

Dissertation

zur Erlangung des Doktorgrades
der Fakultät für Agrarwissenschaften
der Georg-August-Universität Göttingen

Vorgelegt von
Dipl. Math. Friederike Greb
geboren in Düsseldorf

Göttingen, im Dezember 2011

D 7

Gutachter Prof. Dr. Tatyana Krivobokova
Prof. Dr. Stephan von Cramon-Taubadel
Prof. Dr. Axel Munk

Acknowledgements

I am most grateful to Tatyana, Stephan and Axel. It is great to work together with people who are full of ideas and put so much enthusiasm and energy in their projects. And who like to laugh. I feel that I have received much more time and attention than I deserve; measuring Tatyana's commitment to her doctoral children certainly needs asymptotic theory. Not least – thank you very much for encouraging me to take up this endeavor! I would also like to thank my officemates and my colleagues in the Courant Centre for creating such a pleasant working environment; as well as my co-workers from the Chair of Development Economics, the Institute of Mathematical Stochastics and the Department of Agricultural Economics and Rural Development. Finally, thank you to the Courant Centre for providing me with the opportunity to venture into new disciplines.

Summary

Modeling a dependent variable as a linear combination of independent variables with regression coefficients which vary between regimes is known as threshold regression. The choice of regime is determined by a transition function which depends on a transition variable as well as a threshold parameter. There is a wide range of different fields of application of threshold regression models including agricultural economics, but also economics in general, finance, sociology or biostatistics. In this thesis, the application of price transmission analysis and the threshold vector error correction model (TVECM) – a particular threshold regression model specification – as one of its standard tools are the focus of interest.

Threshold parameters are typically estimated by maximizing the profile likelihood function. However, in certain settings, this estimator is biased and characterized by great variance. Estimates are likely to be unreliable when the number of unknown model parameters is large relative to the sample size, there is little difference in coefficients between adjoining regimes, or thresholds leave only few observations in one of the regimes. In the latter case, the profile likelihood estimator can even be inconsistent as it depends on an arbitrary trimming parameter imposing a minimum number of observations to fall into each regime. This becomes critical when modeling spatial price transmission. The framework of a TVECM for the study of price transmission processes is motivated by dynamics resulting from the elimination of spatial arbitrage opportunities; these imply that the situation that almost all of the observations are associated with a single regime is very likely to occur in case of well-integrated markets. The aim of this thesis is to improve threshold estimation, particularly in these difficult settings, in threshold regression models in general, and ultimately to achieve this for the special case of TVECMs.

To this end, in a first paper ("Improved estimation in generalized threshold regression models") attention is limited to a simple generalized threshold regression model. In

this setting, deficits of the profile likelihood estimator are analyzed and an alternative estimator is developed. The idea of this regularized Bayesian estimator is to penalize differences between regimes so as to keep them reasonably small when the data contain little information. The strength of this regularizing penalty, which is fundamental to the estimator, is determined in a data-driven manner employing the so-called empirical Bayes paradigm. The estimator is developed in a Bayesian framework, the penalization a result of the choice of priors. As an important consequence of the regularization, the new estimator does not suffer from the need to choose a trimming parameter. A simulation study shows that it clearly outperforms common estimators, especially in problematic settings.

In a second paper ("The estimation of threshold models in price transmission analysis"), the regularized Bayesian estimator is employed to estimate TVECMs. Even in this more intricate model – contrary to the univariate generalized threshold regression model with a single threshold, it is a multivariate model with multiple thresholds – the new estimator produces better estimates than the profile likelihood estimator. Revisiting the seminal article by Goodwin & Piggott (2001), which established TVECMs for price transmission analysis, regularized Bayesian estimates turn out to be free of several anomalies that characterize the profile likelihood estimates and corroborate Sephton's (2003) doubts that the TVECM with two thresholds is the accurate model specification for the data in question. As a further empirical application, an investigation of spatial price transmission between German and Spanish pork markets is presented. Here, profile likelihood estimates are determined by the trimming parameter and, hence, useless. The alternative estimator presents a method to obtain sensible estimates, and moreover, the new estimating framework yields plausible estimates for the remaining unknown parameters.

Chapters three and four contain the respective manuscripts. To complete this thesis, chapter two, which contextualizes generalized threshold regression models and the TVECM, precedes them. It provides an overview of different threshold regression models, classifies them in terms of asymptotic properties of threshold estimators, and characterizes the TVECM with an emphasis on specifications which are important in price transmission analysis. Chapter one introduces the reader to the research question, chapter five concludes with a discussion of the results.

Table of Contents

List of Figures

List of Tables

1 Introduction

The starting point of this thesis is an estimation problem arising in price transmission analysis. Price transmission refers to the propagation of price shocks between markets, which typically differ in location, stage in the production process or commodity; another important variant of price transmission analysis follows price shocks from the world markets into local markets. Here, the focus is on spatial price transmission, that is, price transmission between markets in different locations. This is a particularly interesting variant of price transmission to study when looking at agricultural commodities. Since these are expensive to transport compared to their own value, but tend to be produced over a wide area, a complicated pattern of price dynamics in space arises (Fackler & Goodwin, 2001).

Price transmission analysis is an important tool to assess market integration. In fact, Fackler & Goodwin (2001) propose "that market integration is best thought of as a measure of the degree to which demand and supply shocks arising in one region are transmitted to another region." The concept of market integration in turn has proven relevant to answer a number of distinct questions. Fackler & Tastan (2008) point to the definition of market boundaries, especially for antitrust regulation and international trade conflicts, as well as the evaluation of the impact of market development and liberalization policies in developing countries. To illustrate this idea, consider interventionist policies to prevent famines. Ravallion (1986) refers to India as an example for conflicting views on the question of whether a government should adopt such a policy. He mentions that the Indian government relied on the effect of the grain trader's response to localized scarcity during most of the nineteenth and beginning twentieth century; and cites the local government of Madras, which during a food shortage in the 1870s put forward that "if time were given to the market, the necessary grain would eventually come, but time was what could not be given". To emphasize the incertitude with respect to the

adequate position for the government to take, he then points to the fact that after independence, the Indian government took on a strongly interventionist position. Ravallion (1986) suggests an empirical analysis of market integration to offer a new perspective, maybe even resolve the debate.

The current workhorse to investigate the transmission of price shocks between various locations is the threshold vector error correction model (TVECM). Time series of prices for the same commodity in different markets tend not to drift too far apart. Elimination of spatial arbitrage opportunities draws them towards an equilibrium. The error correction model incorporates this dynamic by allowing for the correction of part of one period's disequilibrium in the subsequent period. Obviously, the possibility to profit from spatial price differences only exists if these are greater than the transaction costs incurred by moving goods between markets. Consequently, looking at the simplest case of two markets, one of three different situations occurs. Traders carry goods from the first to the second market to profit from spatial arbitrage (the price in the second market exceeds that in the first by more than the transaction costs), no trade takes place (transaction costs exceed the difference in prices), or traders bring goods from the second to the first market (the price in the first market exceeds that in the second by more than the transaction costs). This is reflected by thresholds separating three (a larger number when there is more than two markets involved) model regimes, which are determined by relating the price difference to the transaction costs. In the outer regimes error correction is thought to take place, while prices move independently from another within the middle band, that is, when spatial arbitrage opportunities are lacking.

Against this background, the estimation problem at the outset of this thesis can be defined more precisely as the question how to best estimate the threshold parameters in a TVECM. Histograms resembling the ones in figure 1.1 were the actual starting point. They show the distribution of (the commonly used) profile likelihood threshold estimates for a TVECM with three regimes. Clearly, the estimates tend to be drawn towards zero, the lower threshold is often overestimated, the upper threshold underestimated.

What makes it difficult to estimate the threshold parameters in a TVECM? Challenges are twofold. First, the thresholds are not the only model parameters. Especially when the number of additional unknown parameters is high, the signal-to-noise-ratio tends to be low and, accordingly, threshold estimation very difficult. Second, the TVECM

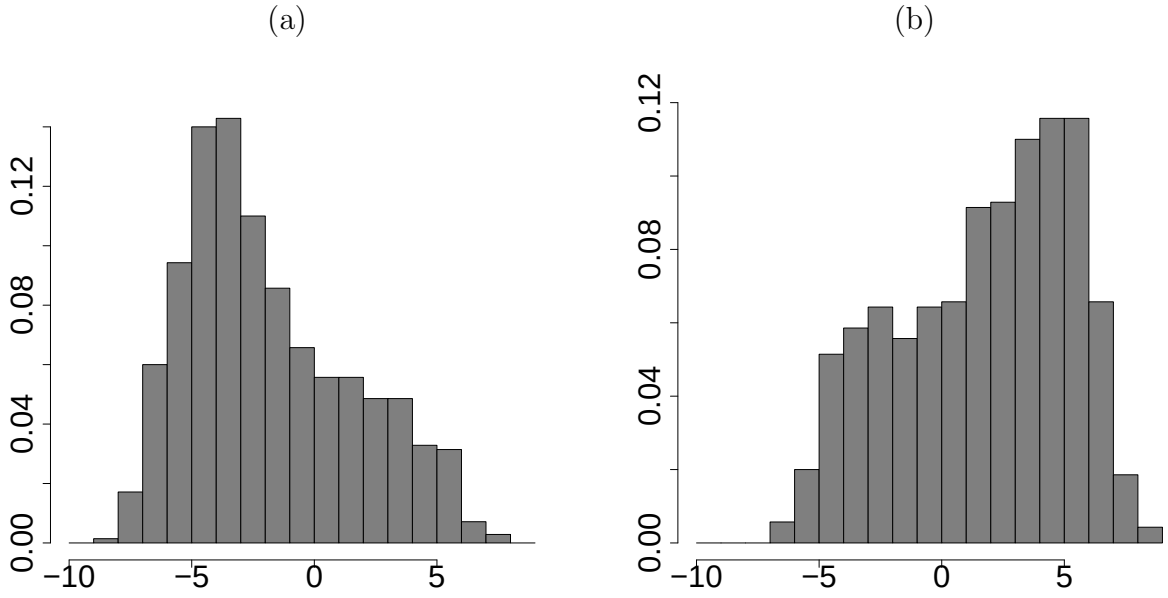

Figure 1.1: Profile likelihood estimates for the lower (a) and upper (b) threshold of a TVECM with three regimes. True values are −4 and 4, respectively.

features several nonregular aspects. The likelihood function is not differentiable with respect to the threshold parameter; the threshold parameter space is not fixed, but randomly increasing; identification becomes critical for certain choices of parameters. More precisely, when there is no difference in coefficients between adjoining regimes or the threshold is at the boundary of its domain, the TVECM turns linear and the threshold parameter as well as part of the coefficients disappear, that is, cannot be identified. When approaching either of these settings (that is, when differences between regimes diminish or few observations are left in one of the regimes), estimation becomes increasingly difficult. Such a situation is especially likely to occur in small samples and when nuisance parameters abound. Clearly, when analyzing price transmission between two well-integrated markets it is very likely that only few observations fall into the outer regimes as spatial price differences exceeding transaction costs are corrected for quickly. Thus, in price transmission analysis, settings in which estimation is complicated are not rare artifacts, but quite frequent.

Confronted with the task to improve estimation in an intricate multivariate model with multiple thresholds and a complicated structure of dependencies among observations,

the natural approach is to first look for a simpler analogue to study. The set-up of a basic univariate threshold regression model with a single breakpoint turns out to be similar enough to share the deficits of commonly used estimators with the TVECM; yet sufficiently simple to considerably facilitate analysis of the source of the difficulties of commonly used estimators and allow to solve the estimation problem. Consequently, the roadmap for this thesis is to improve threshold estimation (i) in so-called generalized threshold regression models (GTRMs) and (ii) in TVCEMs. The idea is to achieve the second aim by extending the strategy developed to reach the first.

The second chapter sets the stage for the investigation of the two models. As typical of regression in general, threshold regression models relate a response variable to a linear combination of explanatory variables. Their defining characteristic are regression coefficients which vary between regimes. The choice of regime is in turn determined by a transition function which depends on a transition variable and a threshold parameter. This vague wording indicates the wide range of models subsumed under the term "threshold regression model"; the TVECM is one of them, the GTRM another. The chapter provides a brief topography of threshold regression models to place these two particular cases in context and clarify the use of the expression.

The third chapter is composed of the manuscript "Regularized Bayesian estimation in generalized threshold regression models". Within the framework of a GTRM, deficits of the common estimators – the profile likelihood estimator and Bayesian estimators with noninformative priors – are analyzed and conditions identified which exacerbate their effect on the estimates. As alternative, the regularized Bayesian estimator is developed. Its superior performance, especially in critical settings, is confirmed in a simulation study. An illustration of the impact of the new estimator in two empirical applications, one from economics, the other from ecology, completes the picture.

The fourth chapter comprises the manuscript "The estimation of threshold models in price transmission analysis". Here, the regularized Bayesian estimator is formulated for the TVECM and its performance assessed for this more complex model. Resembling the findings for GTRMs, regularized Bayesian threshold estimates for the TVECM are clearly more reliable than common estimates. The seminal article by Goodwin & Piggott (2001), which introduced threshold cointegration to price transmission analysis, is revisited, that is, the original dataset is reestimated employing the regularized Bayesian

estimator. The new estimates confirm Sephton's (2003) finding that the two-threshold TVECM is misspecified for the data in question. Moreover, transmission between German and Spanish pork prices is explored. In this application, regularized Bayesian estimates offer a more plausible interpretation of the data than their profile likelihood counterparts. It becomes evident that the new estimator does indeed provide a different perspective in empirical applications. The fifth chapter discusses the results and concludes.

The two articles which form the core of this thesis are

- Greb, F., Krivobokova, T., Munk, A. and von Cramon-Taubadel, S. (2011). Regularized Bayesian estimation in generalized threshold regression models. CRC-PEG Discussion Paper No.99 (submitted to *Bayesian Analysis*)
- Greb, F., von Cramon-Taubadel, S., Krivobokova, T. and Munk, A. (2011). The estimation of threshold models in price transmission analysis. CRC-PEG Discussion Paper No.103 (submitted to *American Journal of Agricultural Economics*)

2 Threshold regression models

2.1 Characterization of threshold regression models

In this section, I define the threshold regression model, outline criteria to categorize different model types and characterize the latter in terms of the limiting distributions of threshold estimators. To complete the picture, I outline further extensions of the model that have appeared in the literature.

2.1.1 Model definition and classification criteria

The term "threshold regression model" is associated with anything but a clearly defined model. The fact that a variety of different labels is used synonymously or for specific types of threshold regression models further complicates the situation. These include change point models, structural change or breaks, two-phase regression, switching regression or threshold switching, time-trending regression, segmented regression, broken-line regression and disequilibrium models.

Possibly the simplest example of a threshold regression model is a sequence of independent normal random variables $y_1, \ldots, y_n$ with a change in mean,

$$y_i = \begin{cases} \mu_1 + \varepsilon_i & i \leq \psi \\ \mu_2 + \varepsilon_i & i > \psi \end{cases} \tag{2.1}$$

where $\mu_1, \mu_2 \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}^+$, and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ independently for $i = 1, \ldots, n$. One classical dataset that lends itself to be described by this model consists of observations of annual volume of discharge from the Nile River at Aswan from 1871 until 1970

(figure 2.1). The measurements indicate an abrupt change in rainfall around the turn of the century, which has been confirmed by records of tropical weather stations (Cobb, 1978). The estimated threshold $\hat{\psi} = 1898$ divides observations into two sets, $\{y_i | i \leq \hat{\psi}\}$ and $\{y_i | i > \hat{\psi}\}$. Until the year 1898, observations y_i are thought to have been generated according to the mechanism governing the first regime, $y_i = \mu_1 + \varepsilon_i$, thereafter according to the one characterizing the second regime, $y_i = \mu_2 + \varepsilon_i$. While model (2.1) illustrates the idea of threshold regression, it is certainly degenerate in the sense that regressors amount to a constant.

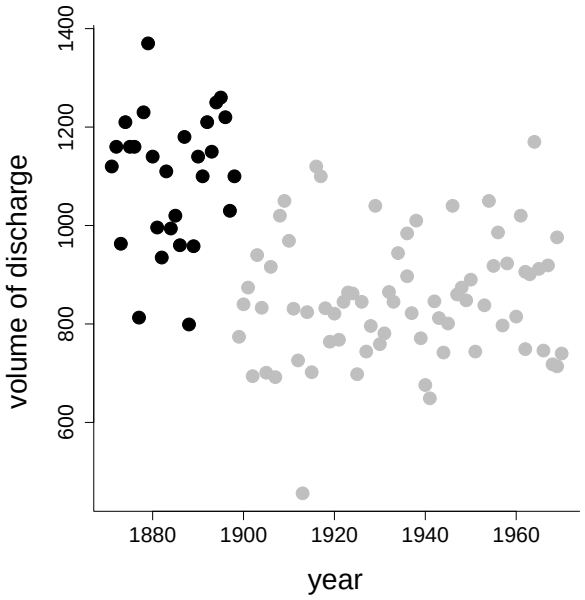

Figure 2.1: Annual volume of discharge from the Nile River at Aswan. Solid black circles represent observations in the first regime (years 1871 – 1898), grey circles those in the second regime (years 1899 – 1971).

The extension of this simple example to a (non-degenerate) regression setting, which in addition allows for a more flexible transition between regimes, arises naturally: A *threshold regression model* describes the relationship among observations $\left(y_i, \boldsymbol{X}_i^T, q_i\right) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, if

$$y_i = \{1 - T(q_i, \psi)\} \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + T(q_i, \psi) \boldsymbol{X}_i^T \boldsymbol{\beta}_2 + \varepsilon_i. \tag{2.2}$$

T denotes the transition function, which depends both on the transition variable q_i and the threshold parameter $\psi \in \mathbb{R}$ and is bounded between zero and one, $\mathbb{R} \ni q_i \mapsto T(q_i, \psi) \in [0, 1]$.

$\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2 \in \mathbb{R}^p$ are the regression coefficients and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$, are independent disturbances. Clearly, model (2.1) is covered by (2.2) with $\beta_1 = \mu_1$ and $\beta_2 = \mu_2$, explanatory variable $X_i = 1$, transition variable $q_i = i$, and transition function $T(i, \psi) = I(i > \psi)$; $I(\cdot)$ denotes the indicator function. It is sometimes advantageous to parameterize the model in terms of the differences between regimes, $\boldsymbol{\delta} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$,

$$y_i = \{1 - T(q_i, \psi)\} \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + T(q_i, \psi) \boldsymbol{X}_i^T \boldsymbol{\beta}_2 + \varepsilon_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + T(q_i, \psi) \boldsymbol{X}_i^T \boldsymbol{\delta} + \varepsilon_i.$$

While the assumption of independent normal disturbances $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$, is often relaxed, i.e. to ε_i satisfying $\mathrm{E}(\varepsilon_i) = 0$ and $\mathrm{E}(\varepsilon_i^2) < \infty$, it allows to embed model (2.2) in the framework of the generalized threshold regression model. This model can be understood as a generalized linear model with a threshold in the "linear" predictor: Observations $\left(y_i, \boldsymbol{X}_i^T, q_i\right) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, follow a *generalized threshold regression model* (GTRM) if the corresponding random variables satisfy

$$\mu_i = \mathrm{E}\left[y_i | \boldsymbol{X}_i^T, q_i\right] = h(\eta_i), \tag{2.3}$$

and

$$\eta_i = \{1 - T(q_i, \psi)\} \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + T(q_i, \psi) \boldsymbol{X}_i^T \boldsymbol{\beta}_2. \tag{2.4}$$

h is a known one-to-one function, the inverse of the link function $g = h^{-1}$. Moreover, conditional on q_i and the design vector $\boldsymbol{X}_i$, the response variables y_i are assumed to be drawn independently from an exponential family distribution with density

$$f(y_i | \psi, \phi, \beta_1, \beta_2) = \exp\left\{\frac{y_i \theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}, \tag{2.5}$$

characterized by known functions b and c, a scale (or dispersion) parameter $\phi \in \mathbb{R}^+$ and the natural parameter $\theta_i = \theta(\mu_i)$.

There exists a vast literature on threshold regression models dating back as far as Quandt (1958). A bibliography compiled three years ago (Khodadadi & Asgharian, 2008) lists more than 600 papers – even though it is limited to a certain kind of threshold regression model (change point model). The variety of fields of application of these models is remarkable. Hansen (2011) devotes an entire article to review the applications of one

specific threshold regression model in different areas of economics. Beckman & Cook (1979) mention per capita expenditure on police service, which first decreases with community size, but then suddenly jumps when policy services are added; release of myelin basic protein into sheep's spinal fluid; amino acid requirements for poultry. van de Geer (1988) motivates her study with an application relating (logarithmic) lifetimes of plastic pipes for transportation of fluids to the ratio of stress to absolute temperature and the inverse of absolute temperature. "The idea is that at high stress and temperature the pipes become brittle and break due to a mechanism different from the one at low stress and temperature" (page 3). Dose-response models in toxicology present a famous exemplar of non-normal threshold regression (Cox, 1987; Calabrese, Baldwin et al., 2003). Ravallion, Chen & Sangraula (2009) determine a global poverty line on the basis of a threshold regression model. Card, Mas & Rothstein (2008) employ a binary threshold regression model to study extreme segregation; they examine the hypothesis that all whites leave a neighborhood when the minority share surpasses a certain threshold level. Worsley (1983) adds several examples including "growth in the number of local telephone calls" and "pollen concentration in lake sediment cores". And nearly any other article on threshold regression opens with a similarly diverse list. Nevertheless, the literature focuses almost exclusively on threshold models such as (2.2) with the identity as the link function (or piecewise linear mean). A threshold regression model (2.3) – (2.5), in which piecewise linearity is not necessarily limited to the mean, has only recently been introduced by Samia & Chan (2011). However, they concentrate on a single type (characterized by a specific transition function and variable) of GTRM and I am not aware of other studies based on the broader framework (2.3) – (2.5). In contrast, numerous model specifications have been studied in the Gaussian case. Hence, I will limit myself to the latter set-up in the following classification of threshold regression models.

A first distinction in order to categorize different types of threshold regression models is based on the transition function T. This thesis concentrates on models with a *step transition function* $\mathbb{R} \ni q_i \mapsto I\left(q_i > \psi\right) \in \{0, 1\}$, i.e. equation (2.2) takes the form

$$y_i = I(q_i \leq \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_1 + I(q_i > \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_2 + \varepsilon_i = \boldsymbol{X}_i^T\boldsymbol{\beta}_1 + I(q_i > \psi)\boldsymbol{X}_i^T\boldsymbol{\delta} + \varepsilon_i. \qquad (2.6)$$

Smooth transition functions – that is, monotonously increasing smooth functions bounded between zero and one and satisfying $\lim_{q_i \to -\infty} T(q_i, \psi) = 0$ and $\lim_{q_i \to \infty} T(q_i, \psi) = 1$ for fixed

ψ – present a popular alternative. With reference to T, the respective models are called smooth transition models. Typically, such T depends on an additional parameter $\gamma \in \mathbb{R}^+$ determining its smoothness. A logistic transition function

$$T_\gamma\left(q_i, \psi\right) = \frac{1}{1 + \exp\left\{-\gamma\left(q_i - \psi\right)\right\}}$$

is often a suitable choice of a smooth transition function. van Dijk, Teräsvirta & Franses (2002) elaborate on this particular T_γ. They point out that T_γ approaches a step function $I\left(q_i > \psi\right)$ as γ becomes large (figure 2.2). For $\gamma \longrightarrow 0$, the smooth transition model merges into a linear model since $T_\gamma \longrightarrow T_0 = 1/2$; model (2.2) becomes $y_i = \boldsymbol{X}_i^T\left(0.5\boldsymbol{\beta}_1 + 0.5\boldsymbol{\beta}_2\right) + \varepsilon_i$ for $\gamma = 0$. It is natural to call ψ a threshold parameter in case of a transition function $I(q_i > \psi)$. However, it can also be considered as such for the logistic transition function $T_\gamma(q_i, \psi)$, as this increases monotonically from zero to one with q_i and $T_\gamma(\psi, \psi) = 0.5$. Lubrano (2000, sections 4.2 and 5.2) treats smooth transition functions in detail and provides further references. An extensive discussion on smooth transition functions in Bacon & Watts (1971) complements these. It is based on a different parameterization of (2.2), hence, puts forward different conditions for T_γ to fulfill.

Focusing on step transition functions $I(q_i > \psi)$, a second division of threshold regression models centers around the continuity of the regression function

$$I(q_i \leq \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_1 + I(q_i > \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_2. \tag{2.7}$$

It is called *continuous* if there is no jump at $q_i = \psi$, more precisely, if $\boldsymbol{X}_i^T\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right) = 0$ for all $\boldsymbol{X}_i$ such that $q_i = \psi$. Obviously, the distinction between continuous and discontinuous models is only meaningful if the transition variable q_i is at the same time one of the explanatory variables. Otherwise, continuity implies that $\boldsymbol{X}_i^T\left(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\right) = 0$ holds for arbitrary $\boldsymbol{X}_i$, and hence, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. The "threshold model" would be linear. More explicitly, for $\boldsymbol{X}_i^T = (1, q_i, X_{i,3}, \ldots, X_{i,p})$, the continuity condition is equivalent to requiring that $\beta_{1,1} + \beta_{1,2}\psi = \beta_{2,1} + \beta_{2,2}\psi$ and $\beta_{1,k} = \beta_{2,k}$ for $k = 3, \ldots, p$. In this case, (2.7) can be written as

$$I(q_i \leq \psi)(1, q_i)\begin{pmatrix}\beta_{1,1}\\ \beta_{1,2}\end{pmatrix} + I(q_i > \psi)(1, q_i)\begin{pmatrix}\beta_{2,1}\\ \beta_{2,2}\end{pmatrix} + (X_{i,3}, \ldots, X_{i,p})(\beta_{1,3}, \ldots, \beta_{1,p})^T.$$

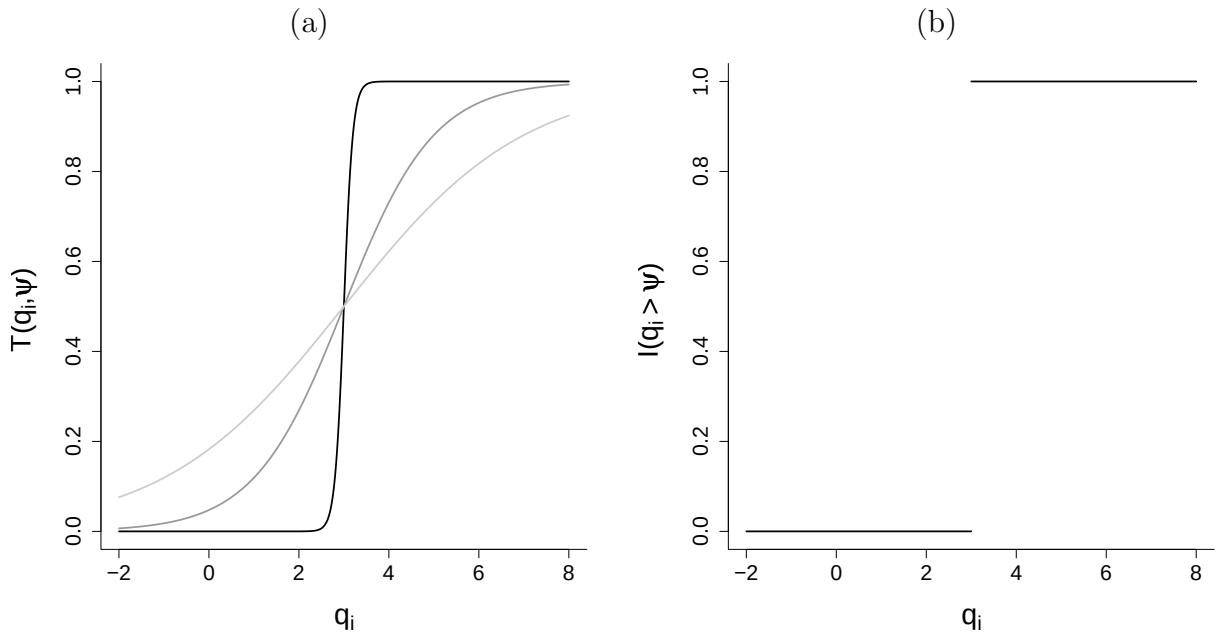

Figure 2.2: (a) Logistic smooth transition function for $\psi = 3$ and $\gamma = 10$ (black line), $\gamma = 1$ (darkgrey line), $\gamma = 0.5$ (lightgrey line). (b) Step transition function for $\psi = 3$.

Threshold regression models are not only classified according to the transition function, but also with respect to the transition variable. q_i can be random or deterministic. Models with a non-random transition variable $q_i = i$ are generally referred to as change point models. I will adopt this terminology. Hence, in the following, a threshold model is associated with a random variable q_i as opposed to a change point model, for which $q_i = i$.

As threshold estimation is at the heart of this thesis, a more thorough consideration of each of these four model types – discontinuous change point models, discontinuous threshold models, continuous change point models and continuous threshold models – will pay special attention to threshold estimators $\hat{\psi}$, hence, characterize distinctions between models in terms of the asymptotic properties of $\hat{\psi}$. Before doing so, one clarification: When dealing with change point problems, it is important to keep in mind that analysis is based one of two fundamentally different data gathering procedures. In a sequential setting, observations continue to arrive while statistical analysis is conducted, whereas in a retrospective setting, the examination is based on a fixed set of observa-

tions gathered beforehand. In this thesis, all analysis is retrospective. The sequential set-up will not be considered; see Lai (2001) for an overview on classical problems and more recent developments. Asymptotic theory in the retrospective set-up is typically based on the assumption that the true threshold $\psi_0 = [n\lambda_0]$ for a proportion $\lambda_0 \in (0,1)$ and $\mathbb{R} \ni s \mapsto [s] = \max\{m \in \mathbb{Z} | m \leq s\} \in \mathbb{Z}$ the greatest integer function. Clearly, this implies that ψ_0 varies with the number of observations n.

2.1.2 Discontinuous change point regression models

As outlined above, in case of a discontinuous change point model equation (2.2) can be further characterized by

$$y_i = I(i \leq \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_1 + I(i > \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_2 + \varepsilon_i = \boldsymbol{X}_i^T\boldsymbol{\beta}_1 + I(i > \psi)\boldsymbol{X}_i^T\boldsymbol{\delta} + \varepsilon_i \tag{2.8}$$

and $\boldsymbol{X}_i^T\boldsymbol{\delta} \neq 0$ for $i = \psi$. Model (2.1) is one example, figures 2.3 (a) and (b) visualize another. Bai (1997b) develops the asymptotic theory for this model. He examines the properties of the least squares estimators $\hat{\boldsymbol{\beta}}_1$, $\hat{\boldsymbol{\delta}}$ and $\hat{\psi}$ under a suitable set of assumptions. These guarantee that the change point ψ is bounded away from 1 and n, that there are enough observations around ψ for it to be identifiable, and that the central limit theorem holds, but are not too restrictive; they permit trending or lagged regressors. With $\boldsymbol{\beta}_{1\,0}$, $\boldsymbol{\delta}_0$ and ψ_0 denoting the true parameter values, he shows that for uncorrelated disturbances $\varepsilon_1, \ldots, \varepsilon_n$ with variance σ^2,

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1\,0} \\ \hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\boldsymbol{V}^{-1}\right)$$

where $\dfrac{1}{n}\begin{pmatrix} \sum\limits_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T & \sum\limits_{i=\psi_0}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T \\ \sum\limits_{i=\psi_0}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T & \sum\limits_{i=\psi_0}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^T \end{pmatrix} \xrightarrow{P} \boldsymbol{V}$; for serially correlated and heteroscedastic error terms $\varepsilon_1, \ldots, \varepsilon_n$,

$$\sqrt{n}\begin{pmatrix} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1\,0} \\ \hat{\boldsymbol{\delta}} - \boldsymbol{\delta}_0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\boldsymbol{V}^{-1}\boldsymbol{U}\boldsymbol{V}^{-1}\right)$$

where $\boldsymbol{U} = \lim_{n\longrightarrow\infty} \frac{1}{n} \begin{Bmatrix} \sum_{i,j\geq 1}^{n} \mathrm{E}\left(\boldsymbol{X}_i \boldsymbol{X}_j^T \varepsilon_i \varepsilon_j\right) & \sum_{i,j\geq \psi_0}^{n} \mathrm{E}\left(\boldsymbol{X}_i \boldsymbol{X}_j^T \varepsilon_i \varepsilon_j\right) \\ \sum_{i,j\geq \psi_0}^{n} \mathrm{E}\left(\boldsymbol{X}_i \boldsymbol{X}_j^T \varepsilon_i \varepsilon_j\right) & \sum_{i,j\geq \psi_0}^{n} \mathrm{E}\left(\boldsymbol{X}_i \boldsymbol{X}_j^T \varepsilon_i \varepsilon_j\right) \end{Bmatrix}$. Regarding the change point estimator $\hat{\psi}$, he proves that

$$\begin{aligned} \hat{\psi} &= \psi_0 + O_P\left(\|\boldsymbol{\delta}_0\|^{-2}\right) \\ \text{and} \qquad \hat{\psi} - \psi_0 &\xrightarrow{d} \arg\max_{m\in\mathbb{Z}} W(m) \end{aligned} \tag{2.9}$$

with

$$W(m) = \begin{cases} -\boldsymbol{\delta}_0^T \sum_{i=1}^{m} \boldsymbol{X}_i \boldsymbol{X}_i^T \boldsymbol{\delta}_0 - 2\boldsymbol{\delta}_0^T \sum_{i=1}^{m} \boldsymbol{X}_i \varepsilon_i & m > 0 \\ 0 & m = 0 \\ -\boldsymbol{\delta}_0^T \sum_{i=m+1}^{0} \boldsymbol{X}_i \boldsymbol{X}_i^T \boldsymbol{\delta}_0 + 2\boldsymbol{\delta}_0^T \sum_{i=m+1}^{0} \boldsymbol{X}_i \varepsilon_i & m < 0, \end{cases}$$

$m \in \mathbb{Z}$, which is a two-sided random walk in case $(\boldsymbol{X}_1, \varepsilon_1), \ldots, (\boldsymbol{X}_n, \varepsilon_n)$ are independent. For the special case of model (2.1), that is, $X_i = 1$,

$$W(m) = \begin{cases} \sum_{i=1}^{m} z_i & m > 0 \\ 0 & m = 0 \\ \sum_{i=m+1}^{0} z_i & m < 0, \end{cases} \tag{2.10}$$

$m \in \mathbb{Z}$, where $z_i \sim \mathcal{N}\left(-{\delta_0}^2, 4{\delta_0}^2\sigma^2\right)$. Hinkley (1970) also looks at this simple situation of a change in means (2.1) and states that $\hat{\psi} - \psi_0$ converges towards the argument maximizing a random walk with increments $\tilde{z}_i \sim \mathcal{N}\left(-2\Delta^2, 4\Delta^2\right)$ for $\Delta = |\delta_0|/2\sigma$. It is easy to see that the two results are equivalent. Since $\arg\max_{m\in\mathbb{Z}} W(m) = \arg\max_{m\in\mathbb{Z}} cW(m)$ for any constant $c \in \mathbb{R}^+$, it holds in particular that $\arg\max_{m\in\mathbb{Z}} W(m) = \arg\max_{m\in\mathbb{Z}} \widetilde{W}(m)$ for $\widetilde{W}(m) = W(m)/(2\sigma^2)$; and $\widetilde{W}$ is a two-sided random walk with normally distributed increments with mean $-2\Delta^2$ and variance $4\Delta^2$.

For an analysis of related models which complements the previous results with an examination of the speeds of estimation for the discontinuous change point model, not only when the true model is discontinuous, but also when it is continuous see van de Geer (1988, in particular examples 6.6 and 6.7).

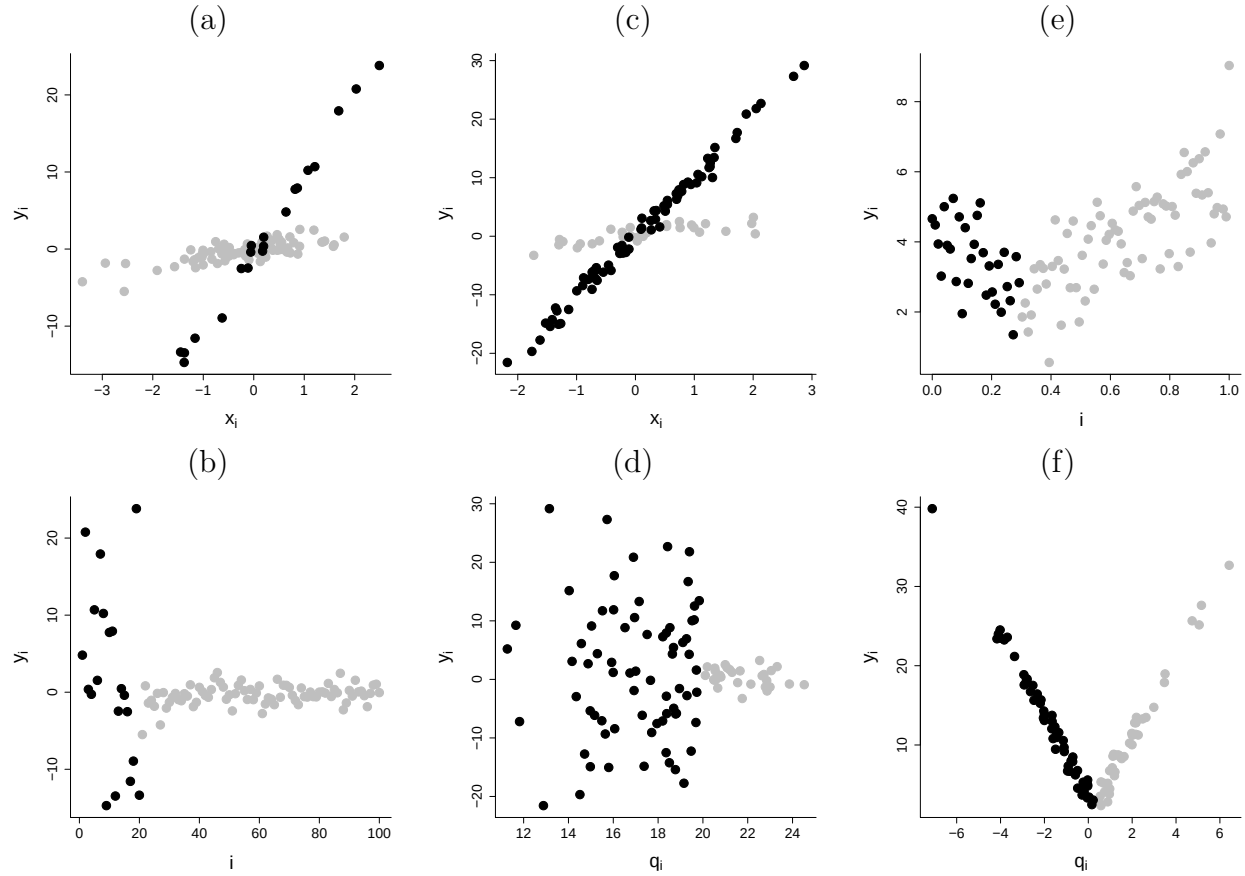

Figure 2.3: Simulated data for different types of threshold regression models. (a)–(d) Discontinuous model $y_i = 10X_iI(q_i \leq 20) + X_iI(q_i > 20) + \varepsilon_i$, $X_i \sim \mathcal{N}(0,1)$, and $q_i = i$ in (a)/(b), $q_i \sim \mathcal{N}(18,8)$ in (c)/(d). (e)–(f) Continuous model $y_i = (4-5q_i)I(q_i \leq 20) + (1+5q_i)I(q_i > 20) + \varepsilon_i$, $q_i = i$ in (e), $q_i \sim \mathcal{N}(0,5)$ in (f). In all examples $\varepsilon_i \sim \mathcal{N}(0,1)$ and $i = 1, \ldots, 100$.

Clearly, the asymptotic distribution of $\hat{\psi}$ not only depends on the true difference between regimes $\boldsymbol{\delta}_0$, but also on the distribution of the disturbances ε_i and the covariates $\boldsymbol{X}_i$, which is usually unknown. A strategy to obtain a limiting distribution that is usable in practice is to assume a diminishing difference between regimes. The resulting asymptotics can then be considered as an approximation for the case of a fixed shift. Bai (1997b) traces this idea back to Picard (1985), Bhattacharya (1987), and Yao (1987), who solved related problems, and transfers it to the discontinuous change point regression model. He develops the asymptotic distribution of the least squares estimator $\hat{\psi}$ given differences that converge to zero with increasing sample size, more explicitly, given

$\boldsymbol{\delta}_n = \boldsymbol{\delta}_0 \nu_n$ for $\nu_n \in \mathbb{R}^+$ such that $\nu_n \longrightarrow 0$ and $n^{0.5-\rho}\nu_n \longrightarrow \infty$ for some $\rho \in (0, 0.5)$ and $\boldsymbol{\delta}_0 \neq 0$. In the literature, this setting is sometimes referred to as "contiguous change". He derives a consistency result,

$$\hat{\psi} = \psi_0 + O_P\left(\|\boldsymbol{\delta}_n\|^{-2}\right).$$

Under additional assumptions, which ensure a functional central limit theorem and stationarity of $(\boldsymbol{X}_i, \varepsilon_i)$ within each regime, he proves that

$$\frac{\left(\boldsymbol{\delta}_n^T \Phi_1 \boldsymbol{\delta}_n\right)^2}{\boldsymbol{\delta}_n^T \boldsymbol{\Omega}_1 \boldsymbol{\delta}_n}\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} \arg\max_{s\in\mathbb{R}} Z(s) \tag{2.11}$$

where

$$Z(s) = \begin{cases} B_1(-s) - |s|/2 & s \leq 0 \\ \sqrt{\phi} B_2(s) - \omega |s|/2 & s > 0, \end{cases}$$

with

$$\phi = \left(\boldsymbol{\delta}_0^T \boldsymbol{\Phi}_2 \boldsymbol{\delta}_0\right) \Big/ \left(\boldsymbol{\delta}_0^T \boldsymbol{\Phi}_1 \boldsymbol{\delta}_0\right) \text{ and } \omega = \left(\boldsymbol{\delta}_0^T \boldsymbol{\Omega}_2 \boldsymbol{\delta}_0\right) \Big/ \left(\boldsymbol{\delta}_0^T \boldsymbol{\Omega}_1 \boldsymbol{\delta}_0\right),$$

$$\boldsymbol{\Phi}_1 = \mathrm{E}(\boldsymbol{X}_i \boldsymbol{X}_i^T) \text{ for } i = 1, \ldots, \psi_0, \text{ and } \boldsymbol{\Phi}_2 = \mathrm{E}(\boldsymbol{X}_i \boldsymbol{X}_i^T) \text{ for } i = \psi_0 + 1, \ldots, n,$$

$$\boldsymbol{\Omega}_1 = \lim_{n\to\infty} \mathrm{E}\left(1/\sqrt{\psi_0} \sum_{i=1}^{\psi_0} \boldsymbol{X}_i \varepsilon_i\right)^2 \text{ and } \boldsymbol{\Omega}_2 = \lim_{n\to\infty} \mathrm{E}\left(1/\sqrt{n-\psi_0} \sum_{i=\psi_0+1}^{n} \boldsymbol{X}_i \varepsilon_i\right)^2.$$

In addition, he underlines two special cases subsumed under the general statement (2.11). If $\boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2$ and $\boldsymbol{\Omega}_1 = \boldsymbol{\Omega}_2$ (which implies $\phi = \omega = 1$) hold, the asymptotic distribution reduces to

$$\frac{\left(\boldsymbol{\delta}_n^T \boldsymbol{\Phi}_1 \boldsymbol{\delta}_n\right)^2}{\boldsymbol{\delta}_n^T \boldsymbol{\Omega}_1 \boldsymbol{\delta}_n}\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} \arg\max_{s\in\mathbb{R}} \left\{B(s) - |s|/2\right\},$$

$B(s)$ a two-sided Brownian motion, $B(s) = B_1(-s)I(s \leq 0) + B_2(s)I(s > 0)$. The result furthers simplifies for uncorrelated errors $\varepsilon_1, \ldots, \varepsilon_n$ with variance σ^2. In this case $\boldsymbol{\Omega}_1 = \sigma^2 \boldsymbol{\Phi}_1$, hence,

$$\frac{\boldsymbol{\delta}_n^T \boldsymbol{\Phi}_1 \boldsymbol{\delta}_n}{\sigma^2}\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} \arg\max_{s\in\mathbb{R}} \left\{B(s) - |s|/2\right\}$$

(see also Hušková, 1996). The distribution function of $\arg\max_{s\in\mathbb{R}} \left\{B(s) - |s|/2\right\}$ is known

(Bhattacharya & Brockwell, 1976).

While Bai (1997b) develops asymptotic theory for the least squares estimator, Bhattacharya (1994) obtains analogous results for the maximum likelihood estimator. Hušková & Antoch (2001) look at M-estimators in general. Both of these articles also assume differences δ_n which approach zero for $n \to \infty$ at a rate slower than $\sqrt{n}$. Although the scaling factor is different, again convergence is towards $\arg\max_{s\in\mathbb{R}} \{B(s) - |s|/2\}$ (Hušková & Antoch, 2001, theorem 2.1).

2.1.3 Discontinuous threshold regression models

In contrast to the discontinuous change point model, the transition variable q_i is a random variable in the discontinuous threshold regression model. Hence, the latter can be written as

$$y_i = I(q_i \leq \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_1 + I(q_i > \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_2 + \varepsilon_i = \boldsymbol{X}_i^T\boldsymbol{\beta}_1 + I(q_i > \psi)\boldsymbol{X}_i^T\boldsymbol{\delta} + \varepsilon_i, \qquad (2.12)$$

where $\boldsymbol{X}_i^T\boldsymbol{\delta} \neq 0$ for $q_i = \psi$, and q_i is assumed to be a continuous random variable with distribution G, $\boldsymbol{X}_i \sim G$, and density g. Panels (c) and (d) of figure 2.3 depict a simple example.

Yu (2012) derives consistency and presents the asymptotic distribution of the maximum likelihood estimators $\hat{\boldsymbol{\beta}}$, $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T\right)^T$, and $\hat{\psi}$ for independently and identically distributed $\left(y_1, \boldsymbol{X}_1^T, q_1\right), \ldots, \left(y_n, \boldsymbol{X}_n^T, q_n\right)$. He presupposes a set of standard assumptions in nonlinear parametric estimation (Yu, 2012, appendix A, remark 1), augmented by a condition to ensure discontinuity. As an example, he points out that all of his assumptions are satisfied when ε_i is independent of $(\boldsymbol{X}_i^T, q_i)$ and $\varepsilon_i \sim \mathcal{N}(0,1)$, $i = 1, \ldots, n$. He shows that

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_P\left(n^{-1/2}\right) \qquad (2.13)$$

and

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\mathcal{J}}^{-1}\right), \qquad (2.14)$$

with $\boldsymbol{\mathcal{J}}$ the information matrix (of the regression coefficients $\boldsymbol{\beta}$.) For the maximum likelihood estimator of the threshold parameter, $\hat{\psi}$, he derives consistency

$$\hat{\psi} = \psi_0 + O_P\left(n^{-1}\right) \tag{2.15}$$

and the asymptotic distribution

$$n\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} M_-, \quad [M_-, M_+) = \arg\max_s D(s). \tag{2.16}$$

$D(s)$ is defined via two compound Poisson processes,

$$D(s) = \begin{cases} \sum_{i=1}^{\mathcal{P}_1(|s|)} z_{1i} & s \leq 0 \\ \sum_{i=1}^{\mathcal{P}_2(s)} z_{2i} & s > 0. \end{cases}$$

$\mathcal{P}_k(s)$, $k = 1, 2$, are two independent Poisson processes with intensity $g(\psi_0)$. z_{1i} and z_{2i} are random variables quantifying the impact of a threshold parameter ψ that is smaller or larger than the true threshold ψ_0, $\psi < \psi_0$ or $\psi > \psi_0$, respectively, on the likelihood ratio process of the model: Denoting the density of the disturbances ε_i by f, random variables $\bar{z}_{1i} = \log\left\{f\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_2\right) \Big/ f\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_1\right)\right\}$ and $\bar{z}_{2i} = \log\left\{f\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_1\right) \Big/ f\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_2\right)\right\}$ are defined. For $\Delta > 0$, z_{1i} is then specified as a random variable following the limiting conditional distribution of $\bar{z}_{1i}$ given $\psi_0 - \Delta < q_i \leq \psi_0$ when $\Delta \to 0$; z_{2i} in turn is distributed as $\bar{z}_{2i}$ given $\psi_0 < q_i \leq \psi_0 + \Delta$ when $\Delta \to 0$.

For a more restricted setting, i.e. for $\boldsymbol{X}_i^T = (1, X_{i2})$ and $q_i = X_{i2}$, Koul & Qian (2002) prove an analogous result. For this simpler model Koul, Qian & Surgailis (2003) generalize findings to cover M-estimators. In a time-series setting, Chan (1993) shows that the asymptotic distribution of the least squares estimator $\hat{\psi}$ involves a compound Poisson process. He considers the so-called self exciting threshold autoregressive model, that is, random variables $X_i, i \in \mathbb{Z}$, generated according to

$$X_i = \begin{cases} \beta_{10} + \beta_{11}X_{i-1} + \cdots + \beta_{1p}X_{i-p} + \varepsilon_n & X_{i-d} \leq \psi \\ \beta_{20} + \beta_{21}X_{i-1} + \cdots + \beta_{2p}X_{i-p} + \varepsilon_n & X_{n-d} > \psi, \end{cases} \tag{2.17}$$

with $\boldsymbol{\beta}_k = (\beta_{k0}, \ldots, \beta_{kp})^T \in \mathbb{R}^{p+1}$, $k = 1, 2$, $\psi \in \mathbb{R}$, $d \in \mathbb{N}$ the parameter specifying the

lag of the transition variable, and $\varepsilon_i, i \in \mathbb{Z}$, a series of errors with zero mean and finite variance. Qian (1998) develops the asymptotic theory for maximum likelihood estimators in Chan's (1993) modeling framework. As mentioned before, little attention has been paid to the GTRM (2.3) - (2.5). However, the case of a discontinuous model with random transition variable and step transition function has been investigated (Samia & Chan, 2011). Findings are analogous to (2.13) - (2.16).

As the discontinuous change point model (2.8) is closely related to the discontinuous threshold model (2.12) with $q_i \sim U[0,1]$, it is interesting to establish the link between the asymptotic distributions (2.9) and (2.16). Yu (2012, section 3.2) comments on this. He argues that in a change point model "q_i essentially follows the uniform distribution on $[0,1]$ which is independent of $(\boldsymbol{X}_i^T, \varepsilon_i)$, although its support is only a set of discrete points $q_i \in \{0, 1/n, \ldots, (n-1)/n, 1\}$". From that he infers that i) for the change point model $\bar{z}_{ki} = z_{ki}$ and ii) the interarrival time τ_{ki} of jumps in the Poisson processes $\mathcal{P}_k(s)$ equals one for $k = 1, 2$ and $i \in \mathbb{N}$. He explains the latter with the fact that the τ_{ki} are independently and identically exponentially distributed with mean $1/g(\psi_0)$, i.e. with mean one for the uniform density $g = I([0,1])$. With the interarrival time identical to that of a random walk, it suffices to take a closer look at z_{ki} to see the correspondence between (2.9) and (2.16). For the simplest case of model (2.1), that is, $\boldsymbol{X}_i^T\boldsymbol{\beta}_1 = \mu_1$, $\boldsymbol{X}_i^T\boldsymbol{\beta}_2 = \mu_2$, and f the normal density (with mean zero and variance σ^2)

$$\begin{aligned} z_{1i} &= \log\left\{ f\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_2\right) \Big/ f\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_1\right)\right\} \\ &= -\frac{1}{2\sigma^2}\left(y_i - \mu_2\right)^2 + \frac{1}{2\sigma^2}\left(y_i - \mu_1\right)^2 \\ &= \frac{1}{2\sigma^2}\left\{2\left(\mu_2 - \mu_1\right) y_i + \mu_1^2 - \mu_2^2\right\}, \end{aligned}$$

which, taking into account that $y_i \sim \mathcal{N}(\mu_1, \sigma^2)$, means that

$$z_{1i} \sim \mathcal{N}\left(\frac{1}{2\sigma^2}\left\{2\left(\mu_2 - \mu_1\right)\mu_1 + \mu_1^2 - \mu_2^2\right\}, \frac{\left(\mu_2 - \mu_1\right)^2}{\sigma^2}\right) = \mathcal{N}\left(-2\Delta^2, 4\Delta^2\right)$$

for $\Delta = \left|\mu_2 - \mu_1\right| / 2\sigma$. With analogous calculations for z_{2i}, this yields Hinkley's (1970) findings (compare with (2.10)).

As in the case of the change point model (2.8) the asymptotic distribution (2.16) for

a discontinuous threshold regression model (2.12) with fixed change between regimes depends on the distributions of ε_i and $\boldsymbol{X}_i$ as well as the unknown regression coefficients $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. It is, thus, not very useful in practice. Seo & Linton (2007) suggest to employ a smoothed estimator to cope with these difficulties. As an alternative to the least squares estimator obtained as the argument maximizing the objective function

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_1 - \boldsymbol{X}_i^T\boldsymbol{\delta}I(q_i > \psi)\right\}^2\\ =&\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_1\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\boldsymbol{X}_i^T\boldsymbol{\delta}\right)^2 - 2\boldsymbol{X}_i^T\boldsymbol{\delta}\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_1\right)\right\}I(q_i > \psi),\end{aligned}$$

they consider an objective function

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_1\right)^2 + \frac{1}{n}\sum_{i=1}^{n}\left\{\left(\boldsymbol{X}_i^T\boldsymbol{\delta}\right)^2 - 2\boldsymbol{X}_i^T\boldsymbol{\delta}\left(y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}_1\right)\right\}\mathcal{K}\left(\frac{q_i - \psi}{h_n}\right). \quad (2.18)$$

They propose to replace $I(q_i > \psi)$ by a smooth and bounded function $\mathbb{R} \ni s \mapsto \mathcal{K}(s) \in [a, b]$ which satisfies $\lim_{s\to-\infty}\mathcal{K}(s) = 0$ and $\lim_{s\to\infty}\mathcal{K}(s) = 1$. $\mathcal{K}(s) = F(s) - sf(s)$, F and f the standard normal cumulative distribution and density functions, respectively, is an example for such $\mathcal{K}$. $h_n \in \mathbb{R}^+$ is a bandwidth parameter. Seo & Linton (2007, section 3) prove that this smoothed least squares estimator is consistent and asymptotically normally distributed, hence, allows for standard inference techniques. The rate of convergence for the smoothed estimator, $\sqrt{nh_n^{-1}}$, is influenced by the bandwidth h_n and slower than n (as in (2.16)).

Another method to obtain a feasible asymptotic distribution for the threshold estimator is to proceed as in section 2.1.2 and assume diminishing differences $\boldsymbol{\delta}_n$. Hansen (2000) employs this paradigm. He assumes differences $\boldsymbol{\delta}_n = \boldsymbol{\delta}_0 n^{-\rho}$, $\boldsymbol{\delta}_0 \neq 0$ and $0 < \rho < 0.5$. For strictly stationary observations $\left(y_1, \boldsymbol{X}_1^T, q_1\right), \ldots, \left(y_n, \boldsymbol{X}_n^T, q_n\right)$ with certain mixing properties (hence, excluding integrated series or trending regressors, but still flexible enough to include the threshold autoregressive model, for example) and continuously distributed q_i, he develops the asymptotic distribution of the least squares estimator $\hat{\psi}$:

$$n^{1-2\rho}\frac{\left(\boldsymbol{\delta}_0^T\boldsymbol{\Phi}\boldsymbol{\delta}_0\right)^2 g(\psi_0)}{\boldsymbol{\delta}_0^T\boldsymbol{\Omega}\boldsymbol{\delta}_0}\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} \arg\max_s\left\{B(s) - |s|/2\right\}, \quad (2.19)$$

where as above $B_k(s)$, $k = 1, 2$, are two independent Brownian motions on $[0, \infty)$, $B_k(0) = 0$, and $B(s) = B_1(-s)I(s \leq 0) + B_2(s)I(s > 0)$. $\boldsymbol{\Phi} = \mathrm{E}(\boldsymbol{X}_i \boldsymbol{X}_i^T | q_i = \psi_0)$ and $\boldsymbol{\Omega} = \mathrm{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^T \varepsilon_i^2 | q_i = \psi_0\right)$.

Substituting $\boldsymbol{\delta}_0 n^{-\rho}$ by $\boldsymbol{\delta}_n$ and multiplying $\boldsymbol{\Omega}$ by $1/n$ to create an "equivalent" to $\boldsymbol{\Omega}_k$ in (2.11) makes the analogy between (2.11) and (2.19) even more evident,

$$n^{1-2\rho} \frac{\left(\boldsymbol{\delta}_0^T \boldsymbol{\Phi} \boldsymbol{\delta}_0\right)^2 g(\psi_0)}{\boldsymbol{\delta}_0^T \boldsymbol{\Omega} \boldsymbol{\delta}_0} = \frac{\left(\boldsymbol{\delta}_n^T \boldsymbol{\Phi} \boldsymbol{\delta}_n\right)^2 g(\psi_0)}{\boldsymbol{\delta}_n^T \left(\boldsymbol{\Omega}/n\right) \boldsymbol{\delta}_n}$$

As Hansen (2000) points out, "the difference is that the asymptotic precision of $\hat{\psi}$ is proportional to the matrix $\mathrm{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^T | q_i = \psi_0\right)$ while in the change point case the asymptotic precision is proportional to the unconditional moment matrix $\mathrm{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^T\right)$." He further notes that the asymptotic distribution of $\hat{\psi}$ becomes less dispersed with larger $g(\psi_0)$, i.e. with an increasing number of observations close to the true threshold, and $\boldsymbol{\delta}_0$, i.e. when the difference between regimes grows, which is very intuitive. Moreover, he underlines the reduced rate of convergence compared with the case of fixed differences.

2.1.4 Continuous change point regression models

In a continuous change point regression model equation (2.2) takes the form

$$y_i = I(i \leq \psi)\left(\beta_{11} + \beta_{12} i\right) + I(i > \psi)\left(\beta_{21} + \beta_{22} i\right) + (X_{i,3}, \ldots, X_{i,p})(\beta_{1,3}, \ldots, \beta_{1,p})^T + \varepsilon_i,$$

$\psi = (\beta_{11} - \beta_{21}) \,/\, (\beta_{22} - \beta_{12})$. Ignoring the regressors not affected by the threshold, $X_{i,3}, \ldots, X_{i,p}$, the model simplifies to

$$y_i = I(i \leq \psi)\left(\beta_{11} + \beta_{12} i\right) + I(i > \psi)\left(\beta_{21} + \beta_{22} i\right) + \varepsilon_i, \quad \psi = \frac{\beta_{11} - \beta_{21}}{\beta_{22} - \beta_{12}}. \tag{2.20}$$

An example is visualized in figure 2.3 (e). The results of Feder's (1975) examination of the more general model

$$y_{ni} = I\left(i \leq n\psi\right) f_1\left(\theta_1, i/n\right) + I(i > n\psi) f_2\left(\theta_2, i/n\right) + \varepsilon_i, \quad i = 1, \ldots, n$$

where $f_k(\theta_k, t)$ are functions that can be represented as $f_k(\theta_k, t) = \sum_j^{p(k)} \theta_{kj} f_{kj}(t)$ for $\theta_k \in \mathbb{R}^{p(k)}$ and suitable $[0,1] \ni t \mapsto f_{kj}(t) \in \mathbb{R}$, $j = 1, \ldots, J(k)$ and $k = 1, 2$, with $f_1(\theta_1, \psi) = f_2(\theta_2, \psi)$ for continuity, apply to the piecewise linear model (2.20) in particular. When comparing this section's findings with those stated in section 2.1.2, it is important to keep in mind that here ψ has been rescaled by a factor $1/n$. It is essentially equal to the proportion $\lambda \in (0,1)$ defined by $\psi = [n\lambda]$ in section 2.1.1. While now observations are taken at time points $t = 1/n, \ldots, (n-1)/n, 1$, that is $\psi \in (0,1)$, time ranges from $t = 1, \ldots, n$ and $\psi \in \{1, \ldots, n\}$ in section 2.1.2 and accordingly $\psi \in \{1, \ldots, n\}$.

Assuming appropriate identifiability conditions, Feder (1975) proves consistency for the least squares estimators of the regression coefficients $\boldsymbol{\beta} = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})^T$,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + O_P\left(n^{-1/2}\sqrt{\log\log n}\right),$$

and the threshold parameter ψ,

$$\hat{\psi} = \psi_0 + O_P\left(n^{-1/2}\sqrt{\log\log n}\right)$$

(Feder, 1975, theorem 3.18). He further states that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\mathcal{J}}^{-1}\right)$$

with $\boldsymbol{\mathcal{J}}$ the information matrix,

$$\boldsymbol{\mathcal{J}} = \frac{\psi_0}{\sigma^2}\begin{pmatrix} 1 & \psi_0/2 & 0 & 0 \\ \psi_0/2 & \psi_0^2/3 & 0 & 0 \\ 0 & 0 & 1 & \psi_0/2 \\ 0 & 0 & \psi_0/2 & \psi_0^2/3 \end{pmatrix},$$

and

$$\sqrt{n}\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} \mathcal{N}\left(0, \boldsymbol{\gamma}\boldsymbol{\mathcal{J}}^{-1}\boldsymbol{\gamma}^T\right),$$

where $\boldsymbol{\gamma} = (1, \psi_0, -1, -\psi_0)/(\beta_{22} - \beta_{12})$ (Feder, 1975, page 77). With reference to an empirical investigation, Hinkley (1969a) remarks that this asymptotic distribution of $\hat{\psi}$ does not present a good approximation in small samples and, consequently, discusses an

alternative (see also Hinkley, 1969b).

2.1.5 Continuous threshold regression models

The specific version of equation (2.2)

$$y_i = I(q_i \leq \psi)\left(\beta_{11} + \beta_{12} q_i\right) + I(q_i > \psi)\left(\beta_{21} + \beta_{22} q_i\right) + (X_{i,3}, \ldots, X_{i,p})(\beta_{1,3}, \ldots, \beta_{1,p})^T + \varepsilon_i$$

characterizes a continuous threshold regression model when $\psi = (\beta_{11} - \beta_{21}) \,/\, (\beta_{22} - \beta_{12})$ and $q_i \sim G$, $i = 1, \ldots, n$, are random variables, which are continuously distributed according to G. Figure 2.3 (f) illustrates the model for a simple example. I am not aware of asymptotic results for this general model formulation. However, Chan & Tsay (1998) study the special case of a continuous self exciting threshold autoregressive model (2.17) restricted to satisfy $\boldsymbol{X}_i^T(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) = 0$ if $X_{i-d} = \psi$. Since continuity implies that $\beta_{1j} = \beta_{2j}$ for $j \neq d$, it is in this case possible to express (2.17) as

$$X_i = \beta_0 + \sum_{j=1, j\neq d}^{p} \beta_j X_{i-j} + \begin{cases} \beta_{1d}\left(X_{i-d} - \psi\right) + \varepsilon_i & X_{i-d} \leq \psi \\ \beta_{2d}\left(X_{i-d} - \psi\right) + \varepsilon_i & X_{i-d} > \psi \end{cases}$$

for $\beta_0 = \beta_{10} + \psi\beta_{1d}$ and $\beta_j = \beta_{1j}$, $j \neq d$. Chan & Tsay (1998, theorem 2.1) show that the least squares estimators for $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{d-1}, \beta_{1d}, \beta_{2d}, \beta_{d+1}, \ldots, \beta_p)$ and ψ are strongly consistent,

$$\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) \longrightarrow (\boldsymbol{\beta}_0, \psi_0) \quad \text{almost surely,}$$

when $\beta_{1d} \neq \beta_{2d}$, $X_i, i \in \mathbb{Z}$, is stationary and ergodic, has finite second moments and the distribution of $(X_1, \ldots, X_p)^T$ is stationary and positive everywhere.

In addition, they derive the asymptotic distribution of the estimators. Assuming that $X_i, i \in \mathbb{Z}$, is stationary, satisfies certain mixing criteria, has a density function that is positive everywhere and bounded over a neighborhood of ψ_0, $\mathrm{E}\left(|X_i|^q\right) < \infty$ for some $q > 2$, and $\beta_{1d} \neq \beta_{2d}$, they show that

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}, \hat{\psi}\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2 \boldsymbol{\mathcal{J}}^{-1}\right),$$

where $\boldsymbol{\mathcal{J}} = \mathrm{E}\left(\boldsymbol{S}_i \boldsymbol{S}_i^T\right)$ and $\boldsymbol{S}_i$ the vector of the derivatives of

$$X_i - \beta_0 - \sum_{j=1, j\neq d}^{p} \beta_j X_{i-j} - \begin{cases} \beta_{1d}\left(X_{i-d} - \psi\right) & X_{i-d} \leq \psi \\ \beta_{2d}\left(X_{i-d} - \psi\right) & X_{i-d} > \psi \end{cases}$$

with respect to $(\boldsymbol{\beta}, \psi)$, i.e.

$$\boldsymbol{S}_i = \Big\{ -1, -X_{i-1}, \ldots, -X_{i-d+1}, -\left(X_{i-d} - \psi\right)_-, -\left(X_{i-d} - \psi\right)_+, \ldots \\ - X_{i-p}, \beta_{1d} I(X_{i-d} \leq \psi) + \beta_{2d} I(X_{i-d} > \psi) \Big\}^T.$$

Since the derivative with respect to ψ is not defined at $X_{i-d} = \psi$, it is by convention set equal to β_{1d} (Chan & Tsay, 1998, theorem 2.2).

2.1.6 Models with multiple thresholds

In this section, I comment on the extension of the models discussed in sections 2.1.2 through 2.1.5 to cover multiple instead of a single threshold. Bai & Perron (1998) study a version of the change point model (2.8) with J breaks,

$$y_i = \sum_{j=0}^{J} \boldsymbol{X}_i^T \boldsymbol{\beta}_j I(\psi_j \leq i < \psi_{j+1}) + \varepsilon_i,$$

$0 = \psi_0 < \psi_1 < \ldots < \psi_J < \psi_{J+1} = n$. They prove a consistency result (Bai & Perron, 1998, proposition 1) and derive asymptotic normality for the least squares estimator of the regression coefficients (proposition 3). Before analyzing the limiting distribution of the least squares estimator $\hat{\psi} = (\hat{\psi}_1, \ldots, \hat{\psi}_J)^T$, they note that "as in the single break case, the usual limiting distribution of the break dates obtained specifying fixed magnitudes of changes depends on the exact distribution of the pair $(\boldsymbol{X}_i^T, \varepsilon_i)$". Consequently, they propose to develop the asymptotic distribution within a framework of diminishing differences between regimes with increasing sample size. In analogy to the case of two regimes, for $j = 1, \ldots, J$ they assume that $\boldsymbol{\beta}_{n,j+1} - \boldsymbol{\beta}_{n,j} = \boldsymbol{\delta}_{n,j} = \boldsymbol{\delta}_{0,j}\nu_n$ for $\boldsymbol{\delta}_{0,j}$ independent of n and $\nu_n \in \mathbb{R}^+$ satisfying $\nu_n \longrightarrow 0$ and $n^{0.5-\rho}\nu_n \longrightarrow \infty$ for some $\rho \in (0, 0.5)$. For suitably defined $\boldsymbol{\Phi}_j$, $\phi_{1,j}$, $\phi_{2,j}$, ω_j and two independent Wiener processes on $[0, \infty)$

starting at zero, $B_k^{(j)}(s)$, $k = 1, 2$ and $j = 1, \ldots, J$, they form

$$Z^{(j)}(s) = \begin{cases} \phi_{1,j} B_1^{(j)}(-s) - |s|/2 & s \leq 0 \\ \phi_{2,j} \sqrt{\omega_j} B_2^{(j)}(s) - \omega_j |s|/2 & s > 0, \end{cases}$$

and show that for each j

$$\left(\boldsymbol{\delta}_{0,j}^T \boldsymbol{\Phi}_j \boldsymbol{\delta}_{0,j}\right) \nu_n^2 \left(\hat{\psi}_j - \psi_{0,j}\right) \xrightarrow{d} \arg\max_s Z^{(j)}(s). \tag{2.21}$$

Hence, the limiting distribution for the single change point case generalizes to the case of multiple change points. Perron & Qu (2006) complement these findings proving that imposing restrictions on the model (for example, that different nonadjacent regimes are identical) does not affect the limiting distribution of the least squares estimator $\hat{\psi}$. However, they stress that all model parameters are estimated more efficiently in small samples when restrictions are taken into account.

Since it can be computationally very expensive to estimate a large number of change points simultaneously, Bai (1997a) investigates the idea of estimating them one at a time. This procedure yields n-consistent estimates (Bai, 1997a, proposition 2). Perron (2006, section 3.5) explains this as follows: "When estimating a single break model in the presence of multiple breaks, the estimate of the break fraction will converge to one of the true break fractions, the one that is dominant in the sense that taking it into account allows for greatest reduction in the sum of squared residuals. Then, allowing for a break at the estimated value, a second one break model can be applied which will consistently estimate the second dominating break and so on." Bai (1997a) finds that the limiting distribution for the sequential estimators is distinct from that of the simultaneous estimator. While the latter depends on the parameters characterizing regimes j and $j + 1$ for the j-th change point, the limiting distribution of the j-th sequentially estimated change point involves more model parameters. He suggests a re-estimation strategy based on initial sequential estimates to overcome this problem (Bai, 1997a, section 6) and obtains the same limiting distribution as for the simultaneous estimator employing this "repartition" method.

Multiple threshold versions of the discontinuous threshold regression model (2.12) have been considered by Ciuperca & Dapzol (2008), Fujii (2008) and Chan & Kutoyants

(2010b). Ciuperca & Dapzol (2008) study the piecewise linear model

$$y_i = \sum_{j=0}^{J} (1, X_i)\beta_j I(\psi_j \leq X_i < \psi_{j+1}) + \varepsilon_i$$

with $\beta_j \in \mathbb{R}^2$, $0 = \psi_0 < \psi_1 < \ldots < \psi_J < \psi_{J+1} = 1$, and $X_i \sim G$ with continuous density g on $(0, 1)$. They impose discontinuity, i.e. require that $\beta_{j,1} - \beta_{j-1,1} \neq (\beta_{j-1,2} - \beta_{j,2})\,\psi_j$ for $j = 1, \ldots, J$. They prove consistency of the maximum likelihood estimator $\hat{\psi} = (\hat{\psi}_1, \ldots, \hat{\psi}_J)^T$,

$$\hat{\psi} = \psi_0 + O_P\left(n^{-1}\right),$$

and show that the limit of the likelihood process amounts to a sum of J independent random processes $D^{(j)}(s) = D_1^{(j)}(-s)I(s \leq 0) + D_2^{(j)}(s)I(s > 0)$, $j = 1, \ldots, n$. Here, $D_k^{(j)}(s)$, $k = 1, 2$, are independent compound Poisson processes with rate $g(\psi_{0j})$ on $[0, \infty)$, which start at zero (Ciuperca & Dapzol, 2008, theorems 3.2 and 3.5). As in the case of a single threshold, the limiting distribution is

$$n\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} M_-. \tag{2.22}$$

However, M_- is now a vector, $M_- = (M_{1-}, \ldots, M_{J-})$, and M_{j-} defined by

$$[M_{j-}, M_{j+}) = \arg\max_s D^{(j)}(s).$$

Chan & Kutoyants (2010b, section 7) touch upon multiple thresholds in the context of the J-regime self-exciting threshold autoregressive model. They outline how to deduce the limit likelihood ratio and come up with a similar result as Ciuperca & Dapzol (2008) involving J compound Poisson processes. While Ciuperca & Dapzol (2008) as well as Chan & Kutoyants (2010b) develop asymptotic results assuming a fixed magnitude of shift between regimes, Fujii (2008) also takes care of a magnitude of shift which converges to zero. He examines the model

$$y_i = \sum_{j=1}^{J+1} f_j(X_i)I(\psi_{j-1} \leq X_i < \psi_j) + \varepsilon_i$$

where $(0, 1) \ni t \mapsto f_j(t) \in \mathbb{R}$ are known L_2-integrable functions, and, as above,

$0 = \psi_0 < \psi_1 < \ldots < \psi_J < \psi_{J+1} = 1$ while $X_i \sim G$ with continuous density g on $(0, 1)$. Although this model does not cover piecewise linear regression (unless all regression coefficients are known and there is a single explanatory variable apart from a constant), I include Fujii's (2008) findings to complete the picture. In addition to investigating fixed differences between regimes and deriving an analogue of (2.22), he studies the limiting distribution of the maximum likelihood estimator $\hat{\psi}$ given a decreasing magnitude of shifts with growing number of observations. Assuming that $f_{j+1}(t) = f_j(t) + \delta_{0,j}\nu_n$ for constant $\delta_{0,j} \in \mathbb{R}$ and ν_n satisfying $\nu_n \longrightarrow 0$ and $n\nu_n^2 \longrightarrow \infty$ as $n \to \infty$, he shows that

$$\hat{\psi}_j - \psi_{0j} \xrightarrow{d} \arg\max_s \left\{ B^{(j)}(\gamma_j s) - \gamma_j |s| / 2 \right\}. \tag{2.23}$$

For $j = 1, \ldots, J$, γ_j varies with $\delta_{0,j}$, $g(\psi_{0j})$ and the distribution of ε_i, and

$$B^{(j)}(s) = B_1^{(j)}(-s)I(s \leq 0) + B_2^{(j)}(s)I(s > 0),$$

$B_1^{(j)}$ and $B_2^{(j)}$ two independent Brownian motions on $[0, \infty)$, $B_1^{(j)}(0) = B_2^{(j)}(0) = 0$.

Clearly, for discontinuous models both with deterministic and random transition variable, asymptotic properties of the threshold estimator for models with a single threshold generalize to the multiple threshold setting in a natural way. Regarding continuous models, Feder (1975) actually develops his asymptotic results presented in section 2.1.4 for the multiple-regime model. Hence, asymptotic normality continuous to hold in case of the continuous multiple change point model.

2.1.7 Further generalizations

This section contains brief remarks on regime-dependent heteroscedasticity, multivariate threshold regression models and generalizations of the transition variable. The purpose is solely to highlight different possibilities to generalize the above models that have been explored in the literature; theoretical properties are not regarded.

One obvious generalization of a model

$$y_i = I(q_i \leq \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_1 + I(q_i > \psi)\boldsymbol{X}_i^T\boldsymbol{\beta}_2 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}\left(0, \sigma^2\right)$$

is to incorporate error terms whose distribution varies according to regime,

$$y_i = I(q_i \leq \psi)\left(\boldsymbol{X}_i^T\boldsymbol{\beta}_1 + \sigma_1^2\varepsilon_i\right) + I(q_i > \psi)\left(\boldsymbol{X}_i^T\boldsymbol{\beta}_2 + \sigma_2^2\varepsilon_i\right), \quad \varepsilon_i \sim \mathcal{N}(0,1).$$

Several of the studies extensively drawn upon in the preceding sections allow for regime-dependent heteroscedasticity (Yu, 2012; Bai, 1997b; Chan & Tsay, 1998), yet not all of them (Hansen, 2000; Feder, 1975). There are also articles explicitly focusing on changes in variance between regimes (Horváth, Hušková & Serbinowska, 1997).

Another natural idea is to try and extend results to cover multivariate models. The most comprehensive treatment of a multivariate version of the discontinuous change point regression model is delivered by Qu & Perron (2007). They look at the model

$$\boldsymbol{y}_i = \sum_{j=0}^{J}\left(\boldsymbol{X}_i^T\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_{ji}\right) I(\psi_j \leq i < \psi_{j+1}) \tag{2.24}$$

and further specify the regressor matrix as $\boldsymbol{X}_i^T = \left(\boldsymbol{I}_m \otimes \boldsymbol{z}_i^T\right)\boldsymbol{S}$. Here, $\boldsymbol{y}_i$ and $\boldsymbol{\varepsilon}_{ji} \in \mathbb{R}^m$, $\boldsymbol{z}_i \in \mathbb{R}^q$ is the vector of regressors, $\boldsymbol{I}_m \in \mathbb{R}^{m\times m}$ the identity matrix and $\boldsymbol{S} \in \mathbb{R}^{mq\times p}$ the matrix selecting the regressors to enter each of the equations. Regime-dependent parameters include $\boldsymbol{\beta}_j \in \mathbb{R}^p$ and the covariance matrix $\boldsymbol{\Sigma}_j$ of the errors. Qu & Perron (2007) name vector autoregressive and linear panel data models as well as partial structural change models as examples of (2.24) which also comply with their further assumptions necessary to allow for a mathematical treatment. Their asymptotic results for a (restricted) quasi maximum likelihood estimator (based on a Gaussian distribution of the disturbances) are qualitatively similar to those seen above (equation (2.21), section 2.1.2). As an example, for differences between regimes of a fixed magnitude and given $\boldsymbol{\Sigma}_j = \sigma^2\boldsymbol{I}_m$ for $j = 1, \ldots, J$,

$$\hat{\psi}_j - \psi_{0j} \xrightarrow{d} \arg\max_{m\in\mathbb{Z}} W^{(j)}(m)$$

with

$$W^{(j)}(m) = \begin{cases} -\sum\limits_{i=\psi_{0j}+1}^{\psi_{0j}+m} \left(\boldsymbol{\delta}_{0j}^T \boldsymbol{X}_i \boldsymbol{X}_i^T \boldsymbol{\delta}_{0j}\right)/(2\sigma^2) + \sum\limits_{i=\psi_{0j}+1}^{\psi_{0j}+m} \left(\boldsymbol{\delta}_{0j}^T \boldsymbol{X}_i \boldsymbol{\varepsilon}_{ji}\right)/\sigma^2 & m > 0 \\ 0 & m = 0 \\ -\sum\limits_{i=\psi_{0j}+m}^{\psi_{0j}} \left(\boldsymbol{\delta}_{0j}^T \boldsymbol{X}_i \boldsymbol{X}_i^T \boldsymbol{\delta}_{0j}\right)/(2\sigma^2) - \sum\limits_{i=\psi_{0j}+m}^{\psi_{0j}} \left(\boldsymbol{\delta}_{0j}^T \boldsymbol{X}_i \boldsymbol{\varepsilon}_{ji}\right)/\sigma^2 & m < 0, \end{cases}$$

$m \in \mathbb{Z}$, $\boldsymbol{\delta}_{0j} = \boldsymbol{\beta}_{0j} - \boldsymbol{\beta}_{0j+1}$, for $j = 1, \ldots, J$. For diminishing change, they prove a limiting distribution resembling (2.23). An important aspect of a multivariate modeling framework is underlined by Perron (2006, section 3.6). He emphasizes potential efficiency gains by examining a system of equations jointly: "The precision of a particular break date in one equation can increase when the system includes other equations even if the parameters of the latter are invariant across regimes. All that is needed is that the correlation between errors is non-zero. While surprising, this result is ex-post fairly intuitive since a poorly estimated break in one regression affects the likelihood function through both the residual variance of that equation and the correlation with the rest of the regressions. Hence, by including ancillary equations without breaks, additional forces are in play to better pinpoint break dates."

A different suggestion to go beyond model (2.6) concentrates on the transition variable q_i. Tishler & Zang (1979) analyze a model

$$y_i = \begin{cases} \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + \varepsilon_i & \boldsymbol{q}_i^T \boldsymbol{\pi} \leq 0 \\ \boldsymbol{X}_i^T \boldsymbol{\beta}_2 + \varepsilon_i & \boldsymbol{q}_i^T \boldsymbol{\pi} > 0, \end{cases} \tag{2.25}$$

that is, allow for an m-vector $\boldsymbol{q}_i$ of transition variables instead of a single scalar variable (apart from a constant). The switch between regimes is then determined by a linear combination $\boldsymbol{q}_i^T \boldsymbol{\pi}$ of these transition variables which depends on an unknown "threshold" parameter $\boldsymbol{\pi} \in \mathbb{R}^m$. Cast this way, the original model (2.6) arises when the transition variable is specified as $(1, q_i)$ and $\boldsymbol{\pi} = (-\psi, 1)$. Another example for the treatment of such a more general model is Seo & Linton (2007). Bauwens, Lubrano & Richard (1999, section 8.6) discuss a restricted version of (2.25), which appears in the economics literature as "disequilibrium model". First, regimes do not have common explanatory variables. This means that for $\boldsymbol{X}_i^T \boldsymbol{\beta}_k$ written as $\boldsymbol{X}_i^T \boldsymbol{\beta}_k = \boldsymbol{X}_{1i}^T \boldsymbol{\beta}_{k1} + \boldsymbol{X}_{2i}^T \boldsymbol{\beta}_{k2}$, $k = 1, 2$,

it is known that $\boldsymbol{\beta}_{12} = \boldsymbol{\beta}_{21} = 0$. Second, $\boldsymbol{q}_i^T \boldsymbol{\pi} = \boldsymbol{X}_{1i}^T \boldsymbol{\beta}_{11} - \boldsymbol{X}_{2i}^T \boldsymbol{\beta}_{22}$. This special case of model (2.25) can equivalently be expressed as

$$y_i = \min\left(\boldsymbol{X}_{1i}^T \boldsymbol{\beta}_{11}, \boldsymbol{X}_{2i}^T \boldsymbol{\beta}_{22}\right) + \varepsilon_i.$$

van de Geer (1988) further generalizes (2.25) to

$$y_i = \begin{cases} \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + \varepsilon_i & \boldsymbol{X}_i \in A \\ \boldsymbol{X}_i^T \boldsymbol{\beta}_2 + \varepsilon_i & \boldsymbol{X}_i \notin A, \end{cases}$$

with A varying in a class $\mathcal{A}$ of subsets of $\mathbb{R}^p$. She analyzes this model for non-random $\boldsymbol{X}_i$ and investigates conditions on $\mathcal{A}$ to guarantee consistency and asymptotic normality of the least squares estimator $\hat{\boldsymbol{\beta}}$. Pole & Smith (1985) similarly allow for a more flexible mechanism to switch between regimes by formulating a model

$$y_i = \begin{cases} \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + \varepsilon_i & g\left(\boldsymbol{\pi}, \boldsymbol{q}_i\right) \leq 0 \\ \boldsymbol{X}_i^T \boldsymbol{\beta}_2 + \varepsilon_i & g\left(\boldsymbol{\pi}, \boldsymbol{q}_i\right) > 0 \end{cases}$$

for some function g, $\mathbb{R}^q \times \mathbb{R}^m \ni (\boldsymbol{\pi}, \boldsymbol{q}_i) \mapsto g\left(\boldsymbol{\pi}, \boldsymbol{q}_i\right) \in \mathbb{R}$.

When taking into account more than just a scalar transition variable, it is natural to think of including both deterministic and random elements. van Dijk, Teräsvirta & Franses (2002, section 3.2) comment on this idea in the context of threshold regression models with smooth transition function (so-called smooth transition autoregressive models). With respect to these time-series models, they point out that "despite a large amount of evidence indicating that both nonlinearity [a random transition variable] and structural change [a time index as transition variable] are relevant for many time series, to date these features have mainly been analyzed in isolation. A reason for this dichotomy may be that nonlinearity, and regime-switching behavior in particular, and structural change can be regarded as competing alternatives to linearity and it might be difficult to distinguish between the two." Within the framework of smooth transition autoregression, the TVSTAR (time-varying smooth transition autoregression) incorporates this type of dynamic. Here, one of the transition variables is random, the other a time-index. van

Dijk, Teräsvirta & Franses (2002) state this model as

$$y_i = \{1 - T_1(q_i, \psi_1)\} \boldsymbol{X}_i^T \boldsymbol{\beta}_1(i) + T_1(q_i, \psi_1) \boldsymbol{X}_i^T \boldsymbol{\beta}_2(i) + \varepsilon_i,$$

and

$$\boldsymbol{\beta}_k(i) = \{1 - T_2(i, \psi_2)\} \boldsymbol{\beta}_{k1} + T_2(i, \psi_2) \boldsymbol{\beta}_{k2}, k = 1, 2.$$

T_1 and T_2 are two transition functions. For a detailed analysis, they refer to Lundbergh, Teräsvirta & Van Dijk (2003).

Lastly, the condition that transition variables be exogenous (which has not been explicitly discussed here, but is assumed in many important contributions such as Hansen, 2000, for example) might be relaxed. A recent article by Kourtellos, Stengos & Tan (2011) examines this possibility.

Note that it is clear from the definition of a threshold regression model in this thesis, that the transition variable q_i is observed. An unobservable q_i opens up a whole new domain of models. Hidden Markov models are a famous example (Zucchini & MacDonald, 2009; Cappé, Moulines & Rydén, 2005).

2.2 Threshold vector error correction model

The threshold vector error correction model is a rather intricate threshold regression model, not only because it is multivariate and a time series model, but because the threshold is only one of its ingredients, cointegration the other. Hence, I begin this section with a brief account on cointegration and error correction. Thereafter, I introduce the concept of threshold cointegration, define the threshold vector error error correction model and comment on some versions of it which are often encountered in empirical applications. Even though asymptotic theory is not a main concern in this section, a remark on the limiting distribution of the threshold estimator is included to establish the link to the first part of the chapter.

2.2.1 Cointegration and error correction

Engle & Granger (1987) begin their seminal paper on cointegration and error correction with the sentences "An individual economic variable, viewed as a time series, can wander extensively and yet some pairs of series may be expected to move so that they do not drift too far apart. Typically, economic theory will propose forces which tend to keep such series together." Prices for the same good observed at spatially different markets are certainly an example of such series. Arbitrage is the mechanism theory suggests to explain their co-movement: If the tomato price goes up in one village, traders from the neighboring villages prefer to sell their tomatoes there. As a result, supply decreases at home and increases in the village with the higher price. This in turn means that the price at home moves up while the grown supply in the place that first experienced the price shock causes it to become smaller again there. Hence, prices are drawn together. To formalize this notion, Engle & Granger (1987) coin the term "cointegration". In the following, I will review some basic concepts used to analyze time series (cf. van der Vaart, 2010) to set the stage for a precise definition of cointegration.

A *time series* $X_t, t \in \mathbb{Z}$, that is, a sequence of random variables with the index interpreted as time, is called *(weakly) stationary* if and only if

(i) there exists $\mu \in \mathbb{R}$ such that $\mathrm{E}(X_t) = \mu$ for all $t \in \mathbb{Z}$,

(ii) there exists a function $\mathbb{Z} \ni \tau \mapsto \gamma(\tau) \in \mathbb{R}$ such that $\mathrm{cov}(X_t, X_{t+\tau}) = \gamma(\tau)$ for all t and $\tau \in \mathbb{Z}$.

An example of a stationary series is a sequence of independent, identically distributed random variables $\varepsilon_t, t \in \mathbb{Z}$, with $\mathrm{var}(\varepsilon_t) = \sigma^2$, in other words, with $\gamma(\tau) = \sigma^2 I(\tau = 0)$. If $\mu = 0$, such a series is called *white noise* (see figure 2.4 (a)). Maybe the most important class of time series models is that of autoregressive moving average models. These can be perceived as linear regression transferred to a time series setting – the present value of the series is expressed as a linear combination of previous values of the series itself as well as the present and lagged values of a white noise series. More formally, a time series $X_t, t \in \mathbb{Z}$, is an *autoregressive moving average series of order* (p, q) (in short, an $ARMA(p, q)$ series) if there exist real numbers $\vartheta_0, \ldots, \vartheta_p$ and $\pi_0, \ldots, \pi_q$ as well as a

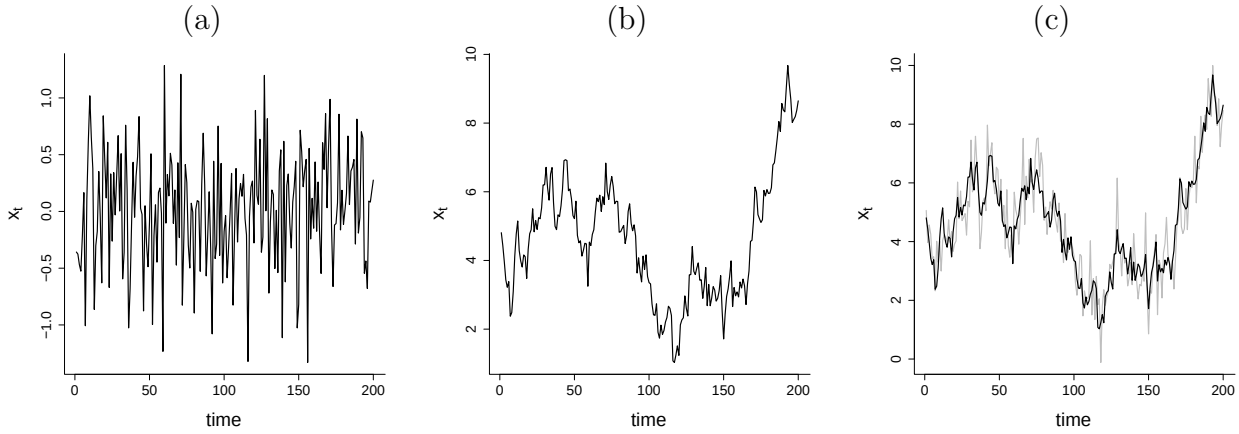

Figure 2.4: Simulated data for different types of times series. (a) Stationary series, (b) integrated series of order 1, and (c) cointegrated series of order $(1, 1)$.

white noise series $\varepsilon_t, t \in \mathbb{Z}$, such that

$$\sum_{i=0}^{p} \vartheta_i X_{t-i} = \sum_{i=0}^{q} \pi_i \varepsilon_{t-i}.$$

van der Vaart (2010) remarks that this equation is meant to hold "pointwise almost surely". It is satisfied for almost every $\omega \in \Omega$, where $(\Omega, \mathcal{A}, P)$ is the probability space on which the random variables X_t and ε_t are defined. $ARMA(p, 0)$ series are also called *autoregressive series* and abbreviated $AR(p)$. For $|\vartheta| < 1$, the $AR(1)$ series $X_t = \vartheta X_{t-1} + \varepsilon_t, t \in \mathbb{Z}$, is another simple example of a stationary time series. An ARMA series X_t is called *invertible*, if ε_t can be represented as

$$\varepsilon_t = \sum_{i=0}^{\infty} \xi_i X_{t-i} \quad \text{for a sequence } \xi_i \text{ such that} \quad \varepsilon_t = \sum_{i=0}^{\infty} |\xi_i| < \infty.$$

Of course, extensively wandering prices typically cannot be described by stationary ARMA models. However, it may be possible to express their increments as such series, that is, that the price series is $I(1)$: A time series $X_t, t \in \mathbb{Z}$, without deterministic component is *integrated of order d*, $X_t \sim I(d)$, if it has a stationary, invertible, ARMA representation after differencing d times. A *random walk* – an $AR(1)$ series with $\vartheta = 1$ – is an example of an $I(1)$ series. Figure 2.4 (b) visualizes a sample random walk. Granger

(1986) nicely highlights the main differences between $I(0)$ and $I(1)$ series: " An $I(0)$ has a mean and there is a tendency for the series to return to the mean, so that it tends to fluctuate around the mean, crossing that value frequently and with extensive excursions. Autocorrelations decline rapidly as lag increases and the process gives low weights to events in the medium to distant past, and thus effectively has a finite memory. An $I(1)$ process without drift will be relatively smooth, will wander widely and will only rarely return to an earlier value. In fact, for a random walk, for a fixed arbitrary value the expected time until the process again passes through this value is infinite. This does not mean that returns do not occur, but that the distribution of the time to return is very long-tailed. Autocorrelations cor $(X_t, X_{t-\tau})$ are all near one in magnitude even for large τ; an innovation to the process affects all later values and the process has indefinitely long memory."

While $I(d)$ captures the "wander extensively" in Engle and Granger's (1987) first sentence, the notion that they "do not drift too far apart" is at the heart of cointegration. The components of a sequence of random k-vectors $\boldsymbol{X}_t, t \in \mathbb{Z}$, are said to be *cointegrated of order* (d, b) if

(i) all components of $\boldsymbol{X}_t$ are $I(d)$,

(ii) there exists a vector $0 \neq \boldsymbol{\gamma} \in \mathbb{R}^k$ so that $\boldsymbol{\gamma}^T \boldsymbol{X}_t \sim I(d-b)$, $b > 0$. The vector $\boldsymbol{\gamma}$ is called the *cointegrating vector*

(Engle & Granger, 1987). A realization of a cointegrated time series is plotted in figure 2.4 (c). In the following, I concentrate on bivariate series and cointegrated series of order $(1, 1)$. In short, I assume tacitly that $k = 2$ and $d = b = 1$. As a simple example of a cointegrated time series, consider $\boldsymbol{X}_t = (X_{t,1}, X_{t,2})^T, t \in \mathbb{Z}$, specified by

$$X_{t,1} = -\gamma W_t + (1-\gamma) z_t \quad \text{and} \quad X_{t,2} = W_t + z_t \tag{2.26}$$

for a random walk $W_t, t \in \mathbb{Z}$, and stationary series $z_t, t \in \mathbb{Z}$. Since the sum of an $I(0)$ and an $I(1)$ series is $I(1)$, whereas the sum of two $I(0)$ series is again $I(0)$ (Engle & Granger, 1991, page 6), both $X_{t,1} \sim I(1)$ and $X_{t,2} \sim I(1)$ while

$$\Delta X_{t,1} = -\gamma \Delta W_t + (1-\gamma) \Delta z_t \sim I(0) \quad \text{and} \quad \Delta X_{t,2} = \Delta W_t + \Delta z_t \sim I(0).$$

Moreover, $X_{t,1} + \gamma X_{t,2} = z_t \sim I(0)$. Consequently, $\boldsymbol{X}_t, t \in \mathbb{Z}$, is cointegrated with cointegrating vector $(1, \gamma)^T$. (To simplify notation, $\boldsymbol{\gamma}$ sometimes denotes the entire cointegrating vector, whereas when its first component is normalized, $\boldsymbol{\gamma}$ stands for the remaining elements of the cointegrating vector.)

Cointegration is closely related to the concept of error correction. Its basic mechanism can be described as follows: If an equilibrium is disturbed in one period, part of this error will be corrected for in the subsequent period. Engle & Granger (1987) define a sequence of random vectors $\boldsymbol{X}_t, t \in \mathbb{Z}$, with values in $\mathbb{R}^2$ to have an *error correction representation* if it can be stated as

$$\Delta \boldsymbol{X}_t = \boldsymbol{\rho}\boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} + \sum_{m=1}^{M} \boldsymbol{\Theta}_m \Delta \boldsymbol{X}_{t-m} + \boldsymbol{\varepsilon}_t, \tag{2.27}$$

where $\boldsymbol{\rho}, \boldsymbol{\gamma} \in \mathbb{R}^2$, $\boldsymbol{\gamma} \neq 0$, $\boldsymbol{\Theta}_m \in \mathbb{R}^{2\times 2}$, $m = 1, \ldots, M$, and $\left\| \sum_{m=1}^{M} \boldsymbol{\Theta}_m \right\|_{\max} < \infty$; $\|\cdot\|_{\max}$ stands for the max norm. $\boldsymbol{\varepsilon}_t, t \in \mathbb{Z}$, denote bivariate stationary disturbances. (The definition of a stationary time series naturally generalizes from the one-dimensional case stated above to bivariate series; replace $\mu \in \mathbb{R}$ by $\boldsymbol{\mu} \in \mathbb{R}^2$ and $\mathbb{Z} \ni \tau \mapsto \operatorname{cov}(X_t, X_{t+\tau}) \in \mathbb{R}$ by $\mathbb{Z} \ni \tau \mapsto \operatorname{cov}(\boldsymbol{X}_t, \boldsymbol{X}_{t+\tau}) \in \mathbb{R}^{2\times 2}$.) $\boldsymbol{\gamma}^T \boldsymbol{X}_t$ is called the *error correction term*. It is associated with the equilibrium error. The strength of the adjustment after disequilibrium is controlled by $\boldsymbol{\rho}$. The Granger Representation Theorem establishes the link between cointegration and error correction: For every cointegrated time series $\boldsymbol{X}_t$, $t \in \mathbb{Z}$, there exists an error correction representation (2.27).

It is easy to see that the cointegrated series $\boldsymbol{X}_t, t \in \mathbb{Z}$, defined in (2.26) with $W_t, t \in \mathbb{Z}$, and $z_t, t \in \mathbb{Z}$, further specified by

$$z_t = \vartheta z_{t-1} + \nu_t, |\vartheta| < 1, \quad \text{and} \quad W_t = W_{t-1} + \eta_t, \tag{2.28}$$

has an error correction representation

$$\Delta \boldsymbol{X}_t = (\vartheta - 1)(1 - \gamma, 1)^T z_{t-1} + (-\gamma, 1)^T \eta_t + (1 - \gamma, 1)^T \nu_t, \tag{2.29}$$

that is, $\Delta \boldsymbol{X}_t = \boldsymbol{\rho} z_{t-1} + \boldsymbol{\varepsilon}_t$ where $\boldsymbol{\rho} = (\vartheta - 1)(1 - \gamma, 1)^T$ and $\boldsymbol{\varepsilon}_t = (-\gamma, 1)^T \eta_t + (1 - \gamma, 1)^T \nu_t$.

2.2.2 Threshold cointegration and the threshold vector error correction model

The reference to the elimination of spatial arbitrage opportunities appears to provide compelling reasoning to model pairs of prices for the same good in spatially separated markets as a cointegrated time series $\boldsymbol{X}_t = (X_{t,1}, X_{t,2})^T, t \in \mathbb{Z}$. However, the specification of the error correction term $z_t, t \in \mathbb{Z}$, as an $AR(1)$ series (as in the example given in (2.26) and (2.28)) is likely to be inadequate. In such a model the speed of adjustment after a deviation from the equilibrium (of equal prices in different locations) does not depend on the magnitude of the latter. It is determined by the parameter ϑ, which is independent of $|z_{t-1}| = |X_{t-1,1} - X_{t-1,2}|$. This is an implausible assumption. It is evident that even if the tomato price in one village exceeds that in the other, there is no incentive for traders to move goods between places unless the difference in prices surpasses transportation costs ψ. Hence, if $|z_{t-1}| < \psi$, prices are expected to move independently, error correction is unlikely to occur, ϑ and $|\rho|$ are anticipated to be close to one and zero, respectively. In the opposite case, for large enough deviations from equilibrium, i.e. $|\boldsymbol{\gamma}^T \boldsymbol{X}_{t-1}| > \psi$, trade is expected to take place and prices to revert to their equilibrium. Threshold cointegration captures this behavior.

Tong (1983) introduces threshold models to time series. They have since attracted plenty of attention (Tong, 2011, for a recent review). I mentioned some asymptotic results for the threshold autoregressive model in section 2.1 of this chapter (Chan, 1993; Chan & Tsay, 1998; Qian, 1998). Balke & Fomby (1997) are the first to examine threshold models in the context of cointegrated time series. They present the idea of threshold cointegration considering a cointegrated time series with an error correction term $z_t, t \in \mathbb{Z}$, characterized by an autoregressive series as in (2.28). However, instead of fixed ϑ, they propose to allow ϑ to vary with z_{t-1}. Recurring to the example of tomatoes in spatially separated markets, while $|z_{t-1}| < \psi$, prices are not expected to return to the state of equilibrium, $z_{t-1} = 0$. Within this band, a random walk describes the series $z_t, t \in \mathbb{Z}$; in other words, if $|z_{t-1}| < \psi$, then $\vartheta = 1$ and $z_t = z_{t-1} + \varepsilon_t$. On the contrary, outside this band deviations from the equilibrium are figured to be corrected for. If $|z_{t-1}| \geq \psi$, then z_t is expected to be zero and $z_t, t \in \mathbb{Z}$, to constitute a stationary series, hence, $|\vartheta| < 1$. More precisely, Balke & Fomby (1997) suggest to specify the error

correction term as follows,

$$z_t = \begin{cases} z_{t-1} + \nu_t & |z_{t-1}| \leq \psi \\ \vartheta z_{t-1} + \nu_t & |z_{t-1}| > \psi, \end{cases} \tag{2.30}$$

for $|\vartheta| < 1$, $\psi > 0$ the threshold parameter and $\nu_t, t \in \mathbb{Z}$, white noise. For the series given by (2.26) and (2.28), incorporating a dependency of ϑ on the error correction term as in (2.30), i.e. $\vartheta(z_{t-1}) = \vartheta I(|z_{t-1}| > \psi) + I(|z_{t-1}| \leq \psi)$, implies a representation

$$\Delta \boldsymbol{X}_t = \begin{cases} \boldsymbol{\rho}_1 z_{t-1} + \boldsymbol{\varepsilon}_t & |z_{t-1}| \leq \psi \\ \boldsymbol{\rho}_2 z_{t-1} + \boldsymbol{\varepsilon}_t & |z_{t-1}| > \psi, \end{cases}$$

where $\boldsymbol{\rho}_1^T = (0, 0)$, $\boldsymbol{\rho}_2^T = (\vartheta - 1)(1 - \gamma, 1)$ and $\boldsymbol{\varepsilon}_t = (-\gamma, 1)^T \eta_t + (1 - \gamma, 1)^T \nu_t$, instead of (2.29).

It is not hard to think of situations in which even more flexibility in the specification of z_t than incorporated through (2.30) is necessary to adequately resemble real dynamics. Transportation costs depending on the direction of trade result in a middle band of no error correction $\psi_1 \leq \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} \leq \psi_2$ for $\psi_1 < 0 < \psi_2$, for example. Thus, Balke & Fomby (1997) put forward a very general model for the error correction term,

$$z_t = \begin{cases} \mu_1 + \sum_{m=1}^{M+1} \vartheta_{1m} z_{t-m} + \nu_{1t} & z_{t-1} \leq \psi_1 \\ \mu_2 + \sum_{m=1}^{M+1} \vartheta_{2m} z_{t-m} + \nu_{2t} & \psi_1 < z_{t-1} \leq \psi_2 \\ \mu_3 + \sum_{m=1}^{M+1} \vartheta_{3m} z_{t-m} + \nu_{3t} & \psi_2 < z_{t-1}, \end{cases} \tag{2.31}$$

with $\mathbb{R} \ni \psi_1 < 0 < \psi_2 \in \mathbb{R}$, $\mu_k, \vartheta_{km} \in \mathbb{R}$ and $\nu_{kt}, t \in \mathbb{Z}$, series of white noise for $k = 1, 2, 3$ and $m = 1, \ldots, M$. A cointegrated series $\boldsymbol{X}_t, t \in \mathbb{Z}$, with such an error correction term $z_t = \boldsymbol{\gamma}' \boldsymbol{X}_t$ can be expressed in terms of a *threshold vector error correction model*

(TVECM),

$$\Delta \boldsymbol{X}_t = \begin{cases} \boldsymbol{\rho}_1 \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} + \boldsymbol{\theta}_1 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{1m} \Delta \boldsymbol{X}_{t-m} + \boldsymbol{\varepsilon}_{1t} & \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} \leq \psi_1 \\ \boldsymbol{\rho}_2 \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} + \boldsymbol{\theta}_2 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{2m} \Delta \boldsymbol{X}_{t-m} + \boldsymbol{\varepsilon}_{2t} & \psi_1 < \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} \leq \psi_2 \\ \boldsymbol{\rho}_3 \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} + \boldsymbol{\theta}_3 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{3m} \Delta \boldsymbol{X}_{t-m} + \boldsymbol{\varepsilon}_{3t} & \psi_2 < \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1}, \end{cases} \tag{2.32}$$

where $\boldsymbol{\rho}_k, \boldsymbol{\theta}_k \in \mathbb{R}^2$, $\boldsymbol{\Theta}_{km} \in \mathbb{R}^{2\times 2}$ and $\boldsymbol{\varepsilon}_{kt}, t \in \mathbb{Z}$, series of white noise for $k = 1, 2, 3$ and $m = 1, \ldots, M$. As exemplified by Balke & Fomby (1997) for a specific bivariate series, a threshold model (2.31) can actually involve regime-dependent differences for the TVECM (2.32) in the adjustment parameters $\boldsymbol{\rho}$, the intercepts $\boldsymbol{\theta}$, the coefficient matrices for the lagged terms $\boldsymbol{\Theta}$ as well as the disturbances $\boldsymbol{\varepsilon}_t$.

Of course, specifying an error correction term (2.31) entails the question of conditions for such $z_t, t \in \mathbb{Z}$, to form a stationary series. To the best of my knowledge, a general answer is still pending. Solutions for special cases include Chan, Petruccelli, Woolford & Tong (1985) among others. For the case $M = 1$, Balke & Fomby (1997) emphasize that "regardless of the behavior of z_t in the interior regime (z_t can display unit root or explosive behavior in the interior) the nature of z_t in the upper and lower regimes determines whether it is stationary. Note that even if the autoregressive coefficient, ϑ, is equal to one everywhere z_t may still be stationary. So long as the drift parameters act to push the series back towards the equilibrium band (i.e. $\mu_1 > 0$ and $\mu_3 < 0$) the series is stationary even though it has a 'unit root'. This suggests that, in general, just examining the autoregressive parameters in the outer regimes is not enough to determine whether the series is stationary."

Two specifications of $z_t, t \in \mathbb{Z}$, which repeatedly occur in the literature are the equilibrium threshold autoregressive model (EQ-TAR) and the band threshold autoregressive model (BAND-TAR). The EQ-TAR is given in (2.30), the BAND-TAR by

$$z_t = \begin{cases} -\psi(1-\vartheta) + \vartheta z_{t-1} + \nu_t & z_{t-1} < -\psi \\ z_{t-1} + \nu_t & |z_{t-1}| \leq \psi \\ \psi(1-\vartheta) + \vartheta z_{t-1} + \nu_t & z_{t-1} > \psi \end{cases} \tag{2.33}$$

When outside the band $[-\psi, \psi]$, an error correction term following an EQ-TAR model reverts to the equilibrium. In contrast, $z_t, t \in \mathbb{Z}$, given by a BAND-TAR model returns to the boundary of this band; the regime-specific mean (if it exists) is $-\psi$ for the lower and ψ for the upper regime. Lo & Zivot (2001) emphasize that the BAND-TAR model (2.33) is most often employed in empirical applications. They call (2.33) a continuous, symmetric threshold, symmetric adjustment BAND-TAR model and point out that the more general version

$$z_t = \begin{cases} \phi_1(1-\vartheta_1) + \vartheta_1 z_{t-1} + \nu_t & z_{t-1} < \psi_1 \\ z_{t-1} + \nu_t & \psi_1 < z_{t-1} \leq \psi_2 \\ \phi_3(1-\vartheta_3) + \vartheta_3 z_{t-1} + \nu_t & z_{t-1} > \psi_2 \end{cases}$$

is implied by a BAND-TVECM model, which they define as a special case of the general TVECM (2.32),

$$\Delta \boldsymbol{X}_t = \begin{cases} \rho_1 \left(\boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} - \lambda_1 \right) + \boldsymbol{\varepsilon}_t & \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} \leq \psi_1 \\ \boldsymbol{\varepsilon}_t & \psi_1 < \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} \leq \psi_2 \\ \rho_3 \left(\boldsymbol{\gamma}^T \boldsymbol{X}_{t-1} - \lambda_3 \right) + \boldsymbol{\varepsilon}_t & \psi_2 < \boldsymbol{\gamma}^T \boldsymbol{X}_{t-1}. \end{cases}$$

They further claim the continuous BAND-TVECM ($\lambda_1 = \psi_1$ and $\lambda_3 = \psi_2$), the symmetric BAND-TVECM ($\psi_1 = -\psi_2$) and the EQ-TVECM ($\lambda_1 = \lambda_3 = 0$) to be TVECMs of special interest. Ihle (2010, table 2.1) lists further special cases of the TVECM (2.32) which have been applied to analyze price transmission and provides the corresponding references. In particular, he identifies the central and non-central AVECM in addition to the models mentioned above.

It is beyond the scope of this thesis to give a detailed account of theoretical results available for the TVECM. Hansen & Seo (2002) is certainly a milestone. Lange & Rahbek (2009) provide an introduction and references; for a brief outline of developments see Gaul et al. (2008, section 2.1). As it is a very complex model, findings are often limited to hold under certain conditions only. Besides, some authors assume a single cointegrating relationship, others allow for more; some concentrate on models with one, others with multiple thresholds; some assume to know the cointegrating vector, others do not. This results in a fragmented literature, which is hard to summarize. Estimation

has only recently been studied systematically. Seo (2011) shows that given certain assumptions (including discontinuity) the least squares estimator $\hat{\psi}$ in a two regime TVECM is consistent,

$$\hat{\psi} = \psi_0 + O_P(n^{-1}).$$

He further looks at the smoothed least squares estimator (Seo & Linton, 2007) defined via the (analog for the TVECM of the) objective function (2.18). He proves consistency for this estimator and shows that – for known cointegrating vector – it converges to a normal distribution with rate $\sqrt{nh_n^{-1}}$ (corollary 2). The asymptotic variance is specified by a rather complicated expression depending (amongst others) on the density of the error correction term at the threshold. I am not aware of any results for the limiting distribution of the least squares or maximum likelihood threshold estimator. However, in view of the results for simpler, but related models, I would expect it to depend on a range of unknown parameters as well as the density of the covariates, and, hence, not to be very useful in practice.

2.3 Summary

To briefly summarize the first part of this chapter, the main criteria to classify threshold regression models are the nature of the transition function (smooth or step function), the regression function (continuous or discontinuous), and the transition variable (random or deterministic). The term "threshold regression model" refers both to any model satisfying (2.2), and, when distinguishing between different kinds of models of this type, to a model with a random transition variable as opposed to change point models, which are characterized by a deterministic transition variable. This ambiguity is the cost for coherence with the tradition in the literature. Assuming a step transition function, the maximum likelihood (least squares) estimator of the threshold parameter, $\hat{\psi}$, is asymptotically normally distributed for continuous models. For discontinuous models, it converges towards a functional of random walks (roughly speaking) in case of a change point model; and a functional of compound Poisson processes in case of a model with random transition variable. Since both of these limiting distributions depend on a host of nuisance parameters including the distribution of the covariates, an alternative framework of diminishing differences between regimes is considered for the limiting dis-

tribution of the threshold estimator in discontinuous models. In this setting, $\hat{\psi}$ converges towards a functional of Brownian motions and the nuisance parameters disappear. The asymptotic distribution of $\hat{\psi}$ for models with multiple thresholds is analogous to that of the estimator in the respective single threshold model. A number of extensions of the specification of a threshold regression model in this thesis have been considered in the literature. These include multivariate models, models featuring regime-dependent heteroscedasticity and models with multi-dimensional transition variables which allow for more flexible mechanisms to switch between regimes.

The second part of the chapter introduces the necessary terminology to formalize the notion of time series that do not drift too far apart from each other as the existence of a stationary linear combination; this is the essence of cointegration. Cointegrated time series have an error correction representation. Such expression reveals the mechanism that keeps the series together, the partial correction of one period's equilibrium error in the following period. Introducing a threshold in the error correction series on some occasions – and price transmission analysis is one of them – allows to more adequately model reality. This results in threshold cointegration and a model that can be represented as a TVECM. The EQ-TVECM and BAND-TVECM are particularly popular in empirical applications.

3 Regularized Bayesian estimation in generalized threshold regression models

Abstract

Estimation of threshold parameters in (generalized) threshold regression models is typically performed by maximizing the corresponding profile likelihood function. Certain Bayesian techniques based on non-informative priors have also been developed and are widely used. This article draws attention to finite-sample settings (not rare in practice) in which these standard estimators perform poorly or even fail. In particular, if estimation of the regression coefficients is associated with high uncertainty, the profile likelihood for the threshold parameters and thus the corresponding estimators can be strongly affected. We suggest an alternative regularized Bayesian estimator that circumvents the deficiencies of standard estimators in small samples. The new estimator can be obtained employing the empirical Bayes paradigm and, hence, requires little additional numerical effort compared with commonly used estimators. Simulations confirm excellent finite sample properties of the suggested estimator, especially in the critical settings. The practical relevance of our approach is illustrated by two real-data examples already analyzed in the literature.

Key words and phrases: empirical Bayes, nuisance parameter, threshold estimation.

3.1 Introduction

Modeling a response variable as a linear combination of some covariates with regression coefficients that vary between (possibly several) regimes is known as threshold regression. The choice of regime is determined by a transition function which depends on a transition variable as well as a threshold parameter. Transition functions can be either smooth (van Dijk, Teräsvirta & Franses, 2002, provide a comprehensive overview) or step functions. In the following, we restrict attention to the latter. In principle, the response variable can follow any distribution from the exponential family. However, such generalized threshold regression models have only recently been formally introduced by Samia & Chan (2011) and most of the literature on threshold regression deals with models with piecewise linear mean. In this article we concentrate on generalized regression models with regimes controlled by a step transition function and refer to such models as generalized threshold regression models. Generalized threshold regression models are employed in a wide range of different fields of application. Hansen (2011) provides an overview of the extensive use of generalized threshold regression models in economic applications including e.g. models of output growth, forecasting, and the term structure of interest rates or stock returns. Samia, Chan & Stenseth (2007) employ a generalized threshold regression model to analyze plague outbreaks and Lee, Seo & Shin (2011) complement these applications with examples in finance, sociology, and biostatistics among others.

Threshold estimation in generalized threshold regression models is typically performed in two stages: the estimation of the regression coefficients is followed by the maximization of the profile likelihood for the threshold parameters using a grid search, as the likelihood function is not differentiable with respect to the threshold parameter. This estimation procedure has two intrinsic problems. First, the profile likelihood is not defined for thresholds that leave fewer observations in one of the regimes than are necessary to estimate the regression coefficients. Hence, in practice it is unavoidable to restrict the domain of the threshold parameters depending on the dimension of the regression coefficients. The literature offers arbitrary constraints including one observation per dimension of the regression coefficient (Samia & Chan, 2011) or 15% of the observations (Andrews, 1993) to give just two examples. This restriction can be problematic in small samples, especially if the true threshold is close to the boundary of its domain. The second problem is due to the direct impact of the uncertainty inherent in the regression

coefficients' estimates on the profile likelihood for the threshold parameters. In an unfavorable setting the profile likelihood becomes jagged with multiple extrema, increasing the uncertainty of the threshold estimator. Large variance of the regression coefficients' estimator is again likely to be found in small samples and for the true threshold at the boundary of its domain, but also if the signal-to-noise ratio is low or the residual variance is misspecified (overdispersion in the generalized regression setting). We are not aware of any work that points out these deficiencies of the common threshold estimator even though the problematic settings frequently occur in empirical applications. Macro-economic data are often only available for a small sample, e.g. if observations correspond to different countries. Spatial arbitrage modeling is another example (Greb, von Cramon-Taubadel, Krivobokova & Munk, 2011).

Bayesian methods are also popular to estimate thresholds. In the literature, a Bayesian threshold estimator is typically based on non-informative priors; we refer to it as the non-informative Bayesian estimator. For the case of a threshold regression model with piecewise linear mean, Yu (2012) shows that, regardless of the choice of priors, Bayesian threshold estimators are asymptotically efficient among all estimators in the locally asymptotically minimax sense. However, in the critical small sample settings described above, the non-informative Bayesian estimator shares all the drawbacks of the profile likelihood estimator and can completely fail in certain cases, as we discuss in section 3.3.2.

In this article, we suggest an alternative threshold estimator, which we call the regularized Bayesian estimator. If regression coefficients were known, none of the problems outlined above would exist. This suggests that stabilizing their estimates might help to prevent them from distorting the threshold estimates. In addition, regularization of regression coefficient estimates allows us to obtain a posterior density which is well-defined on the entire domain of the threshold parameters. This is highlighted in the right plot of figure 3.1, which contrasts three estimation techniques. Simulations confirm that the regularized Bayesian estimator yields good results even in settings in which profile likelihood and non-informative Bayesian estimator are highly susceptible to faults.

In this article, we suggest an alternative threshold estimator, which we call the regularized Bayesian estimator. Contrary to previous work on estimation in threshold regression (Samia & Chan, 2011; Yu, 2012), we focus on the estimator's performance in critical

small sample situations. Simulations confirm that it yields good results even in settings in which profile likelihood and non-informative Bayesian estimator are highly susceptible to faults. To summarize the intuition behind this new estimator: If regression coefficients were known, none of the problems outlined above would exist. This suggests that stabilizing their estimates might help to prevent them from distorting the threshold estimates. In addition, regularization of regression coefficient estimates allows us to obtain a posterior density which is well-defined on the entire domain of the threshold parameters. We achieve regularization by a particular specification of priors. While it proves to be beneficial in the critical small sample situations, the choice of priors does not have an impact asymptotically (as Yu, 2012, shows for a threshold regression model with piecewise linear mean and independent observations). We further derive an explicit (approximate) expression of the posterior density, which allows us to utilize existing functions for mixed models in standard software to easily compute the estimator.

The rest of this article is organized as follows. We specify the generalized threshold regression model in the second section. In the third section, we review existing threshold estimators and point out their deficiencies. The regularized Bayesian estimator is introduced in the fourth section. In the fifth section, we briefly look at inference about the threshold parameter. Simulation results are presented in the sixth section. We use the last section to discuss two empirical applications. The appendix contains some technical details.

3.2 Model

Observations $\left(y_i, \boldsymbol{X}_i^T, q_i\right) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$, are assumed to be realizations of random variables that follow a generalized threshold regression model with threshold parameter $\psi \in \mathbb{R}$, regression coefficients $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$ and scale (or dispersion) parameter $\phi \in \mathbb{R}^+$, that is

$$\mu_i = \mathrm{E}\left(y_i | \boldsymbol{X}_i^T, q_i\right) = h(\eta_i) \tag{3.1}$$

where h is a known one-to-one function, the inverse of the link function $g = h^{-1}$, and

$$\eta_i = I\left(q_i \leq \psi\right) \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + I\left(q_i > \psi\right) \boldsymbol{X}_i^T \boldsymbol{\beta}_2, \tag{3.2}$$

with $I(\cdot)$ as the indicator function. Moreover, conditional on the design vector $\boldsymbol{X}_i^T$ and the transition variable q_i, the response variables y_i are independently drawn from an exponential family distribution with density

$$f(y_i|\psi, \phi, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}, \tag{3.3}$$

characterized by known functions b and c together with the natural parameter $\theta_i = \theta(\mu_i)$. Above and in the following, the same symbol denotes both a random variable and its realization; the context should eliminate ambiguities. To use matrix notation, we define vectors $\boldsymbol{\mu}$, $\boldsymbol{\eta}$, $\boldsymbol{y}$, $\boldsymbol{q}$, $\boldsymbol{I}(\boldsymbol{q} \leq \psi)$ and $\boldsymbol{I}(\boldsymbol{q} > \psi)$ by stacking μ_i, η_i, y_i, q_i, $I(q_i \leq \psi)$ and $I(q_i > \psi)$, respectively, and create an $n \times p$ matrix $\boldsymbol{X}$ with rows $\boldsymbol{X}_i^T$, $i = 1, \ldots, n$. With $\operatorname{diag}\{\boldsymbol{I}(\cdot)\}$ the diagonal matrix with entries $\boldsymbol{I}(\cdot)$ along the diagonal and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, we can write

$$\boldsymbol{\eta} = \operatorname{diag}\{\boldsymbol{I}(\boldsymbol{q} \leq \psi)\} \boldsymbol{X}\boldsymbol{\beta}_1 + \operatorname{diag}\{\boldsymbol{I}(\boldsymbol{q} > \psi)\} \boldsymbol{X}\boldsymbol{\beta}_2 = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 = \boldsymbol{X}_\psi\boldsymbol{\beta}.$$

We consider generalized threshold regression models with one threshold to keep the exposition simple; extension to generalized threshold regression models with more thresholds is straightforward.

Naturally, our model covers $y_i = I\left(q_i \leq \psi\right) \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + I\left(q_i > \psi\right) \boldsymbol{X}_i^T \boldsymbol{\beta}_2 + \varepsilon_i$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $i = 1, \ldots, n$. This is by far the most frequently encountered generalized threshold regression model in the literature. It is broad enough to comprise the popular threshold autoregressive model in which the transition variable q_i is an element of $\boldsymbol{X}_i$ (Tong & Lim, 1980; Tong, 2011, for a review of the development of the model).

Depending on the assumptions on the data generating process, model (3.1) – (3.3) can have different asymptotic behavior. A first differentiation regards the transition variable q_i. Change point models are characterized by deterministic $q_i = i$, while for threshold models q_i is a random variable which follows any continuous distribution. This is reflected in distinct limit likelihood ratio processes and, hence, asymptotic behavior of

the maximum likelihood estimators for ψ in the two models. The limiting likelihood ratio process involves a functional of random walks for change point models and of compound Poisson processes for threshold models. Check Bai (1997b) for more details on the asymptotic properties in the former, and Samia & Chan (2011) for the limiting behavior of the profile log-likelihood and the asymptotic distribution of the profile likelihood threshold estimator in the latter case. If the transition variable coincides with one of the covariates and the regression function is continuous at the threshold, least squares estimates are known to be normally distributed (for threshold models see Chan & Tsay, 1998; Feder, 1975, treats change-point models), which simplifies inference. Clearly, once the data is sampled, the estimation procedure in both change point and threshold models is the same. Referring to a threshold regression model with piecewise linear mean, Hansen (2000) points out that "if the observed values of q_i are distinct, the parameters can be estimated by sorting the data based on q_i, and then applying known methods for change point problems". However, as the focus of this article is on estimation problems that arise in small samples, we do not further differentiate models. In the real-data examples, we concentrate on discontinuous threshold models since they are frequently encountered in applications and have not been studied as extensively as change point models due to their more intricate limiting behavior.

3.3 Commonly used threshold estimators

3.3.1 The profile likelihood estimator

As noted in the introduction, the prevalent threshold estimator in the literature is the profile likelihood estimator, see e.g. Samia & Chan (2011) or Hansen (2000). Splitting all model parameters into a parameter of interest and nuisance parameters, the profile likelihood function $\mathcal{L}_p$ is constructed from the likelihood function $\mathcal{L}$ by replacing nuisance parameters with their maximum likelihood estimates at given values of the parameter of interest. In generalized threshold regression models, our parameter of interest is the threshold parameter ψ and its domain is restricted to a random set $\Psi = \left\{\psi \in \mathbb{R} | q_{(1)} \leq \psi \leq q_{(n)}\right\} \subseteq \mathbb{R}$, where $q_{(i)}$ denotes the ith order statistics. The nuisance parameters are $\boldsymbol{\beta}^T \in \mathbb{R}^{2p}$ and $\phi \in \mathbb{R}$. Hence, we work with the conditional profile

likelihood function given $\boldsymbol{X}$ and $\boldsymbol{q}$,

$$\mathcal{L}_p(\psi) = \prod_{i=1}^{n} f(y_i|\psi, \hat{\phi}_\psi, \hat{\boldsymbol{\beta}}_\psi) = \exp\left[\sum_{i=1}^{n}\left\{\frac{y_i\hat{\theta}_i - b(\hat{\theta}_i)}{\hat{\phi}_\psi} + c(y_i, \hat{\phi}_\psi)\right\}\right],$$

where $\hat{\theta}_i = \theta \circ h(\hat{\eta}_i) = \theta \circ h\left\{I(q_i \leq \psi)\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}_{1_\psi} + I(q_i > \psi)\boldsymbol{X}_i^T\hat{\boldsymbol{\beta}}_{2_\psi}\right\}$ and $\hat{\boldsymbol{\beta}}_\psi$ and $\hat{\phi}_\psi$ are maximum likelihood estimators at a fixed ψ. In the following, we assume a canonical link, that is, $\theta_i = \eta_i$. All developments still hold approximately if this assumption is not given. We denote the profile log-likelihood with $\ell_p(\psi) = \log \mathcal{L}_p(\psi)$.

To measure the proximity of a threshold ψ to the boundary of its domain Ψ, we introduce $\mathrm{d}(\psi) = \min(j, n-j)/p$ with j such that $q_{(j)} \leq \psi < q_{(j+1)}$. $\mathrm{d}(\psi)$ quantifies the distance between ψ and Ψ's boundary in terms of the number of observations between them relative to the dimension of the regression coefficients, $p = \dim(\boldsymbol{\beta}_k)$, $k = 1, 2$. When $\mathrm{d}(\psi) = 1$, ψ assigns at least p observations to each of the regimes. The allocation of 5% of the observations into one of the regimes can be expressed as $\mathrm{d}(\psi) = 0.05\, n/p$. Clearly, $\mathcal{L}_p(\psi)$ is not defined for $\mathrm{d}(\psi) < 1$, since in this case ψ does not leave enough observations for the estimation of $\boldsymbol{\beta}_k$ in one of the regimes. Hence, in practice it is inevitable to restrict Ψ to $\Psi^*(c) = \{\Psi|\ \mathrm{d}(\psi) > c\}$ for some $c \geq 1$. In the literature different heuristic suggestions for the choice of c have been proposed. For example, Hansen & Seo (2002) propose $c = 0.05\, n/p$, we find $c = 0.15\, n/p$ in Andrews (1993) and Samia & Chan (2011) even use $c = 0.25\, n/p$ for their application.

The profile likelihood threshold estimator is then given by

$$\hat{\psi}_{pL} = \underset{\psi \in \Psi^*(c)}{\operatorname{argmax}}\ \mathcal{L}_p(\psi).$$

This definition based on the restricted domain $\Psi^*(c)$ immediately suggests that in settings in which $\mathrm{d}(\psi_0) < c$ for a true threshold ψ_0, $\hat{\psi}_{pL}$ is inconsistent. The left panel of figure 3.1 illustrates this showing the profile log-likelihood for a sample run of a generalized threshold regression model corresponding to the simulation setting 1C detailed in section 3.6. If $\Psi^*(1) = [0.3, 0.7]$ would be restricted any further, e.g. to be $[0.31, 0.69]$, then the true threshold $\psi_0 = 0.3$ would be excluded from the threshold domain and $\hat{\psi}_{pL}$ would move to the next extremum. For small n, large p and ψ_0 close to the boundary of Ψ, $\mathrm{d}(\psi_0) < c$ is likely to be the case. Altogether, subjective restriction of the threshold

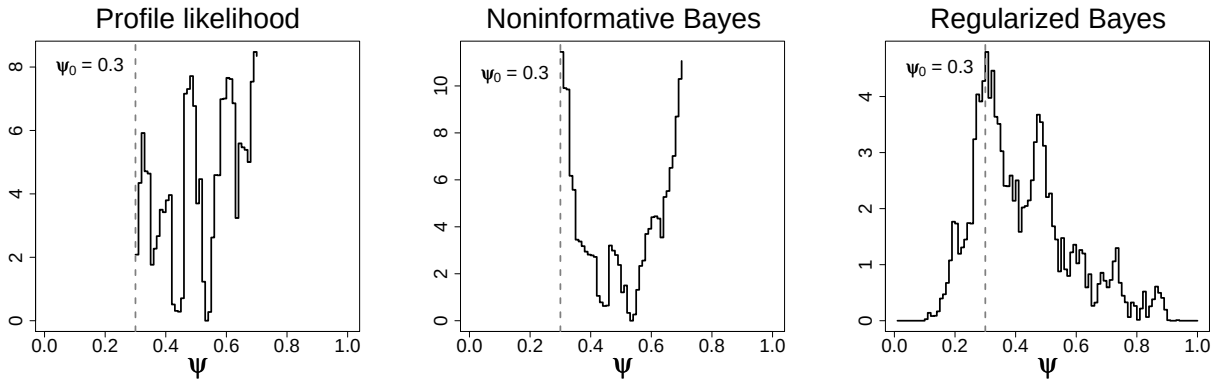

Figure 3.1: For a sample run corresponding to setting 1C of section 3.6, $\ell_p(\psi)$ is shown on the left, $\log p_{nB}(\psi|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ in the middle and $\log p_{rB}(\psi|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ on the right.

domain is an undesirable property of threshold estimation based on the profile likelihood.

The same plot in figure 3.1 also exemplifies that in certain small-sample settings the profile (log-)likelihood can be jagged and have multiple extrema, leading to a very variable threshold estimator. To shed light on this behavior of $\ell_p(\psi)$, we contrast it with its analogue for known $\boldsymbol{\beta}$. Accordingly, we compare

$$-\ell_p(\psi) \propto -\frac{2}{\hat{\phi}_\psi} \sum_{i=1}^{n} \{y_i \hat{\theta}_i - b(\hat{\theta}_i)\} \approx (\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi \hat{\boldsymbol{\beta}}_\psi)^T \boldsymbol{W} (\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi \hat{\boldsymbol{\beta}}_\psi)$$

with its analogue for known $\boldsymbol{\beta}$. Here, $\hat{\boldsymbol{z}} = \boldsymbol{X}_\psi \hat{\boldsymbol{\beta}}_\psi + \boldsymbol{G}(\boldsymbol{y} - \hat{\boldsymbol{\mu}})$ is the working variable, $\boldsymbol{G} = \operatorname{diag}\{g'(\mu_i)\}$ and $\boldsymbol{W}^{-1} = \operatorname{diag}\{\phi b''(\theta_i) g'(\mu_i)^2\}$. The estimated $\boldsymbol{W}$ for fixed ψ is assumed to vary little or not at all as a function of the mean so we use $\boldsymbol{W}$ evaluated at the true $\boldsymbol{\beta}$ directly. This is a typical assumption in the literature on generalized linear models. The same applies to $\boldsymbol{G}$. We focus on the case of $\psi \leq \psi_0$, but the same arguments hold for $\psi > \psi_0$. Denoting $\boldsymbol{z} = \boldsymbol{X}_\psi \boldsymbol{\beta} + \boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu})$, $\boldsymbol{X}_{[\psi,\psi_0]} = \operatorname{diag}\{\boldsymbol{I}(\psi < \boldsymbol{q} \leq \psi_0)\}\boldsymbol{X}$ and $\boldsymbol{H} = \boldsymbol{W}^{1/2} \boldsymbol{X}_2 (\boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_{[\psi,\psi_0]}$, both of which disappear for $\psi = \psi_0$, we

find

$$\left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right)^T \boldsymbol{W}\left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right) = (\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T \boldsymbol{W}(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta}) + \mathcal{O}_p(2p) \tag{3.4}$$

$$+ \left\{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^T\boldsymbol{H}^T - 2(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T\boldsymbol{W}^{1/2}\right\}\boldsymbol{H}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \tag{3.5}$$

$$+ \left\{2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^T\boldsymbol{H}^T - 2(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T\boldsymbol{W}^{1/2}\right\}^T \mathcal{O}_p\left(\boldsymbol{W}^{1/2}\sqrt{\operatorname{var}\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right) \tag{3.6}$$

where $\mathcal{O}_p(2p)$ is bounded in probability for fixed p and $n \to \infty$; $\operatorname{var}(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi)$ is a row vector containing the diagonal elements of $\boldsymbol{X}_\psi(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi)^{-1}\boldsymbol{X}_\psi^T$. Taking a closer look at $(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi)^T\boldsymbol{W}(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi)$ as a function of ψ, we note that replacing the true regression coefficients $\boldsymbol{\beta}$ by their maximum likelihood estimators influences $(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})$ in several ways. The $\mathcal{O}_p$-term in (3.4) is independent of ψ and simply shifts the profile likelihood by a constant. The deterministic term in (3.5) equals zero for $\psi = \psi_0$, but starts growing as $|\psi - \psi_0|$ increases. That is, even if there is no uncertainty due to estimation of $\boldsymbol{\beta}$, that is, $\operatorname{var}(\hat{\boldsymbol{\beta}}_\psi) = 0$, the true least squares is inflated for ψ away from ψ_0, making the extremum less pronounced. The most important term is the last one in (3.6). This random term depends on $\operatorname{var}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right)$ and can have a strong deforming effect on the true least squares even for ψ close to ψ_0. Large variance of $\hat{\boldsymbol{\beta}}_\psi$ is associated with settings characterized by small n relative to p, but can also be due to low signal-to-noise ratio, model misspecifications (e.g. overdispersion) or if the threshold is close to the boundary of its domain. This is exposed in the left as compared with the middle plot of figure 3.2; the log-likelihoods depicted in these plots belong to models which only differ in one aspect: in the plot on the left-hand side, the residual standard deviation is 0.75, while in the middle plot it is 1.5. Clearly, the log-likelihood in the middle plot is highly distorted over the whole range of Ψ, triggering multiple extrema and a highly variable estimator for ψ. Moving the true threshold closer to the boundary, as shown in the right plot of figure 3.2, leads to an even stronger deformation of the log-likelihood. In summary, in small samples and particular settings exemplified above, the profile likelihood threshold estimator can perform poorly, being very sensitive to inappropriate estimates of the nuisance parameters and relying on a subjective restriction of its domain.

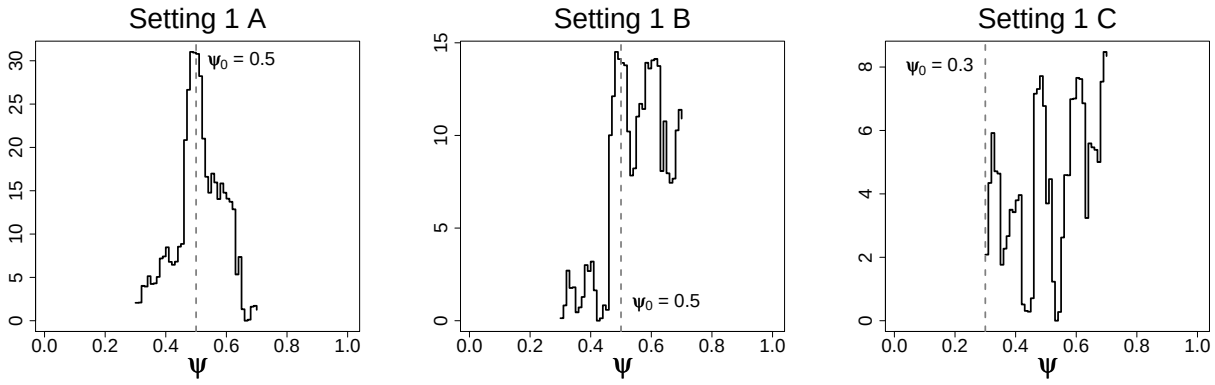

Figure 3.2: Sample (log) profile likelihood functions $\ell_p(\psi)$ for different simulation settings.

3.3.2 The Bayesian estimator

For threshold estimation in regression models with piecewise linear mean, there is a long tradition of using Bayesian techniques in applied work beginning with Bacon & Watts (1971) and including Geweke & Terui (1993) among many others. This popularity can be at least partially attributed to practical advantages, since the Bayesian approach offers a natural framework for inference and accounts for the variability of the nuisance parameters. The theoretical properties of Bayesian threshold estimators in certain generalized threshold regression models have been investigated by Yu (2012). He shows that for independently and identically distributed observations Bayesian threshold estimators are asymptotically efficient among all estimators in the locally asymptotically minimax sense and strictly more efficient than the maximum likelihood estimator. In a related paper, Chan & Kutoyants (2010a) examine asymptotic properties of Bayesian estimators in threshold autoregression models. They note that the limit variance of the Bayesian estimator is smaller than that of the maximum likelihood estimator.

Without any prior knowledge of possible parameter values, it is natural to assume a uniform prior for the threshold parameter and non-informative priors for the regression coefficients; these choices are (almost) omnipresent in the Bayesian literature on generalized threshold regression models with piecewise linear mean. While the priors do not have an impact asymptotically, it turns out that they do affect the performance of the Bayesian threshold estimator in finite samples. We show that non-informative priors can

distort estimates, especially in small samples. It is straightforward to obtain an approximation of a generalized threshold regression model's posterior density $p_{nB}(\psi|\phi, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ associated with non-informative (improper) priors $p(\boldsymbol{\beta}) \propto 1$ and $p(\psi|\boldsymbol{q}) \propto I(\psi \in \Psi)$ based on a Laplace approximation (Shun & McCullagh, 1995; Severini, 2000) of the integral for fixed $p \ll n$

$$\int\limits_{\mathbb{R}^{2p}} p(y|\psi, \phi, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{q}) d\boldsymbol{\beta} = \mathcal{L}_p(\psi)(2\pi)^p \left| -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \left(\psi, \phi, \hat{\boldsymbol{\beta}}_\psi\right) \right|^{-1/2} + \mathcal{O}\left(n^{-1}\right),$$

with $\ell(\psi, \phi, \boldsymbol{\beta}) = \log \mathcal{L}(\psi, \phi, \boldsymbol{\beta})$. As $\left| -\partial^2 \ell / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T \left(\psi, \phi, \hat{\boldsymbol{\beta}}_\psi\right) \right| = \left| \boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi \right|$, we get

$$p_{nB}(\psi|\phi, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q}) = \mathcal{L}_p(\psi)(2\pi)^p \left| \boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi \right|^{-1/2} I(\psi \in \Psi)/p(\boldsymbol{y}) + \mathcal{O}\left(n^{-1}\right).$$

With this, the prevalent Bayesian threshold estimator in the literature is the posterior mean $\hat{\psi}_{nB} = \int\limits_{\Psi^*} \psi p_{nB}(\psi|\phi, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q}) d\psi$. Comparing $p_{nB}(\psi|\phi, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ with $\mathcal{L}_p(\psi)$, we note that they differ by a term proportional to $\left| \boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi \right|^{-1/2}$. In the case of Gaussian observations, $\boldsymbol{W} = \boldsymbol{I}_n/\sigma^2$. Since $\left| \boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi \right| = \left| \boldsymbol{X}_1^T \boldsymbol{W} \boldsymbol{X}_1 \right| \cdot \left| \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2 \right| \to 0$ for $\mathrm{d}(\psi) \to 0$, $p_{nB}(\psi|\phi, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ becomes very large for ψ close to the boundary of Ψ. Moreover, as the profile likelihood function requires $\mathrm{d}(\psi) \geq 1$ to be well-defined, so does the calculation of the posterior density. Again, the only solution in the literature is to restrict the parameter space Ψ (which in our Bayesian framework is equivalent to working with a uniform prior $\psi \sim U[\Psi^*]$ instead of $\psi \sim U[\Psi]$). In this case, however, $p_{nB}(\psi|\phi, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ becomes largest exactly for values of ψ which are arbitrarily included or excluded from Ψ^* by varying c. Consequently, expanding or reducing Ψ^* critically affects the Bayesian threshold estimate, whether it is calculated as the posterior mode, mean or median. The middle plot in figure 3.1 illustrates this problem.

3.4 The regularized Bayesian estimator

When rethinking the threshold estimator, there are good arguments for continuing to pursue Bayesian options. In general, Bayesian estimators naturally incorporate the variability of nuisance parameters and there are reasons to expect them to be (at least

asymptotically) the most efficient estimators, as discussed in section 3.3.2. Our idea now is to exploit understanding of when reliable estimation becomes particularly difficult in order to regularize the posterior density. We observe that both profile likelihood and posterior density become increasingly distorted as ψ approaches the boundary of Ψ (or the farther it is away from the true threshold ψ_0). Using the notation introduced in section 3.2, we define

$$\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 = (\boldsymbol{X}_1 + \boldsymbol{X}_2)\boldsymbol{\beta}_1 + \boldsymbol{X}_2(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1) = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\delta}. \tag{3.7}$$

While maintaining a non-informative constant prior for $\boldsymbol{\beta}_1$, we pick a normal prior with zero mean for $\boldsymbol{\delta}$, $\boldsymbol{\delta} \sim \mathcal{N}(0, \sigma_\delta^2 \boldsymbol{I}_p)$. When σ_δ^2 tends towards infinity, this prior becomes non-informative. However, for small values σ_δ^2, we introduce prior knowledge suggesting that $\boldsymbol{\delta}$ takes values close to zero. The most important characteristic of this new choice of priors is that it regularizes the posterior density for ψ close to the boundary of Ψ. Putting priors on σ_δ^2 (e.g. an inverse Gamma distribution) and ψ specifies a full Bayesian model and allows for estimation with Markov chain Monte Carlo techniques.

Alternatively, we suggest to use a Laplace approximation to get the approximate posterior $p(\psi|\phi, \sigma_\delta^2, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$. This accelerates estimation and enables us to illustrate the regularizing effect. To evaluate the posterior density

$$p(\psi|\phi, \sigma_\delta^2, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q}) = \frac{p(\psi|\boldsymbol{q})}{p(\boldsymbol{y}|\phi, \sigma_\delta^2, \boldsymbol{X}, \boldsymbol{q})} \int\limits_{\mathbb{R}^p} \int\limits_{\mathbb{R}^p} p(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{\delta}, \psi, \phi, \sigma_\delta^2, \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{\delta}|\sigma_\delta^2) d\boldsymbol{\delta} d\boldsymbol{\beta}_1,$$

we use a Laplace approximation and follow a line of reasoning closely resembling Breslow & Clayton (1993) to obtain

$$\begin{aligned} &\int\limits_{\mathbb{R}^p} \int\limits_{\mathbb{R}^p} p(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{\delta}, \psi, \phi, \sigma_\delta^2, \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{\delta}|\sigma_\delta^2) d\boldsymbol{\delta} d\boldsymbol{\beta}_1 \\ &\qquad = (2\pi)^{p/2} \exp\left\{ -\frac{1}{2}(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1)^T \boldsymbol{V}^{-1} (\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1) + \sum_{i=1}^n c(y_i, \phi) \right\} \\ &\qquad\qquad \cdot \left| \sigma_\delta^2 \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2 + \boldsymbol{I}_p \right|^{-1/2} \left| \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X} \right|^{-1/2} + \mathcal{O}\left(n^{-1}\right), \end{aligned} \tag{3.8}$$

with the working variable $\tilde{\boldsymbol{z}}$ defined as $\tilde{\boldsymbol{z}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\delta}} + \boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu})$, $\boldsymbol{G} = \operatorname{diag}\{g'(\mu_i)\}$, and $\boldsymbol{V} = \boldsymbol{W}^{-1} + \sigma_\delta^2 \boldsymbol{X}_2 \boldsymbol{X}_2^T$ for $\boldsymbol{W}^{-1} = \operatorname{diag}\{\phi b''(\theta_i) g'(\mu_i)^2\}$. Here, $\boldsymbol{\mu}$, $\boldsymbol{G}$, $\boldsymbol{W}$ and $\boldsymbol{V}$ are

evaluated at the (approximate) posterior mode $(\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\delta}}) = \arg\max_{(\boldsymbol{\beta}_1, \boldsymbol{\delta}) \in \mathbb{R}^{2p}} p(\boldsymbol{\beta}_1, \boldsymbol{\delta} | \psi, \phi, \sigma_\delta^2, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$, that is, $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \tilde{\boldsymbol{z}}$ and $\hat{\boldsymbol{\delta}} = \sigma_\delta^2 \boldsymbol{X}_2^T \boldsymbol{V}^{-1} (\tilde{\boldsymbol{z}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_1)$. Details on the derivation of (3.8) are provided in the appendix. In contrast to the posterior based on non-informative priors, the term $|\boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi|$ disappears, and with it the deteriorations near the boundary of Ψ observed for $p_{nB}(\psi | \phi, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$. Moreover, $p(\psi | \phi, \sigma_\delta^2, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ is well-defined for all $\psi \in \Psi$, independent of $\mathrm{d}(\psi)$. It is easy to see that $\hat{\boldsymbol{\delta}} \to 0$ and $\hat{\boldsymbol{\beta}}_1 \to (\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \tilde{\boldsymbol{z}}$ at the boundary of Ψ, for $\boldsymbol{X}_2 = 0$ or $\boldsymbol{X}_2 = \boldsymbol{X}$. We do not encounter the ill-posed problem of estimating p nuisance parameters from $m < p$ observations, or calculating $\hat{\boldsymbol{\beta}}_\psi$ when $\mathrm{d}(\psi) < 1$, as in profile likelihood or non-informative Bayesian estimation. Consequently, there is no need to subjectively restrict the parameter space. Considering

$$\hat{\boldsymbol{\delta}} = \sigma_\delta^2 \boldsymbol{X}_2^T \boldsymbol{V}^{-1} (\tilde{\boldsymbol{z}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_1) = \arg\min_{\boldsymbol{\delta} \in \mathbb{R}^p} (\tilde{\boldsymbol{z}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2 \boldsymbol{\delta})^T \boldsymbol{W} (\tilde{\boldsymbol{z}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2 \boldsymbol{\delta}) + \frac{1}{\sigma_\delta^2} \boldsymbol{\delta}^T \boldsymbol{\delta}, \tag{3.9}$$

it becomes evident that the proposed prior leads to the strategy of turning an ill-posed into a well-posed problem tracing back to Tikhonov, Arsenin & John (1977). For small values of the regularization parameter $1/\sigma_\delta^2$, the first term of the functional to be minimized in (3.9) will drive the resulting $\hat{\boldsymbol{\delta}}$, for large values it is the latter. For the nuisance parameter estimates $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\delta}}$, basic matrix algebra reveals that $\hat{\boldsymbol{\beta}}_1 \to (\boldsymbol{X}_1^T \boldsymbol{W} \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{W} \tilde{\boldsymbol{z}}$ and $\hat{\boldsymbol{\beta}}_2 \to (\boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^T \boldsymbol{W} \tilde{\boldsymbol{z}}$ for $\sigma_\delta^2 \to \infty$, while for $\sigma_\delta^2 \to 0$, both $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ converge to $(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{W} \tilde{\boldsymbol{z}}$.

Clearly, the choice of the regularization parameter σ_δ^2 is essential to any estimate based on $p(\psi | \phi, \sigma_\delta^2, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$. It can naturally be estimated in the full Bayesian framework. However, pursuing our approximate approach further we rather make use of the empirical Bayes paradigm. In general, the empirical Bayes approach to modeling observations $\boldsymbol{y}$ differs from the usual Bayesian setup in that the hyperparameters for the highest level in the model's hierarchy are replaced by their maximum likelihood estimates. In our case, we obtain $\hat{\sigma}_\delta^2$ for fixed $\boldsymbol{X}$, $\boldsymbol{q}$ and ψ by maximizing

$$p(\boldsymbol{y} | \psi, \phi, \sigma_\delta^2, \boldsymbol{X}, \boldsymbol{q}) = \int_{\mathbb{R}^p} \int_{\mathbb{R}^p} p(\boldsymbol{y} | \boldsymbol{\beta}_1, \boldsymbol{\delta}, \psi, \phi, \sigma_\delta^2, \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{\delta} | \sigma_\delta^2) d\boldsymbol{\delta} d\boldsymbol{\beta}_1,$$

so as to base threshold estimation on

$$p_{rB}(\psi|\boldsymbol{y},\boldsymbol{X},\boldsymbol{q}) = p(\psi|\boldsymbol{y},\boldsymbol{X},\boldsymbol{q},\hat{\phi}_\psi,\hat{\sigma}^2_\delta) \propto \left|\hat{\sigma}^2_\delta \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2 + \boldsymbol{I}_p\right|^{-1/2} \left|\boldsymbol{X}^T \hat{\boldsymbol{V}}^{-1} \boldsymbol{X}\right|^{-1/2} \cdot \exp\left\{-\frac{1}{2}(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1)^T \hat{\boldsymbol{V}}^{-1}(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1) + \sum_{i=1}^{n} c\left(y_i, \hat{\phi}_\psi\right)\right\} I(\psi \in \Psi),$$

with $\hat{\boldsymbol{V}}$ evaluated at $\hat{\sigma}^2_\delta$. The right plot in figure 3.1 shows log of this posterior density for a sample run corresponding to the simulation setting 1 C of section 3.6. It is clearly well-defined over the whole domain of the threshold and its values are regularized at the boundary regions, making the extremum more pronounced.

Once the posterior density is obtained, one can calculate $\hat{\psi}_{rB}$. We observed that in the critical small-sample settings the posterior density is often characterized by multiple modes. Thus, an estimate based on maximization (the posterior mode) is likely to suffer from this. The posterior mean presents a more robust alternative. However, when the true threshold is located close to the boundary of Ψ, the posterior distribution is skewed towards this boundary. As a result, the posterior mean tends to be drawn towards the middle of Ψ (Doodson, 1917; Kendall, 1943, page 35). Hence, we opt for the posterior median as a compromise between the latter two. Accordingly, we suggest to calculate a regularized Bayesian threshold estimator $\hat{\psi}_{rB}$ as

$$\int_{q_{(1)}}^{\hat{\psi}_{rB}} p_{rB}(\psi|\boldsymbol{y},\boldsymbol{X},\boldsymbol{q},\phi)d\psi = 0.5$$

assuming a prior $p(\psi|\boldsymbol{q}) \propto I(\psi \in \Psi)$ for ψ.

By definition, the restricted (or residual) likelihood function (Harville, 1977) of a generalized linear mixed model is the approximate posterior (3.8). Hence, the function `glmmPQL` in the R-package *MASS* readily provides us with the desired estimate $\hat{\sigma}^2_\delta$. Moreover, the function simultaneously produces an estimate $\hat{\phi}_\psi$. For the Gaussian case, we can employ the function `lme` directly (with its parameter `method` left at the default value `REML`). It is part of the R-package *nlme*. This possibility to take advantage of existing functions implemented for mixed models greatly facilitates computation of our proposed estimator, which can be performed in seconds.

3.5 Inference about the threshold parameter

In our Bayesian framework it is natural to form confidence regions for ψ as credible sets; an equi-tailed credible set C of level $1 - 2\alpha$ is defined as

$$C = \int_{q_p(\alpha)}^{q_p(1-\alpha)} p(\psi|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q}, \phi)d\psi, \qquad q_p(\alpha) = \inf_{x \in \Psi} \left\{ x \Big| \int_{\psi \leq x} p(\psi|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q}, \phi)d\psi \geq \alpha \right\}.$$

These credible sets are valid for change-point and threshold models, both continuous and discontinuous. In the frequentist framework it is straightforward to obtain confidence intervals for continuous models. For discontinuous models the asymptotic distribution does not readily provide a feasible way to construct confidence intervals as it depends on (a possibly large number of) nuisance parameters. As a strategy to circumvent this problem, it has been suggested to base asymptotic developments on the assumption of a diminishing difference in coefficients between regimes, that is, to work with $\boldsymbol{\delta} = \boldsymbol{\delta}(n)$ and $\boldsymbol{\delta}(n) \to 0$ as $n \to \infty$ (Hansen, 2000). However, this approach has only been applied in the context of models with piecewise linear mean so far. To the best of our knowledge there does not exist previous work on confidence sets for the threshold parameter in discontinuous generalized threshold regression models with a non-identity link.

To test for a threshold effect, a natural approach is to take advantage of the link to generalized linear mixed models. Understanding a threshold effect as the presence of a random effect $\boldsymbol{\delta}$ allows us to draw on existing methods for mixed models, more explicitly, on tests for a zero random effect variance $\sigma_\delta^2 = 0$. For the Gaussian case of a linear mixed model, this theory has been developed by Crainiceanu & Ruppert (2004) and Scheipl, Greven & Küchenhoff (2008), who also implemented the approach in the R package `RLRsim`. An extension to generalized linear mixed models might possibly provide the basis for a unified test in generalized threshold regression models. Yet, it is beyond the scope of this paper to pursue this thought further.

3.6 Simulations

To assess the performance of the suggested estimator $\hat{\psi}_{rB}$ we performed a simulation study. We report results for eight different settings, covering both situations in which common estimators produce reliable results and others in which they are prone to be distorted. The difference between setting 1 and setting 2 is in the conditional distribution of y_i: in the first case, $y_i|\boldsymbol{X}_i^T, q_i$ is normally distributed, in the second case it follows a Poisson distribution. The design matrix $\boldsymbol{X}$ is random, each entry $x_{ij} \sim U[0,1]$ for setting 1, $x_{ij} \sim U[0,0.01]$ for setting 2. The transition variable follows a uniform distribution $q_i \sim U[0,1]$. As this implies $\text{pr}\{\text{d}(\psi_0) < 1\} \approx 0.46$ for setting C, we base our simulations on a fixed sample of transition variables $q_i = i/n$, $i = 1, \ldots, n$. This way, we ensure that $\text{d}(\psi_0) = 1$, hence, that $\mathcal{L}_p(\psi_0)$ is always well-defined. While settings A and B differ from setting C in the threshold ($\psi_0 = 0.5$ for A and B; $\psi_0 = 0.3$ for C), setting A is distinct from settings B and C in the signal-to-noise-ratio, which we control by the choice of $\boldsymbol{\delta} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ relative to the variance of the observations. For setting 1 A – C, the difference $\boldsymbol{\delta} \sim U[-0.5, 0.5]$ and random variables are simulated with variances $\text{var}(y_i) = 0.75^2$ (setting A) and $\text{var}(y_i) = 1.5^2$ (settings B and C). For

	Normal response (1)			
	A	B	C	D
ψ_0	0.5	0.5	0.3	0.3
$\boldsymbol{\delta}$	$U[-0.5, 0.5]$	$U[-0.5, 0.5]$	$U[-0.5, 0.5]$	$U[-0.25, 0.25]$
$\text{var}(y_i)$	0.75^2	1.5^2	1.5^2	0.25^2
x_{ij}	$U[0,1]$	$U[0,1]$	$U[0,1]$	$U[0,1]$
p	30	30	30	10
	Poisson response (2)			
	A	B	C	D
ψ_0	0.5	0.5	0.3	0.3
$\boldsymbol{\delta}$	$U[10, 20]$	$U[0, 10]$	$U[0, 10]$	$U[10, 20]$
x_{ij}	$U[0, 0.01]$	$U[0, 0.01]$	$U[0, 0.01]$	$U[0, 0.01]$
p	30	30	30	10

Table 3.1: Differences between simulation settings.

setting 2 A the difference $\boldsymbol{\delta} \sim U[10, 20]$, whereas $\boldsymbol{\delta} \sim U[0, 10]$ for settings 2 B and C. Setting D features less nuisance parameters than A – C; $p = \dim(\boldsymbol{\beta}_1) = \dim(\boldsymbol{\beta}_2) = 10$ for D, $p = 30$ for A – C. The sample size is $n = 100$. Table 3.1 sums up differences between settings. Regression coefficients $\boldsymbol{\beta}_1$ are drawn from a Poisson distribution with mean 10. To be unambiguous, parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_1$ are fixed; we randomly generate them once at the beginning of the simulation according to the distributions specified. Our Monte Carlo sample contains 1000 replications.

All three estimators $\hat{\psi}_{pL}$, $\hat{\psi}_{nB}$ and $\hat{\psi}_{rB}$ perform well given a high signal-to-noise-ratio and ψ_0 in the middle of Ψ (setting A). Lowering the signal-to-noise-ratio (setting B) alters the results: we observe nearly unbiased estimates $\hat{\psi}_{pL}$, $\hat{\psi}_{nB}$ and $\hat{\psi}_{rB}$, but due to its very small variance the latter stands out by its small mean square error. When we

		bias			mean square error		
Setting	ψ_0	$\hat{\psi}_{pL}$	$\hat{\psi}_{nB}$	$\hat{\psi}_{rB}$	$\hat{\psi}_{pL}$	$\hat{\psi}_{nB}$	$\hat{\psi}_{rB}$
1 A	0.5	-0.002	0.001	0.000	0.002	0.009	0.000
		(0.045)	(0.095)	(0.000)			
1 B	0.5	-0.003	-0.001	-0.001	0.010	0.023	0.006
		(0.100)	(0.152)	(0.077)			
1 C	0.3	0.110	0.087	0.031	0.024	0.024	0.008
		(0.110)	(0.126)	(0.084)			
1 D	0.3	0.064	0.080	0.059	0.036	0.066	0.017
		(0.032)	(0.060)	(0.014)			
2 A	0.5	0.001	0.026	0.000	0.000	0.001	0.000
		(0.000)	(0.000)	(0.000)			
2 B	0.5	0.004	−0.111	−0.004	0.003	0.029	0.001
		(0.055)	(0.126)	(0.032)			
2 C	0.3	0.054	0.049	−0.002	0.007	0.010	0.001
		(0.071)	(0.089)	(0.032)			
2 D	0.3	0.025	-0.045	0.015	0.013	0.032	0.003
		(0.013)	(0.030)	(0.003)			

Table 3.2: Simulation results. Standard errors are reported in parentheses below the bias.

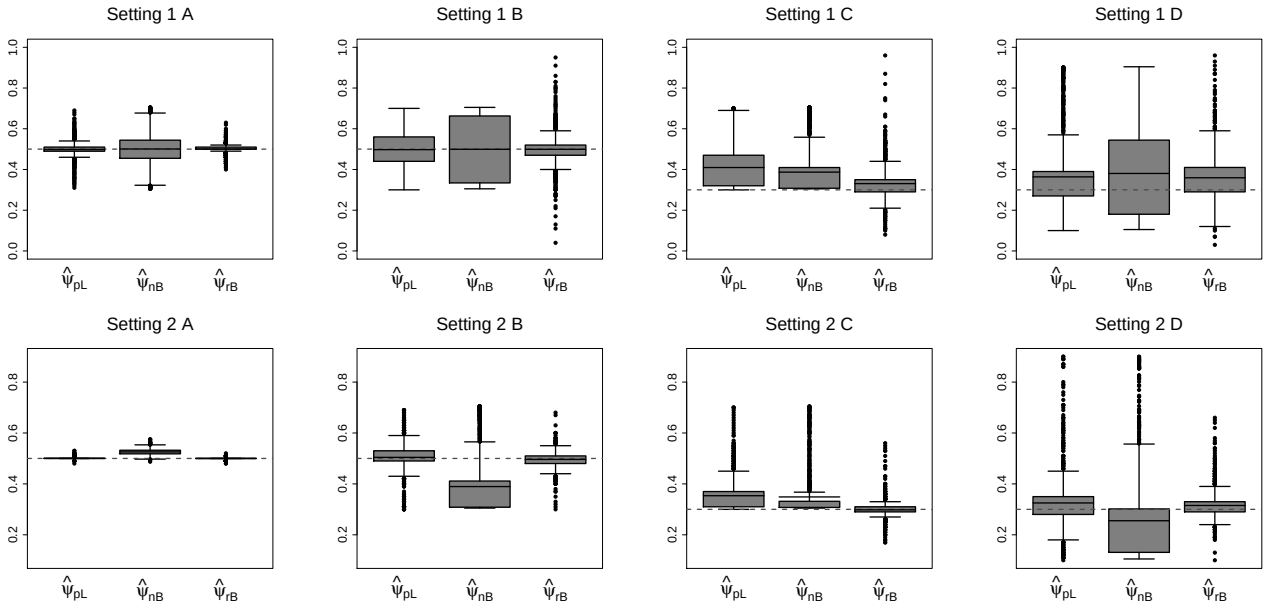

Figure 3.3: Simulation results (boxplots). Dashed lines indicate the true threshold ψ_0, black lines in the boxes are sample means.

shift the true threshold towards the boundary of Ψ (setting C), $\hat{\psi}_{rB}$ clearly outperforms both $\hat{\psi}_{pL}$ and $\hat{\psi}_{nB}$ in terms of mean square error, bias and variance. The differences in bias, variance and mean squared error are more pronounced with a greater number of nuisance parameters p, but still visible in simulations with smaller ratio p/n (setting D). Results are summarized in figure 4.1 and table 4.1. The effects of increasing the signal-to-noise-ratio and shifting ψ_0 on $\ell_p(\psi)$ are illustrated in figure 3.2. The mode of $\ell_p(\psi)$ is less pronounced in setting 1B than in 1A. Further, the number of local maxima rises and they become more distinctive as we move to setting 1B and then to 1C.

3.7 Applications

This work is originally motivated by the application of threshold vector error correction models in price transmission analysis. Such models are rather involved and we refer to Greb, von Cramon-Taubadel, Krivobokova & Munk (2011) for more details and two more real data examples using our regularized Bayesian estimator.

3.7.1 Cross-country growth behaviour

As another application of the regularized Bayesian threshold estimator, we consider the case of economic growth modeling. Durlauf & Johnson (1995) estimate a standard growth model using cross-sectional data on a sample of 96 countries and investigate whether the coefficients of this model differ across sub-sets of countries depending on their initial conditions. Their analysis is based on the so-called regression tree methodology (Breiman, Friedman, Olshen & Stone, 1984), which suggests three thresholds based on two different transition variables for this application. Hansen (2000) revisits their paper. Using the Durlauf and Johnson data he estimates a regression

$$\begin{aligned}&\log\left(GDP\right)_{i,1985} - \log\left(GDP\right)_{i,1960} \\ &\quad = \zeta + \beta \log\left(GDP\right)_{i,1960} + \pi_1 \log\left(INV\right)_i + \pi_2 \log(n_i + g + \delta) + \pi_3 \log\left(SCHOOL\right)_i + \varepsilon_i\end{aligned}$$

which explains real GDP growth between 1960 and 1985 in country i, $\log\left(GDP\right)_{i,1985} - \log\left(GDP\right)_{i,1960}$, using real GDP in 1960 $GDP_{i,1960}$, the investment to GDP ratio INV_i, the growth rate of the working-age population n_i, the rate of technological change g, the rate of depreciation of physical and human capital stocks δ, and the fraction of working-age population enrolled in secondary school $\left(SCHOOL\right)_i$. With reference to Durlauf & Johnson (1995), he sets $g + \delta = 0.05$. He tests for a threshold effect based on either one of transition variables they propose. He only finds evidence based on the transition variable $\log\left(GDP\right)_{i,1960}$ and calculates the profile likelihood (or, equivalently, least squares) estimate as $\hat{\psi}_{pL} = 6.76$ together with an asymptotic 95% confidence interval $[6.39, 7.49]$. This corresponds to an estimate of \$863 per capita GDP in 1960 with an associated confidence interval of [\$594, \$1794]. Hansen (2000) acknowledges that while the confidence interval seems rather tight (given observations for $GDP_{i,1960}$ ranging from \$383 to \$12362), it effectively contains 40 of the 96 countries in the sample. This is in line with the number of local maxima in the profile likelihood function which hints at the uncertainty inherent in this method (figure 3.4). In addition, the fact that $\hat{\psi}_{pL}$ leaves only 18 observations in the first regime gives rise to concern that the threshold might be located close to the boundary of Ψ. We know that the profile likelihood is typically distorted if this is the case.

Hence, we reestimate the model with the regularized Bayesian estimator. The latter

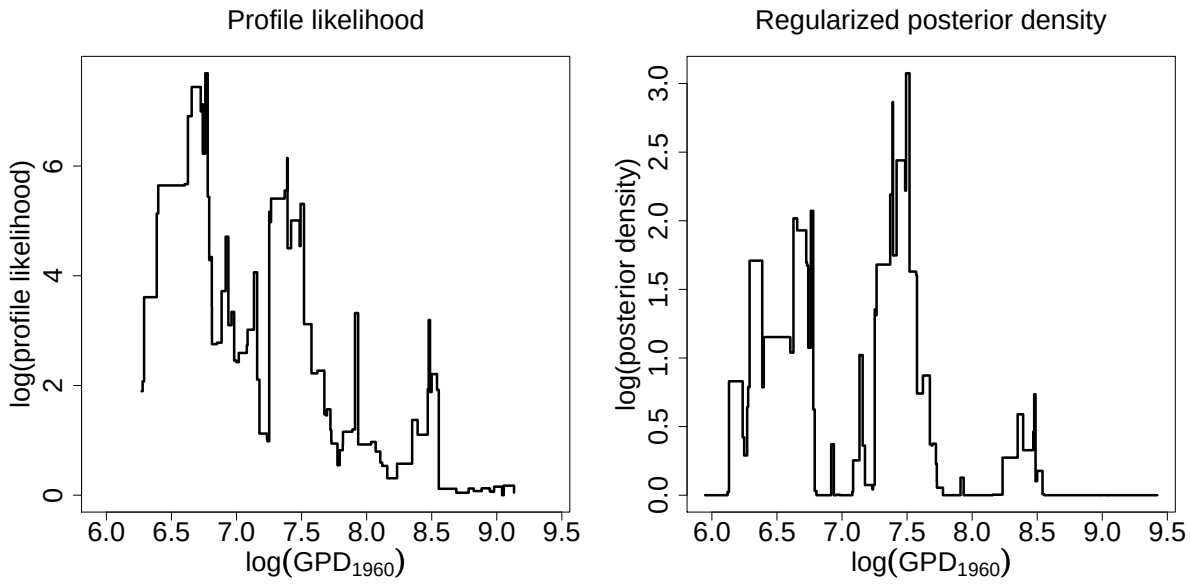

Figure 3.4: Profile likelihood and regularized posterior density for a threshold based on the transition variable $q_i = \log(GDP)_{i,1960}$.

depends on the parameterization of the transition variable. As $\log(GDP)_{i,1960}$ is an explanatory variable, we choose the parameterization $q_i = \log(GDP)_{i,1960}$. Figure 3.4 shows that the resulting posterior density differs considerably from the profile likelihood function and that the location of the maximum shifts. This is not surprising given the deformations often observed for the profile likelihood function close to the boundary of the threshold parameter space. The posterior median is located at $\hat{\psi}_{rB} = 7.37$ compared with Hansen's (2000) $\hat{\psi}_{pL} = 6.76$. It implies that for the 43 poorest countries coefficients for the growth model are distinct from the rest, whereas the profile likelihood estimate implicates that this is only the case for the poorest 18 countries. While it is not possible to state conclusively that the regularized Bayesian estimate is more appropriate from an economic perspective, the shapes of the likelihoods in figure 3.4 and the fact that the profile likelihood estimate is near the boundary of its domain suggests that the latter may be distorted by the weaknesses of the profile likelihood method discussed above.

Comparing profile likelihood estimates for the regression coefficients with their regularized Bayesian counterparts, we note that there is much less difference between regimes according to regularized Bayesian than profile likelihood estimates (see table 3.7.1). The difference between the two regimes as estimated within the regularized Bayesian frame-

work is negligible. This is in line with Hansen's finding that the null hypothesis of no threshold is not rejected at the 5%-level (Hansen, 2000, page 587). The example demonstrates the effect of using the suggested regularized Bayesian estimator instead of the profile likelihood estimator in small samples with a multi-modal profile likelihood and high uncertainty attached to the estimate $\hat{\psi}_{pL}$ obtained by maximizing it.

	1st regime					2nd regime				
	$\hat{\zeta}$	$\hat{\beta}$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$	$\hat{\zeta}$	$\hat{\beta}$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$
pL	4.31	-0.66	0.23	-0.29	0.02	3.66	-0.32	0.50	-0.49	0.36
	(3.21)	(0.33)	(0.14)	(0.92)	(0.11)	(0.85)	(0.07)	(0.11)	(0.30)	(0.07)
rB	3.36	-0.41	0.47	-0.60	0.22	3.37	-0.38	0.47	-0.62	0.20
	(0.85)	(0.08)	(0.09)	(0.28)	(0.06)	(0.85)	(0.07)	(0.09)	(0.28)	(0.07)

Table 3.3: Regressions coefficient estimates. "pL" refers to the profile likelihood, "rB" to the regularized Bayesian framework. Standard errors in parentheses below the estimates.

3.7.2 Effects of climate on snowshoe hare survival

In addition, we study a famous dataset of snowshoe hare abundance in the main drainage of Hudson Bay in Canada. It consists of annual observations starting in the 19th century. A preeminent feature of the data is cyclical fluctuations in the hare population, see figure 3.5. These have been ascribed to the predator-prey relationship between lynx and snowshoe hares. Samia & Chan (2011) highlight selected references and further investigate one strand of the discussion focusing on the effect of snow conditions on hunting efficiency in different phases of the cylce. To this end, they estimate a generalized threshold regression model with the hare count y_t as a Poisson distributed response whose mean is related to the explanatory variables via a log-link,

$$\log(\mu_t) = \beta_0 + \beta_1 D_t + \begin{cases} \sum_{i=1}^{3} \beta_{1,i} \log(y_{t-i}+1) + \beta_{1,4} w_{t-1} & y_{t-d} \leq \psi, \\ \sum_{i=1}^{3} \beta_{2,i} \log(y_{t-i}+1) + \beta_{2,4} w_{t-1} & y_{t-d} > \psi \end{cases}$$

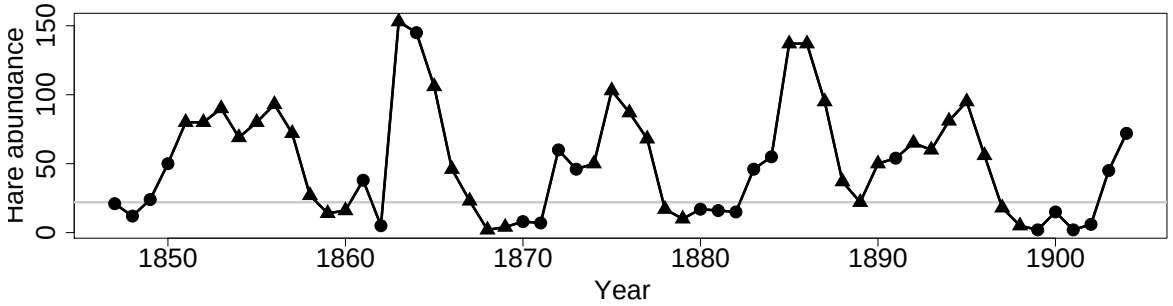

Figure 3.5: Annual hare abundance. Observations estimated to belong to the lower regime are plotted as dots, observations estimated to belong to the upper regime as triangles. The horizontal grey line indicates the location of the estimated threshold, $\hat{\psi}_{rB} = 22$.

for the years $t = 1844, \ldots, 1904$. Apart from the regression coefficients and the threshold, the delay of the transition variable d is included as an additional parameter, $d \in \{1, 2, 3\}$. As the count for the year $t = 1863$ is considered an outlier, the model contains a dummy variable $D_t = I(t = 1863)$. The covariate w_t denotes the detrended annual winter climate index of the North Atlantic Oscillation. We follow them in estimating this model. Our analysis is based on the series of hare abundance initially presented graphically by MacLulich (1937), which we calibrate with data available online; it is included in the supplementary material to this paper. We further use the North Atlantic Oscillation index published at www.cru.uea.ac.uk/cru/data/nao.

The series of 61 observations is rather short and maximizing out regression coefficients leaves us with a profile likelihood function for (d, ψ) which is characterized by various local maxima; it is displayed in the upper row of figure 3.6 for $d = 1, 2, 3$ and $\psi \in \Psi^*(1)$. In addition, we cannot rule out overdispersion. Hence, we are confronted with a setting in which the regularized Bayesian estimate can be more reliable than the profile likelihood estimate. This becomes evident in the second row of figure 3.6, which shows the posterior densities for ψ corresponding to $d = 1, 2, 3$. While we obtain a profile likelihood estimate $(\hat{d}_{pL}, \hat{\psi}_{pL}) = (3, 55)$, the regularized Bayesian estimator yields $(\hat{d}_{rB}, \hat{\psi}_{rB}) = (2, 22)$ with $\hat{d}_{rB}$ calculated as the posterior median based on a flat prior on $\{1, 2, 3\}$.

When referring to Samia & Chan (2011) we have to keep in mind that their results diverge slightly from ours and are not directly comparable as we were not able to obtain the data

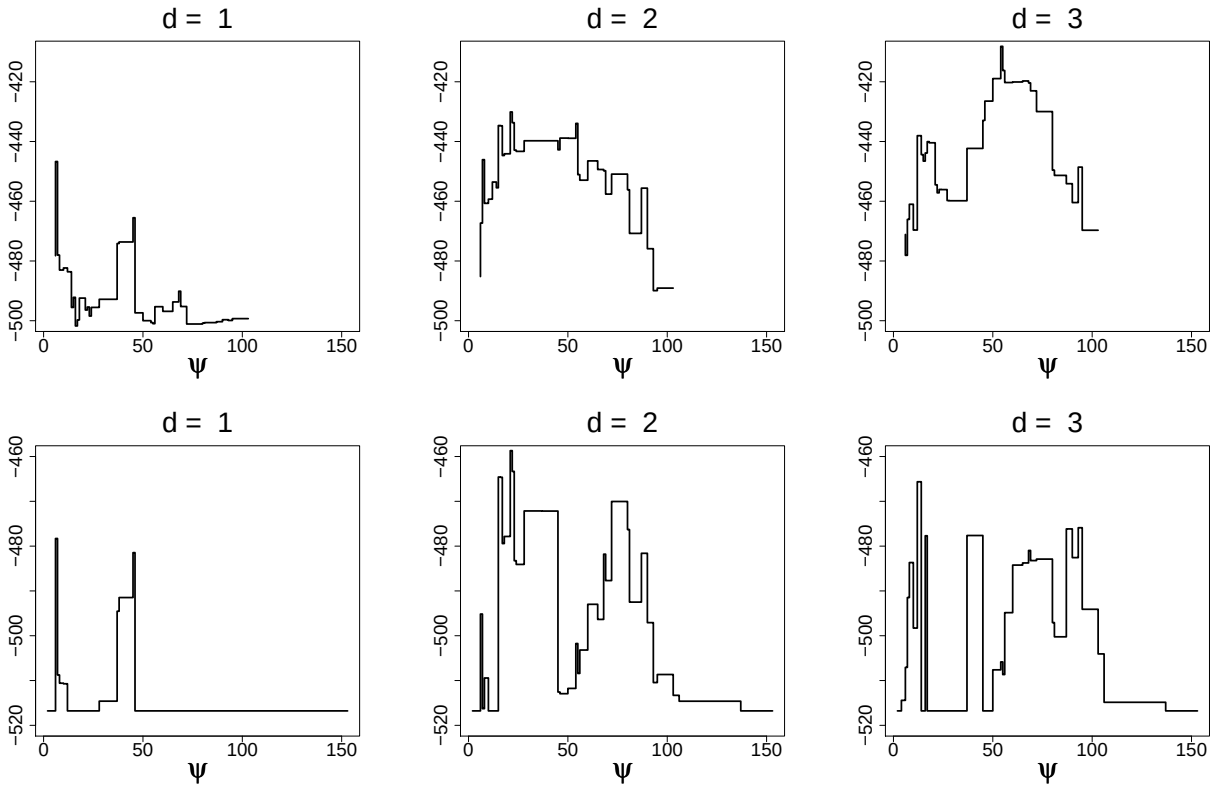

Figure 3.6: Log-likelihood functions (upper row) and log-posterior densities (lower row) for different delays of the transition variable.

they used. Yet, their profile likelihood estimate is still very close, $(\hat{d}_{pL}, \hat{\psi}_{pL}) = (3, 69)$. However, they discard this estimate in favor of $(\hat{d}, \hat{\psi}) = (2, 25)$, giving heuristic arguments based on residual analysis. The latter also allows for a very plausible interpretation. Apparently, our regularized Bayesian estimate $(\hat{d}_{rB}, \hat{\psi}_{rB}) = (2, 22)$ is close to the preferred estimate in Samia & Chan (2011). In fact, the difference in estimated thresholds only has implications for a single observation ($t = 1869$). Except for this, thresholds induce identical allocations of observations to regimes (in the respective datasets), as is clearly visible when comparing our figure 3.5 with figure 1 in Samia & Chan (2011). Hence, the regularized Bayesian estimator enables us to attain a meaningful estimate directly avoiding any arbitrary modification of the suggested estimation method as done by Samia & Chan (2011). Coefficient estimates are similar in both modeling frameworks.

3.8 Conclusions

In this work we describe settings in which estimation of generalized threshold regression models can be problematic. We suggest a new regularized Bayesian estimator which outperforms standard estimators. In particular, the suggested threshold estimator is defined on the whole parameter space and thus circumvents the subjective and often misleading restriction of the threshold domain which standard estimators require. Moreover, regularizing the posterior density at the boundary of its domain helps to improve estimation, especially if the true threshold is close to this boundary. Employing the empirical Bayes approach, we can use built-in functions for generalized linear mixed models in statistics software and obtain estimates with little additional numerical effort and without the use of Markov chain Monte Carlo or other sampling techniques. Inference about the estimated parameter can be carried out in the standard Bayesian manner. Simulation studies and a real-data example confirm the effectiveness and relevance of our method.

Acknowledgements

The authors acknowledge the support of the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the Institutional Strategy of the University of Göttingen and FOR 916.

3.9 Appendix (technical details)

3.9.1 Derivation of equations (3.4) – (3.6)

We first approximate the profile likelihood,

$$
\begin{aligned}
-\ell_p(\psi) \propto -\frac{1}{\hat{\phi}_\psi}\sum_{i=1}^{n} y_i\hat{\theta}_i - b(\hat{\theta}_i) \approx \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\phi}_\psi b''(\hat{\theta}_i)} = \frac{1}{2}\sum_{i=1}^{n}\frac{\{\hat{z}_i - (\boldsymbol{X}_\psi)_i\hat{\boldsymbol{\beta}}_\psi\}^2}{\hat{\phi}_\psi g'(\hat{\mu}_i)^2 b''(\hat{\theta}_i)} \\
= \frac{1}{2}\left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right)^T \widehat{\boldsymbol{W}}\left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right) \approx \frac{1}{2}\left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right)^T \boldsymbol{W}\left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right)
\end{aligned}
$$

where

$$\hat{\boldsymbol{z}} = \boldsymbol{X}_\psi \hat{\boldsymbol{\beta}}_\psi + \widehat{\boldsymbol{G}}\left(\boldsymbol{y} - \hat{\boldsymbol{\mu}}\right), \ \widehat{\boldsymbol{G}} = \operatorname{diag}\left\{g'(\hat{\mu}_i)\right\} \text{ and } \widehat{\boldsymbol{W}} = \operatorname{diag}\left\{\hat{\phi}_\psi b''(\hat{\theta}_i) g'(\hat{\mu}_i)^2\right\}^{-1}.$$

The estimated $\widehat{\boldsymbol{W}}$ for fixed ψ is assumed to vary little or not at all as a function of the mean so we use $\boldsymbol{W}$ evaluated at the true $\boldsymbol{\beta}$ directly. This is a typical assumption in the literature on generalized linear models (e.g. Breslow & Clayton, 1993). We assume the same for $\widehat{\boldsymbol{G}}$.

To compare $\left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi \hat{\boldsymbol{\beta}}_\psi\right)^T \boldsymbol{W} \left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi \hat{\boldsymbol{\beta}}_\psi\right)$ with $\left(\boldsymbol{z} - \boldsymbol{X}_\psi \boldsymbol{\beta}\right)^T \boldsymbol{W} \left(\boldsymbol{z} - \boldsymbol{X}_\psi \boldsymbol{\beta}\right)$, where $\boldsymbol{z} = \boldsymbol{X}_\psi \boldsymbol{\beta} + \boldsymbol{G}\left(\boldsymbol{y} - \boldsymbol{\mu}\right)$, we then note that

$$\operatorname{bias}\left(\hat{\boldsymbol{\beta}}_\psi\right) = \mathrm{E}\left(\hat{\boldsymbol{\beta}}_\psi\right) - \boldsymbol{\beta} = \left\{0, (\boldsymbol{X}_2' \boldsymbol{W} \boldsymbol{X}_2)^{-1} \boldsymbol{X}_{[\psi,\psi_0]}^T \boldsymbol{W} \boldsymbol{X}_{[\psi,\psi_0]} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)\right\}^T. \quad (3.10)$$

To see this, we approximate $\hat{\boldsymbol{\beta}}_\psi$ via Fisher-scoring as $\hat{\boldsymbol{\beta}}_\psi = \left(\boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi\right)^{-1} \boldsymbol{X}_\psi^T \boldsymbol{W} \hat{\boldsymbol{z}}$, then, exploiting that $\hat{\boldsymbol{z}}$ is the working variable obtained in the last (m-th) iteration, calculate

$$\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{z}}) = \mathrm{E}(\boldsymbol{z}_m) &= \mathrm{E}\left\{\boldsymbol{G}_m\left(\boldsymbol{y} - \boldsymbol{\mu}_m\right) + \boldsymbol{X}_\psi(\boldsymbol{\beta}_\psi)_m\right\} = \boldsymbol{G}_m\left(\mathrm{E}(\boldsymbol{y}) - \boldsymbol{\mu}_m\right) + \boldsymbol{X}_\psi(\boldsymbol{\beta}_\psi)_m \\
&= \boldsymbol{G}_m\left\{h(\boldsymbol{\eta}) - h(\boldsymbol{\eta}_m)\right\} + \boldsymbol{X}_\psi(\boldsymbol{\beta}_\psi)_m \approx \boldsymbol{G}_m\left[\frac{\partial h}{\partial \boldsymbol{\eta}}(\boldsymbol{\eta}_m)\left\{\boldsymbol{X}_{\psi_0}\boldsymbol{\beta} - \boldsymbol{X}_\psi(\boldsymbol{\beta}_\psi)_m\right\}\right] + \boldsymbol{X}_\psi(\boldsymbol{\beta}_\psi)_m \\
&= \boldsymbol{G}_m \boldsymbol{G}_m^{-1}\left\{\boldsymbol{X}_{\psi_0}\boldsymbol{\beta} - \boldsymbol{X}_\psi(\boldsymbol{\beta}_\psi)_m\right\} + \boldsymbol{X}_\psi(\boldsymbol{\beta}_\psi)_m \\
&= \boldsymbol{X}_{\psi_0}\boldsymbol{\beta},
\end{aligned}$$

and get

$$\begin{aligned}
E\left(\hat{\boldsymbol{\beta}}_\psi\right) &= \left(\boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi\right)^{-1} \boldsymbol{X}_\psi^T \boldsymbol{W}\, \mathrm{E}(\hat{\boldsymbol{z}}) = \begin{Bmatrix} (\boldsymbol{X}_1^T \boldsymbol{W} \boldsymbol{X}_1)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & (\boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2)^{-1} \end{Bmatrix} \boldsymbol{X}_\psi^T \boldsymbol{W}\, \mathrm{E}(\hat{\boldsymbol{z}}) \\
&= \begin{Bmatrix} (\boldsymbol{X}_1^T \boldsymbol{W} \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{W} \\ (\boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^T \boldsymbol{W} \end{Bmatrix} \left\{(\boldsymbol{X}_1 + \boldsymbol{X}_{[\psi,\psi_0]})\boldsymbol{\beta}_1 + (\boldsymbol{X}_2 - \boldsymbol{X}_{[\psi,\psi_0]})\boldsymbol{\beta}_2\right\} \\
&= \begin{Bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 + (\boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_{[\psi,\psi_0]} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \end{Bmatrix}.
\end{aligned}$$

It is helpful to further keep in mind that

$$\begin{aligned}\mathrm{E}\,\hat{\boldsymbol{\mu}} =& \mathrm{E}\left\{h\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right)\right\} \approx \mathrm{E}\left\{h\left(\boldsymbol{X}_\psi\boldsymbol{\beta}\right) + \boldsymbol{G}^{-1}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi - \boldsymbol{X}_\psi\boldsymbol{\beta}\right)\right\} \\ =& h(\boldsymbol{X}_\psi\boldsymbol{\beta}) + \boldsymbol{G}^{-1}\boldsymbol{X}_\psi\,\mathrm{E}\left(\hat{\boldsymbol{\beta}}_\psi - \boldsymbol{\beta}\right) = h(\boldsymbol{X}_\psi\boldsymbol{\beta}) + \boldsymbol{G}^{-1}\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi\end{aligned}$$

and

$$\mathrm{var}\,\hat{\boldsymbol{\mu}} \approx \mathrm{var}\left\{h\left(\boldsymbol{X}_\psi\boldsymbol{\beta}\right) + \boldsymbol{G}^{-1}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi - \boldsymbol{X}_\psi\boldsymbol{\beta}\right)\right\} = \left(\boldsymbol{G}^{-1}\right)^2\mathrm{var}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right),$$

consequently,

$$\hat{\boldsymbol{\mu}} = \mathrm{E}\,\hat{\boldsymbol{\mu}} + \mathcal{O}_p\left(\sqrt{\mathrm{var}\,\hat{\boldsymbol{\mu}}}\right) \approx h(\boldsymbol{X}_\psi\boldsymbol{\beta}) + \boldsymbol{G}^{-1}\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi + \boldsymbol{G}^{-1}\mathcal{O}_p\left(\sqrt{\mathrm{var}\,\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right). \tag{3.11}$$

We now define $\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{X}_2(\boldsymbol{X}_2^T\boldsymbol{W}\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^T\boldsymbol{W}\boldsymbol{X}_{[\psi,\psi_0]}$ and use (3.10) and (3.11) to obtain

$$\begin{aligned}-\ell_p(\psi) \propto& \left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right)^T\boldsymbol{W}\left(\hat{\boldsymbol{z}} - \boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi\right) \\ \approx& (\boldsymbol{y} - \hat{\boldsymbol{\mu}})^T\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}\,(\boldsymbol{y} - \hat{\boldsymbol{\mu}}) \\ =& \{\boldsymbol{y} - h(\boldsymbol{X}_\psi\boldsymbol{\beta})\}^T\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}\,\{\boldsymbol{y} - h(\boldsymbol{X}_\psi\boldsymbol{\beta})\} \\ &+ 2\,\{\boldsymbol{y} - h(\boldsymbol{X}_\psi\boldsymbol{\beta})\}^T\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}\,\{h(\boldsymbol{X}_\psi\boldsymbol{\beta}) - \hat{\boldsymbol{\mu}}\} + \{h(\boldsymbol{X}_\psi\boldsymbol{\beta}) - \hat{\boldsymbol{\mu}}\}^T\boldsymbol{G}\boldsymbol{W}\boldsymbol{G}\,\{h(\boldsymbol{X}_\psi\boldsymbol{\beta}) - \hat{\boldsymbol{\mu}}\} \\ \approx& (\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T\boldsymbol{W}\,(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta}) \\ &- 2\,(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T\boldsymbol{W}\left\{\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi + \mathcal{O}_p\left(\sqrt{\mathrm{var}\,\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right)\right\} \\ &+ \left\{\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi + \mathcal{O}_p\left(\sqrt{\mathrm{var}\,\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right)\right\}^T\boldsymbol{W}\left\{\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi + \mathcal{O}_p\left(\sqrt{\mathrm{var}\,\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right)\right\} \\ =& (\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T\boldsymbol{W}\,(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta}) \\ &- 2\,(\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T\boldsymbol{W}\left\{\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi + \mathcal{O}_p\left(\sqrt{\mathrm{var}\,\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right)\right\} \\ &+ \left(\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi\right)^T\boldsymbol{W}\left(\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi\right) + 2\left(\boldsymbol{X}_\psi\,\mathrm{bias}\,\hat{\boldsymbol{\beta}}_\psi\right)^T\boldsymbol{W}\mathcal{O}_p\left(\sqrt{\mathrm{var}\,\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right) \\ &+ \mathcal{O}_p\left(\sqrt{\mathrm{var}\,\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right)^T\boldsymbol{W}\mathcal{O}_p\left(\sqrt{\mathrm{var}\,\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_\psi}\right)\end{aligned}$$

$$
\begin{aligned}
&= (\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T \boldsymbol{W} (\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta}) \\
&\quad + \left\{ (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^T \boldsymbol{H}^T - 2 (\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T \boldsymbol{W}^{1/2} \right\} \boldsymbol{H} (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) \\
&\quad + \left\{ 2(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^T \boldsymbol{H}^T - 2 (\boldsymbol{z} - \boldsymbol{X}_\psi\boldsymbol{\beta})^T \boldsymbol{W}^{1/2} \right\}^T \mathcal{O}_p \left(\boldsymbol{W}^{1/2} \sqrt{\operatorname{var} \boldsymbol{X}_\psi \hat{\boldsymbol{\beta}}_\psi} \right) \\
&\quad + \mathcal{O}_p(2p),
\end{aligned}
$$

3.9.2 Derivation of equation (3.8)

We obtain the approximate posterior (3.8) as follows. Laplace approximation produces

$$
\begin{aligned}
&\int_{\mathbb{R}^p} \int_{\mathbb{R}^p} p(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{\delta}, \psi, \phi, \sigma_\delta^2, \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{\delta}|\sigma_\delta^2) d\boldsymbol{\delta} d\boldsymbol{\beta}_1 \\
&\qquad = (2\pi)^{-p/2} |\sigma_\delta^2 \boldsymbol{I}_p|^{-1/2} \int_{\mathbb{R}^p} \int_{\mathbb{R}^p} \exp\left\{ -\kappa\left(\boldsymbol{\delta}, \boldsymbol{\beta}_1\right) \right\} d\boldsymbol{\delta} d\boldsymbol{\beta}_1 \\
&\qquad = (2\pi)^{p/2} |\sigma_\delta^2 \boldsymbol{I}_p|^{-1/2} \exp\left\{ -\kappa\left(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}_1\right) \right\} \left| \frac{\partial^2 \kappa}{\partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1) \partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)^T} (\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}_1) \right|^{-1/2} + \mathcal{O}\left(n^{-1}\right)
\end{aligned}
$$

for $\kappa\left(\boldsymbol{\delta}, \boldsymbol{\beta}_1\right) = -\sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi} - c(y_i, \phi) + \frac{1}{2\sigma_\delta^2} \boldsymbol{\delta}^T \boldsymbol{\delta}$ and $\left(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}_1\right) = \underset{(\boldsymbol{\delta}, \boldsymbol{\beta}_1) \in \mathbb{R}^{2p}}{\operatorname{argmax}} -\kappa\left(\boldsymbol{\delta}, \boldsymbol{\beta}_1\right)$.

Given the derivatives

$$
\frac{\partial \kappa}{\partial \boldsymbol{\delta}} (\boldsymbol{\delta}) = -\sum_{i=1}^{n} \frac{(y_i - \mu_i)(\boldsymbol{X}_2)_i}{\phi b''(\theta_i) g'(\mu_i)} + \frac{1}{\sigma_\delta^2} \boldsymbol{\delta} = -\boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{G} (\boldsymbol{y} - \boldsymbol{\mu}) + \frac{1}{\sigma_\delta^2} \boldsymbol{\delta},
$$

$$
\frac{\partial \kappa}{\partial \boldsymbol{\beta}_1} (\boldsymbol{\beta}_1) = -\sum_{i=1}^{n} \frac{(y_i - \mu_i)(\boldsymbol{X})_i}{\phi b''(\theta_i) g'(\mu_i)} = -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{G} (\boldsymbol{y} - \boldsymbol{\mu}),
$$

and

$$
\partial^2 \kappa \Big/ \partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1) \partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)^T = \begin{pmatrix} \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2 + (1/\sigma_\delta^2) \boldsymbol{I}_p & \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X} \\ \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}_2 & \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \end{pmatrix} \tag{3.12}
$$

for $\boldsymbol{W}^{-1} = \operatorname{diag}\left\{ \phi b''(\theta_i) g'(\mu_i)^2 \right\}$ and $\boldsymbol{G} = \operatorname{diag}\left\{ g'(\mu_i) \right\}$, we obtain

$$
\left| \partial^2 \kappa \Big/ \partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1) \partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)^T \right| = \left| \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2 + \left(1/\sigma_\delta^2\right) \boldsymbol{I}_p \right| \left| \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X} \right|
$$

using basic matrix algebra.

To find $\hat{\boldsymbol{\delta}}$ and $\hat{\boldsymbol{\beta}}_1$, we iteratively solve

$$\boldsymbol{X}_2^T\boldsymbol{W}\boldsymbol{G}(\boldsymbol{y}-\boldsymbol{\mu}) = \frac{1}{\sigma_\delta^2}\boldsymbol{\delta} \text{ and } \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{G}(\boldsymbol{y}-\boldsymbol{\mu}) = 0$$

via Fisher-scoring: Starting at $\hat{\boldsymbol{\delta}} = \boldsymbol{\delta}_0$ and $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{\beta}_1)_0$, we solve

$$\mathcal{I}(\boldsymbol{\delta}_m, \boldsymbol{\beta}_m)\begin{pmatrix}\boldsymbol{\delta}_{m+1}\\ (\boldsymbol{\beta}_1)_{m+1}\end{pmatrix} = \mathcal{I}(\boldsymbol{\delta}_m, \boldsymbol{\beta}_m)\begin{pmatrix}\boldsymbol{\delta}_m\\ (\boldsymbol{\beta}_1)_m\end{pmatrix} + s(\boldsymbol{\delta}_m, (\boldsymbol{\beta}_1)_m),$$

$\mathcal{I} = \partial^2\kappa\Big/\partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)\partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)^T$ and $s = -\partial\kappa\Big/\partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)$, or, more explicitely,

$$\Big\{\boldsymbol{X}_2^T\boldsymbol{W}_m\boldsymbol{X}_2 + \frac{1}{\sigma_\delta^2}\boldsymbol{I}_p\Big\}\boldsymbol{\delta}_{m+1} + \boldsymbol{X}_2^T\boldsymbol{W}_m\boldsymbol{X}(\boldsymbol{\beta}_1)_{m+1} = \boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{z}_m$$

and

$$\boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{X}_2\boldsymbol{\delta}_{m+1} + \boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{X}(\boldsymbol{\beta}_1)_{m+1} = \boldsymbol{X}^T\boldsymbol{W}_m\boldsymbol{z}_m,$$

where $\boldsymbol{z}_m = \boldsymbol{X}_2\boldsymbol{\delta}_m + \boldsymbol{X}(\boldsymbol{\beta}_1)_m + \boldsymbol{G}_m(\boldsymbol{y}-\boldsymbol{\mu}_m)$. This yields

$$\hat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\tilde{\boldsymbol{z}} \quad \text{and} \quad \hat{\boldsymbol{\delta}} = \sigma_\delta^2\boldsymbol{X}_2^T\boldsymbol{V}^{-1}(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1),$$

where $\boldsymbol{V} = \boldsymbol{W}^{-1} + \sigma_\delta^2\boldsymbol{X}_2\boldsymbol{X}_2^T$ and $\tilde{\boldsymbol{z}} = \boldsymbol{X}_2^T\hat{\boldsymbol{\delta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 + \boldsymbol{G}(\boldsymbol{y}-\boldsymbol{\mu})$, with $\boldsymbol{W}$, $\boldsymbol{G}$ and $\boldsymbol{\mu}$ evaluated at $\boldsymbol{\delta} = \hat{\boldsymbol{\delta}}$ and $\boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_1$ (Harville, 1977).

With this, we can now further simplify the posterior. Following Breslow & Clayton (1993) in replacing

$$-2\sum_{i=1}^n\{y_i\theta_i - b(\theta_i)\} \quad \text{by the chi-squared statistic} \quad \sum_{i=1}^n\frac{(y_i-\mu_i)^2}{b''(\theta_i)}$$

we can exploit the identity

$$\boldsymbol{V}^{-1}\left(\tilde{\boldsymbol{z}} - \hat{\boldsymbol{\beta}}_1\right) = \boldsymbol{W}\left(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\hat{\boldsymbol{\delta}}\right),$$

which results in

$$\left(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\hat{\boldsymbol{\delta}}\right)^T \boldsymbol{W}\left(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\hat{\boldsymbol{\delta}}\right) = \left(\tilde{\boldsymbol{z}} - \hat{\boldsymbol{\beta}}_1\right)^T \boldsymbol{V}^{-1}\left(\tilde{\boldsymbol{z}} - \hat{\boldsymbol{\beta}}_1\right) - \frac{1}{\sigma_\delta^2}\hat{\boldsymbol{\delta}}^T\hat{\boldsymbol{\delta}},$$

and, hence,

$$\begin{aligned}
&\exp\left\{\sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi) - \frac{1}{2\sigma_\delta^2}\hat{\boldsymbol{\delta}}^T\hat{\boldsymbol{\delta}}\right\} \\
&\qquad\approx \exp\left\{-\frac{1}{2}\left(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\hat{\boldsymbol{\delta}}\right)^T \boldsymbol{W}\left(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2\hat{\boldsymbol{\delta}}\right) + \sum_{i=1}^{n} c(y_i,\phi) - \frac{1}{2\sigma_\delta^2}\hat{\boldsymbol{\delta}}^T\hat{\boldsymbol{\delta}}\right\} \\
&\qquad= \exp\left\{-\frac{1}{2}\left(\tilde{\boldsymbol{z}} - \hat{\boldsymbol{\beta}}_1\right)^T \boldsymbol{V}^{-1}\left(\tilde{\boldsymbol{z}} - \hat{\boldsymbol{\beta}}_1\right) + \sum_{i=1}^{n} c(y_i,\phi)\right\}.
\end{aligned}$$

Alltogether, this leaves us with

$$\begin{aligned}
&\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} p(\boldsymbol{y}|\boldsymbol{\beta}_1,\boldsymbol{\delta},\psi,\phi,\sigma_\delta^2,\boldsymbol{X},\boldsymbol{q})p(\boldsymbol{\delta}|\sigma_\delta^2)d\boldsymbol{\delta}d\boldsymbol{\beta}_1 \\
&\qquad= (2\pi)^{p/2}|\sigma_\delta^2\boldsymbol{I}_p|^{-1/2}\exp\left\{\sum_{i=1}^{n}\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi) - \frac{1}{2\sigma_\delta^2}\hat{\boldsymbol{\delta}}^T\hat{\boldsymbol{\delta}}\right\}\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right|^{-1/2} \\
&\qquad\quad\cdot\left|\boldsymbol{X}_2^T\boldsymbol{W}\boldsymbol{X}_2 + \left(1/\sigma_\delta^2\right)\boldsymbol{I}_p\right|^{-1/2} + \mathcal{O}\left(n^{-1}\right) \\
&\qquad\approx (2\pi)^{p/2}\exp\left\{-\frac{1}{2}\left(\tilde{\boldsymbol{z}} - \hat{\boldsymbol{\beta}}_1\right)^T \boldsymbol{V}^{-1}\left(\tilde{\boldsymbol{z}} - \hat{\boldsymbol{\beta}}_1\right) + \sum_{i=1}^{n} c(y_i,\phi)\right\}\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right|^{-1/2} \\
&\qquad\quad\cdot\left|\sigma_\delta^2\boldsymbol{X}_2^T\boldsymbol{W}\boldsymbol{X}_2 + \boldsymbol{I}_p\right|^{-1/2} + \mathcal{O}\left(n^{-1}\right).
\end{aligned}$$

4 The estimation of threshold models in price transmission analysis

Abstract

The threshold vector error correction model is a popular tool for the analysis of spatial price transmission and market integration. In the literature, the profile likelihood estimator is the preferred choice for estimating this model. Yet, in certain settings this estimator performs poorly. In particular, if the true thresholds are such that one or more regimes contain only a small number of observations, if unknown model parameters are numerous or if parameters differ little between regimes, the profile likelihood estimator displays large bias and variance. Such settings are likely when studying price transmission. For simpler, but related threshold models Greb, Krivobokova, Munk & von Cramon-Taubadel (2011) have developed an alternative estimator, the regularized Bayesian estimator, which does not exhibit these weaknesses. We explore the properties of this estimator for threshold vector error correction models. Simulation results show that it outperforms the profile likelihood estimator, especially in situations in which the profile likelihood estimator fails. Two empirical applications – a reassessment of the the seminal paper by Goodwin & Piggott (2001), and an analysis of price transmission between German and Spanish markets for pork – demonstrate the relevance of the new approach for spatial price transmission analysis.

Key words and phrases. Bayesian estimator, market integration, spatial arbitrage, TVECM.

4.1 Introduction

When assessing the integration of spatially separated markets, agricultural economists typically analyze the transmission of price shocks between these markets (Fackler & Goodwin, 2001). The law of one price (LOP) states that prices for a homogeneous good at different locations should differ by no more than the transaction costs of trading the good between these locations. Otherwise traders will engage in spatial arbitrage, which increases the price at the low-price location and reduces the price at the high-price location until the LOP is restored. In spatial equilibrium, the manner in which price shocks are transmitted between two locations will therefore depend on the magnitude of the price difference between these locations (Goodwin & Piggott, 2001; Stephens, Mabaya, Cramon-Taubadel & Barrett, 2011). Shocks that increase the price difference so that it exceeds the costs of trade between the two locations will lead to arbitrage and price transmission. However, if the price difference remains less than these transaction costs, arbitrage will not be profitable and there will be no price transmission. The result is referred to in the literature as "regime-dependent" price transmission. Specifically, the spatial equilibrium model described above will lead to three regimes delineated by two threshold values that equal the transaction costs of trade in one and the other direction, respectively. In the outer regimes where the price difference is greater than the transaction costs of trade in the one or the other direction, arbitrage will lead to the transmission of price shocks. If the price difference lies within the "band of inaction" between these transaction costs, prices can evolve independently of one another. The costs of trade between two locations need not be symmetric; for example, river transport might be more expensive going upstream than it is going downstream. Hence, the thresholds that define the boundaries of the spatial price transmission regimes will have opposite signs and possibly different magnitudes.

Threshold vector error correction models (TVECMs) are frequently used to model this regime-dependent spatial price transmission process. TVECMs became popular with Balke & Fomby's (1997) article on threshold cointegration. Goodwin & Piggott's (2001) seminal paper established TVECMs in price transmission analysis, and dozens of applications have followed. As an indication of the ongoing popularity of the TVECM, a search of the AgEconSearch website (`www.ageconsearch.umn.edu`) on November 15, 2011 with the keywords "price transmission" and "threshold" produced 11 papers posted

in 2010 and 2011.

Typically, and as we explain in greater detail below, thresholds in TVECMs are estimated by maximizing the profile likelihood (Hansen & Seo, 2002). However, in many settings, this estimator is biased and has a high variance. Lo & Zivot (2001) as well as Balcombe, Bailey & Brooks (2007) acknowledge this problem. Profile likelihood estimates are especially prone to be unreliable in situations characterized by large numbers of unknown model parameters besides the thresholds, when there is little difference between adjoining regimes, and when the location of the thresholds leaves only few observations in one of the regimes (which is inevitable in small samples). These problems are generic and emerge in many econometric settings, but they are particularly acute when profile likelihood is used to estimate TVECMs. To cope with these shortcomings, several strategies are proposed in the literature. Perhaps the most well-known of these is the modified profile likelihood function introduced by Barndorff-Nielsen (1983). However, the proposed modifications are usually based on regularity assumptions that do not hold for the TVECM. A further weakness of the profile likelihood estimator is that it depends on an arbitrary trimming parameter that ensures that each regime contains a minimum number of observations and, thus, that estimation of the model parameters in that regime is possible. This can be a problematic restriction when modeling spatial price transmission. If market integration is strong, differences in prices between two locations that exceed the transaction cost thresholds – and therefore fall into one of the outer regimes – will be corrected quickly. If this is the case, there will be few observations in the outer regimes, and a trimming parameter which forces more observations into these regimes will inevitably lead to unreliable estimates of both the threshold values and the model parameters in each regime. Estimation is not necessarily easier if the price data originate from markets that are poorly integrated because in this case the weak price transmission displayed in the outer regimes may be observationally quite similar to the independent price movements in the inner "band of inaction". Finally, the non-differentiability of the TVECM's likelihood function with respect to the thresholds exacerbates computation of its maximum, which can also be a source of imprecise estimates.

These problems with the profile likelihood estimator suggest that there is a need to rethink the estimation of TVECMs in price transmission analysis. In this article we investigate the suitability of an alternative threshold estimator developed for generalized

threshold regression models (Greb, Krivobokova, Munk & von Cramon-Taubadel, 2011). Among its advantages, this alternative estimator does not require a trimming parameter. We demonstrate using Monte Carlo experiments that this so-called regularized Bayesian estimator clearly outperforms the profile likelihood estimator not only for generalized threshold regression models, but also specifically for TVECMs, even in settings in which the profile likelihood estimator is highly biased and variable. We also show that although employing the regularized Bayesian estimator is technically easy, careful numerical implementation – even if it is computationally intensive – can be decisive. Of course, it is important to go beyond the demonstration of the superior statistical properties of the regularized Bayesian threshold estimator, and to consider as well its implications for empirical price transmission analysis using TVECMs. Here, it is crucial to interpret not only the estimated threshold parameters, but also the parameters that describe the dynamics of price transmission within each regime. We draw on two empirical applications to illustrate this.

The rest of this article is organized as follows. In the next section, we specify the TVECM, discuss existing threshold estimators and their deficiencies, present the alternative estimator, and comment on computational pitfalls in threshold estimation. Subsequently, we illustrate the performance of the new estimator by means of a simulation study. As empirical applications we first revisit the analysis of spatial market integration for four corn and soybean markets in North Carolina detailed in the seminal contribution by Goodwin & Piggott (2001), and second analyze spatial price transmission between German and Spanish pork markets. The last section concludes.

4.2 Theory

4.2.1 The model

Observations $\boldsymbol{p}_t = (p_{1,t}, p_{2,t})^T$, $t = 1 \dots n$, of a two-dimensional time series generated by a TVECM with three different regimes, which are characterized by parameters $\boldsymbol{\rho}_k, \boldsymbol{\theta}_k \in \mathbb{R}^2$

and $\boldsymbol{\Theta}_{km} \in \mathbb{R}^{2\times 2}$ for $k = 1, 2, 3$ and $m = 1, \ldots, M$, can be written as

$$\Delta \boldsymbol{p}_t = \begin{cases} \boldsymbol{\rho}_1 \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} + \boldsymbol{\theta}_1 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{1m} \Delta \boldsymbol{p}_{t-m} + \boldsymbol{\varepsilon}_t & , \quad \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} \leq \psi_1 \quad \text{(Regime 1)} \\ \boldsymbol{\rho}_2 \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} + \boldsymbol{\theta}_2 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{2m} \Delta \boldsymbol{p}_{t-m} + \boldsymbol{\varepsilon}_t & , \quad \psi_1 < \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} \leq \psi_2 \quad \text{(Regime 2)} \\ \boldsymbol{\rho}_3 \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} + \boldsymbol{\theta}_3 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{3m} \Delta \boldsymbol{p}_{t-m} + \boldsymbol{\varepsilon}_t & , \quad \psi_2 < \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} \quad \text{(Regime 3).} \end{cases} \tag{4.1}$$

We assume that $\boldsymbol{p}_t$ forms an $I(1)$ time series with cointegrating vector $\boldsymbol{\gamma} \in \mathbb{R}^2$ and error-correction term $\boldsymbol{\gamma}^T \boldsymbol{p}_t$. We further assume that the errors denoted by ε_t have expected value $\mathrm{E}(\boldsymbol{\varepsilon}_t) = 0$ and covariance matrix $\mathrm{cov}(\boldsymbol{\varepsilon}_t) = \sigma^2 \boldsymbol{I}_2 \in (\mathbb{R}^+)^{2\times 2}$; $\boldsymbol{I}_2 \in \mathbb{R}^{2\times 2}$ denotes the identity matrix. We call ψ_1, ψ_2 the threshold parameters and define the threshold parameter space Ψ to include all $\psi = (\psi_1, \psi_2)$ such that $\min(\boldsymbol{\gamma}^T \boldsymbol{p}_t) < \psi_1 < \psi_2 < \max(\boldsymbol{\gamma}^T \boldsymbol{p}_t)$. Although all of the coefficients in equation (4.1) can vary across regimes, some of them can remain constant.

In the spatial equilibrium setting, $p_{1,t}$ and $p_{2,t}$ are prices at different locations and $\boldsymbol{\gamma}$ is often taken to equal $(1, -1)^T$ so that the error correction term $\boldsymbol{\gamma}^T \boldsymbol{p}_t$ measures the difference between p_1 and p_2 at time t. The threshold ψ_1 (ψ_2) corresponds to the transaction costs of trade from location 1 to location 2 (location 2 to location 1). Regimes 1 and 3 are the outer regimes in which the violation of spatial equilibrium leads to arbitrage and price transmission, and regime 2 represents the inner "band of inaction". For economic interpretation, not only the estimates of the threshold parameters are of interest. The estimates of $\boldsymbol{\rho}_k$ ($k = 1, 2, 3$) (often referred to as the "adjustment parameter") are also of interest as they measure the speed with which violations of spatial equilibrium between two locations are corrected in the respective regimes.

To express the model in matrix notation, we define vectors $\Delta \boldsymbol{p}_i$ and $\boldsymbol{\varepsilon}_i$ by stacking the ith components of $\Delta \boldsymbol{p}_t$ and $\boldsymbol{\varepsilon}_t$, respectively; and $\boldsymbol{I}\left(\boldsymbol{\gamma}^T \boldsymbol{p} \leq \psi_1\right)$, $\boldsymbol{I}\left(\psi_1 < \boldsymbol{\gamma}^T \boldsymbol{p} \leq \psi_2\right)$, and $\boldsymbol{I}\left(\psi_2 < \boldsymbol{\gamma}^T \boldsymbol{p}\right)$ by stacking $I\left(\boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} \leq \psi_1\right)$, $I\left(\psi_1 < \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} \leq \psi_2\right)$ and $I\left(\psi_2 < \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1}\right)$, respectively. $I(\cdot)$ denotes the indicator function. For observations at n time points, an $n \times d$ matrix $\boldsymbol{X}$ is constructed by stacking rows $\boldsymbol{X}_t^T = (\boldsymbol{\gamma}^T \boldsymbol{p}_{t-1}, 1, \Delta \boldsymbol{p}_{t-1}^T, \ldots, \Delta \boldsymbol{p}_{t-M}^T)$ of length $d = 2M + 2$. $\boldsymbol{\beta}_{i,k}$ is the ith column of the matrix $(\boldsymbol{\rho}_k, \boldsymbol{\theta}_k, \boldsymbol{\Theta}_{k1}, \ldots, \Theta_{kM})^T$, $i = 1, 2$ and $k = 1, 2, 3$. With $\mathrm{diag}\{\boldsymbol{I}(\cdot)\}$ defined as the diagonal matrix with entries $\boldsymbol{I}(\cdot)$ in the

diagonal, we can write

$$\begin{aligned}\Delta \boldsymbol{p}_i = &\operatorname{diag}\left\{\boldsymbol{I}\left(\boldsymbol{\gamma}^T\boldsymbol{p} \leq \psi_1\right)\right\}\boldsymbol{X}\beta_{i,1} + \operatorname{diag}\left\{\boldsymbol{I}\left(\psi_1 < \boldsymbol{\gamma}^T\boldsymbol{p} \leq \psi_2\right)\right\}\boldsymbol{X}\beta_{i,2} \\ &+ \operatorname{diag}\left\{\boldsymbol{I}\left(\psi_1 < \boldsymbol{\gamma}^T\boldsymbol{p} \leq \psi_2\right)\right\}\boldsymbol{X}\beta_{i,3} + \varepsilon_i \\ = &\boldsymbol{X}_1\boldsymbol{\beta}_{i,1} + \boldsymbol{X}_2\boldsymbol{\beta}_{i,2} + \boldsymbol{X}_3\boldsymbol{\beta}_{i,3} + \boldsymbol{\varepsilon}_i\end{aligned} \tag{4.2}$$

for $i = 1, 2$. This leads to the a compact representation of model (4.1),

$$\Delta\boldsymbol{p} = \begin{pmatrix}\Delta\boldsymbol{p}_1 \\ \Delta\boldsymbol{p}_2\end{pmatrix} = (\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)\boldsymbol{\beta}_1 + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_2)\boldsymbol{\beta}_2 + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)\boldsymbol{\beta}_3 + \boldsymbol{\varepsilon}, \tag{4.3}$$

where $\boldsymbol{\beta}_k^T = \left(\boldsymbol{\beta}_{1,k}^T, \boldsymbol{\beta}_{2,k}^T\right)$ for $k = 1, 2, 3$, and $\boldsymbol{X} = \boldsymbol{X}_1 + \boldsymbol{X}_2 + \boldsymbol{X}_3$.

A variety of modifications and restrictions of the general TVECM (4.1) have been implemented in price transmission studies. Lo & Zivot (2001) and Ihle (2010, table 2.1) provide details on a number of important specifications. We limit attention to the general TVECM. Restrictions of the model imply further information about the parameters (or relations between them) and, hence, facilitate estimation. The most general case is thus the most challenging. Although the TVECM can be generalized to include r thresholds and $r + 1$ regimes, we focus on a TVECM with two thresholds and three regimes as this is the version of the TVECM that is grounded in spatial equilibrium theory as outlined above. Generalization is straightforward.

4.2.2 Commonly used threshold estimators

The most frequently used threshold estimator in the econometrics literature is the profile likelihood estimator (Hansen & Seo, 2002; Lo & Zivot, 2001). According to this method, for each possible pair of the threshold parameters $\psi = (\psi_1, \psi_2)$ the remaining parameters in the likelihood function corresponding to (4.1) are replaced by their maximum likelihood estimates. The pair of thresholds that maximizes the resulting profile likelihood function is selected as the estimate. More precisely, denoting the log-likelihood function of (4.1) by $\ell\left(\psi, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \sigma^2\right)$, the profile likelihood estimator is defined as

$$\hat{\psi}_{pL} = \arg\max \ell_p(\psi) \qquad \text{with} \qquad \ell_p(\psi) = \ell\left(\psi, \hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \hat{\boldsymbol{\beta}}_3, \hat{\sigma}^2\right) \tag{4.4}$$

and $\hat{\boldsymbol{\beta}}_k$ and $\hat{\sigma}^2$ the maximum likelihood estimates of $\boldsymbol{\beta}_k$ and σ^2. Hence,

$$\begin{aligned}\ell_p(\psi) \propto -\Big\{\Delta\boldsymbol{p} - (\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)\hat{\boldsymbol{\beta}}_1 - (\boldsymbol{I}_2 \otimes \boldsymbol{X}_2)\hat{\boldsymbol{\beta}}_2 - (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)\hat{\boldsymbol{\beta}}_3\Big\}^T \\ \Big\{\Delta\boldsymbol{p} - (\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)\hat{\boldsymbol{\beta}}_1 - (\boldsymbol{I}_2 \otimes \boldsymbol{X}_2)\hat{\boldsymbol{\beta}}_2 - (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)\hat{\boldsymbol{\beta}}_3\Big\}\end{aligned} \tag{4.5}$$

and $\hat{\boldsymbol{\beta}}_k = \left\{(\boldsymbol{I}_2 \otimes \boldsymbol{X}_k)^T(\boldsymbol{I}_2 \otimes \boldsymbol{X}_k)\right\}^{-1}(\boldsymbol{I}_2 \otimes \boldsymbol{X}_k)^T\Delta\boldsymbol{p}$, $k = 1, 2, 3$. Since the profile likelihood function is not differentiable with respect to the threshold parameters, the thresholds that maximize the profile likelihood are determined by calculating (4.5) for each point on a two-dimensional grid of possible threshold values, which is why the literature often refers to the "grid search" method.

The bias and high variance of the profile likelihood threshold estimator are mentioned but not further pursued in the literature on TVECMs (see table 4 and figure 1 in Lo & Zivot, 2001). The simulation results we present below confirm the existence of these weaknesses (see table 4.1 and figures 4.1 and 4.2). Greb, Krivobokova, Munk & von Cramon-Taubadel (2011) provide a detailed analysis of the problems associated with the profile likelihood approach to threshold estimation. In summary, there are two principal problems: i) the dependence on an arbitrary trimming parameter; and ii) the uncertainty inherent in the $\hat{\boldsymbol{\beta}}_k$ which are estimated for each combination of possible threshold values. The problems can be very pronounced in small samples.

In spatial arbitrage modeling, the first issue can be decisive. ψ places each of the observations into one of three regimes. In order to compute $\hat{\boldsymbol{\beta}}_k$, it is essential that at least $d = \dim(\boldsymbol{\beta}_{i,k})$ observations fall into the k-th regime. To ensure this, ψ_1 must be greater than or equal to $\boldsymbol{\gamma}^T\boldsymbol{p}_{(d)}$, where $\boldsymbol{\gamma}^T\boldsymbol{p}_{(1)}, \ldots, \boldsymbol{\gamma}^T\boldsymbol{p}_{(n)}$ is the ordered sequence of error correction terms, and ψ_2 must be correspondingly less than or equal to $\boldsymbol{\gamma}^T\boldsymbol{p}_{(n-d)}$. The trimming parameter restricts ψ accordingly. A variety of trimming parameters are suggested in the literature. Goodwin & Piggott (2001) specify that each regime in the TVECM that they estimate must include at least 25 observations. Balcombe, Bailey & Brooks (2007) impose the restriction that each regime must include at least 20% of the observations in their sample, while Andrews (1993) proposes a minimum proportion of 15%. However, if markets are well-integrated, then arbitrage will lead to rapid correction of any price differences that exceed the thresholds, and the outer regimes will contain correspondingly few observations. Especially in small samples, this

can lead to a situation in which the outer regimes actually contain fewer observations than imposed by the chosen trimming parameter. In this case, the resulting estimator cannot be consistent as the threshold parameter space Ψ (and, hence, the grid that is searched) excludes the true thresholds. Despite its potential impact on threshold estimation, the literature only offers a variety of arbitrary suggestions for the trimming parameter.

The second problem naturally becomes more pronounced as the number of parameters in the model (i.e. the dimension of $\hat{\boldsymbol{\beta}}_k$) increases. Each additional lag included in a bivariate TVECM with three regimes adds 12 coefficients. Hence, the number of coefficients to be estimated can grow rapidly relative to the potentially few observations in the outer regimes. If there is also little difference in coefficients between regimes, pinpointing the location of the thresholds becomes increasingly difficult.

As an alternative to profile likelihood, Bayesian estimators have been employed in some price transmission studies (Balcombe, Bailey & Brooks, 2007; Balcombe & Rapsomanikis, 2008). As explained in Greb, Krivobokova, Munk & von Cramon-Taubadel (2011), the performance of a Bayesian estimator in generalized threshold regression models crucially depends on the selected priors. In the absence of any prior knowledge of potential parameter values, so-called noninformative priors are the natural choice. However, these can distort estimates. In particular, the posterior density associated with noninformative priors for the $\hat{\boldsymbol{\beta}}_k$ inherits the dependence on a trimming parameter from the profile likelihood. Due to an extra term in the likelihood function, which grows rapidly as fewer observations are left in one of the regimes, the posterior density takes its largest values exactly for those threshold values that are arbitrarily included or excluded from the threshold parameter space Ψ when the trimming parameter is varied. Consequently, the trimming parameter strongly affects the threshold estimate. Nevertheless, Balcombe, Bailey & Brooks (2007) and Balcombe & Rapsomanikis (2008) base their Bayesian estimators on noninformative priors. Chen (1998) suggests a Bayesian estimator based on a normal prior with known hyper-parameters for the $\hat{\boldsymbol{\beta}}_k$ and a uniform prior for the threshold parameter. However, she designs the latter to assign zero probability to threshold values that do not leave a minimum number of observations in each regime, which is equivalent to assuming an arbitrary trimming parameter.

4.2.3 Regularized Bayesian estimator

Given the deficiencies of profile likelihood and Bayesian estimation with noninformative priors, we explore the properties of an alternative threshold estimator (Greb, Krivobokova, Munk & von Cramon-Taubadel, 2011) in the context of TVCEMs. This regularized Bayesian estimator (RBE) was developed for univariate generalized threshold regression models with one threshold. The idea of the estimator is to penalize differences between regimes so as to keep these differences reasonably small when the data contain little information. The strength of this regularizing penalty is fundamental to the estimator. It is determined in a data-driven manner employing the so-called empirical Bayes paradigm. The estimator is developed in a Bayesian framework and the penalization is a result of the choice of priors. As an important consequence of the regularization, the posterior density is well-defined on the entire threshold parameter space Ψ. Hence, there is no need to choose a trimming parameter and no risk of excluding the true threshold from Ψ. In the setting of generalized threshold regression models, the RBE outperforms commonly used estimators, especially when the threshold leaves only few observations in one of the regimes or there is little difference in coefficients between regimes.

Extension of the theory detailed in Greb, Krivobokova, Munk & von Cramon-Taubadel (2011) to the TVECM with two thresholds in equation (4.1) is straightforward. It involves reparametrizing the model in equation (4.3),

$$\begin{aligned}
\Delta \boldsymbol{p} &= (\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)\boldsymbol{\beta}_1 + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_2)\boldsymbol{\beta}_2 + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)\boldsymbol{\beta}_3 + \boldsymbol{\varepsilon} \\
&= (\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) + \{(\boldsymbol{I}_2 \otimes \boldsymbol{X}_1) + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_2) + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)\}\,\boldsymbol{\beta}_2 \\
&\quad + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)(\boldsymbol{\beta}_3 - \boldsymbol{\beta}_2) + \boldsymbol{\varepsilon} \\
&= (\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2) + (\boldsymbol{I}_2 \otimes \boldsymbol{X})\boldsymbol{\beta}_2 + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)(\boldsymbol{\beta}_3 - \boldsymbol{\beta}_2) + \boldsymbol{\varepsilon} \\
&= (\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)\boldsymbol{\delta}_1 + (\boldsymbol{I}_2 \otimes \boldsymbol{X})\boldsymbol{\beta}_2 + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)\boldsymbol{\delta}_3 + \boldsymbol{\varepsilon},
\end{aligned} \tag{4.6}$$

and specifying a noninformative constant prior for $\boldsymbol{\beta}_2$ and normal priors for $\boldsymbol{\delta}_i$, $\boldsymbol{\delta}_i \sim \mathcal{N}(0, \sigma^2_{\delta_i}\boldsymbol{I}_{2d})$, $i = 1, 3$. The empirical Bayes strategy amounts to replacing σ^2, $\sigma^2_{\delta_1}$, and $\sigma^2_{\delta_3}$ by their maximum likelihood estimates $\tilde{\sigma}^2$, $\tilde{\sigma}^2_{\delta_1}$, and $\tilde{\sigma}^2_{\delta_3}$. As illustrated in the appendix, this

yields a log posterior density

$$p(\psi|\Delta\boldsymbol{p},\boldsymbol{X}) \propto -\frac{1}{2}\Bigg\{(2n-2d)\log\tilde{\sigma}^2 + \log|\boldsymbol{V}| + \log|\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}| + \frac{1}{\tilde{\sigma}^2}(\Delta\boldsymbol{p}-\boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)^T\boldsymbol{V}^{-1}(\Delta\boldsymbol{p}-\boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)\Bigg\} \tag{4.7}$$

with $\tilde{\boldsymbol{\beta}}_2 = (\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\Delta\boldsymbol{p}$ and $\boldsymbol{V} = \boldsymbol{I}_{2n} + \tilde{\sigma}^2_{\delta_1}/\tilde{\sigma}^2\boldsymbol{Z}_1\boldsymbol{Z}_1^T + \tilde{\sigma}^2_{\delta_3}/\tilde{\sigma}^2\boldsymbol{Z}_3\boldsymbol{Z}_3^T$ for $\boldsymbol{Z} = \boldsymbol{I}_2 \otimes \boldsymbol{X}$, $\boldsymbol{Z}_1 = \boldsymbol{I}_2 \otimes \boldsymbol{X}_1$ and $\boldsymbol{Z}_3 = \boldsymbol{I}_2 \otimes \boldsymbol{X}_3$. A comparison of $\ell_p(\psi)$ in equation (4.5) with $p(\psi|\Delta\boldsymbol{p},\boldsymbol{X})$ in equation (4.7) shows that unlike the former, the latter does not depend on $\hat{\boldsymbol{\beta}}_k$, $k = 1,2,3$, which are not well-defined unless ψ leaves a minimum of d observations in each regime. Accordingly, $p(\psi|\Delta\boldsymbol{p},\boldsymbol{X})$ is defined on the entire threshold parameter space $\Psi = \big\{(\psi_1,\psi_2)$ such that $\min(\boldsymbol{\gamma}^T\boldsymbol{p}_t) < \psi_1 < \psi_2 < \max(\boldsymbol{\gamma}^T\boldsymbol{p}_t)\big\}$.

The regularized Bayesian threshold estimator $\hat{\psi}_{rB} = \left(\hat{\psi}_{1rB},\hat{\psi}_{2rB}\right)$ is computed as the posterior median

$$\int_{\min(\boldsymbol{\gamma}^T\boldsymbol{p}_t)}^{\hat{\psi}_{irB}} p(\psi_i|\Delta\boldsymbol{p},\boldsymbol{X})d\psi_i = 0.5, \quad i = 1,2, \tag{4.8}$$

assuming a prior $p(\psi|\boldsymbol{X}) \propto I(\psi \in \Psi)$ for ψ. Here, $p(\psi_i|\Delta\boldsymbol{p},\boldsymbol{X})$ denotes the i-th threshold's marginal posterior density. We choose the median of the posterior distribution because it is more robust than the mode and yields more reliable results than the mean when this density is skewed (which tends to be the case when the true threshold is close to the boundary of the threshold parameter space Ψ).

4.2.4 Computational aspects

Any two threshold values which produce the same allocation of observations into regimes produce identical values of the profile likelihood function $\mathcal{L}_p(\psi)$. Hence, $\mathcal{L}_p(\psi)$ is a step function and not differentiable. The same holds for the posterior density $p(\psi|\Delta\boldsymbol{p},\boldsymbol{X})$. However, searching a grid that includes all of the observed error-correction terms yields the exact maximum of $\mathcal{L}_p(\psi)$ and also makes it possible to calculate the precise value of the integral of $p(\psi|\Delta\boldsymbol{p},\boldsymbol{X})$. Obviously, this can be computationally intensive in

large samples. Hence, in practice, profile likelihood functions are often evaluated on a coarser grid. For example, some authors (e.g. Goodwin & Piggott, 2001) employ evenly spaced grids that divide the threshold parameter space Ψ into a chosen number of equal steps and that therefore do not necessarily include each of the observed error-correction terms. In the absence of local maxima and large jumps between subsequent steps, such a simplified grid will provide a reasonable approximation of the maximum/integral. However, when the dimension of $\hat{\boldsymbol{\beta}}_k$ is high or the thresholds leave few observations in one of the regimes, $\mathcal{L}_p(\psi)$ and $p(\psi|\Delta\boldsymbol{p}, \boldsymbol{X})$ tend to be jagged and display several local maxima. In such a case, even a fairly dense grid can produce a poor approximation of the true maximum and, consequently, poor estimates, if it does not include all function values. We demonstrate this effect of an inappropriate grid choice in one of the empirical applications below.

Computation of the RBE is greatly simplified by taking advantage of functions for mixed models available in statistical software packages. Again, we refer to Greb, Krivobokova, Munk & von Cramon-Taubadel (2011) for details. R code for calculating RB estimates (for the general TVECM in equation (4.1) and for restricted models such as the BAND-TVECM) is available from the authors.

4.3 Simulations

In a simulation study, we generate data using model (4.1) with the following parameters: thresholds are set to $\psi_1 = -4$ and $\psi_2 = 4$; adjustment coefficients $\boldsymbol{\rho}_1 = \boldsymbol{\rho}_3 = (-0.25, 0)^T$ and $\boldsymbol{\rho}_2 = (0, 0)^T$; intercepts $\boldsymbol{\theta}_1 = (-1, 0)^T$, $\boldsymbol{\theta}_2 = (0, 0)^T$, $\boldsymbol{\theta}_3 = (1, 0)^T$; and $\boldsymbol{\Theta}_{11} = \boldsymbol{\Theta}_{31} = \left(\begin{smallmatrix} 0.2 & 0.2 \\ 0 & 0 \end{smallmatrix}\right)$, $\boldsymbol{\Theta}_{21} = \left(\begin{smallmatrix} 0 & 0 \\ 0 & 0 \end{smallmatrix}\right)$. The cointegrating vector $\boldsymbol{\gamma} = (1, -1)^T$ is assumed to be known; this implies an error correction term $\boldsymbol{\gamma}^T \boldsymbol{p}_t = p_{1,t} - p_{2,t}$ that is simply equal to the difference between p_1 and p_2. Errors are normally distributed, $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_2)$ with $\sigma^2 = 1$. The length of the series is $n = 200$. We have selected the parameters to take on values that are plausible in real data applications. They imply that in most simulations about one half of the data belongs to the inner and one fourth to each of the outer regimes.

We estimate thresholds by applying the profile likelihood and RB estimators to a Monte Carlo sample of 300 replications of the data generating process defined above. We show profile likelihood estimates for three different trimming parameters. These are, first,

the least restrictive trimming parameter possible ($d = 2M + 2$, which ensures that each regime contains at least exactly the minimum number of observations necessary to estimate all model parameters), second, 15%, and third, 20% of the sample size. Results are summarized in figures 4.1 and 4.2 together with table 4.1. The RBE clearly outperforms the profile likelihood estimator. We observe a considerable reduction in both bias and variance and, consequently, mean squared error. In contrast to the profile likelihood estimates, the RB estimates are not drawn towards zero. The histograms show that the distribution of the RB estimates is also less skewed. Further simulations (including restricted models) confirm these findings. Altogether, the results indicate that the RBE is not only superior for generalized threshold regression models, but also for TVECMs specifically.

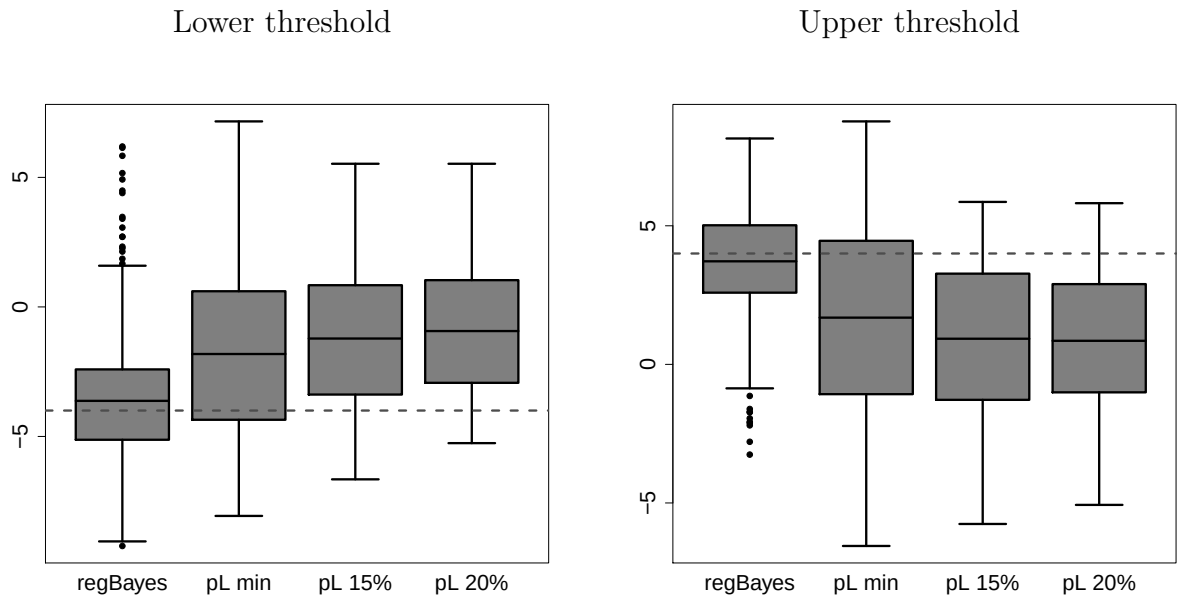

Figure 4.1: Simulation results (boxplots). Note: The horizontal dashed gray line indicates the true threshold. The dark lines in the shaded boxes are the respective sample means. "pL min", "pL 15%", and "pL 20%" denote profile likelihood estimates with trimming parameters equal to the smallest possible value ($d = 2M + 2$), 15% of the sample size, and 20% of the sample size, respectively.

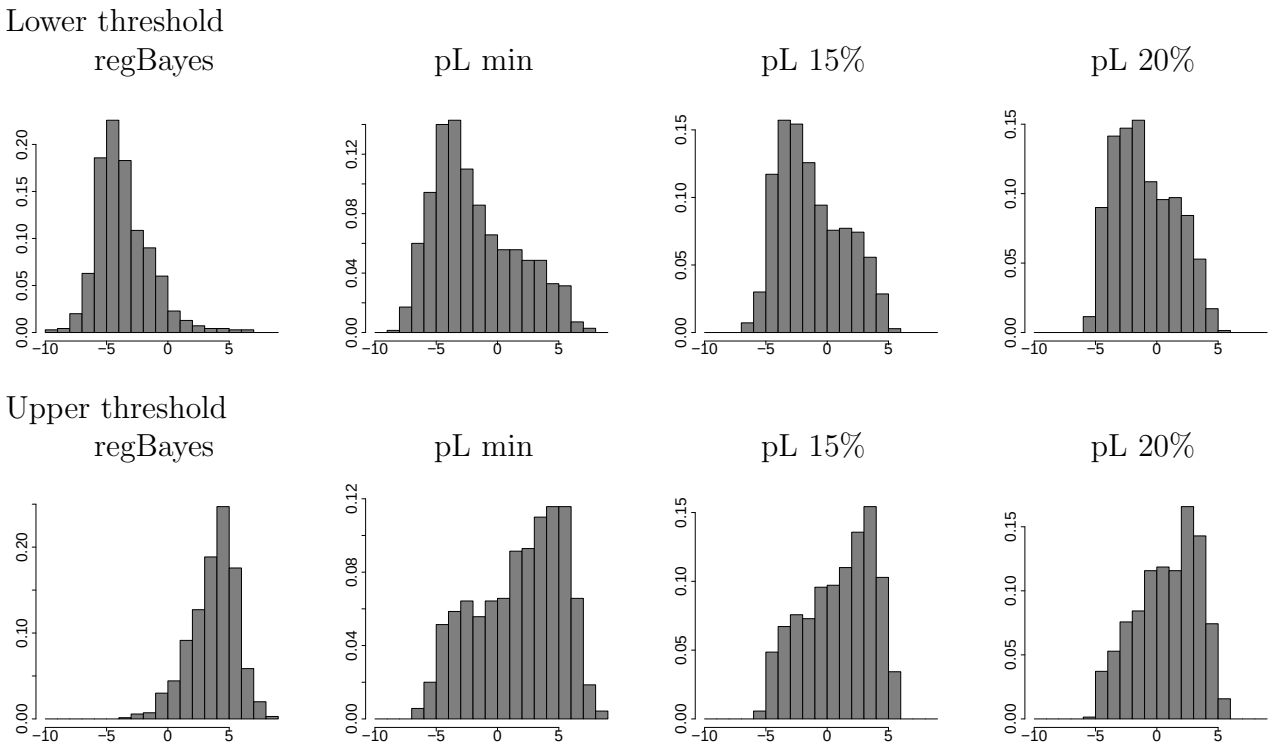

Figure 4.2: Simulation results (histograms). Note: "pL min", "pL 15%", and "pL 20%" denote profile likelihood estimates with trimming parameters equal to the smallest possible value ($d = 2M + 2$), 15% of the sample size, and 20% of the sample size, respectively.

	Regularized Bayesian estimator		Profile likelihood estimator					
	lower threshold	upper threshold	lower threshold			upper threshold		
			min	15%	20 %	min	15%	20 %
true	-4	4	-4	-4	-4	4	4	4
mean	-3.63	3.72	-1.82	-1.22	-0.93	1.69	0.92	0.85
	(2.23)	(1.88)	(3.40)	(2.67)	(2.45)	(3.50)	(2.79)	(2.52)
MSE	5.10	3.62	16.31	14.86	15.40	17.57	17.21	16.24

Table 4.1: Simulation Results. Note: Standard errors are reported in parentheses below the mean. "min", "15%", and "20%" denote trimming parameters equal to the smallest possible value ($d = 2M + 2$), 15% of the sample size, and 20% of the sample size, respectively.

4.4 Empirical applications

4.4.1 Goodwin & Piggott (2001) revisited

In the first empirical application, we revisit Goodwin & Piggott's (2001) seminal analysis of spatial price transmission with TVECMs. We apply the RBE to their dataset and compare the results with their profile likelihood estimates. Goodwin & Piggott (2001) explore daily corn and soybean prices at important North Carolina terminal markets (figure 4.3). These are Williamston, Candor, Cofield, and Kinston for corn, and Fayetteville, Raleigh, Greenville, and Kinston for soybeans. Observations range from 2 January 1992 until 4 March 1999. For each commodity, Goodwin & Piggott (2001) evaluate pairs consisting of a central market – Williamston for corn and Fayetteville for soybeans – and each of the other markets in turn. They estimate the TVECM in equation (4.1) with logarithmic prices by maximizing the profile likelihood function $\mathcal{L}_p(\psi)$ under the assumption of Gaussian errors (or, equivalently, minimizing the sum of squared errors). In accordance with spatial equilibrium theory they assume that $\psi_1 \leq 0$ and $\psi_2 \geq 0$ and search for the maximum of $\mathcal{L}_p(\psi)$ among those ψ that meet this condition. To obtain comparable results, we also incorporate this information in the RBE; we specify a prior on ψ which is zero for any ψ such that $\psi_1 > 0$ or $\psi_2 < 0$, and uniform otherwise. Goodwin & Piggott (2001) evaluate the estimating function at 100 equally spaced grid points for each threshold. In contrast, we compute the RB estimates exactly, that is, the posterior density is evaluated on a complete grid (that includes all observed values of the error-correction term).

We report RB estimates together with Goodwin & Piggott's (2001) original profile likelihood estimates in table 4.2. It is evident that relative to the profile likelihood estimates, the RB estimates for both thresholds tend to be of greater magnitude. This is confirmed by the results reported in the last three columns of the same table, which show (in square brackets) for each pair of markets the number of observations assigned to each of the three regimes by the respective estimation method. Since the thresholds estimated by the regularized Bayesian method are farther from zero, this method assigns correspondingly less (more) observations to the outer (inner) regimes. The only exceptions are found in regime 3 for Cofield – Williamston (corn) and Greenville – Fayetteville (soybeans).

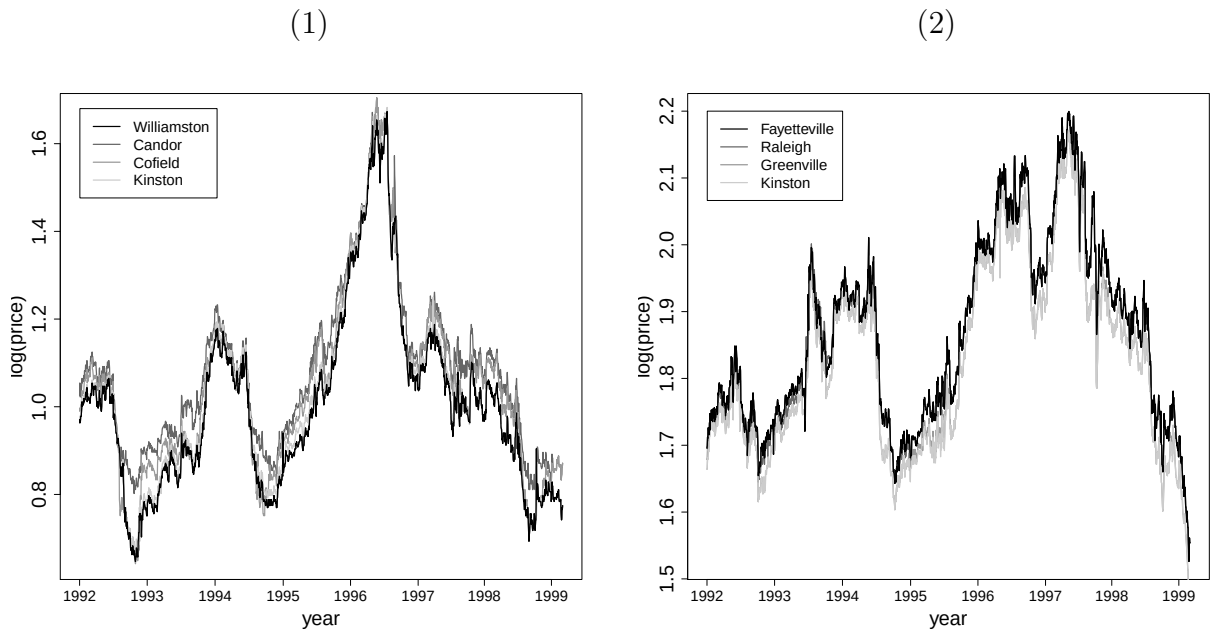

Figure 4.3: Daily corn (1) and soybean (2) (log)prices at four North Carolina terminal markets. Source: Goodwin & Piggott (2001), who kindly made this data available.

In the last three columns of table 4.2 we also illustrate the effect of using a complete rather than a uniform grid on the allocation of observations into regimes. For the profile likelihood results, the first number in square brackets is the number of observations allocated to the respective regime when Goodwin and Piggott's uniform grid is employed, and the second number is the corresponding number of observations when a complete grid is employed. If both grids lead to similar estimates of the thresholds ψ_1 and ψ_2, then they will also lead to similar allocations of observations into regimes. While this is the case for some market pairs, the cases of Raleigh – Fayetteville and Greenville – Fayetteville in particular illustrate that a complete grid is necessary to ensure correct identification of the global maximum of the likelihood function.

What are the economic implications of these results? Several points can be made. First, the fact that the regularized Bayesian threshold estimates are father apart can be interpreted as evidence of greater market integration. It implies that more observations are in the inner "band of inaction", and correspondingly fewer are in the outer bands

where spatial equilibrium is violated, triggering trade and price adjustments.[1] However, if thresholds are estimates of the transaction costs of trade between two locations, then the RBE suggests that these costs are higher than indicated by the profile likelihood estimates (see O'Connell & Wei, 2002). Hence, the regularized Bayesian threshold estimates suggest that the markets in question are more integrated in the sense that they display fewer violations of spatial equilibrium, but also that they are separated by higher transactions costs which must be overcome before arbitrage becomes profitable.

Second, market integration is reflected not only in how often violations of spatial equilibrium occur, but also in the speed with which such violations are corrected. According to the two-market spatial equilibrium theory discussed above, the outer regimes should be characterized by more rapid error correction than the inner regime, within which prices can move independently and no error correction is expected. The profile likelihood and regularized Bayesian estimates of the adjustment parameters presented in table 4.2 generally confirm this expectation. However, the profile likelihood estimates are surprising in two cases. First, for corn in Kinston – Williamston, the estimated adjustment parameters in regime 1 are both greater than one in magnitude, which is implausible as it would suggest that errors are amplified and not corrected. The total adjustment implied by these two parameters (-0.166 in the third to last column of table 4.2) is negative, which confirms that this regime is not consistent with error correction and cointegration. This may be a reflection of the "weaker" evidence for cointegration between corn prices in Kinston and Williamston reported by Goodwin & Piggott (2001, page 306). Second, total adjustment in the inner regime (regime 2) in the case of corn in Candor and Williamston (-0.015 in the second to last column of table 4.2) is also negative, which suggests that price differences in this regime will also be amplified rather than corrected. This result is incompatible with spatial equilibrium theory, which does not predict that prices will be driven apart in the inner regime. However, it is not incompatible with market integration between Candor and Williamston overall, because outer bands for this pair of markets are characterized by error correction that will drive prices back towards equilibrium whenever they leave the inner regime.

The regularized Bayesian estimates of the adjustment parameters presented in table 4.2

[1] The only major exception to this pattern is Fayetteville – Greenville, for which the inner regime is more than twice as wide according to profile likelihood as it is according to the RBE. We discuss this exception below.

do not display any anomalies of this nature and therefore provide stronger evidence of market integration. However, for many market pairs the adjustment coefficients in the outer bands are smaller according to the RBE compared with profile likelihood. For corn in Candor and Williamston, for example, total adjustment amounts to 0.130 in regime 1 and 0.120 in regime 3 according to profile likelihood, compared with 0.043 in both regimes according to the RBE. Hence, while 13% (12%) of any difference between the two prices is corrected per period in regime 1 (3) according to the profile likelihood results, only 4.3% is corrected in either regime according to the RBE.[2] Hence, the RBE results point to slower transmission of price shocks than the profile likelihood results.

One other aspect of the results in table 4.2 deserves mention. For one of the corn market pairs (Candor – Williamston) and all three of the soybean market pairs, the regularized Bayesian estimates of the adjustment parameters are very similar or identical across all three regimes. These results might indicate that the two-threshold, three-regime model of price transmission is misspecified. As data on trade in corn and soybeans between the markets in question are not available, it is not clear whether a model with two thresholds, which includes a regime for trade from market 1 to market 2, but also a regime for trade in the opposite direction, is correctly specified. If trade only flows in one direction, then a model with one threshold and two regimes would be more appropriate. Sephton (2003), who also revisits the Goodwin & Piggott (2001) data, finds that a one-threshold model is indicated for four of the six pairs, and that the pairs Raleigh-Fayetteville and Greenville-Fayetteville display little evidence of any threshold effects whatsoever. Our regularized Bayesian estimates of very similar or identical adjustment coefficients across regimes for some market pairs appear to corroborate Sephton's (2003) finding.

[2] While most of the adjustment coefficients reported in table 4.2 are quite small, regardless of the method used to estimate them, they are estimated using daily prices. Hence, in most cases the adjustment half-life is in the range of 1 – 2 weeks.

Est.	Dep.var.	$\boldsymbol{\rho}_1$	$\sigma(\boldsymbol{\rho}_1)$	ψ_1	$\boldsymbol{\rho}_2$	$\sigma(\boldsymbol{\rho}_2)$	ψ_2	$\boldsymbol{\rho}_3$	$\sigma(\boldsymbol{\rho}_3)$	Total($\boldsymbol{\rho}_1$) [#obs.]	Total($\boldsymbol{\rho}_2$) [#obs.]	Total($\boldsymbol{\rho}_3$) [#obs.]
Corn: Candor-Williamston												
PL	$\Delta \boldsymbol{p}^{CAN}$	0.003	(0.061)	-0.025	0.006	(0.053)	0.003	-0.030	(0.040)	0.130	-0.015	0.120
	$\Delta \boldsymbol{p}^{WIL}$	**0.133**	(0.061)		-0.009	(0.053)		**0.090**	(0.040)	[295/298]	[761/670]	[716/797]
RBE	$\Delta \boldsymbol{p}^{CAN}$	0.008	(0.019)	-0.069	0.002	(0.013)	0.030	0.002	(0.013)	0.043	0.043	0.043
	$\Delta \boldsymbol{p}^{WIL}$	**0.051**	(0.019)	(0.011)	**0.045**	(0.013)	(0.016)	**0.045**	(0.013)	[12]	[1545]	[208]
Corn: Cofield-Williamston												
PL	$\Delta \boldsymbol{p}^{COF}$	-0.083	(0.063)	-0.057	**0.028**	(0.012)	0.065	-0.351	(0.558)	0.136	0.007	1.074
	$\Delta \boldsymbol{p}^{WIL}$	0.053	(0.063)		**0.035**	(0.012)		0.723	(0.558)	[69/68]	[1669/1686]	[35/11]
RBE	$\Delta \boldsymbol{p}^{COF}$	0.007	(0.011)	-0.056	**0.024**	(0.013)	0.034	**0.020**	(0.012)	0.045	0.022	0.024
	$\Delta \boldsymbol{p}^{WIL}$	**0.052**	(0.011)	(0.026)	**0.046**	(0.013)	(0.021)	**0.044**	(0.012)	[73]	[1409]	[283]
Corn: Kinston-Williamston												
PL	$\Delta \boldsymbol{p}^{KIN}$	**-2.619**	(0.773)	-0.013	**0.162**	(0.053)	0.0190	0.954	(0.686)	-0.166	0.028	0.204
	$\Delta \boldsymbol{p}^{WIL}$	**-2.785**	(0.773)		**0.190**	(0.053)		**1.158**	(0.686)	[249/197]	[1469/1558]	[55/10]
RBE	$\Delta \boldsymbol{p}^{KIN}$	-0.011	(0.273)	-0.020	**0.092**	(0.038)	0.0192	**0.087**	(0.041)	0.384	0.035	0.039
	$\Delta \boldsymbol{p}^{WIL}$	0.373	(0.273)	(0.004)	**0.127**	(0.038)	(0.005)	**0.126**	(0.041)	[7]	[1753]	[5]
Soybeans: Raleigh-Fayetteville												
PL	$\Delta \boldsymbol{p}^{RAL}$	-0.352	(0.277)	-0.001	-0.091	(0.108)	0.010	0.257	(0.465)	0.417	-0.002	0.095
	$\Delta \boldsymbol{p}^{FAY}$	0.065	(0.277)		-0.093	(0.108)		0.352	(0.465)	[166/492]	[1559/1226]	[47/47]

RBE	$\Delta \boldsymbol{p}^{RAL}$	-0.090	(0.063)	-0.022	-0.090	(0.063)	0.014	-0.090	(0.063)	0.128	0.123	0.123
	$\Delta \boldsymbol{p}^{FAY}$	0.038	(0.063)	(0.002)	0.033	(0.063)	(0.003)	0.033	(0.063)	[11]	[1714]	[40]
Soybeans: Greenville-Fayetteville												
PL	$\Delta \boldsymbol{p}^{GRE}$	-0.040	(0.048)	-0.009	0.042	(0.047)	0.012	0.083	(0.587)	0.064	0.029	0.370
	$\Delta \boldsymbol{p}^{FAY}$	0.024	(0.048)		0.071	(0.047)		0.453	(0.587)	[410/435]	[1292/1026]	[70/304]
RBE	$\Delta \boldsymbol{p}^{GRE}$	0.014	(0.022)	-0.008	0.014	(0.022)	0.006	0.015	(0.022)	0.053	0.053	0.053
	$\Delta \boldsymbol{p}^{FAY}$	**0.067**	(0.022)	(0.026)	**0.067**	(0.022)	(0.011)	**0.068**	(0.022)	[462]	[558]	[745]
Soybeans: Kinston-Fayetteville												
PL	$\Delta \boldsymbol{p}^{KIN}$	**-0.071**	(0.043)	-0.006	0.029	(0.182)	0.007	-0.104	(0.093)	0.094	0.115	0.200
	$\Delta \boldsymbol{p}^{FAY}$	0.023	(0.043)		0.144	(0.182)		0.096	(0.093)	[544/550]	[508/502]	[721/713]
RBE	$\Delta \boldsymbol{p}^{KIN}$	-0.008	(0.021)	-0.097	-0.003	(0.022)	0.021	-0.003	(0.022)	0.070	0.064	0.064
	$\Delta \boldsymbol{p}^{FAY}$	**0.062**	(0.021)	(0.029)	**0.061**	(0.022)	(0.010)	**0.061**	(0.022)	[9]	[1691]	[65]

Table 4.2: Estimates for the Data in Figure 4.3 – TVECM with three Regimes. Notes:
- PL is the profile likelihood estimator; RBE is the regularized Bayesian estimator.
- Standard errors of the estimated adjustment parameters ($\boldsymbol{\rho}_k$) are provided in brackets. These must be interpreted with care because they are computed without accounting for the variability of the threshold estimate. Estimates that are significant at the 10% level are in **bold**. Standard errors for regularized Bayesian threshold estimates (in brackets below the estimate) are calculated in the customary Bayesian manner as their posterior standard deviation. To the best of our knowledge, it is an open question how to compute standard errors for PL threshold estimates in TVECMs.
- The error correction term is normalized so that the first adjustment parameter in each pair is expected to be negative, and the second positive. For example, for soybeans, the market pair Kinston – Fayetteville, and the profile likelihood (PL) estimator, the $\boldsymbol{\rho}_1$-values (-0.071 and 0.023) have the expected signs.
- Total($\boldsymbol{\rho}_k$) measures the total error-correction of price differences in regime k as the sum of the second adjustment parameter in each pair minus the first. For example, for soybeans, the market pair Kinston – Fayetteville, and the profile likelihood (PL) estimator, Total($\boldsymbol{\rho}_1$) = 0.094 = 0.023-(-0.071).
- The number in square brackets below Total($\boldsymbol{\rho}_k$) is the estimated number of observations in regime k. For PL, the first number corresponds to Goodwin & Piggot's estimates, the second to PL estimates based on a complete grid.

4.4.2 Price transmission between German and Spanish pork prices

As a second empirical application, we analyze transmission between German and Spanish pork prices. The analysis is carried out using the data presented in figure 4.4, which are average weekly prices of grade E pig carcasses for Germany and Spain in Euro per 100 kg between May 21, 1989 and October 17, 2010 (1091 observations). We specify a TVECM with three regimes,

$$\Delta \boldsymbol{p}_t = \begin{cases} \boldsymbol{\rho}_1 \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} + \boldsymbol{\theta}_1 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{1m} \Delta \boldsymbol{p}_{t-m} + \boldsymbol{\varepsilon}_t & , \quad \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} < \psi_1 \quad \text{(Regime 1)} \\ \boldsymbol{\rho}_2 \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} + \boldsymbol{\theta}_2 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{2m} \Delta \boldsymbol{p}_{t-m} + \boldsymbol{\varepsilon}_t & , \quad \psi_1 \leq \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} \leq \psi_2 \quad \text{(Regime 2)} \\ \boldsymbol{\rho}_3 \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} + \boldsymbol{\theta}_3 + \sum_{m=1}^{M} \boldsymbol{\Theta}_{3m} \Delta \boldsymbol{p}_{t-m} + \boldsymbol{\varepsilon}_t & , \quad \psi_2 < \boldsymbol{\gamma}^T \boldsymbol{p}_{t-1} \quad \text{(Regime 3).} \end{cases} \tag{4.9}$$

with $\Delta \boldsymbol{p}_t = \left(\Delta \boldsymbol{p}_t^{Germany}, \Delta \boldsymbol{p}_t^{Spain} \right)^T$ and $M = 3$. We apply profile likelihood and the RBE with the error correction term $\boldsymbol{\gamma}^T \boldsymbol{p}_{t-1}$ defined as the difference between the Spanish and the German prices, $\boldsymbol{\gamma}^T \boldsymbol{p}_t = \Delta \boldsymbol{p}_t^{Germany} - \Delta \boldsymbol{p}_t^{Spain}$.

We plot the profile likelihood for the upper threshold (ψ_2) in figure 4.5. To generate this figure, the lower threshold (ψ_1) has been fixed at its profile likelihood estimate. We see that the profile likelihood reaches its maximum at the boundary of the range defined by the smallest possible trimming parameter (i.e. the requirement that each regime contains at least one observation per parameter to be estimated). Hence, any more restrictive trimming parameter (such as requiring that each regime contain at least 2.5 or 5% of all observations) strongly influences the profile likelihood estimate (see figure 4.5), rendering it arbitrary and unreliable. Compared with an estimate $\hat{\psi}_2 = 26.07$ for the least restrictive trimming parameter, requiring 2.5% (5%) of the observations to fall into each regime produces the estimate $\hat{\psi}_2 = 21.83$ ($\hat{\psi}_2 = 14.01$).

The RBE does not require an arbitrary trimming parameter. It produces threshold estimates $(-36.41, 34.76)$ that are considerably larger in magnitude than the profile likelihood estimates $(-27.80, 26.07)$. Furthermore, the RBE produces estimates of the adjustment parameters that are more plausible than their profile likelihood counterparts (table 4.3). In regime 1, where the difference between the German and Spanish prices

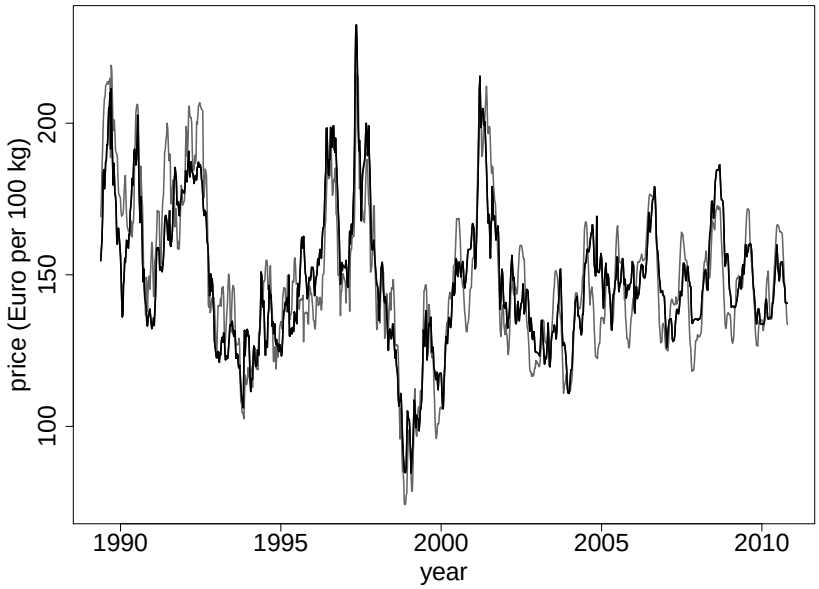

Figure 4.4: Weekly prices for grade E pig carcasses in Germany and Spain (Euro per 100 kg). Source: European Commission: `http://ec.europa.eu/agriculture/markets/pig/porcs.pdf`.

is less than the lower threshold value, the profile likelihood estimate of the adjustment parameter for the Spanish price is significant and of relatively high magnitude (-0.665), but with an implausible sign. Both magnitude and sign are implausible for the corresponding parameter estimate in regime 3 (-1.193), where the difference between the German and the Spanish prices exceeds the upper threshold. The corresponding estimated adjustment parameters for the German price in regimes 1 and 3 (-0.198 and -0.334) have the expected negative signs, but they are insignificant. Altogether, the total adjustments for regimes 1 and 3 are negative according to the profile likelihood method (see the third-to-last and last colums of table 4.3). Hence, the profile likelihood estimates suggest that there is no mechanism that returns German and Spanish prices to their long run equilibrium when shocks drive them apart. In comparison, the regularized Bayesian estimates of the adjustment parameters make considerably more sense. All of the regularized Bayesian estimates that are significant, have the expected sign, and together they indicate that when the difference between the German and the Spanish prices exceeds one of the thresholds, adjustments are triggered that return these prices to their long run equilibrium (total adjustment equals 0.318 in regime 1 and 0.348 in regime 3).

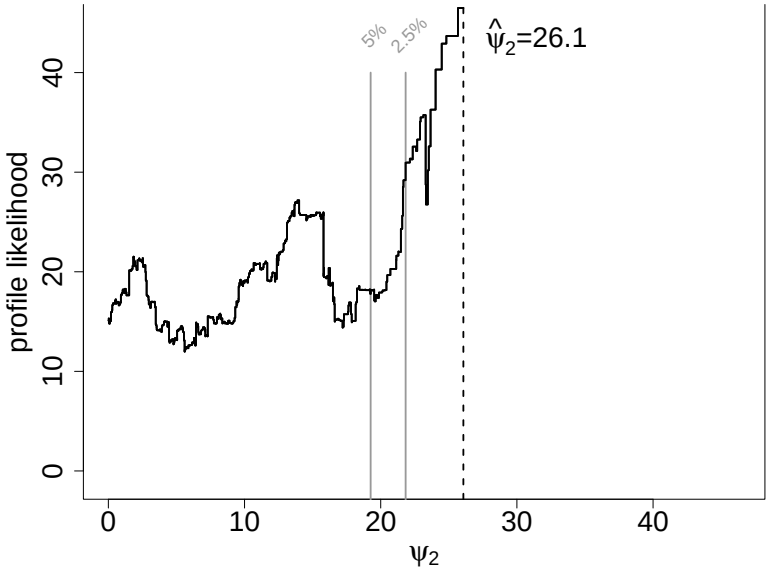

Figure 4.5: Profile likelihood function for the upper threshold, ψ_2, estimated with the pig price data in figure 4.4. The dashed vertical line indicates the profile likelihood estimate for the upper threshold, $\hat{\psi}_2$, estimated using the least restrictive possible trimming parameter. Solid grey lines indicate how the threshold parameter space is restricted when 2.5% (5%) of the observations are required to fall into each regime. The lower threshold is fixed at its profile likelihood estimate, $\hat{\psi}_1 = -27.8$.

In summary, the empirical applications illustrate the advantages of the RBE in the context of spatial price transmission analysis. The RBE does not depend on a trimming parameter that arbitrarily influences the profile likelihood results in the application with Spanish and German pork prices. Furthermore, in both applications the RB estimates of the adjustment parameters are more plausible. In the application with the Goodwin & Piggott (2001) data they appear to confirm Sephton's (2003) finding that the two-threshold TVECM is misspecified. In the application with Spanish and German pork prices they are, unlike the profile likelihood estimates, consistent with spatial equilibrium theory and price transmission between the markets in question.

Est.	Dep.var.	$\boldsymbol{\rho}_1$	$\sigma(\boldsymbol{\rho}_1)$	ψ_1	$\boldsymbol{\rho}_2$	$\sigma(\boldsymbol{\rho}_2)$	ψ_2	$\boldsymbol{\rho}_3$	$\sigma(\boldsymbol{\rho}_3)$	Total($\boldsymbol{\rho}_1$) [#obs.]	Total($\boldsymbol{\rho}_2$) [#obs.]	Total($\boldsymbol{\rho}_3$) [#obs.]
PL	$\Delta \boldsymbol{p}^{Germany}$	-0.198	(0.354)	-27.8	**-0.028**	(0.012)	26.1	-0.334	(1.498)	-0.467	0.080	-0.859
	$\Delta \boldsymbol{p}^{Spain}$	**-0.665**	(0.354)		**0.052**	(0.012)		-1.193	(1.498)	[21]	[1058]	[8]
RBE	$\Delta \boldsymbol{p}^{Germany}$	**-0.286**	(0.104)	-36.4	**-0.029**	(0.011)	34.8	**-0.355**	(0.116)	0.318	0.092	0.348
	$\Delta \boldsymbol{p}^{Spain}$	0.031	(0.104)	(28.3)	**0.063**	(0.011)	(149.9)	-0.007	(0.116)	[2]	[1084]	[1]

Table 4.3: Estimates for the Data in Figure 4.3 – TVECM with three Regimes. Note: The notes below table 4.2 apply.

4.5 Conclusions

We discuss the estimation of TVCEMs in spatial price transmission analysis. We point out shortcomings of the common (profile likelihood) estimation procedure and emphasize the relevance of these problems for applied price transmission studies. As an alternative, we suggest employing a regularized Bayesian estimator (Greb, Krivobokova, Munk & von Cramon-Taubadel, 2011), and we demonstrate this estimator's superior performance in a simulation study. Revisiting the empirical analysis in Goodwin & Piggott's (2001) influential paper on TVECMs in price transmission analysis, we find that the RB estimates are free of several anomalies that characterise the profile likelihood estimates, and appear to corroborate Sephton's (2003) finding that the two-threshold, three-regime TVECM is misspecified for the data in question. A second application, with German and Spanish pork prices, confirms the advantages of the RBE in spatial price transmission modeling, producing results that are more consistent with the theory of spatial equilibrium than the corresponding profile likelihood results. Future work could move beyond the pairwise consideration of markets to study multivariate sets of prices and the more complex multiple-threshold relationships that exist between them. Another extension would be to investigate time-varying thresholds, since especially for longer time-series the assumption of constant transaction costs is questionable.

Acknowledgements

We acknowledge the support of the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the Institutional Strategy of the University of Göttingen. We would like to thank Barry Goodwin and Nicholas Piggott for sharing their dataset with us.

4.6 Appendix (technical details)

Our aim is to compute the posterior density $p\left(\psi|\Delta p, \boldsymbol{X}\right)$ for the model

$$\Delta \boldsymbol{p} = (\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)\boldsymbol{\delta}_1 + (\boldsymbol{I}_2 \otimes \boldsymbol{X})\boldsymbol{\beta}_2 + (\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)\boldsymbol{\delta}_3 + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I}_{2n})$$

with a normal prior $\boldsymbol{\delta}_1 \sim \mathcal{N}(0, \sigma^2_{\delta_1}\boldsymbol{I}_{2d})$, where $d = 2M + 2$ with M the number of lags included in the model; a uniform prior $\boldsymbol{\beta}_2 \sim U\left(\mathbb{R}^{2d}\right)$; a normal prior $\boldsymbol{\delta}_3 \sim \mathcal{N}(0, \sigma^2_{\delta_3}\boldsymbol{I}_{2d})$; and a uniform prior $\psi \sim U(\psi \in \Psi)$.

To this end, we first calculate $p\left(\Delta\boldsymbol{p}|\psi, \boldsymbol{X}\right)$, since

$$p\left(\psi|\Delta\boldsymbol{p}, \boldsymbol{X}\right) = p\left(\Delta\boldsymbol{p}|\psi, \boldsymbol{X}\right) p(\psi|\boldsymbol{X}) \Big/ p(\Delta\boldsymbol{p}|\boldsymbol{X}) \propto p\left(\Delta\boldsymbol{p}|\psi, \boldsymbol{X}\right)$$

given a constant prior $p(\psi|\boldsymbol{X})$. Employing an empirical Bayes approach, it suffices to compute $p\left(\Delta\boldsymbol{p}|\psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3}\right)$: parameters σ^2, $\sigma^2_{\delta_1}$, and $\sigma^2_{\delta_3}$ are replaced by their maximum likelihood estimates $\tilde{\sigma}^2$, $\tilde{\sigma}^2_{\delta_1}$, and $\tilde{\sigma}^2_{\delta_3}$. Given our specification of priors,

$$\begin{aligned} p\left(\Delta\boldsymbol{p}|\psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3}\right) &= \int p\left(\Delta\boldsymbol{p}, \boldsymbol{\beta}_2|\psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3}\right) d\boldsymbol{\beta}_2 \\ &= \int p\left(\Delta\boldsymbol{p}|\boldsymbol{\beta}_2, \psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3}\right) p\left(\boldsymbol{\beta}_2|\psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3}\right) d\boldsymbol{\beta}_2 \\ &= \int p\left(\Delta\boldsymbol{p}|\boldsymbol{\beta}_2, \psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3}\right) d\boldsymbol{\beta}_2 \end{aligned}$$

and

$$\begin{aligned} &\Delta\boldsymbol{p}|\boldsymbol{\beta}_2, \psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3} \sim \\ &\qquad \mathcal{N}\left\{(\boldsymbol{I}_2 \otimes \boldsymbol{X})\boldsymbol{\beta}_2, \sigma^2\boldsymbol{I}_{2n} + \sigma^2_{\delta_1}(\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)(\boldsymbol{I}_2 \otimes \boldsymbol{X}_1)^T + \sigma^2_{\delta_3}(\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)(\boldsymbol{I}_2 \otimes \boldsymbol{X}_3)^T\right\}. \end{aligned}$$

To simplify notation, define $\boldsymbol{Z} = \boldsymbol{I}_2 \otimes \boldsymbol{X}$, $\boldsymbol{Z}_1 = \boldsymbol{I}_2 \otimes \boldsymbol{X}_1$, $\boldsymbol{Z}_3 = \boldsymbol{I}_2 \otimes \boldsymbol{X}_3$, and $\boldsymbol{V} = \boldsymbol{I}_{2n} + \sigma^2_{\delta_1}/\sigma^2\boldsymbol{Z}_1\boldsymbol{Z}_1^T + \sigma^2_{\delta_3}/\sigma^2\boldsymbol{Z}_3\boldsymbol{Z}_3^T$, and write

$$\Delta\boldsymbol{p}|\boldsymbol{\beta}_2, \psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3} \sim \mathcal{N}\left(\boldsymbol{Z}\boldsymbol{\beta}_2, \sigma^2\boldsymbol{V}\right).$$

Consequently,

$$\begin{aligned} &p\left(\Delta\boldsymbol{p}|\psi, \boldsymbol{X}, \sigma^2, \sigma^2_{\delta_1}, \sigma^2_{\delta_3}\right) \\ &\quad = \int \left(\frac{1}{2\pi\sigma^2}\right)^{2n/2} \frac{1}{\sqrt{|\boldsymbol{V}|}} \exp\left\{-\frac{1}{2\sigma^2}(\Delta\boldsymbol{p} - \boldsymbol{Z}\boldsymbol{\beta}_2)^T\boldsymbol{V}^{-1}(\Delta\boldsymbol{p} - \boldsymbol{Z}\boldsymbol{\beta}_2)\right\} d\boldsymbol{\beta}_2 \end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{1}{2\pi\sigma^2}\right)^{2n/2} \frac{1}{\sqrt{|\boldsymbol{V}|}} \exp\left\{-\frac{1}{2\sigma^2}(\Delta\boldsymbol{p} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)^T\boldsymbol{V}^{-1}(\Delta\boldsymbol{p} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)\right\} \cdot \\
&\qquad \int \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{\beta}_2 - \tilde{\boldsymbol{\beta}}_2)^T\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}(\boldsymbol{\beta}_2 - \tilde{\boldsymbol{\beta}}_2)\right\} d\boldsymbol{\beta}_2 \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{2n/2} \frac{1}{\sqrt{|\boldsymbol{V}|}} \exp\left\{-\frac{1}{2\sigma^2}(\Delta\boldsymbol{p} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)^T\boldsymbol{V}^{-1}(\Delta\boldsymbol{p} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)\right\} (2\pi\sigma^2)^{2d/2} \cdot \\
&\qquad \frac{1}{\sqrt{|\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}|}} \\
&= \left(\frac{1}{2\pi\sigma^2}\right)^{2(n-d)/2} \frac{1}{\sqrt{|\boldsymbol{V}||\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}|}} \exp\left\{-\frac{1}{2\sigma^2}(\Delta\boldsymbol{p} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)^T\boldsymbol{V}^{-1}(\Delta\boldsymbol{p} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)\right\}
\end{aligned}
$$

with $\tilde{\boldsymbol{\beta}}_2 = (\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{V}^{-1}\Delta\boldsymbol{p}$. Substituting $\tilde{\sigma}^2$, $\tilde{\sigma}^2_{\delta_1}$, and $\tilde{\sigma}^2_{\delta_3}$ for σ^2, $\sigma^2_{\delta_1}$, and $\sigma^2_{\delta_3}$ respectively yields a log posterior density

$$
\begin{aligned}
p(\psi|\Delta\boldsymbol{p}, \boldsymbol{X}) \propto p\left(\Delta\boldsymbol{p}|\psi, \boldsymbol{X}\right) \propto -\frac{1}{2}\Bigg\{&(2n - 2d)\log\tilde{\sigma}^2 + \log|\boldsymbol{V}| + \log|\boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}| \\
&+ \frac{1}{\tilde{\sigma}^2}(\Delta\boldsymbol{p} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)^T\boldsymbol{V}^{-1}(\Delta\boldsymbol{p} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}_2)\Bigg\}.
\end{aligned}
$$

Note that for ease of notation we use the same letter $\boldsymbol{V}$ to denote the covariance matrix based on $\tilde{\sigma}^2$, $\tilde{\sigma}^2_{\delta_1}$, and $\tilde{\sigma}^2_{\delta_3}$ or on σ^2, $\sigma^2_{\delta_1}$, and $\sigma^2_{\delta_3}$. Here $\boldsymbol{V} = \boldsymbol{I}_{2n} + \tilde{\sigma}^2_{\delta_1}/\tilde{\sigma}^2\boldsymbol{Z}_1\boldsymbol{Z}_1^T + \tilde{\sigma}^2_{\delta_3}/\tilde{\sigma}^2\boldsymbol{Z}_3\boldsymbol{Z}_3^T$.

5 Discussion

The introduction to this thesis puts forward the research question, contextualizes and motivates it. It finishes with a rough sketch of the path suggested to resolve the problem. This discussion forms its counterpart. It summarizes the solution, indicates in what sense it goes beyond answering the original question and brings up issues which might merit further investigation.

The analysis of the deficits of the commonly used profile likelihood estimator reveals that it suffers from two drawbacks. First, it is necessary to restrict the domain of the threshold parameter for the profile likelihood function to be well-defined. This is due to identification issues. The literature offers no more than the ad hoc solution of an arbitrary trimming parameter. Second, profile likelihood estimators generally do not account for variability introduced by replacing nuisance parameters with estimators. This becomes a serious problem in certain critical settings, namely, when differences between regimes diminish or only few observations are left in one of them. Both a low signal-to-noise-ratio and a small number of observations in one of the regimes cause the variability of the nuisance parameters' estimators to grow very large.

To put this latter issue into perspective – of course, likelihood estimation in the presence of nuisance parameters is not a new problem. It has been discussed in the literature for decades (Neyman & Scott, 1948; Lancaster, 2000, for a survey) and a variety of strategies to correct the profile likelihood function for this deficiency have been suggested (Cox & Reid, 1987; Kalbfleisch & Sprott, 1973; McCullagh & Tibshirani, 1990; Severini, 2007, among others). The most well-known attempt is perhaps the modified profile likelihood function introduced by Barndorff-Nielsen (1983). Yet, the proposed modifications are usually based on insights drawn from an asymptotic expansion of the score function. As the profile likelihood function of a threshold regression model (assuming a step transition function) is not differentiable with respect to the threshold parameter, these corrections

are not helpful. Hence, the interplay between the two features of the model blamed to complicate threshold estimation – nuisance parameters and nonregularities – also hinders ready application of available solutions.

Apart from the profile likelihood (or least squares) estimator Bayesian estimators with noninformative priors are frequently used to estimate the threshold. However, since the posterior density is proportional to the profile likelihood function, the former inherits the shortcomings of the latter.

The alternative regularized Bayesian estimator is based on the idea of penalizing large differences between regimes when little information is available. This regularization is achieved within a Bayesian framework by means of an appropriate choice of priors. The strength of the penalty is determined in a data-driven manner employing the empirical Bayes paradigm. This estimator does not depend on a trimming parameter. One aspect of the new estimator which deserves to be accentuated is that it is elementary to compute. As it is possible to take advantage of existing methods for mixed models (which are already implemented in standard software packages), it takes barely more effort to calculate the regularized Bayesian than the profile likelihood estimator. Given that the estimation problem disappears asymptotically, simulation studies constitute the adequate way to assess the performance of the regularized Bayesian estimator and compare it with the profile likelihood estimator's. For both the GTRM and the TVECM, simulation results evidence superior quality of the newly developed estimator. Figure 5.1 shows the histograms for the regularized Bayesian threshold estimates for a TVECM, it is figure 1.1's counterpart.

The effect of using the regularized Bayesian estimator in empirical applications is exemplified in four different cases. Novel estimates differ from conventional ones on all occasions. However, in the instance of the study of the effects of climate on snowshoe hare populations it furthermore occurs that, contrary to the profile likelihood estimator, the regularized Bayesian estimator produces meaningful results without depending on additional ad hoc arguments. Thus, there is a clear benefit to employing the latter estimator. Likewise, when analyzing transmission between German and Spanish pork prices, the profile likelihood estimates are entirely driven by the arbitrary trimming parameter used to restrict the domain of the threshold parameter; hence, they are essentially worthless. In contrast, the regularized Bayesian estimates are immune to this defect.

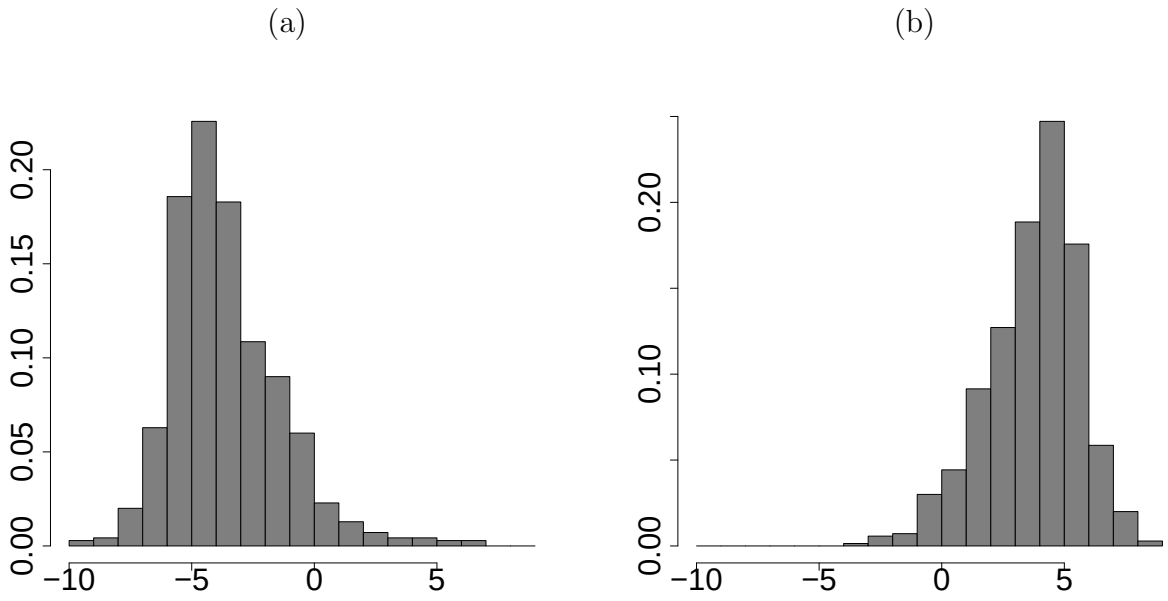

Figure 5.1: Regularized Bayesian estimates for the lower (a) and upper (b) threshold of a TVECM with three regimes. True values are -4 and 4, respectively.

Moreover, the associated estimates for the adjustment coefficients are considerably more plausible than those obtained via the standard estimation procedure. Accordingly, this application presents another example in which the use of the regularized Bayesian estimator has definite advantages. These examples confirm the practical relevance of the new estimator.

Improving threshold estimation in GTRMs in some respect exceeds what is necessary to solve the estimation problem for TVECMs. In the introduction, I present the GTRM as a simplification of the TVECM; the idea is to emphasize the strategy of boiling the problem down to its core by getting rid of complexities unrelated to the question at hand. However, this is only half the truth. The GTRM is not a simplification in every respect. The TVECM is characterized by a piecewise linear conditional mean structure, while the GTRM allows for a link function – which is not necessarily the identity – to relate the piecewise linear predictor to the mean of the observations. It further covers observations following any distribution belonging to the exponential family, for example the Poisson or the binomial distribution. In short, the GTRM goes beyond the essential reduction in model complexity to the extent by which a generalized linear model broadens a linear

model.

To appreciate the significance of the findings in chapter three detached from the context of the TVECM it is helpful to bring to mind the prominence of the model. Limiting attention only to the Gaussian threshold regression model (as in (2.6)), there exist numerous applications in a host of different fields of science – a cross-country growth regression is discussed in chapter three and the second chapter points to further examples (see section 2.1.1). Statements like van de Geer's (1988, page 3) "We shall take two-phase regression models [...] as the major illustration of the theory we develop for general regression models. In this way, we hope to provide some insight into the significance of our results." further establish the importance of the model. As the Gaussian threshold regression model is but a particular GTRM, the latter is at least equally significant. As a matter of fact, Samia & Chan (2011) stress that "nonnormal data are far more abundant than normal ones, for example, time series of counts and positive time series". The threshold regression model's popularity is generally attributed to a combination of parsimony and flexibility in functional form without being susceptible to curse of dimensionality (Kourtellos, Stengos & Tan, 2011). Samia & Chan (2011) add "relative ease of tractability and interpretation".

Which potential directions for further research open up during the course this thesis? As touched upon in chapter three, it suggests itself to explore the possibility to create a test based on the same idea as the estimator. The approximate regularized posterior can be thought of as the restricted likelihood function of a generalized linear mixed model, the difference between regimes interpreted as a random effect. Put into this perspective, a test for a threshold translates into a test for a random effect. For the Gaussian case of a linear mixed model, such test has been developed by Crainiceanu & Ruppert (2004) and Scheipl, Greven & Küchenhoff (2008) and implemented in the R package `RLRsim`. An extension to generalized linear mixed models might possibly provide the basis for a unified test in GTRMs. The literature on tests for a threshold is large (Davies, 1977; Davies, 1987; Andrews, 1993; Andrews & Ploberger, 1994; Hansen, 1996; Lee, Seo & Shin, 2011). However, resembling estimation, most tests concentrate on Gaussian threshold regression models (or models with piecewise linear mean). Moreover, a test within the mixed model setting might fare better than existing tests in the critical situations of low signal-to-noise ratio and few observations in one of the regimes.

Related to this the question arises whether it is even possible to exploit the idea of a test for random effects in the more general situation of models with unidentified parameters under the null hypothesis. Several of the articles cited above are actually based on this broader framework. Hansen (1996), for example, focuses on regression models with additive nonlinearity,

$$y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + H(\boldsymbol{Z}_i, \boldsymbol{\gamma})^T\boldsymbol{\theta} + \varepsilon_i$$

for observations $(y_i, \boldsymbol{X}_i^T, \boldsymbol{Z}_i^T) \in \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^q$, parameters $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\gamma} \in \mathbb{R}^m$, $\boldsymbol{\theta} \in \mathbb{R}^k$, and disturbances ε_i, with zero mean and finite variance, $i = 1, \ldots, n$. The null hypothesis of a model $y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \varepsilon_i$, that is of $\boldsymbol{\theta} = 0$, is tested against the alternative that the term $H(\boldsymbol{Z}_i, \boldsymbol{\gamma})^T\boldsymbol{\theta}$ enters. The parameter $\boldsymbol{\gamma}$ is not identified under the null hypothesis. Apart from tests for a threshold, that is, $H(\boldsymbol{Z}_i^T, \gamma) = I(q_i \leq \gamma)\boldsymbol{X}_i$ for $\boldsymbol{Z}_i^T = (q_i, \boldsymbol{X}_i^T)$, or a change point, that is $H(\boldsymbol{X}_i^T, \gamma) = I(i/n \leq \gamma)\boldsymbol{X}_i$, Hansen (1996) enumerates "Box-Cox transformations: $H(Z_i, \gamma) = (Z_i^\gamma - 1)/\gamma$; [...] Bierens's (1990) consistent tests of functional form: $H(\boldsymbol{Z}_i, \boldsymbol{\gamma}) = \exp(\boldsymbol{\gamma}^T\boldsymbol{Z}_i)$; White's (1989) neural network tests of functional form: $H(Z_i, \gamma) = 1/\{1 + \exp(\gamma - Z_i)\}$"as further potential applications.

Another vague thought along these lines is to explore the connection between the problem considered in this thesis and the embedded model problem (Smith, 1989; Cheng & Iles, 1990). Embedded models arise as limiting cases of an original model. In this context, the idea of the difference between regimes as a random effect amounts to phrasing the parameter θ as a random effect with mean μ if the embedded model emerges for $\theta \to \mu$.

Not directly related to the findings of this thesis, but maybe still worth mentioning, an observation regarding the literature summarized in chapter two. To my mind, the number of articles on the limiting properties of threshold estimators that cover models that are only slightly distinct from one another (or even identical, but with estimators varying among least squares, maximum likelihood and M-estimators) is remarkable. Besides, results closely resemble each other or are natural extensions of one another. Without having studied the derivations of the limiting distributions in detail, it is apparently very easy to generalized findings at times, whereas on other occasions the opposite is the case. For example, Ciuperca & Dapzol (2008) omit the proof of their theorem 3.5 (see section 2.1.6), which establishes the limit likelihood process for a multiple-threshold regression model as a sum of independent compound Poisson processes, for its similarity to the single-threshold case derived by Koul & Qian (2002). Contrary to this, He & Severini

(2010) who look at observations $\boldsymbol{X}_i \in \mathbb{R}^p$ drawn independently from a distribution

$$\sum_{j=1}^{J+1} f_j\left(\boldsymbol{\theta}, \boldsymbol{\xi}_j; \boldsymbol{X}_i\right) I(\psi_{j-1} + 1 \leq i \leq \psi_j),$$

where f_j are continuous density functions characterized by a parameters $\boldsymbol{\theta} \in \mathbb{R}^m$ and $\boldsymbol{\xi}_j \in \mathbb{R}^{q(j)}$, $j = 1, \ldots, J$, and $\psi_1 < \ldots < \psi_J \in \mathbb{N}$ are the change points, $\psi_0 = 0$ and $\psi_{J+1} = n$, establish consistency of the maximum likelihood estimator of the change points and determine its rate of convergence, but leave the limiting distribution for further research. They remark that "unfortunately, there are problems in extending the approach used in Hinkley (1970, 1972) to the setting considered here. The method used in Hinkley (1970, 1972) is based on considering the relative locations of a candidate change point and the true change point. When there is only a single change point, there are only three possibilities: the candidate change point is either greater than, less than or equal to the true change point. However, in models with J change points, the relative positions of the candidate change points and the true change points can become quite complicated and the simplicity and elegance of the single change point argument is lost" (He & Severini, 2010, page 760). Similarly, Tsay (1998) notes that extending results for the univariate continuous threshold regression model (Chan & Tsay, 1998) to a multivariate model "is yet to be rigorously investigated", hence, apparently not straightforward. van de Geer (1988) takes a unified approach by embedding the two-phase regression model, which initiated her study, within the context of nonlinear regression, but admits that she "did not succeed in avoiding ad hoc arguments for this model". Taking these statements into account, it might be an interesting endeavor to try and identify the actual scope of the results regarding the limiting properties of threshold estimators, which I outlined in chapter two.

With regard to the specific case of threshold estimation in TVECMs, and especially when thinking of empirical applications of the latter in price transmission analysis, it is disturbing to see how hard it can be to produce reliable estimates. Unlike in the real world, in the Monte Carlo studies conducted, no errors are introduced through data collection and a perfectly specified model is estimated – this is not likely to be the case in empirical studies. Still, in some situations it is not easy to obtain trustworthy threshold estimates. Obviously, the TVECM with two thresholds (and three regimes) suffers from

some serious oversimplifications. The two most obvious are that (i) transaction costs are assumed to be constant over time; (ii) while the reasoning for the model set-up is very persuasive when considering just two markets in isolation, it is unclear whether it is the appropriate model specification when they form part of a larger network of interconnected markets. Thus, it is natural to try and incorporate more flexibility to better represent real conditions. However, against the background that estimation can already be problematic in the "simple" TVECM studied in this thesis, the question of a proper balance between flexibility and feasibility of reliable estimation comes to the fore. Increased model flexibility influences parameter estimators in two counteracting ways. It means that there is a larger number of parameters to estimate, hence, the estimators tend to have greater variability. However, at the same time it implies that model misspecification is less likely, hence, facilitates adequate estimation. Consequently, to my mind it would be interesting to examine the optimal balance of these competing forces given constraints by the data available. Yet, an answer which goes beyond very specific cases might be impossible to give.

References

Andrews, D. (1993). Tests for parameter instability and structural change with unknown change point. *Econometrica 61*(4), 821–856.

Andrews, D. and Ploberger, W. (1994). Optimal tests when a nuisance parameter is present only under the alternative. *Econometrica 62*(6), 1383–1414.

Bacon, D. and Watts, D. (1971). Estimating the transition between two intersecting straight lines. *Biometrika 58*(3), 525–534.

Bai, J. (1997a). Estimating multiple breaks one at a time. *Econometric Theory 13*(1), 315–352.

Bai, J. (1997b). Estimation of a change point in multiple regression models. *Review of Economics and Statistics 79*(4), 551–563.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica 66*(1), 47–78.

Balcombe, K., Bailey, A., and Brooks, J. (2007). Threshold effects in price transmission: the case of Brazilian wheat, maize, and soya prices. *American Journal of Agricultural Economics 89*(2), 308–323.

Balcombe, K. and Rapsomanikis, G. (2008). Bayesian estimation and selection of nonlinear vector error correction models: the case of the sugar-ethanol-oil nexus in Brazil. *American Journal of Agricultural Economics 90*(3), 658–668.

Balke, N. and Fomby, T. (1997). Threshold cointegration. *International Economic Review 38*(3), 627–45.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika 70*(2), 343.

Bauwens, L., Lubrano, M., and Richard, J. (1999). *Bayesian inference in dynamic econometric models*. Oxford University Press, USA.

Beckman, R. and Cook, R. (1979). Testing for two-phase regressions. *Technometrics 21*(1), 65–69.

Bhattacharya, P. (1987). Maximum likelihood estimation of a change-point in the distribution of independent random variables: general multiparameter case. *Journal of Multivariate Analysis 23*(2), 183–208.

Bhattacharya, P. (1994). Some aspects of change-point analysis. *IMS Lecture Notes-Monograph Series* **23**, 28–56.

Bhattacharya, P. and Brockwell, P. (1976). The minimum of an additive process with applications to signal estimation and storage theory. *Probability Theory and Related Fields 37*(1), 51–75.

Bierens, H. (1990). A consistent conditional moment test of functional form. *Econometrica 58*(6), 1443–1458.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, California.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association 88*(421), 9–25.

Calabrese, E., Baldwin, L., et al. (2003). Toxicology rethinks its central belief. *Nature 421*(6924), 691–692.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Verlag.

Card, D., Mas, A., and Rothstein, J. (2008). Tipping and the dynamics of segregation. *The Quarterly Journal of Economics 123*(1), 177.

Chan, K. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *The Annals of Statistics 21*(1), 520–533.

Chan, K., Petruccelli, J., Woolford, S., and Tong, H. (1985). A multiple threshold AR (1) model. *Journal of applied probability 22*(2), 267–279.

Chan, K. and Tsay, R. (1998). Limiting properties of the least squares estimator of a continuous threshold autoregressive model. *Biometrika 85*(2), 413–426.

Chan, N. and Kutoyants, Y. (2010a). On parameter estimation of threshold autoregressive models. *Arxiv preprint arXiv:1003.3800*.

Chan, N. and Kutoyants, Y. (2010b). Recent developments of threshold estimation for nonlinear time series. *J. Japan Statist. Soc 40*(2), 277–303.

Chen, C. (1998). A Bayesian analysis of generalized threshold autoregressive models. *Statistics & probability letters 40*(1), 15–22.

Cheng, R. C. H. and Iles, T. C. (1990). Embedded models in three-parameter distribu-

tions and their estimation. *Journal of the Royal Statistical Society, Series B 52*(1), 135–149.

Ciuperca, G. and Dapzol, N. (2008). Maximum likelihood estimator in a multi-phase random regression model. *Statistics 42*(4), 363–381.

Cobb, G. (1978). The problem of the nile: conditional solution to a changepoint problem. *Biometrika 65*(2), 243–251.

Cox, C. (1987). Threshold dose-response models in toxicology. *Biometrics 43*(3), 511–523.

Cox, D. and Reid, N. (1987). Parameter orthogonality and approximate conditional inference. *Journal of the Royal Statistical Society, Series B 49*(1), 1–39.

Crainiceanu, C. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of the Royal Statistical Society, Series B 66*(1), 165–185.

Davies, R. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 64*(2), 247–254.

Davies, R. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika 74*(1), 33–43.

van Dijk, D., Teräsvirta, T., and Franses, P. (2002). Smooth transition autoregressive models–a survey of recent developments. *Econometric Reviews 21*(1), 1–47.

Doodson, A. (1917). Relation of the mode, median and mean in frequency curves. *Biometrika 11*(4), 425–429.

Durlauf, S. and Johnson, P. (1995). Multiple regimes and cross-country growth behaviour. *Journal of Applied Econometrics 10*(4), 365–384.

Engle, R. and Granger, C. (1987). Co-integration and error correction: representation, estimation, and testing. *Econometrica 55*(2), 251–276.

Engle, R. and Granger, C. (1991). *Long-run economic relationships: Readings in cointegration*. Oxford University Press.

Fackler, P. and Goodwin, B. (2001). Spatial price analysis. *Handbook of agricultural economics* **1**, 971–1024.

Fackler, P. and Tastan, H. (2008). Estimating the degree of market integration. *American Journal of Agricultural Economics 90*(1), 69.

Feder, P. (1975). On asymptotic distribution theory in segmented regression problems–identified case. *The Annals of Statistics 3*(1), 49–83.

Fujii, T. (2008). On weak convergence of the likelihood ratio process in multi-phase regression models. *Statistics & Probability Letters 78*(14), 2066–2074.

Gaul, J. et al. (2008). *Three Essays on Unit Roots and Nonlinear Co-Integrated Processes*. Ph. D. thesis, Universitäts-und Landesbibliothek Bonn.

van de Geer, S. (1988). *Regression analysis and empirical processes*. Amsterdam.

Geweke, J. and Terui, N. (1993). Bayesian threshold autoregressive models for nonlinear time series. *Journal of Time Series Analysis 14*(5), 441–454.

Goodwin, B. and Piggott, N. (2001). Spatial market integration in the presence of threshold effects. *American Journal of Agricultural Economics 83*(2), 302–317.

Granger, C. (1986). Developments in the study of cointegrated economic variables. *Oxford Bulletin of economics and statistics 48*(3), 213–228.

Greb, F., von Cramon-Taubadel, S., Krivobokova, T., and Munk, A. (2011). Threshold estimation in price transmission analysis. CRC-PEG Discussion Paper No.103, Courant Research Centre "Poverty, Equity and Growth in Developing Countries", Georg-August-Universität Göttingen.

Greb, F., Krivobokova, T., Munk, A., and von Cramon-Taubadel, S. (2011). Regularized Bayesian estimation in generalized threshold regression models. CRC-PEG Discussion Paper No.99, Courant Research Centre "Poverty, Equity and Growth in Developing Countries", Georg-August-Universität Göttingen.

Hansen, B. (2000). Sample splitting and threshold estimation. *Econometrica 68*(3), 575–603.

Hansen, B. (2011). Threshold autoregression in economics. *Statistics and Its Interface 4*(2), 123–128.

Hansen, B. and Seo, B. (2002). Testing for two-regime threshold cointegration in vector error-correction models. *Journal of Econometrics 110*(2), 293–318.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica 64*(2), 413–430.

Harville, D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association 72*(358), 320–338.

He, H. and Severini, T. (2010). Asymptotic properties of maximum likelihood estimators in models with multiple change points. *Bernoulli 16*(3), 759–779.

Hinkley, D. (1969a). Inference about the intersection in two-phase regression.

Biometrika 56(3), 495–504.

Hinkley, D. (1969b). On the ratio of two correlated normal random variables. *Biometrika 56*(3), 635–639.

Hinkley, D. (1970). Inference about the change-point in a sequence of random variables. *Biometrika 57*(1), 1–17.

Hinkley, D. (1972). Time-ordered classification. *Biometrika 59*(3), 509–523.

Horváth, L., Hušková, M., and Serbinowska, M. (1997). Estimators for the time of change in linear models. *Statistics 29*(2), 109–130.

Hušková, M. (1996). Estimation of a change in linear models. *Statistics & probability letters 26*(1), 13–24.

Hušková, M. and Antoch, J. (2001). M-estimators of structural changes in regression models. *Tatra Mt. Math. Publ* **22**, 1–12.

Ihle, R. (2010). *Models for Analyzing Nonlinearities in Price Transmission*. Ph. D. thesis, Niedersächsische Staats-und Universitätsbibliothek Göttingen.

Kalbfleisch, J. and Sprott, D. (1973). Marginal and conditional likelihoods. *Sankhyā: The Indian Journal of Statistics, Series A 35*(3), 311–328.

Kendall, M. G. (1943). *The advanced theory of statistics, Vol. 1.* J.B. Lippincott company.

Khodadadi, A. and Asgharian, M. (2008). Change-point problem and regression: An annotated bibliography. *COBRA Preprint Series* **44**.

Koul, H. and Qian, L. (2002). Asymptotics of maximum likelihood estimator in a two-phase linear regression model. *Journal of Statistical Planning and Inference 108*(1-2), 99–119.

Koul, H., Qian, L., and Surgailis, D. (2003). Asymptotics of M-estimators in two-phase linear regression models. *Stochastic processes and their applications 103*(1), 123–154.

Kourtellos, A., Stengos, T., and Tan, C. (2011). Structural threshold regression. University of cyprus working papers in economics, University of Cyprus Department of Economics.

Lai, T. (2001). Sequential analysis: some classical problems and new challenges. *Statistica Sinica 11*(2), 303–350.

Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of econometrics 95*(2), 391–413.

Lange, T. and Rahbek, A. (2009). An introduction to regime switching time series models. *Handbook of Financial Time Series*, 871–887.

Lee, S., Seo, M., and Shin, Y. (2011). Testing for threshold effects in regression models. *Journal of the American Statistical Association 106*(493), 220–231.

Lo, M. and Zivot, E. (2001). Threshold cointegration and nonlinear adjustment to the law of one price. *Macroeconomic Dynamics 5*(4), 533–576.

Lubrano, M. (2000). Bayesian analysis of nonlinear time series models with a threshold. In *Nonlinear Econometric Modeling in Time Series: Proceedings of the Eleventh International Symposium in Economic Theory.*

Lundbergh, S., Teräsvirta, T., and Van Dijk, D. (2003). Time-varying smooth transition autoregressive models. *Journal of Business and Economic Statistics 21*(1), 104–121.

MacLulich, D. (1937). *Fluctuations in the numbers of the varying hare (Lepus americanus)*. University of Toronto Press.

McCullagh, P. and Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society, Series B 52*(2), 325–344.

Neyman, J. and Scott, E. (1948). Consistent estimates based on partially consistent observations. *Econometrica 16*(1), 1–32.

O'Connell, P. and Wei, S. (2002). "The bigger they are, the harder they fall": Retail price differences across US cities. *Journal of International Economics 56*(1), 21–53.

Perron, P. (2006). Dealing with structural breaks. *Palgrave handbook of econometrics* **1**, 278–352.

Perron, P. and Qu, Z. (2006). Estimating restricted structural change models. *Journal of Econometrics 134*(2), 373–399.

Picard, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability 17*(4), 841–867.

Pole, A. M. and Smith, A. F. M. (1985). A Bayesian analysis of some threshold switching models. *Journal of econometrics 29*(1-2), 97–119.

Qian, L. (1998). On maximum likelihood estimators for a threshold autoregression. *Journal of Statistical Planning and Inference 75*(1), 21–46.

Qu, Z. and Perron, P. (2007). Estimating and testing structural changes in multivariate regressions. *Econometrica 75*(2), 459–502.

Quandt, R. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association 53*(284), 873–880.

Ravallion, M. (1986). Testing market integration. *American Journal of Agricultural Economics 68*(1), 102–109.

Ravallion, M., Chen, S., and Sangraula, P. (2009). Dollar a day revisited. *The World Bank Economic Review 23*(2), 163–184.

Samia, N. and Chan, K. (2011). Maximum likelihood estimation of a generalized threshold stochastic regression model. *Biometrika 98*(2), 433.

Samia, N., Chan, K., and Stenseth, N. (2007). A generalized threshold mixed model for analyzing nonnormal nonlinear time series, with application to plague in Kazakhstan. *Biometrika 94*(1), 101–118.

Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis 52*(7), 3283–3299.

Seo, M. and Linton, O. (2007). A smoothed least squares estimator for threshold regression models. *Journal of econometrics 141*(2), 704–735.

Seo, M. H. (2011). Estimation of nonlinear error correction models. *Econometric Theory 27*(02), 201–234.

Sephton, P. (2003). Spatial market arbitrage and threshold cointegration. *American Journal of Agricultural Economics 85*(4), 1041–1046.

Severini, T. (2000). *Likelihood methods in statistics*. Oxford University Press, USA.

Severini, T. (2007). Integrated likelihood functions for non-Bayesian inference. *Biometrika 94*(3), 529–542.

Shun, Z. and McCullagh, P. (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society, Series B 57*(4), 749–760.

Smith, R. (1989). A survey of nonregular problems. In *Proceedings of International Statistical Institute*, Volume 47, pp. 353–372.

Stephens, E., Mabaya, E., Cramon-Taubadel, S., and Barrett, C. (2011). Spatial price adjustment with and without trade. *Oxford Bulletin of Economics and Statistics*.

Tikhonov, A., Arsenin, V., and John, F. (1977). *Solutions of ill-posed problems*. Vh Winston Washington, DC.

Tishler, A. and Zang, I. (1979). A switching regression method using inequality con-

ditions. *Journal of Econometrics 11*(2-3), 259–274.

Tong, H. (1983). *Threshold models in non-linear time series analysis. Lecture notes in statistics, No. 21*. Springer-Verlag.

Tong, H. (2011). Threshold models in time series analysis–30 years on. *Statistics and its Interface 4*(2), 107–118.

Tong, H. and Lim, K. (1980). Threshold autoregression, limit cycles and cyclical data. *Journal of the Royal Statistical Society, Series B*, 245–292.

Tsay, R. (1998). Testing and modeling multivariate threshold models. *Journal of the American Statistical Association 93*(443), 1188–1202.

van der Vaart, A. W. (2010). Times series. Lecture notes, Vrije Universiteit Amsterdam.

White, H. (1989). An additional hidden unit test for neglected nonlinearity in multilayer feedforward networks. In *Neural Networks, 1989. IJCNN., International Joint Conference on*, pp. 451–455. IEEE.

Worsley, K. (1983). Testing for a two-phase multiple regression. *Technometrics 25*(1), 35–42.

Yao, Y. (1987). Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *The Annals of Statistics 15*(3), 1321–1328.

Yu, P. (2012). Likelihood estimation and inference in threshold regression. *Journal of Econometrics 167*(1), 274–294.

Zucchini, W. and MacDonald, I. (2009). *Hidden Markov models for time series: an introduction using R*, Volume 110. Chapman & Hall/CRC.