Regularized Bayesian estimation of generalized threshold regression models

Friederike Greb [1, 2] Tatyana Krivobokova [2, 3] Axel Munk [3, 4]
Stephan von Cramon-Taubadel [1]
Georg-August-Universität Göttingen

July 7, 2013

Abstract

In this article we discuss estimation of generalized threshold regression models in settings when the threshold parameter lacks identifiability. In particular, if estimation of the regression coefficients is associated with high uncertainty and/or the difference between regimes is small, estimators of the threshold and, hence, of the whole model can be strongly affected. A new regularized Bayesian estimator for generalized threshold regression models is proposed. We derive conditions for superiority of the new estimator over the standard likelihood one in terms of mean squared error. 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, regularization, threshold identification.

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 et al., 2002, provide a comprehensive overview) or step functions. In the

[1]Department of Agricultural Economics and Rural Development, Georg-August-Universität Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany
[2]Courant Research Centre "Poverty, Equity and Growth in Developing Countries", Georg-August-Universität Göttingen, Wilhelm-Weber-Str. 2, 37073 Göttingen, Germany
[3]Institute for Mathematical Stochastics, Georg-August-Universität Göttingen, Goldschmidstrasse 7, 37077 Göttingen, Germany
[4]Max Planck Institute for Biophysical Chemistry, Am Faßberg 11, 37077 Göttingen, Germany

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 and Chan (2011), and most of the literature on threshold regression deals with models with a 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 et al. (2007) employ a generalized threshold regression model to analyze plague outbreaks, and Lee et al. (2011) complement these applications with examples in finance, sociology, and biostatistics among others.

Obviously, a good threshold estimator is crucial for the entire threshold regression model estimation. In this paper we discuss settings in which threshold identification becomes difficult. Typically, threshold parameters are estimated by the maximization of the corresponding profile likelihood using a grid search, as the likelihood function is not differentiable with respect to the threshold parameter. This estimation procedure itself has an intrinsic problem: 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 and 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.

Another problem occurs, if the threshold parameter itself lacks identifiability. In particular, if differences between regimes are small and/or the regression coefficients' estimators are highly variable, the uncertainty of the threshold estimator increases. Note that the large variance of the regression coefficients' estimator is likely to be found in small samples, for the true threshold at the boundary of its domain and also if the signal-to-noise ratio is low. 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 et al., 2013).

Bayesian methods are also popular to estimate threshold regression models. In the literature Bayesian estimation is typically based on non-informative priors, leading to what we refer to as the non-informative Bayesian estimator. For the threshold estimator in 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 standard likelihood estimator and can completely fail in certain cases, as we discuss in Section 3.2.

In this article, we suggest an alternative estimator, which we call the regularized Bayesian estimator. Contrary to previous work on estimation in threshold regression (Samia and 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 likelihood and non-informative Bayesian estimator are highly susceptible to faults. Given the threshold parameter's crucial function within the model, our idea is to improve estimation of the whole model by improving estimation of this essential parameter.

To summarize the intuition for the new threshold estimator: If regression coefficients were known, none of the problems in threshold estimation 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 that 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 threshold estimator and simultaneously obtain estimates for the remaining model parameters.

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 estimators for threshold regression models and point out their deficiencies. Here, we concentrate on estimators for the crucial threshold parameter. The regularized Bayesian estimator is introduced in the fourth section. In the fifth section, we derive conditions under which the regularized Bayesian estimates fare better than their likelihood counterparts. Simulation results are presented in the sixth section. We use the last section to discuss two empirical applications. The appendix contains some technical details.

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

$$\mu_i = \mathrm{E}\left(y_i | \boldsymbol{X}_i^T, q_i\right) = h(\eta_i) \tag{1}$$

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

$$\eta_i = I\left(q_i \leq \psi\right) \boldsymbol{X}_i^T \boldsymbol{\beta}_1 + I\left(q_i > \psi\right) \boldsymbol{X}_i^T \boldsymbol{\beta}_2, \tag{2}$$

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

$$f(y_i | \psi, \phi, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}, \tag{3}$$

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 $\mathrm{diag}\left\{\boldsymbol{I}(\cdot)\right\}$ 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

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

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 (see e.g. Greb et al., 2013).

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 and Lim, 1980; Tong, 2011, for a review of the development of the model).

Depending on the assumptions on the data generating process, inferences (or estimators) for model (1) – (3) can take on 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 (1997) for more details on the asymptotic properties in the former, and Samia and 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 and 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".

As the focus of this article is on estimation problems that arise in small samples, we do not further differentiate between 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 Estimation of threshold regression models

3.1 The likelihood estimator

As noted in the introduction, the prevalent estimator of threshold regression models is the likelihood estimator, see e.g. Samia and Chan (2011) or Hansen (2000). Thereby, the threshold parameter is estimated from the corresponding profile likelihood $\mathcal{L}_p$, which is constructed from the likelihood function $\mathcal{L}$, by replacing nuisance parameters $\boldsymbol{\beta}^T \in \mathbb{R}^{2p}$ and $\phi \in \mathbb{R}$ with their maximum likelihood estimates at given values of ψ (which are just standard (weighted) least squares estimators). More specifically, 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 does not hold. We denote the profile log-likelihood with $\ell_p(\psi) = \log \mathcal{L}_p(\psi)$.

In generalized threshold regression models, the domain of the threshold parameter ψ 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 statistic. 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 $q_{(j)} \leq \psi < q_{(j+1)}$. The quantity $\mathrm{d}(\psi)$ is 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

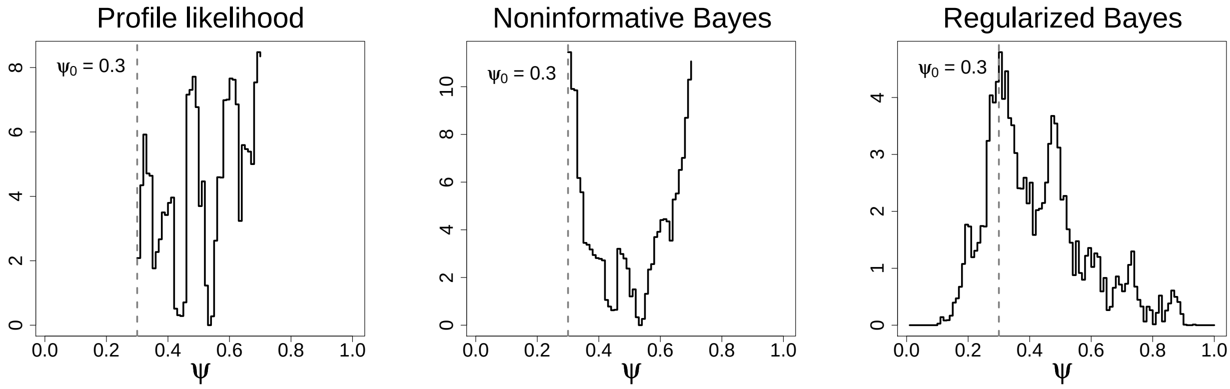

Figure 1: For a sample run corresponding to setting 1C of Section 6, $\ell_p(\psi)$ is shown on the left, $\log p_{nB}(\psi|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ in the middle and $\log p_{rB}(\psi|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{q})$ on the right.

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 and Seo (2002) propose $c = 0.05\, n/p$, we find $c = 0.15\, n/p$ in Andrews (1993) and Samia and 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 Fig. 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 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 domain is an undesirable property of threshold estimation based on the profile likelihood.

The same plot in Fig. 1 also exemplifies that in certain small-sample settings the profile (log-)likelihood can be jagged and have multiple extrema, leading to an estimated

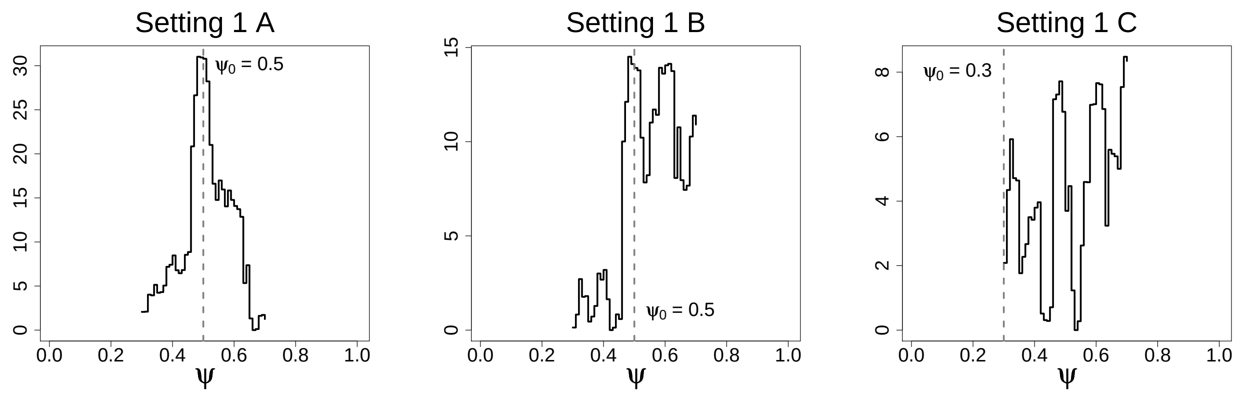

Figure 2: Sample (log) profile likelihood functions $\ell_p(\psi)$ for different settings.

threshold that is very sensitive to the initialization of the search. Large variance of $\hat{\boldsymbol{\beta}}_\psi$ and/or small differences between regimes compared to the noise level can have a strong distorting effect on the profile (log-)likelihood and are 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 a threshold that is close to the boundary of its domain. This is exposed in the left as compared with the middle plot of Fig. 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, increasing the signal-to-noise ratio and $\mathrm{var}(\hat{\boldsymbol{\beta}}_\psi)$. 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 Fig. 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.

3.2 The Bayesian estimator

For threshold regression models with piecewise linear mean, there is a long tradition of using Bayesian techniques in applied work beginning with Bacon and Watts (1971) and including Geweke and 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 uncertainty of the nuisance parameters. The Bayesian regression coefficients estimators coincide with the maximum likelihood ones for non-informative priors. 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 and Kutoyants (2012) examine asymptotic properties of Bayesian estimators in threshold autoregression models. They note that in the limit, the 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 and

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 Fig. 1 illustrates this problem.

4 The regularized Bayesian estimator

When rethinking the threshold regression estimation, there are good arguments for continuing to pursue Bayesian options. In general, Bayesian estimators naturally incorporate the uncertainty of nuisance parameters and there are reasons to expect the threshold estimators to be (at least asymptotically) the most efficient estimators, as discussed in Section 3.2.

Our idea now is to exploit understanding of when reliable estimation becomes particularly difficult in order to regularize the posterior density. First, we define

$$\boldsymbol{\eta} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 = (\boldsymbol{X}_1 + \boldsymbol{X}_2)\boldsymbol{\beta}_1 + \boldsymbol{X}_2(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1) = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\delta}. \tag{4}$$

Here, $\boldsymbol{X}$ is independent of ψ, while $\boldsymbol{X}_2 = \boldsymbol{X}_2(\psi) = \operatorname{diag}\{\boldsymbol{I}(\boldsymbol{q} > \psi)\}\,\boldsymbol{X}$. Hence, if $\boldsymbol{\delta}$ is small and/or its estimators are highly variable, it becomes hard to identify the threshold ψ. We, therefore, suggest to regularize the estimator for $\boldsymbol{\delta}$. In a Bayesian framework the natural approach is to assume $\boldsymbol{\delta} \sim \mathcal{N}(0, \sigma_\delta^2 \boldsymbol{I}_p)$. 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, that is there is no threshold in the model. 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 fully 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

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

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

$$\begin{aligned}
&\int_{\mathbb{R}^p}\int_{\mathbb{R}^p} p(\boldsymbol{y}|\boldsymbol{\beta}_1, \boldsymbol{\delta}, \psi, \phi, \sigma_\delta^2, \boldsymbol{X}, \boldsymbol{q}) p(\boldsymbol{\delta}|\sigma_\delta^2) d\boldsymbol{\delta} d\boldsymbol{\beta}_1 \\
&\quad = (2\pi)^{p/2} \exp\left\{ -\frac{1}{2}(\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1)^T \boldsymbol{V}^{-1} (\tilde{\boldsymbol{z}} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1) + \sum_{i=1}^n c(y_i, \phi) \right\} \\
&\qquad \cdot \left|\sigma_\delta^2 \boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2 + \boldsymbol{I}_p\right|^{-1/2} \left|\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}\right|^{-1/2} + \mathcal{O}\left(n^{-1}\right),
\end{aligned} \tag{5}$$

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 $\hat{\boldsymbol{\delta}} = \sigma_\delta^2 \boldsymbol{X}_2^T \boldsymbol{V}^{-1} (\tilde{\boldsymbol{z}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_1)$. Note that these regression parameter estimators are regularized and are different from usual likelihood estimators. Details on the derivation of (5) 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

$$\hat{\boldsymbol{\delta}} = \sigma_\delta^2 \boldsymbol{X}_2^T \boldsymbol{V}^{-1} (\tilde{\boldsymbol{z}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_1) = \arg \min_{\boldsymbol{\delta} \in \mathbb{R}^p} (\tilde{\boldsymbol{z}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2 \boldsymbol{\delta})^T \boldsymbol{W} (\tilde{\boldsymbol{z}} - \boldsymbol{X} \hat{\boldsymbol{\beta}}_1 - \boldsymbol{X}_2 \boldsymbol{\delta}) + \frac{1}{\sigma_\delta^2} \boldsymbol{\delta}^T \boldsymbol{\delta}, \tag{6}$$

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 et al. (1977). For small values of the regularization parameter $1/\sigma_\delta^2$, the first term of the functional to be minimized in (6) 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 fully Bayesian framework. However, pursuing our approximate approach further we prefer to 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

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

with $\hat{\boldsymbol{V}}$ evaluated at $\hat{\sigma}_\delta^2$. The right plot in Fig. 1 shows the log of this posterior density for a sample run corresponding to simulation setting 1 C of Section 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 critical small-sample settings the posterior density is often characterized by multiple modes. Thus, obtaining an estimate based on numerical maximization (the posterior mode) is likely to be challenging. 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 calculating 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 (5). 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.

Inference about all of the model parameters naturally follows in this Bayesian framework. In particular, confidence regions for ψ are formed as credible sets; an equi-tailed credible set C of level $1 - 2\alpha$ is defined as

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

These credible sets are valid for change-point and threshold models, both continuous and discontinuous. By contrast, 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.

5 Comparison of regularized Bayesian and maximum likelihood estimation

Our new estimation procedure results in new regularized regression coefficients estimators, whose properties have not been investigated so far. In the following, we compare regularized Bayesian and maximum likelihood approaches to estimation of threshold regression models in terms of mean squared error under the frequentist model. Thereby, we treat the threshold as fixed and known, but allow for any, not necessarily true threshold ψ.

A natural measure for comparing coefficient estimates is the mean squared error $\mathrm{M}(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}) = \mathrm{E}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}} - \boldsymbol{X}_\psi\boldsymbol{\beta}\right)^T\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}} - \boldsymbol{X}_\psi\boldsymbol{\beta}\right)$, where E denotes the conditional expectation without averaging over the prior assumptions, i.e. expectation with respect to the distribution of $\boldsymbol{Y}$ given $\boldsymbol{\delta}$, which corresponds to the usual frequentist framework. In the context of ridge regression, this approach has been critized for indiscriminately putting together the mean squared errors of the components (Nelder, 1972; Theobald, 1974). As an alternative, Theobald (1974) suggested to consider a weighted sum $\mathrm{M_A}(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}) = \mathrm{E}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}} - \boldsymbol{X}_\psi\boldsymbol{\beta}\right)^T\boldsymbol{A}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}} - \boldsymbol{X}_\psi\boldsymbol{\beta}\right)$ for a non-negative definite matrix $\boldsymbol{A}$. Here, ψ is an arbitrary, fixed threshold. Of course, a comparison between $\mathrm{M}(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}})$ (or $\mathrm{M_A}(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}})$) for different $\hat{\boldsymbol{\beta}}$ is both interesting for such general ψ as well as the true threshold ψ_0. With this in mind, we state the following result.

Theorem 1 For maximum likelihood estimates $\hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi)^{-1}\boldsymbol{X}_\psi\boldsymbol{W}\boldsymbol{z}$ and regularized Bayesian estimates $\hat{\boldsymbol{\beta}}_{rB} = (\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi + \boldsymbol{H})^{-1}\boldsymbol{X}_\psi\boldsymbol{W}\boldsymbol{z}$ of $\boldsymbol{\beta}$ based on a threshold $\psi \le \psi_0$, ψ_0 the true threshold,

(i) $$\mathrm{M_A}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{ML}\right) - \mathrm{M_A}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{rB}\right) \ge 0 \text{ for all non-negative definite matrices } \boldsymbol{A}$$

$$\Leftrightarrow \quad \boldsymbol{D}\left\{\left(\boldsymbol{I}+\boldsymbol{C}\right)\boldsymbol{H} - \left(\boldsymbol{B}+\boldsymbol{H}\right)\boldsymbol{\beta}\boldsymbol{\beta}^T\left(\boldsymbol{B}^T+\boldsymbol{H}\right) + \boldsymbol{C}\boldsymbol{B}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{B}^T\boldsymbol{C}^T\right\}\boldsymbol{D}^T$$

is non-negative definite.

(ii) $$\mathrm{M}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{ML}\right) - \mathrm{M}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{rB}\right) \ge 0$$

$$\Leftrightarrow \quad \mathrm{tr}\left\{\boldsymbol{H}\boldsymbol{D}^T\boldsymbol{D}\left(\boldsymbol{I}+\boldsymbol{C}\right)\right\} - \boldsymbol{\beta}^T\left\{\left(\boldsymbol{B}^T+\boldsymbol{H}\right)\boldsymbol{D}^T\boldsymbol{D}\left(\boldsymbol{B}+\boldsymbol{H}\right) + \boldsymbol{B}^T\boldsymbol{D}_0^T\boldsymbol{D}_0\boldsymbol{B}\right\}\boldsymbol{\beta} \ge 0.$$

Here, $\boldsymbol{W}^{-1} = \mathrm{diag}\left\{\phi b''(\theta_i)g'(\mu_i)^2\right\}$, $\boldsymbol{G} = \mathrm{diag}\left\{g'(\mu_i)\right\}$, and $\boldsymbol{z} = \boldsymbol{X}_\psi\boldsymbol{\beta} + \boldsymbol{G}(\boldsymbol{y} - \boldsymbol{\mu})$ as before, $\boldsymbol{H} = 1/\sigma_\delta^2\begin{pmatrix}\boldsymbol{I}_p & -\boldsymbol{I}_p\\ -\boldsymbol{I}_p & \boldsymbol{I}_p\end{pmatrix}$, $\boldsymbol{B} = \begin{pmatrix}0 & 0\\ -\boldsymbol{X}_{[\psi,\psi_0]}^T\boldsymbol{W}\boldsymbol{X}_{[\psi,\psi_0]} & \boldsymbol{X}_{[\psi,\psi_0]}^T\boldsymbol{W}\boldsymbol{X}_{[\psi,\psi_0]}\end{pmatrix}$ with $\boldsymbol{X}_{[\psi,\psi_0]} = \mathrm{diag}\{\boldsymbol{I}(\psi < \boldsymbol{q} \le \psi_0)\}\boldsymbol{X}$, $\boldsymbol{C} = \boldsymbol{I} + \boldsymbol{H}\left(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi\right)^{-1}$, $\boldsymbol{D} = \boldsymbol{X}_\psi\left(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi + \boldsymbol{H}\right)^{-1}$, and $\boldsymbol{D}_0 = \boldsymbol{X}_\psi\left(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi\right)^{-1}$.

Remark 1 For the Gaussian model with $\boldsymbol{W} = 1/\sigma^2\boldsymbol{I}_n$ and at the true threshold $\psi = \psi_0$, equivalence (i) reduces to

$$\begin{aligned} & \mathrm{M_A}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{ML}\right) - \mathrm{M_A}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{rB}\right) \geq 0 \text{ for all non-negative definite matrices } \boldsymbol{A} \\ \Leftrightarrow \quad & \boldsymbol{\delta}^T(2\sigma_\delta^2/\sigma^2\boldsymbol{I} + \boldsymbol{Z})^{-1}\boldsymbol{\delta} \leq \sigma^2, \end{aligned} \tag{7}$$

where $\boldsymbol{Z} = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^T\boldsymbol{X}_2)^{-1}$, while equivalence (ii) reduces to

$$\begin{aligned} & \mathrm{M}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{ML}\right) - \mathrm{M}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{rB}\right) \geq 0 \\ \Leftrightarrow \quad & \boldsymbol{\delta}^T\boldsymbol{Z}\left(\sigma^2/\sigma_\delta^2\boldsymbol{I}_p + \boldsymbol{Z}\right)^{-2}\boldsymbol{\delta} \leq \sigma^2\left\{p - \mathrm{tr}\left(\boldsymbol{I}_p + \sigma^2/\sigma_\delta^2\boldsymbol{Z}\right)^{-2}\right\} \end{aligned} \tag{8}$$

Remark 2 Using a singular value decomposition $\boldsymbol{Z} = \boldsymbol{U}\,\mathrm{diag}\,(\eta_1, \ldots, \eta_p)\,\boldsymbol{U}^T$ and writing $\boldsymbol{U}^T\boldsymbol{\delta} = \boldsymbol{\alpha}$, inequality (8) is equivalent to

$$\sum_{i=1}^{p} \frac{\eta_i\left(2\sigma_\delta^2/\sigma^2 + \eta_i - \alpha_i^2/\sigma^2\right)}{\left(\sigma_\delta^2/\sigma^2 + \eta_i\right)^2} \geq 0,$$

which holds in particular if

$$\frac{\alpha_{max}^2 - \eta_{min}\sigma^2}{2} \leq \sigma_\delta^2 \tag{9}$$

with $\alpha_{max} = \max\limits_{1\leq i\leq p} \alpha_i$ and $\eta_{min} = \min\limits_{1\leq i\leq p} \eta_i$. Analogously, we obtain

$$\frac{p\alpha_{max}^2 - \eta_{min}\sigma^2}{2} \leq \sigma_\delta^2 \tag{10}$$

as a condition for inequality (7) to be satisfied.

Remark 3 The left-hand side of inequalities (7) – (10) decreases when $\delta_1, \ldots, \delta_p$ diminish in magnitude, while the right-hand side increases with growing variance σ^2, that is, when the signal-to-noise ratio becomes smaller. Hence, it is reasonable to expect regularized Bayesian regression coefficient estimates to be particularly superior to their profile likelihood counterparts in settings previously identified as problematic.

Remark 4 The regularized Bayesian estimator for the regression coefficients $\hat{\boldsymbol{\beta}}_{rB} = (\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi + \boldsymbol{H})^{-1}\boldsymbol{X}_\psi\boldsymbol{W}\boldsymbol{z}$ closely resembles the ridge estimator. However, the special form of the penalty matrix $\boldsymbol{H} = \sigma_\delta^{-2}\begin{pmatrix}\boldsymbol{I}_p & -\boldsymbol{I}_p \\ -\boldsymbol{I}_p & \boldsymbol{I}_p\end{pmatrix}$ (instead of just $\sigma_\delta^{-2}\boldsymbol{I}_{2p}$ in the ridge regression) has considerable implications for the estimator.

6 Simulations

To assess the performance of the suggested approach and the estimator $\hat{\psi}_{rB}$ in particular 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 $P\left(\mathrm{d}\left(\psi_0\right) < 1\right) \approx 0.46$ 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 $\mathrm{d}\left(\psi_0\right) = 1$, hence, that $\mathcal{L}_p\left(\psi_0\right)$ 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

	Normal response (1)			
	A	B	C	D
ψ_0	0.5	0.5	0.3	0.3
$\boldsymbol{\delta}$	$U[-0.5, 0.5]$	$U[-0.5, 0.5]$	$U[-0.5, 0.5]$	$U[-0.25, 0.25]$
$\mathrm{var}(y_i)$	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 1: Differences between simulation settings.

	MSE($\hat{\psi}$)			MSE($\boldsymbol{X}_{\hat{\psi}}\hat{\boldsymbol{\beta}}$)	
	pL	nB	rB	pL	rB
1 A	0.006	0.035	0.002	0.00002	0.00001
1 B	0.040	0.093	0.024	0.00009	0.00005
1 C	0.272	0.264	0.089	0.00009	0.00005
1 D	0.401	0.738	0.191	0.00001	0.00001
2 A	0.000	0.003	0.000	0.05953	0.01947
2 B	0.013	0.115	0.004	0.07625	0.02916
2 C	0.083	0.116	0.014	0.57250	0.02266
2 D	0.146	0.358	0.036	0.72387	0.18669

Table 2: Simulation results.

$\boldsymbol{\delta} = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ 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). The effects of increasing the signal-to-noise ratio and shifting ψ_0 on $\ell_p(\psi)$ are illustrated in Fig. 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. For 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 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 $R = 1000$ replications.

With regard to the threshold parameter, we summarize simulation results in Fig. 3, where the boxplots of the threshold estimators are shown and in the left half of Table 2, where $\text{MSE}\left(\hat{\psi}\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\psi}^{(r)}\big/\psi - 1\right)^2$ are reported. 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).

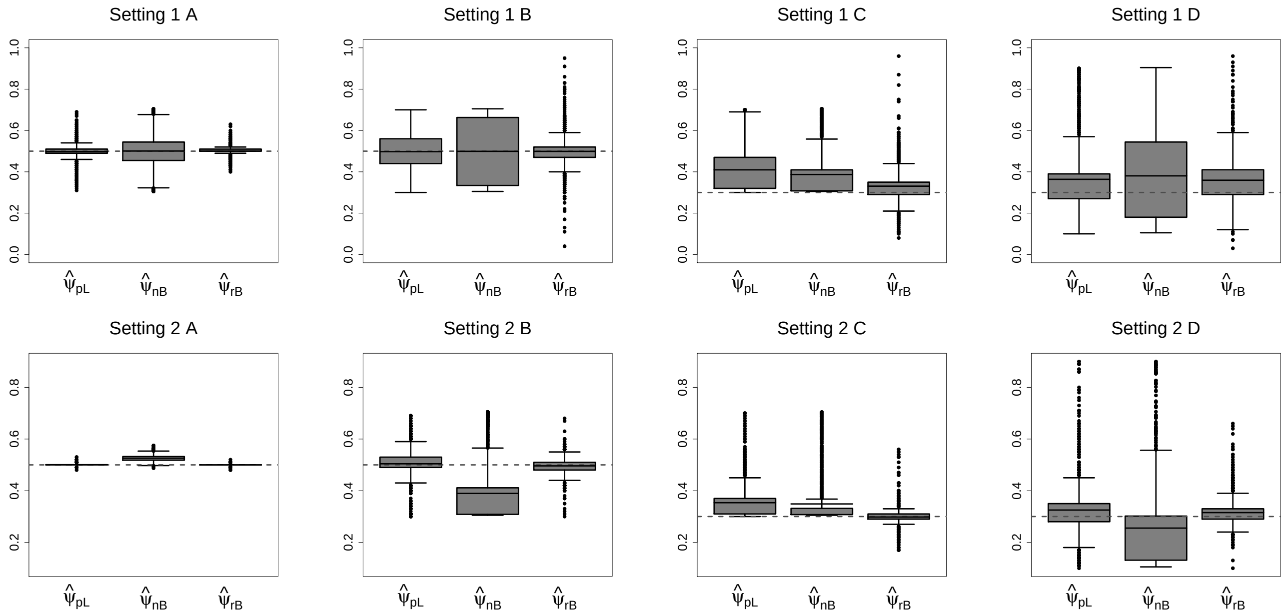

Figure 3: Boxplots for different threshold estimators and selected simulations. Dashed lines indicate the true threshold ψ_0, black lines in the boxes are sample means.

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 squared error. When we shift the true threshold towards the boundary of Ψ (setting C), $\hat{\psi}_{rB}$ clearly outperforms both $\hat{\psi}_{pL}$ and $\hat{\psi}_{nB}$. The differences in mean squared error are more pronounced with a greater number of nuisance parameters p, but are still visible in simulations with smaller ratio p/n (setting D).

To complement findings for the threshold estimators with results concerning estimation of the model as a whole, in particular including the regression coefficients' estimator, we consider the mean squared error for the entire model. The regularized Bayesian approach fares better in general. While the mean squared error is much lower for simulations with normal than with Poisson response, differences between the likelihood and regularized Bayesian framework are more marked for the latter. The right half of Table 2 contains details. We denote $\mathrm{MSE}\left(\boldsymbol{X}_{\hat{\psi}}\hat{\boldsymbol{\beta}}\right) = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n}\left(\boldsymbol{X}^{(r)}_{\hat{\psi}^{(r)}}\hat{\boldsymbol{\beta}}^{(r)}\Big/\boldsymbol{X}^{(r)}_{\psi}\boldsymbol{\beta} - \mathbf{1}\right)^T\left(\boldsymbol{X}^{(r)}_{\hat{\psi}^{(r)}}\hat{\boldsymbol{\beta}}^{(r)}\Big/\boldsymbol{X}^{(r)}_{\psi}\boldsymbol{\beta} - \mathbf{1}\right)$ with the division $\boldsymbol{X}^{(r)}_{\hat{\psi}^{(r)}}\hat{\boldsymbol{\beta}}^{(r)}\Big/\boldsymbol{X}^{(r)}_{\psi}\boldsymbol{\beta}$ defined elementwise and $\mathbf{1} = (1, \ldots, 1)^T \in \mathbb{R}^n$. Note that in settings 2 the Fisher scoring algorithm for the estimation of generalized regression models can be unstable for small sample sizes, sometimes leading to a false convergence. Therefore, we excluded such outliers (5% of the Monte Carlo sample) from the calculation of MSE($\boldsymbol{X}_{\hat{\psi}}\hat{\boldsymbol{\beta}}$) for settings 2 A – D.

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, but one important characteristic in this context is that they contain a large number of parameters besides the threshold and available data series are typically short in relation to the complexity of the model. Greb et al. (2013) investigates the merits of the regularized Bayesian approach for this particular model; simulations demonstrate the superiority of the regularized Bayesian threshold estimator (see figure 1, figure 2, and table 1 in Greb et al., 2013) and two real data examples confirm its relevance in practice.

7.1 Cross-country growth behavior

As another application of the regularized Bayesian threshold estimator, we consider the case of economic growth modeling. Durlauf and 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 et al., 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

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

which explains real GDP growth between 1960 and 1985 in country i, $\log(GDP)_{i,1985} - \log(GDP)_{i,1960}$, 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 $(SCHOOL)_i$. With reference to Durlauf and 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 $\log(GDP)_{i,1960}$ 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 (Fig. 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 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 4

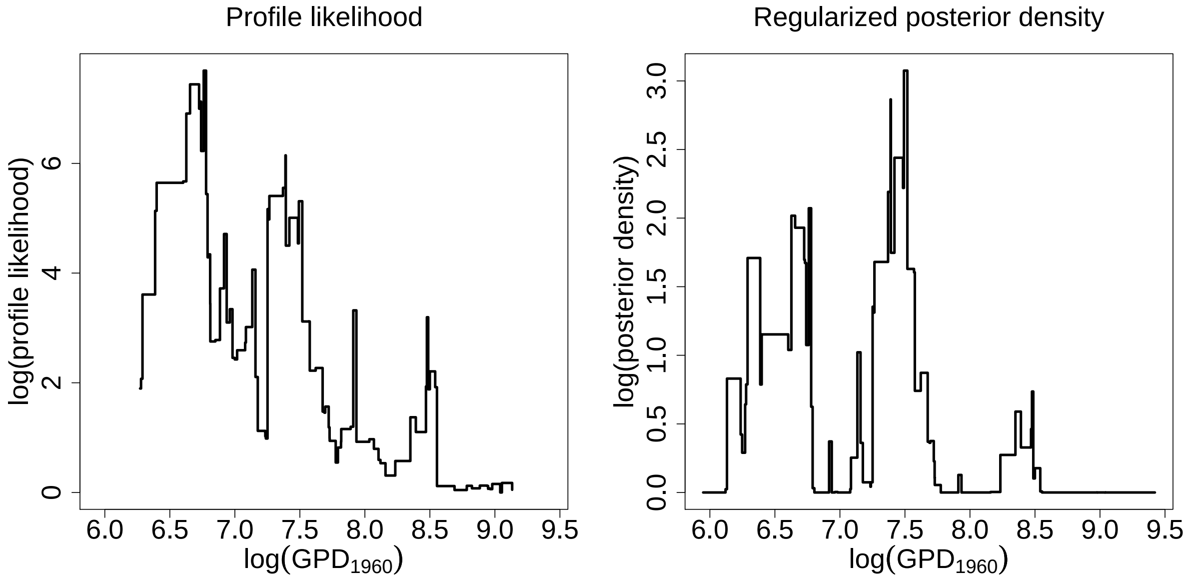

Figure 4: Profile likelihood and regularized posterior density for a threshold based on the transition variable $q_i = \log{(GDP)}_{i,1960}$.

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 Fig. 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 (see table 7.1). Moreover, the difference between the two regimes as estimated within the regularized Bayesian framework is negligible. This is in line with Hansen's (2000) 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: Regressions coefficient estimates. "pL" refers to the profile likelihood, "rB" to the regularized Bayesian framework. Standard errors in parentheses below the estimates.

7.2 Effects of climate on snowshoe hare survival

In our final example, 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 Fig. 5. These have been ascribed to the predator-prey relationship between lynx and snowshoe hares. Samia and 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 cycle. 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,

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

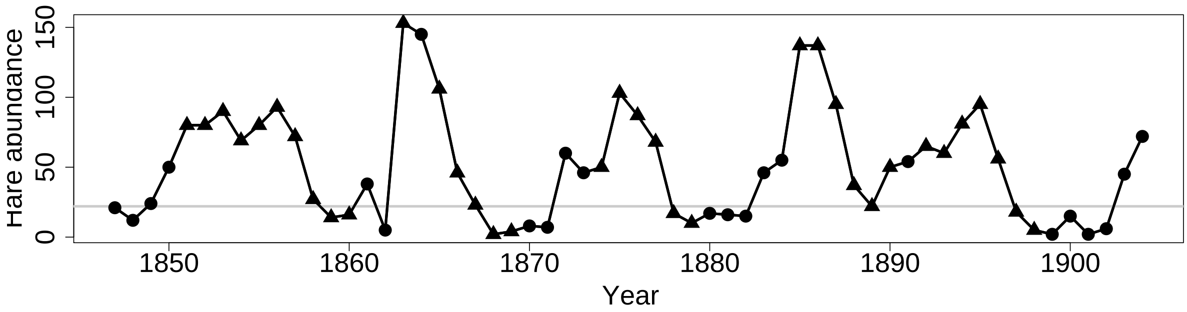

Figure 5: Annual hare abundance. Observations estimated to belong to the lower regime are plotted as dots, observations estimated to belong to the upper regime as triangles. The horizontal grey line indicates the location of the estimated threshold, $\hat{\psi}_{rB} = 22$.

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, published at www.cru.uea.ac.uk/cru/data/nao.

We follow Samia and Chan (2011) 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.

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 Fig. 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 Fig. 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 $\hat{d}_{rB}$

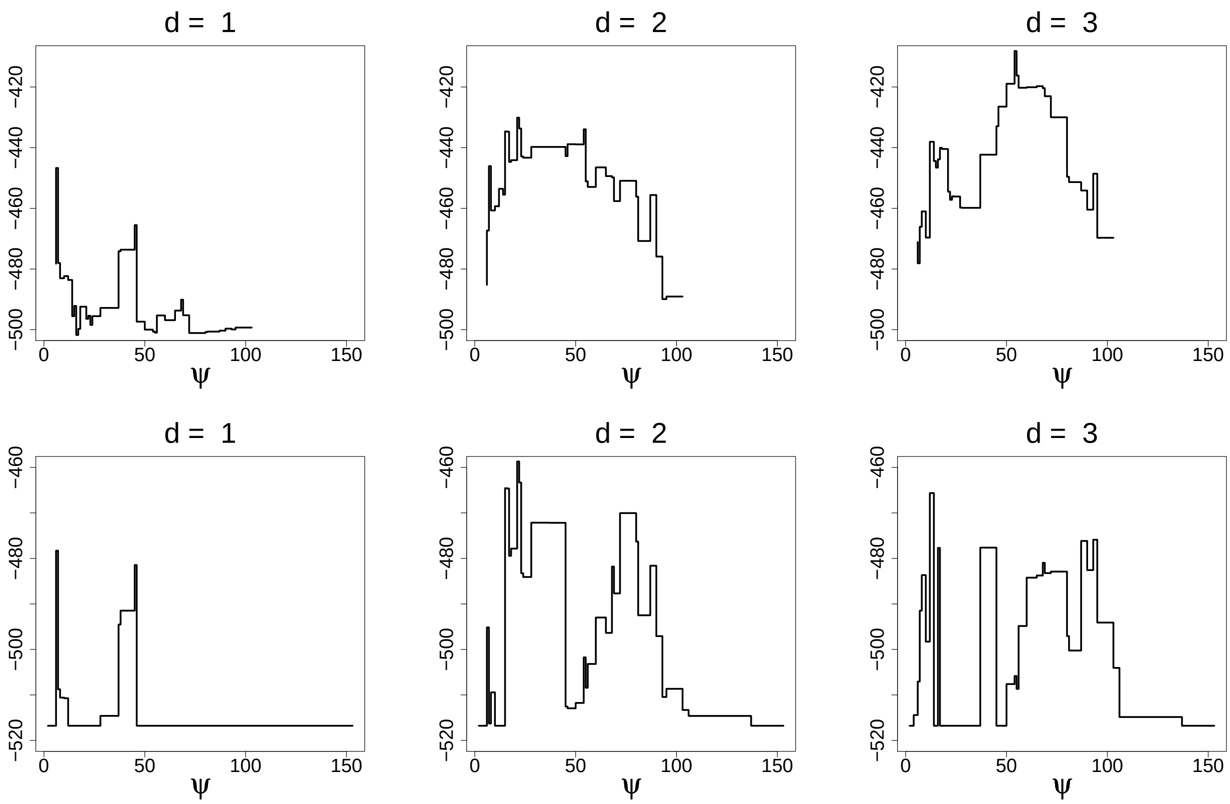

Figure 6: Log-likelihood functions (upper row) and log-posterior densities (lower row) for different delays of the transition variable.

calculated as the posterior median based on a flat prior on $\{1, 2, 3\}$.

When referring to Samia and 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 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 and 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 Fig. 5 with Fig. 1 in Samia and 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 and Chan (2011). Coefficient estimates are similar in both modeling frameworks.

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.

Acknowledgments

The authors are very grateful to the responsible editor, the associate editor and two anonymous referees for numerous constructive comments, which have greatly improved the paper. The support of the German Research Foundation (Deutsche Forschungsgemeinschaft) as part of the Institutional Strategy of the University of Göttingen is acknowledged by all the authors. The third author also acknowledges funding through FOR 916 and the VW foundation.

Appendix

Derivation of equation (5)

We obtain the approximate posterior (5) as follows. Laplace approximation produces

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

$$= (2\pi)^{-p/2} |\sigma_\delta^2 \boldsymbol{I}_p|^{-1/2} \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} \exp\left\{-\kappa\left(\boldsymbol{\delta}, \boldsymbol{\beta}_1\right)\right\} d\boldsymbol{\delta} d\boldsymbol{\beta}_1$$

$$= (2\pi)^{p/2} |\sigma_\delta^2 \boldsymbol{I}_p|^{-1/2} \exp\left\{-\kappa\left(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}_1\right)\right\} \left|\frac{\partial^2 \kappa}{\partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)\partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)^T}(\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\beta}}_1)\right|^{-1/2} + \mathcal{O}\left(n^{-1}\right)$$

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

$$\partial^2 \kappa \Big/ \partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)\partial(\boldsymbol{\delta}, \boldsymbol{\beta}_1)^T = \begin{pmatrix} \boldsymbol{X}_2^T \boldsymbol{W}\boldsymbol{X}_2 + \left(1/\sigma_\delta^2\right)\boldsymbol{I}_p & \boldsymbol{X}_2^T \boldsymbol{W}\boldsymbol{X} \\ \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X}_2 & \boldsymbol{X}^T \boldsymbol{W}\boldsymbol{X} \end{pmatrix} \tag{11}$$

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 $\hat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{\beta}_1\right)_0$, we solve

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

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

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

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 and 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

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

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

Details for Theorem 1

Basic matrix algebra yields a representation of the regularized Bayesian estimators $\hat{\boldsymbol{\beta}}_1 = \left(\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{z}$ and $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_1 + \hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\beta}}_1 + \sigma_\delta^2\boldsymbol{X}_2^T\boldsymbol{V}^{-1}\left(\boldsymbol{z} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_1\right)$, where $\boldsymbol{V} = \boldsymbol{W}^{-1} + \sigma_\delta^2\boldsymbol{X}_2\boldsymbol{X}_2^T$, as $\hat{\boldsymbol{\beta}}_{rB} = \left(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi + \boldsymbol{H}\right)^{-1}\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{z}$. To obtain equivalence (i), we employ a theorem by Theobald (1974, theorem 1) stating that for two estimators $\hat{\boldsymbol{\beta}}_\star$ and $\hat{\boldsymbol{\beta}}_{\star\star}$

$$\begin{aligned}
&\mathrm{M_A}(\hat{\boldsymbol{\beta}}_\star) - \mathrm{M_A}(\hat{\boldsymbol{\beta}}_{\star\star}) \geq 0 \text{ for all non-negative definite matrices } \boldsymbol{A} \\
\Leftrightarrow\quad &\mathrm{E}\big(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}\big)\big(\hat{\boldsymbol{\beta}}_\star - \boldsymbol{\beta}\big)^T - \mathrm{E}\left(\hat{\boldsymbol{\beta}}_{\star\star} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{\star\star} - \boldsymbol{\beta}\right)^T \text{ is non-negative definite.}
\end{aligned}$$

The equivalence then follows from

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{rB} - \boldsymbol{X}_\psi\boldsymbol{\beta}\right)\left(\boldsymbol{X}_\psi\hat{\boldsymbol{\beta}}_{rB} - \boldsymbol{X}_\psi\boldsymbol{\beta}\right)^T &= \boldsymbol{X}_\psi\left(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi + \boldsymbol{H}\right)^{-1}\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi\left(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi + \boldsymbol{H}\right)^{-1}\boldsymbol{X}_\psi^T \\
&+ \boldsymbol{X}_\psi\left(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi + \boldsymbol{H}\right)^{-1}(\boldsymbol{B} + \boldsymbol{H})\boldsymbol{\beta}\boldsymbol{\beta}^T\left(\boldsymbol{B}^T + \boldsymbol{H}\right)\left(\boldsymbol{X}_\psi^T\boldsymbol{W}\boldsymbol{X}_\psi + \boldsymbol{H}\right)^{-1}\boldsymbol{X}_\psi^T.
\end{aligned}$$

Using $\mathrm{E}\left(\hat{\boldsymbol{\beta}}_{rB} - \boldsymbol{\beta}\right)^T\left(\hat{\boldsymbol{\beta}}_{rB} - \boldsymbol{\beta}\right) = \mathrm{tr}\left\{\mathrm{E}\left(\hat{\boldsymbol{\beta}}_{rB} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}_{rB} - \boldsymbol{\beta}\right)^T\right\}$ then yields equivalence (ii).

For remark 1, $\psi = \psi_0$ implies $\boldsymbol{B} = 0$. Consequently,

$$\boldsymbol{D}\left\{(\boldsymbol{I} + \boldsymbol{C})\boldsymbol{H} - (\boldsymbol{B} + \boldsymbol{H})\boldsymbol{\beta}\boldsymbol{\beta}^T\left(\boldsymbol{B}^T + \boldsymbol{H}\right) + \boldsymbol{C}\boldsymbol{B}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{B}^T\boldsymbol{C}^T\right\}\boldsymbol{D}^T \geq 0$$

reduces to

$$\boldsymbol{D}\left\{(\boldsymbol{I} + \boldsymbol{C})\boldsymbol{H} - \boldsymbol{H}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{H}\right\}\boldsymbol{D}^T \geq 0.$$

Assuming that $\operatorname{rank}(\boldsymbol{X}) = p$, this is equivalent to

$$\begin{aligned} & \left(\boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi + \boldsymbol{H}\right)^{-1} \left\{(\boldsymbol{I} + \boldsymbol{C})\, \boldsymbol{H} - \boldsymbol{H}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{H}\right\} \left(\boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi + \boldsymbol{H}\right)^{-1} \geq 0 \\ \Leftrightarrow \quad & (\boldsymbol{I} + \boldsymbol{C})\, \boldsymbol{H} - \boldsymbol{H}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{H} = 2\boldsymbol{H} + \boldsymbol{H} \left(\boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi\right)^{-1} \boldsymbol{H} - \boldsymbol{H}\boldsymbol{\beta}\boldsymbol{\beta}^T \boldsymbol{H} \geq 0 \end{aligned}$$

since $\boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi + \boldsymbol{H}$ is positive definite and symmetric. Taking advantage of a result by Gruber (1990, theorem 2.5.3), this amounts to

$$\begin{aligned} & \boldsymbol{\beta}^T \boldsymbol{H} \left(2\sigma_\delta^2 \boldsymbol{H} + \sigma_\delta^2 \boldsymbol{H} \left(\boldsymbol{X}_\psi^T \boldsymbol{W} \boldsymbol{X}_\psi\right)^{-1} \boldsymbol{H}\right)^+ \boldsymbol{H}\boldsymbol{\beta} \leq 1/\sigma_\delta^2 \\ \Leftrightarrow \quad & \boldsymbol{\delta}^T \left\{2\sigma_\delta^2 \boldsymbol{I} + (\boldsymbol{X}_1^T \boldsymbol{W} \boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^T \boldsymbol{W} \boldsymbol{X}_2)^{-1}\right\}^{-1} \boldsymbol{\delta} \leq 1. \end{aligned}$$

For $\boldsymbol{W} = 1/\sigma_\delta^2 \boldsymbol{I}$ this is equilvalent to

$$\boldsymbol{\delta}^T \left\{2\sigma_\delta^2/\sigma^2 \boldsymbol{I} + (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} + (\boldsymbol{X}_2^T \boldsymbol{X}_2)^{-1}\right\}^{-1} \boldsymbol{\delta} \leq \sigma^2.$$

Basic matrix calculations suffice to obtain the rest of this as well as the following remarks.

References

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

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

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

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.

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. H. and Kutoyants, Y. A. (2012). On parameter estimation of threshold autoregressive models. *Statistical inference for stochastic processes*, 15(1):81–104.

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.

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

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

Greb, F., von Cramon-Taubadel, S., Krivobokova, T., and Munk, A. (2013). The estimation of threshold models in price transmission analysis. *American Journal of Agricultural Economics*. First published online: March 28, 2013.

Gruber, M. H. (1990). *Regression estimators: A comparative study*. Academic Press (Boston).

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.

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

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

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.

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

Nelder, J. (1972). Discussion of a paper by D.V. Lindley and A.F.M. Smith. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34:18–20.

Samia, N. and Chan, K. (2011). Maximum likelihood estimation of a generalized threshold stochastic regression model. *Biometrika*, 98(2):433–448.

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.

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

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

Theobald, C. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 103–106.

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

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 (Methodological)*, 42(3):245–292.

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

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