Three Essays on Regularization and Machine Learning

Dissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschaftswissenchaften Dr.rer.pol.

vorgelegt von
Jana Marečková

an der

Politik-Recht-Wirtschaft
Wirtschaftswissenschaften

Konstanz, 2019

Tag der mündlichen Prüfung: 12.7.2019

1. Referent: Prof. Dr. Winfried Pohlmeier

2. Referent: Prof. Dr. Ralf Brüggemann

3. Referent: Prof. Dr. Martin Spindler

Contents

1 Introduction

Regularization and machine learning became more present in theoretical and applied econometrics in the last decades. Variable selection based on L_1-norm regularization called least absolute shrinkage and selection operator (LASSO) became a useful tool in an econometrician's toolbox as introduced in Tibshirani (1996). The properties of regularization approaches yielding stable estimates and sparse solutions in the case of L_1-norm regularization led to many methodological contributions in econometrics. Among many, some remarkable are Chernozhukov et al. (2015) who introduce L_1-norm selection on many confounders and instruments yielding a valid statistical inference for the treatment effect which is further generalized to the selection on observables by flexible learners in Chernozhukov et al. (2017). Tutz and Oelker (2017) use an L_1-norm penalty to detect fixed and random effects in a data-driven way for panel models. Athey et al. (2019) develop a causal random forest procedure to detect heterogeneous treatment effects contributing to causal machine learning techniques. There are also many other contributions in financial econometrics and other fields. Fan et al. (2011) provides a concise overview for high-dimensional problems in economics.

Empirical contributions using regularization or machine learning techniques are recently emerging. A straightforward application of L_1-norm regularization can be found in Brodie et al. (2009) who select stocks for a Markowitz portfolio with a LASSO yielding a stable and sparse portfolio. In the machine learning strand of the literature, causal machine learning methods are applied in to estimate heterogeneous treatment effects. Bertrand et al. (2017) evaluate impacts of public working programs in Côte d'Ivoire. Knaus et al. (2017) estimate heterogeneous effects of job market programs in Switzerland combining causal machine learning models with regularization techniques. Application of regularization and machine learning techniques in economics is still rather exceptional. Nevertheless, the popularity increases and more papers appear for both, predictive and causal economic modeling. Varian (2014), Mullainathan and Spiess (2017) and Athey (2018) are valuable selected references addressing general value of machine learning in economics.

This thesis follows this recent trend in econometrics research and consists of three chapters contributing to the methodology and economic applications of regularization and machine learning techniques. The second chapter introduces a flexible regularization approach that reduces point estimation risk of group means stemming from e.g. categorical regressors, (quasi-)experimental data or panel data models. The loss function is penalized by adding weighted squared

L_2-norm differences between group location parameters and informative first-stage estimates which then lead to an optimal aggregation of information across the groups. Under quadratic loss, the penalized estimation problem has a simple interpretable closed-form solution that nests methods established in the literature on ridge regression, discretized support smoothing kernels and model averaging methods. The infeasible risk-optimal penalty parameters are derived and a plug-in approach is proposed for estimation. The large sample properties are analyzed in an asymptotic local to zero framework. The proposed plug-in estimator uniformly dominates the ordinary least squares in terms of asymptotic risk if the number of groups is larger than three. Simulations reveal robust improvements over standard methods in finite samples. The method is then applied for estimating time trends in a panel and group means in a difference-in-differences study to illustrate potential applications.

In the third chapter an L_1-norm fusion penalty is added to detect structural breaks in a linear regression with time-varying parameters. The method is a more flexible extension of the method introduced in Qian and Su (2016). The idea of the fusion penalty is to start with the most general model and shrink the differences of parameters consecutive in time to zero. At points, at which the differences between the estimated parameters are non-zero, structural breaks are detected. The main advantage of the fusion penalty over the standard statistical tests is the flexibility of the model and allowing number of breaks to grow with the length of the time series. Regarding the estimation, an important issue is the optimal choice of shrinkage otherwise the number of breaks is over- or underestimated and/or the estimates are extremely biased. To select the optimal shrinkage, two criteria are taken from the literature and one new is introduced in the paper. In the simulation study, the criteria are compared to each other and to standard tests in terms of the correct number of detected breaks, the closeness to the true position of the break and the bias of the estimates. The results show that the relative position of the break can be consistently estimated. In small samples the criteria from the literature tend to underestimate the number of breaks and the parameters are strongly biased. The newly introduced criterion performs better regarding the bias and has better or comparable performance regarding the detection of the correct number of breaks in comparison to the other chosen criteria and the standard tests.

In the fourth chapter, the goal is to analyze and improve predictive quality of noncognitive skill measures by means of machine learning techniques. Unlike previous empirical approaches centering around the in-sample explanatory power of noncognitive skills, our approach focuses on the performance of predicting individual unemployment over a long-run horizon of 20 years and more. Our machine

learning approach can cope not only with the challenge of selecting the most relevant factors from data with a large number of skill measures but also leads to a sparse set of predictive skill measures which is economically and psychologically interpretable. Using data from the British Cohort Study (BCS), we compare the predictive power of different noncognitive skill measures and illustrate, how our estimates can be used to optimize the assignment mechanisms for manpower training programs and psychological intervention schemes for youths and young adults. Moreover, the chapter tackles the question of the objective function for tuning parameters of a classification model. An economic alternative is proposed to the standard statistical tuning.

The thesis then concludes with an outlook of the potential of regularization and machine learning techniques in economics. The previous and recent development of these methods brings new value added for applied economic research as outlined in the review studies mention above. This thesis aims to follow this endeavor in the following three chapters.

2 Shrinkage for Categorical Regressors

2.1 Introduction

In general, estimation of conditional mean functions with categorical regressors can be very challenging. Even models with a moderate number of parameters lead to substantial estimation risk if the numbers of observations per group or cell that are determined by the explanatory variables are small. Regression models with multiple interactions, (quasi-)experimental designs or panel data models with time trends and fixed effects naturally fall into that category.

We propose a flexible penalization approach called pairwise cross-smoothing (PCS) to improve on the issue of point estimation risk. The method penalizes the loss function by adding sums of weighted squared L_2-norm differences between group location (reference) parameters and informative first-stage estimates (targets). It nests existing smoothing and averaging methods for orthogonal regressors, has favorable computational cost due to closed-form solutions for both estimator and penalty or smoothing parameters and can easily be extended to the case of mixed data.

Nonparametric methods in the fashion of Aitchison and Aitken (1976) are originally intended to deal with the small to empty cell problem in the context of multivariate discrete distributions (Hall, 1981; Simonoff, 1996) or mixed data distributions (Li and Racine, 2003; Hall et al., 2004). In the nonparametric regression framework Hall et al. (2007) and Ouyang et al. (2009) propose kernel methods with particular emphasis on cross-validated smoothing parameters and their behavior under the presence of irrelevant regressors. In a Bayesian sense, these methods shrink a multivariate mean towards a target value such as the global mean. The smoothing parameters depend only on a specific target covariate and are independent of the reference group. This is similar to (generalized) ridge regression (GRR) (Hoerl and Kennard, 1970). Smoothing a multivariate mean in the GRR context yields an optimization problem in which every location parameter k is effectively shrunk towards a "leave the k-th group out average". Thus for any group, GRR effectively pushes the location parameter towards a joint target. In contrast to kernel regression, smoothing parameters depend only on the reference group and not the target. Pairwise cross-smoothing on the other hand allows nonhomogeneous smoothing for both reference and target groups. Therefore, in the context of estimating group means, both kernel and (general-

ized) ridge regression can be seen as different restricted versions of PCS.

For probability distribution functions there is also a literature on empirical Bayes methods with data driven shrinkage parameters under appropriate priors for multinomial data, see e.g. Fienberg and Holland (1973), Titterington and Bowman (1985) or Simonoff (1995) for a comprehensive review with particular focus on sparse asymptotics.

The question of how to aggregate across distinctive groups can also be rephrased from a model or variable selection perspective, i.e. which groups deserve their own location parameter and which groups can be merged into one? In terms of a regression framework, one would like to know whether a more or less saturated model in terms of group dummy variables is appropriate. Classical model selection aims at selecting a single best model among a set of candidates by an appropriate criterion such as the Akaike Information Criterion (AIC, Akaike, 1973), Mallows C_p (Mallows, 1973), the Schwarz-Bayes Criterion (BIC, Schwarz, 1978) or traditional multivariate testing procedures. There is no particular reason, why these discrete model selection approaches should always yield an optimal solution. In particular, if groups or parameters are different but close to each other, averaging parameter estimates across different models could serve as a superior model selection strategy. Hjort and Claeskens (2003) consider frequentist model averaging estimators and their distributional theory in a general maximum likelihood framework with a local to zero $n^{-1/2}$-asymptotic framework. See also Claeskens et al. (2008) for a comprehensive overview. Buckland et al. (1997) and Burnham and Anderson (2003) consider smooth variants of the AIC by applying exponential weighting structures. Hansen (2007) introduces a weighting procedure for least squares estimates based on Mallows Criterion. Liang et al. (2011) consider optimal weighting schemes in terms of the mean squared error for the linear model and general likelihood models. Zhang et al. (2011) propose a focused information criterion and a model averaging estimator for generalized additive partially linear model with polynomial splines.

These smooth model averaging or shrinkage methods often have superior asymptotic risk properties over their non-shrunken counterparts. Hansen and Racine (2012) develop a jackknife model averaging estimator using cross-validation for conditional mean functions under potential misspecification of the submodels. They allow for heteroskedastic errors and non-nested models and show asymptotic optimality in the class of averaging estimators with weights in the unit simplex or a constrained subset thereof. Hansen (2014) derives conditions for asymptotic dominance of the averaging estimator in a nested least squares setup in a local to zero $n^{-1/2}$ framework, i.e. weak partial correlations of additional regressors beyond a correctly specified base model. Liu (2015) derives distributional

theory for least squares averaging estimators in the linear framework under different data-dependent weighting schemes and generalized error term structures. He considers a local to zero $n^{-1/2}$-asymptotic framework for subsets of regressors and shows the nonstandard distributional behavior of the averaging estimators. Cheng et al. (2015) consider averaging between two general method of moments estimators under potential misspecification of the second, overidentified model. They show that the averaging estimator using estimated mean squared error optimal weights can dominate the asymptotic risk of the base estimator uniformly over all degrees of misspecification. Hansen (2016a) considers shrinkage of parametric models towards restricted parameter spaces under locally quadratic loss functions and provides conditions for risk dominance. If applied to cell means or selection of orthogonal dummies, many of these methods become special cases of PCS estimators, i.e. PCS estimators with a more restricted shrinkage subspace and thus behave qualitatively similar in terms of asymptotic distribution and estimation risk. They are also closely related to classical shrinkage estimators that shrink parametric estimates towards constant vectors or restricted subspaces (Stein, 1956; Oman, 1982).

In the literature on regularization methods, coefficient estimates are enhanced by adding L_1-norm penalties of pairwise differences which allow for partial and complete fusion of groups, see e.g. Tibshirani et al. (2005) for linear models and Tutz and Oelker (2017) for group-specific generalized linear models. The main differences to the other methods are nonsmooth aggregation, i.e. groups are set to be identical, and estimation that is done in a single, one-step procedure while e.g. model averaging directly and nonparametric smoothing implicitly use first-stage estimates such as submodels or averages. These regularization methods are more suited for sparse high-dimensional applications but suffer from similar criticism as pre-testing or superefficient estimators, i.e. in finite samples actual risk gains can be inferior to standard likelihood or least squares approaches and heavily depend on the size of the coefficients (Hansen, 2016b).

The direct or implicit aggregation that is introduced by all of these methods for regression models leads to the question of what an "optimal" aggregation rule is. Our framework allows for almost any linear aggregation based on a set of smoothing parameters. Using a simple projection as a first-stage estimate, we derive mean squared error optimal smoothing parameters and propose a plug-in approach for estimation. The additional flexibility provides a substantial decrease in the oracle risk bounds compared to more restrictive aggregation methods such as (generalized) ridge regression or kernel smoothing and thus also serves as a benchmark for future research.

We further contribute to the literature by analyzing the behavior of both es-

timated smoothing parameters as well as the PCS estimator in an asymptotic local to zero framework. We introduce a class of sequences for close and distant systems of locations that covers a wide range of data generating processes. We derive the asymptotic distribution of the PCS estimator under fixed, theoretically optimal and estimated smoothing parameters. In addition, we show that the feasible PCS dominates the OLS estimation risk uniformly over the class of sequences if the number of groups is larger than three.

Monte Carlo evidence suggests that the asymptotic uniform dominance property of the feasible PCS translates into superior finite sample performance over the OLS. More often than not, the method compares favorably to alternative shrinkage and model selection approaches such as (generalized) ridge regression, kernel smoothing and Mallows C_p.

The method is applied to the estimation of time trends in a short panel based on the field experiment in private day-care centers for children in Haifa by Gneezy and Rustichini (2000a) and to the difference-in-differences study about the effect of minimum wages on employment by Card and Krueger (1994) illustrating potential applications.

Section 2.2 introduces the model, the pairwise cross-smoothing estimator and its connection to established smoothing and regularization methods. Section 2.3 presents the MSE optimal smoothing parameters and the plug-in estimator. Section 2.4 introduces the local asymptotic framework and provides the distributional properties of the PCS estimator under fixed, optimal and plug-in weights. Section 2.5 discusses the asymptotic risk properties of the feasible PCS. Section 2.6 provides some Monte Carlo evidence on estimation risk in finite samples. Section 2.7 contains the applications. Section 2.8 concludes. The proofs and technical details are collected in Appendix A.1. The extended Appendix A.2 contains additional information which is helpful for a deeper understanding of the results.

2.2 Pairwise Cross-Smoothing

2.2.1 The Model

In this section we introduce and discuss the model, the penalization strategy and the pairwise cross-smoothing estimator. Column vectors are denoted in boldface letters. Consider independent and identically distributed data $(Y_i, \mathbf{X}_i')$, $i = 1, \ldots, n$, where Y_i is a real-valued random variable and $\mathbf{X}_i$ contains ordered and/or unordered discrete random variables[1]. These always uniquely determine J orthogonal groups. For example, two binary discrete random variables determine

[1] For an extension to mixed data consider Extended Appendix A.2.1.

four orthogonal groups. Let $J \times 1$ vector $\mathbf{D}_i$ indicate whether an observation i belongs to a group $j \in \{1, \ldots, J\}$. In such a case, the j-th entry of the vector $\mathbf{D}_i$ contains one, $D_{ij} = 1$, and the remaining entries are equal to zero, $D_{ij'} = 0$ for all $j' \neq j$, i.e. $\mathbf{D}_i$'s have realizations in $\{\mathbf{e}_j, 1 \leq j \leq J\}$. We assume for the remainder that the groups are asymptotically non-empty, i.e. $P(D_{ij} = 1) \equiv p_j > 0$ for all j.

Within this framework, a regression model for the conditional mean of Y_i looks as follows:

$$Y_i = \mathbf{D}_i' \boldsymbol{\mu} + \varepsilon_i \tag{2.2.1}$$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)'$, $E[\varepsilon_i | \mathbf{D}_i] = 0$ and $V[\varepsilon_i | \mathbf{D}_i] = \sigma^2(\mathbf{D}_i)$ allowing for heteroskedasticity. Let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \ldots, \hat{\mu}_J)'$ be a consistent first-stage estimator for the group means. We propose to estimate the model for the conditional mean of Y_i as a penalized least squares problem:

$$\begin{aligned} (\hat{\mu}_1^{PCS}, \ldots, \hat{\mu}_J^{PCS}) &= \underset{\mu_1, \ldots, \mu_J}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{D}_i' \boldsymbol{\mu})^2 + Q(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) \\ Q(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) &= \sum_{k=1}^{J} \sum_{j=1}^{J} \lambda_{kj} (\mu_k - \hat{\mu}_j)^2, \end{aligned} \tag{2.2.2}$$

where $\boldsymbol{\Lambda} = (\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1J}, \ldots, \lambda_{JJ})'$ are given smoothing or penalty parameters with $\lambda_{jj} = 0$ for all $j \in \{1, \ldots, J\}$. PCS stands for pairwise cross-smoothing since the penalty term $Q(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ can be geometrically seen as a set of smooth real quadric iso-hypersurfaces in $\mathbb{R}^J$ orthogonally crossing each other.[2] The idea behind the penalty is to improve the conditional group mean estimates by using information from other groups which is collected in the first-stage estimate. By allowing for reference and target dependent penalty parameters λ_{kj}, the penalty provides maximal flexibility for smoothing.

Regarding the choice of the smoothing parameters, the more informative group j is for group k, the larger the smoothing parameter λ_{kj} should be and vice versa. In the special case of $\lambda_{kj} = 0$ for all pairs (k, j), none of the groups uses information from the other groups and the optimization is identical to the ordinary least squares problem. By choosing a large λ_{kj}, $\hat{\mu}_k^{PCS}$ is shrunk towards $\hat{\mu}_j$. Setting all λ_{kj}'s to large values pushes $\hat{\mu}_k^{PCS}$ towards the mean of all $\hat{\mu}_j$ where $j \neq k$. We discuss the issue of selecting optimal smoothing parameters in Section 2.3.

Let $n_k := \sum_{i=1}^{n} D_{ik}$ denote the number of observations within group k. Existence and uniqueness of the solution to (2.2.2) are guaranteed if $\sum_{l \neq k} \lambda_{kl} > -n_k$

[2]For $J = 3$ under the restriction of $\lambda_{jj} = 0$, the quadric iso-hypersurfaces are elliptic or hyperbolic cylinders and each of the quadric iso-hypersurfaces $j \in \{1, 2, 3\}$ is centered around a point $[\hat{\mu}_j, \hat{\mu}_j, \hat{\mu}_j]$.

for all $k \in \{1, \ldots, J\}$[3]. Under this condition, the k-th group estimate is given by

$$\hat{\mu}_k^{PCS}(\mathbf{\Lambda}_k) = \frac{n_k \bar{Y}_k}{n_k + \sum_{l \neq k} \lambda_{kl}} + \sum_{j \neq k} \frac{\lambda_{kj} \hat{\mu}_j}{n_k + \sum_{l \neq k} \lambda_{kl}}, \tag{2.2.3}$$

with $\mathbf{\Lambda}_k = (\lambda_{k1}, \ldots, \lambda_{kJ})'$ and $\bar{Y}_k$ denoting the sample mean of group k. One can see that the k-th group location estimator is a linear combination of its own cell mean and the first-stage group estimates.

A possible choice for $\hat{\boldsymbol{\mu}}$ is the linear (cell-based) projection of Y_i on $\mathbf{D}_i$, i.e. $\hat{\boldsymbol{\mu}} = (\sum_{i=1}^n \mathbf{D}_i \mathbf{D}_i')^{-1} \sum_{i=1}^n \mathbf{D}_i Y_i$, the vector of cell means. The cell-based projection is also referred to as frequency approach in the literature since it weighs the outcomes only according to cell probabilities to form an estimate for a mean. The k-th mean PCS estimator can then be written as:

$$\hat{\mu}_k^{PCS}(\mathbf{\Lambda}_k) = \frac{n_k \bar{Y}_k}{n_k + \sum_{l \neq k} \lambda_{kl}} + \sum_{j \neq k} \frac{\lambda_{kj} \bar{Y}_j}{n_k + \sum_{l \neq k} \lambda_{kl}}, \tag{2.2.4}$$

which is a linear combination of cell means.

It is noteworthy that the smoothing parameters and therefore also the implicit weights $\lambda_{kj}/(n_k + \sum_{l \neq k} \lambda_{kl})$ are not all restricted to be larger or equal than zero. Just the overall smoothing for one reference category cannot be too negative. This fundamentally differentiates our approach from discretized support kernel approaches that are built as weighted averages using probability mass functions (Hall et al., 2004). They lead to weights which are restricted to be larger than zero. Shrinking simultaneously to different targets demands high flexibility from the smoothing parameters. Imposing strict positivity might not necessarily be optimal since smoothing away from distant groups can help to increase the smoothing to closer groups. The actual signs then depend on the absolute distances between group locations. For further discussion of the presence of negative smoothing parameters consider Section 2.3.

We next show that the penalty function can be considered a generalization of both generalized ridge regression (Hoerl and Kennard, 1970) and nonparametric kernel regression in the case of orthogonal binary regressors (Aitchison and Aitken, 1976; Ouyang et al., 2009). The generalized ridge estimator can be obtained by imposing equivalent shrinkage intensities $\lambda_{kj} = \lambda_k$ within all reference groups, i.e.

$$Q_{GRR}(\mathbf{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{k=1}^{J} \sum_{j \neq k} \lambda_k (\mu_k - \hat{\mu}_j)^2 \tag{2.2.5}$$

[3]For a complete proof see Appendix A.1.1.

$$\hat{\mu}_k^{GRR}(\lambda_k) = \frac{n_k \bar{Y}_k}{n_k + (J-1)\lambda_k} + \lambda_k \sum_{j \neq k} \frac{\hat{\mu}_j}{n_k + (J-1)\lambda_k}. \tag{2.2.6}$$

Thus, the GRR smooths every location parameter heterogeneously towards the corresponding shrinkage targets $\frac{1}{J-1}\sum_{j\neq k}\hat{\mu}_j$ that can be interpreted as "leave the k-th group out" averages.

The nonparametric smoothing kernel estimator can be obtained by imposing homogeneous shrinkage intensities $\lambda_{kj} = \lambda_j$ across all reference groups, i.e.

$$Q_{Kernel}(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{k=1}^{J} \sum_{j \neq k} \lambda_j (\mu_k - \hat{\mu}_j)^2 \tag{2.2.7}$$

$$\hat{\mu}_k^{Kernel}(\boldsymbol{\Lambda}) = \frac{n_k \bar{Y}_k}{n_k + \sum_{l \neq k} \lambda_l} + \frac{\sum_{j \neq k} \lambda_j \hat{\mu}_j}{n_k + \sum_{l \neq k} \lambda_l}. \tag{2.2.8}$$

In this case, the estimator effectively smooths to a "weighted leave the k-th group out" average with homogeneous smoothing parameters for identical components across reference categories k.

A further restriction of the shrinkage intensities $\lambda_{kj} = \lambda$ to be equal for all reference groups and targets yields an ordinary ridge regression with a nonzero target, i.e.

$$Q_{RR}(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \lambda \sum_{k=1}^{J} \sum_{j \neq k} (\mu_k - \hat{\mu}_j)^2. \tag{2.2.9}$$

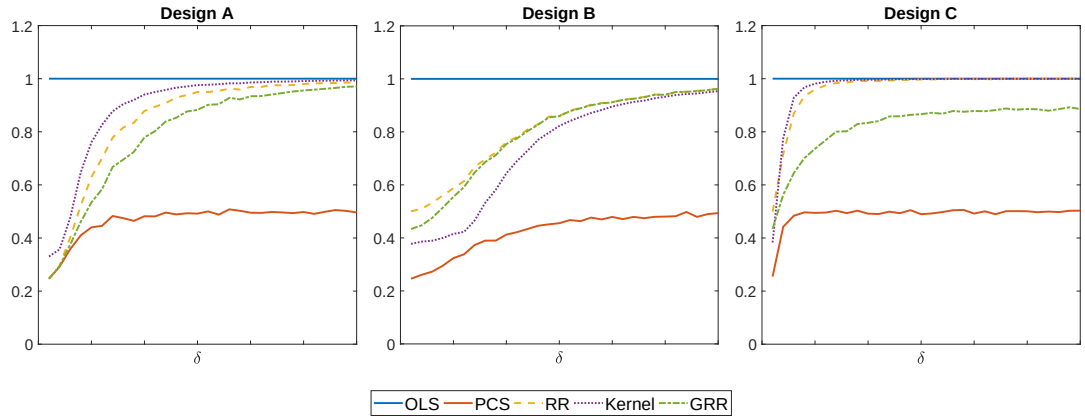
$$\hat{\mu}_k^{RR}(\lambda) = \frac{n_k \bar{Y}_k}{n_k + (J-1)\lambda} + \lambda \frac{\sum_{j \neq k} \hat{\mu}_j}{n_k + (J-1)\lambda}. \tag{2.2.10}$$

Ridge regression in this case smooths homogeneously towards the unweighted "leave the k-th group out" averages. It is reasonable to assume that allowing for more flexible shrinkage should be beneficial in terms of statistical risk if the smoothing parameters are chosen appropriately. This should be particularly pronounced if the groups are heterogeneous in terms of their size and variance.

Figure 2.2.1 depicts the mean squared error of the different estimators relative to the OLS using mean squared error optimal smoothing parameters[4] in the case of four groups and different levels of heterogeneity regarding both means and variances. One can see that all of the approaches compare favorably to the OLS if the differences in group locations are not too big. The theoretically optimal PCS not only dominates all other approaches but also qualitatively behaves closer to a correctly chosen restricted estimator that has superior risk properties even when locations are not close to identical. In addition, for the very heterogeneous design C, the differences compared to the optimal kernel and optimal (G)RR are

[4]The PCS optimal smoothing parameters can be found in Section 2.3. For the alternative methods consider Appendix A.1.5.

Figure 2.2.1: Oracle Estimators - Relative Mean Squared Errors



The figure depicts the simulated mean squared error relative to OLS for the estimators with risk-optimal (oracle) shrinkage parameters under normally distributed errors for $n = 400$, equal selection probabilities, and different parameter values δ. The parameter vectors are $\boldsymbol{\mu}_A = \boldsymbol{\mu}_B = (0, 0, 0, \delta)'$, $\boldsymbol{\mu}_C = (0, 3\delta, -2\delta, \delta)'$. The group variances are $\boldsymbol{\sigma}_A^2 = (1, 1, 1, 1)'$ and $\boldsymbol{\sigma}_B^2 = \boldsymbol{\sigma}_C^2 = (1, 1, 1, 10)'$. Simulations are based on 5000 replications.

most pronounced. Thus, the flexibility of the PCS penalty should in principle be able to generate substantial risk improvements compared to the alternative shrinkage methods.

2.3 Oracle Risk and Plug-In Estimation

In the following, we provide a simple mean squared error criterion for evaluating the estimation risk, derive the optimal smoothing parameters and introduce a plug-in approach as a feasible counterpart. For the remainder we use a modified cell average (OLS) as a first-stage $\hat{\boldsymbol{\mu}}$ to assure existence, i.e.

$$\hat{\mu}_k = \frac{\sum_{i=1}^n D_{ik} Y_i}{\sum_{i=1}^n D_{ik} + \mathbb{1}(\sum_{i=1}^n D_{ik} = 0)} \tag{2.3.1}$$

for all k. Under given assumptions, a first order approximation for the first-stage is given by[5]

$$\hat{\mu}_k - \mu_k = \frac{1}{np_k} \sum_{i=1}^n D_{ik}(Y_i - \mu_k) + o_p(n^{-1/2}). \tag{2.3.2}$$

[5]See Appendix A.1.2.

Note that rewriting (2.2.3) yields a weight-based representation of the PCS

$$\hat{\mu}_k^{PCS}(\boldsymbol{\Lambda}_k) \equiv \hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k) = (1 - \sum_{j \neq k} \omega_{kj})\bar{Y}_k + \sum_{j \neq k} \omega_{kj}\hat{\mu}_j \tag{2.3.3}$$

with $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kJ})'$ and $\omega_{kj} = \lambda_{kj}/(n_k + \sum_{l \neq k} \lambda_{kl})$ being a one-to-one correspondence[6]. While the penalized regression representation is insightful for comparison with alternative methods, using the weighted version will simplify further analysis. Using (2.3.2), the approximate parameter mean squared error is given by the following proposition:

Proposition 2.3.1 *Let $E[Y_i|D_{ij} = 1] = \mu_j$, $V[\varepsilon_i|D_{ij} = 1] = \sigma_j^2$ be finite, $P(D_{ij} = 1) \equiv p_j > 0$ and $\hat{\boldsymbol{\mu}}$ be chosen according to (2.3.1). The MSE of the leading term of $\hat{\mu}_k(\boldsymbol{\omega}_k)$ for $k = \{1, \ldots, J\}$ is then given by*

$$\begin{aligned} MSE(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k)) &= \left(\sum_{j \neq k} \omega_{kj}(\mu_k - \mu_j) \right)^2 + (1 - \sum_{j \neq k} \omega_{kj})^2 \frac{\sigma_k^2}{np_k} + \sum_{j \neq k} \omega_{kj}^2 \frac{\sigma_j^2}{np_j} \\ &= \boldsymbol{\omega}_k' \Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k' \boldsymbol{\omega}_k + n^{-1} \boldsymbol{\omega}_k' diag(\boldsymbol{\gamma})^{-1} \boldsymbol{\omega}_k \end{aligned} \tag{2.3.4}$$

with $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kJ})'$ s.t. $\boldsymbol{\omega}_k' \boldsymbol{\iota}_J = 1$, Δ_k being the k-th $J \times J$ dimensional partition of $\Delta = (I_J \otimes \boldsymbol{\iota}_J) - (\boldsymbol{\iota}_J \otimes I_J)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)'$ being vector of inverse OLS first stage variances with $\gamma_j = p_j/\sigma_j^2$.

Note that due to the diagonal structure of the Gram matrix of the set of $\mathbf{D}_i$'s, the PCS estimator only depends on the smoothing parameters within its own reference category but is independent from the remaining smoothing parameters. Therefore, optimization of the parameter MSEs can be done group by group contrary to kernel smoothing and ridge regression. The following theorem establishes the MSE optimal smoothing parameters:

Theorem 2.3.1 *For given $k = \{1, \ldots, J\}$, the criterion in Proposition 2.3.1 is minimized at $\boldsymbol{\omega}_k^* = (\omega_{k1}^*, \ldots, \omega_{kJ}^*)$, where*

$$\begin{aligned} \omega_{kj}^* &= \frac{\frac{p_j}{\sigma_j^2} + n \sum_{m \neq k} (\mu_k - \mu_m)(\mu_j - \mu_m) \frac{p_m p_j}{\sigma_m^2 \sigma_j^2}}{\sum_{l=1}^J \frac{p_l}{\sigma_l^2} + n \sum_{l=1}^J \sum_{m \neq k} (\mu_k - \mu_m)(\mu_l - \mu_m) \frac{p_l p_m}{\sigma_l^2 \sigma_m^2}} \\ &= \frac{\gamma_j (1 + n \boldsymbol{\mu}' \Delta_k' diag(\boldsymbol{\gamma}) \Delta_j \boldsymbol{\mu})}{\boldsymbol{\gamma}' \boldsymbol{\iota}_J + \frac{n}{2} \boldsymbol{\mu}' \Delta' M_1 \Delta \boldsymbol{\mu}} \end{aligned} \tag{2.3.5}$$

with $M_1 = diag(\boldsymbol{\gamma}) \otimes diag(\boldsymbol{\gamma})$.

The solution is always unique[7]. The minimizers corresponding to the λ_{kj}'s can

[6] See Extended Appendix A.2.2.

[7] The optimal MSE smoothing parameters satisfy the existence and uniqueness condition for $\hat{\boldsymbol{\mu}}^{PCS}$. For more details consider Extended Appendix A.2.3 and A.2.4.

be found in the Extended Appendix. Note that if the first-stage estimates are all identical for a reference group, the corresponding optimal smoothing parameters become strictly positive. This is in line with Hoerl and Kennard (1970) who show for a generalized ridge regression that MSE optimal smoothing parameters have to be positive in the case of a common target. In the general case, negative smoothing parameters can be optimal due to different group specific targets.

The properties of PCS using oracle weights can be found in Section 2.4. One can construct the oracle weights for the restricted PCS, i.e. (G)RR and kernel regression in a similar fashion, however (weighted) MSE optimal ridge regression and kernel weights will in general not have a closed-form solution for larger J, see Appendix A.1.5.

While the oracle weights are theoretically appealing, they depend on unknown quantities through cell means, variances and probabilities and are generally infeasible. To construct a feasible counterpart we propose to replace the unknown quantities by consistent estimators. The plug-in weights are then given by

$$\begin{aligned}\hat{\omega}_{kj} &= \frac{\frac{\hat{p}_j}{\hat{\sigma}_j^2} + n\sum_{m\neq k}(\hat{\mu}_k - \hat{\mu}_m)(\hat{\mu}_j - \hat{\mu}_m)\frac{\hat{p}_m\hat{p}_j}{\hat{\sigma}_m^2\hat{\sigma}_j^2}}{\sum_{l=1}^{J}\frac{\hat{p}_l}{\hat{\sigma}_l^2} + n\sum_{l=1}^{J}\sum_{m\neq k}(\hat{\mu}_k - \hat{\mu}_m)(\hat{\mu}_l - \hat{\mu}_m)\frac{\hat{p}_l\hat{p}_m}{\hat{\sigma}_l^2\hat{\sigma}_m^2}} \\ &= \frac{\hat{\gamma}_j(1 + n\hat{\boldsymbol{\mu}}'\Delta_k' diag(\hat{\boldsymbol{\gamma}})\Delta_j\hat{\boldsymbol{\mu}})}{\hat{\boldsymbol{\gamma}}'\boldsymbol{\iota}_J + \frac{n}{2}\hat{\boldsymbol{\mu}}'\Delta'\hat{M}_1\Delta\hat{\boldsymbol{\mu}}}\end{aligned} \tag{2.3.6}$$

with $\hat{p}_k = n_k/n$, $\hat{\sigma}_k^2 = \frac{1}{n_k - 1}\sum_{i=1}^{n} D_{ik}(Y_i - \hat{\mu}_k)^2$ and equivalently for $\hat{\boldsymbol{\gamma}} = (\hat{\gamma}_1, \ldots, \hat{\gamma}_J)$, $\hat{\gamma}_k = \hat{p}_k/\hat{\sigma}_k^2$ and $\hat{M}_1 = diag(\hat{\boldsymbol{\gamma}}) \otimes diag(\hat{\boldsymbol{\gamma}})$ with $n_k \geq 2$ for all k. The feasible or plug-in PCS is then given by

$$\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}) = \sum_{j=1}^{J}\hat{\omega}_{kj}\hat{\mu}_j \tag{2.3.7}$$

The idea is that a first step is sufficiently informative for the optimal weights such that using a plug-in estimate will yield an estimated weighting scheme that improves on the actual performance of the resulting estimator. This approach is very close in spirit to other approaches based on MSE optimal averaging, focused information criteria and corresponding averaging estimators such as Hjort and Claeskens (2003), Liu (2015) and Cheng et al. (2015). In Section 2.5 we show that while oracle performance cannot be obtained for arbitrary data generating processes, the plug-in estimator still uniformly dominates the ordinary least squares in terms of (weighted) mean squared error.

Note that in contrast to the model averaging literature (Hansen and Racine, 2012; Liu, 2015) the weights are not restricted to lie in the unit simplex. Under fixed weights, the model averaging and the PCS estimator are linear in the

outcome. For admissibility of linear estimators of the mean of a multivariate normal distribution, Cohen (1966) shows that symmetry and nonnegative eigenvalue bounds have to be met by the linear operator that maps outcomes to predictions, see also Li (1987) in a regression context. Hansen and Racine (2012) show that in the case of nested linear regression models, positivity of the model weights is a necessary condition for admissibility under mean squared error loss. However, if data dependent weights are used, the resulting estimator is no longer linear in the outcome. Furthermore, the less restrictive weighting scheme of the PCS can contradict the nesting requirement by Hansen and Racine (2012) despite the fact that the submodels are effectively linear. As a consequence, the overall shrinkage sum and not each shrinkage parameter is bounded from below and thus inadmissibility of the PCS does not follow. Interestingly, the eigenvalue conditions of Cohen (1966) for admissibility still hold with probability one for both optimal and feasible PCS[8].

2.4 Large Sample Theory

2.4.1 Local Parameterization

In the following, we derive and discuss the large sample properties of the different weighting schemes and of the PCS estimator over a sufficient class of data generating processes relevant in the context of group means. In particular, we would like to distinguish between systems of locations in which the differences between group means are small (*close* systems) and large (*distant* systems) for a given sample size. For simplicity, instead of invoking standard assumptions to assure consistency and asymptotic normality of the first stage through moment conditions or similar, we start from a less rigorous point[9]. Let $\mathcal{F}$ be the set of distribution functions.

Definition 1 *A sequence of data generating processes* $\{F_n\}$ *is close with local parameter* $\boldsymbol{\delta} \in \mathbb{R}^J$ *if*[10]

$$\{F_n\} \in S(\boldsymbol{\delta}, V_0)$$
$$S(\boldsymbol{\delta}, V_0) = \{\{F_n\} : F_n \in \mathcal{F}, \sqrt{n}\Delta\boldsymbol{\mu} \to \Delta\boldsymbol{\delta} \in \mathbb{R}^{J^2},\ \sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z},\ \hat{V} \xrightarrow{p} V_0\}.$$

[8] See Extended Appendix A.2.6.

[9] Note that standard regularity conditions usually imply asymptotic normality for estimated cell probabilities and variances as well. However, this is not required for any of the results in this and the following subsection.

[10] Regarding the notation in Definition 1, the mean vector $\boldsymbol{\mu}$ should have a subscript n as it becomes sample size dependent in this case. We omit the subscript for readability.

Definition 2 *A sequence of data generating processes* $\{F_n\}$ *is distant if*

$$\{F_n\} \in S(\infty, V_0)$$
$$S(\infty, V_0) = \{\{F_n\} : F_n \in \mathcal{F},\ \sup_{k,j} \sqrt{n}|\mu_k - \mu_j| \to \infty,\ \sqrt{n}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z},\ \hat{V} \xrightarrow{p} V_0\}$$

with $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_J)'$, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, V_0)$, $\hat{V} = diag(n\hat{\sigma}_j^2/n_j)$ and $V_0 = diag(\sigma_j^2/p_j)$. Close systems require that all scaled pairwise differences do not diverge, i.e. their differences depend on the local parameters $\delta_k - \delta_j$.[11] This nests the case in which all means are exactly identical and the local parameters are zero. For distant systems we require the scaled differences to go to infinity for at least one pair in the system. The union of these systems is sufficiently rich to describe a wide range of data generating processes.

To further motivate these classes of sequences and in particular the rate at which the differences converge to the local parameters, consider J locations that are estimated via least squares. Assume that the asymptotic variances are known. Let Z_n be a random variable that converges in distribution to a standard normal. A simple test for equality of two means μ_k and μ_j can be rewritten as follows:

$$T_n = \sqrt{n}\frac{\hat{\mu}_k - \hat{\mu}_j}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} = \sqrt{n}\frac{\mu_k - \mu_j}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} + Z_n \tag{2.4.1}$$

Using the local parameterization it follows that

$$T_n(F_n) \xrightarrow{d} \mathcal{N}\left((\delta_k - \delta_j) \Big/ \sqrt{\sigma_k^2/p_k + \sigma_j^2/p_j}, 1\right) \text{ if } \{F_n\} \in S(\boldsymbol{\delta}, V_0), \tag{2.4.2}$$
$$P(|T_n(F_n)| > c) \to 1 \text{ for all } c > 0 \qquad \text{if } \{F_n\} \in S(\infty, V_0). \tag{2.4.3}$$

Therefore, depending on the local parameter difference $\delta_k - \delta_j$, one can obtain a small, moderate or even large mean of the test statistics' distribution. In the special case of $\delta_k - \delta_j$ being exactly equal to zero, the local parameterization does no longer affect the asymptotic distribution and classical inference can be conducted using the standard normal distribution. It is apparent that in any other case, choosing a model based on such a test might be misleading. If the local parameter is at a size that centers the limiting distribution e.g. around the critical value used for rejection of the null hypothesis, rejection would occur with probability one half. If this pretest is used for model selection, it is likely to suggest an underparameterized model that translates into higher parameter risk. The PCS estimator can be considered as a smooth variant of such a classical

[11] Note that since the system effectively depends only on differences in local parameters, constant shifts to $\boldsymbol{\delta}$ do not affect the analysis.

pretesting based estimator. Hence, we expect it to perform better exactly in these regions in which type-II errors are relatively large. Standard asymptotic analysis, however, will always favor the more parameterized model except if parameters are exactly equal. Thus, the approximations based on the local asymptotic framework should be closer to the actual finite sample behavior. The intuition can directly be translated to simultaneous tests of equality in locations for more than two groups[12].

2.4.2 Distributional Theory

For investigation of the large sample properties of the PCS, the behavior of the smoothing parameters along the sequences of DGPs is crucial. We consider PCS with weights ω_{kj}^f that correspond to fixed penalty parameters λ_{kj} in (2.2.2), MSE optimal ω_{kj}^* and plug-in parameters $\hat{\omega}_{kj}$. The following lemma demonstrates the behavior of the different weighting schemes in large samples.

Lemma 2.4.1 *Let ω_{kj}^f, ω_{kj}^* and $\hat{\omega}_{kj}$ denote the PCS weights in (2.3.7) corresponding to fixed*[13]*, MSE optimal according to (2.3.5) and plug-in solutions according to (2.3.6). Their limiting behavior along the local parameterization is then given by:*

$$\omega_{kj}^f = O_p(n^{-1}) \text{ if } k \neq j \text{ and } \omega_{kk}^f = 1 + O_p(n^{-1}) \qquad \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0)$$

$$\omega_{kj}^* \to \bar{w}_{kj} = \frac{\gamma_j(1 + \boldsymbol{\delta}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\delta})}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}\boldsymbol{\delta}'\Delta' M_1 \Delta\boldsymbol{\delta}} \qquad \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0)$$

$$\omega_{kj}^* \to \bar{w}_{kj} = 2\gamma_j \frac{\boldsymbol{\mu}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\mu}}{\boldsymbol{\mu}'\Delta' M_1 \Delta\boldsymbol{\mu}} \qquad \text{if } \{F_n\} \in S(\infty, V_0)$$

$$\hat{\omega}_{kj} \xrightarrow{d} w_{kj}^a = \frac{\gamma_j(1 + (\mathbf{Z} + \boldsymbol{\delta})'\Delta_k diag(\boldsymbol{\gamma})\Delta_j(\mathbf{Z} + \boldsymbol{\delta}))}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}(\mathbf{Z} + \boldsymbol{\delta})'\Delta' M_1 \Delta(\mathbf{Z} + \boldsymbol{\delta})} \qquad \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0)$$

$$\hat{\omega}_{kj} \xrightarrow{p} \bar{w}_{kj} = 2\gamma_j \frac{\boldsymbol{\mu}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\mu}}{\boldsymbol{\mu}'\Delta' M_1 \Delta\boldsymbol{\mu}} \qquad \text{if } \{F_n\} \in S(\infty, V_0)$$

with Δ_k being the k-th $J \times J$ dimensional partition of $\Delta = (I_J \otimes \boldsymbol{\iota}_J) - (\boldsymbol{\iota}_J \otimes I_J)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_J)'$ being vector of inverse OLS first stage variances with $\gamma_j = p_j/\sigma_j^2$.

Note that the MSE optimal smoothing parameters do not in general vanish asymptotically, i.e. there is potential aggregation even in the limit. This, however, does not exclude the possibility to completely smooth out uninformative groups in large samples. It is qualitatively different from the smoothing kernel approach where uninformative, i.e. conditionally independent, regressors are

[12] Extended Appendix A.2.5 contains an example using a Wald test for equality of all means.

[13] Please note that the smoothing parameters λ_{kj}'s are fixed. The weights ω_{kj}^f's can then be seen as a function of the fixed smoothing parameters and the sample size.

always smoothed to a global average with smoothing parameters converging to their upper bound (Hall et al., 2004, 2007). The estimated smoothing parameters converge in distribution to a function of a normal random vector if groups are locally close while under a distant system, they converge in probability to the oracle parameters. Thus, adding a single distant parameter to a locally close system is sufficient to obtain convergence in probability. This is due to the fact that the effective shrinkage target in the PCS are weighted leave the k-th group out averages. If the k-th group is the distant one, a weighted combination of the set of locally close locations will be distant enough to pin down the optimal weights in probability. If the k-th group is within the set of the locally close groups, the leave the k-th group out average will contain the distant group which is sufficient in large samples to distinguish the shrinkage target from the reference mean and thus lead to probabilistic convergence. Only in the case of all groups being locally close, the differences are not sufficient such that the limiting behavior is governed by a continuous function of a random normal vector. However, due to the rate of convergence of the estimated smoothing parameters in the case of distant systems, they will have an effect on the first order term determining the limiting distribution of the PCS under estimated smoothing parameters compared to the oracle distribution.

The following theorem establishes the distributional behavior of the different PCS variants.

Theorem 2.4.1 *Let $\boldsymbol{\omega}_k^f$, $\boldsymbol{\omega}_k^*$ and $\hat{\boldsymbol{\omega}}_k$ denote the vector of fixed, MSE optimal according to (2.3.5) and plug-in weights according to (2.3.6) for the PCS. The asymptotic distributions of the PCS estimators are given by*

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}^f) - \mu_k - B_{1k}(\boldsymbol{\omega}^f)) \xrightarrow{d} Z_k \sim \mathcal{N}\left(0, \frac{\sigma_k^2}{p_k}\right) \quad \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0)$$

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}^*) - \mu_k - B_{2k}(\boldsymbol{\omega}^*)) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=1}^{J} \bar{\omega}_{kj}^2 \frac{\sigma_j^2}{p_j}\right) \quad \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0)$$

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}) - \mu_k) \xrightarrow{d} \sum_{j=1}^{J} \omega_{kj}^a Z_j + \sum_{j=1}^{J} \omega_{kj}^a (\delta_j - \delta_k) \quad \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0)$$

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}) - \mu_k - B_{3k}(\bar{\boldsymbol{\omega}})) \xrightarrow{d} Z_k \sim \mathcal{N}\left(0, \frac{\sigma_k^2}{p_k}\right) \quad \text{if } \{F_n\} \in S(\infty, V_0)$$

with $B_{1k}(\boldsymbol{\omega}^f) = \sum_{j \neq k} \omega_{kj}^f (\mu_j - \mu_k), B_{2k}(\boldsymbol{\omega}^*) = \sum_{j \neq k} w_{kj}^* \ (\mu_j - \mu_k)$, $B_{3k}(\bar{\boldsymbol{\omega}}) = \sum_{j \neq k} \bar{\omega}_{kj} (\mu_j - \mu_k)$.

Theorem 2.4.1 contains the asymptotic distributions of the different PCS estimators. There are some results that clearly parallel the literature on (generalized) ridge regression. In particular, a fixed penalty is asymptotically negligible for the

distribution and thus the efficiency of the estimator, i.e. PCS with fixed weights converges in distribution to the corresponding OLS limit. The PCS under optimal weights differs in terms of its distribution from the OLS and thus the improvements in MSE in general do not disappear even for large samples. The behavior of the PCS under estimated smoothing parameters is particularly noteworthy. Note that the distribution to the normal is not uniform along all sequences of DGPs. In particular, the PCS with estimated smoothing parameters under locally close systems converges in distribution to a sum of normal random variables and local parameters multiplied by the limiting weights that are themselves functions of the same random normal variables, local parameters and other features of the DGP. Thus, the limiting distribution in close systems is in general different from the normal. Assessing or estimating that limiting distribution has to be done with caution as the local parameters cannot be estimated consistently due to the $\sqrt{n}$ multiplier. This is similar to other shrinkage and model averaging methods that rely on smooth aggregation methods in the spirit of James-Stein shrinkage and frequentist model averaging (Hjort and Claeskens, 2003; Liu, 2015; Cheng et al., 2015; Hansen, 2016a).

2.5 Asymptotic Risk

Theorem 2.4.1 shows that when evaluating the risk of the feasible PCS, one has to take the additional variation of the weights under locally close systems into account. In the following, we will focus on the risk under close systems as from there distant systems can be obtained as a special case by letting $||\Delta\boldsymbol{\delta}||_\infty \to \infty$. For derivation of the oracle risk, recall that due to the flexibility of the PCS, optimization can be done separately for each individual group. When evaluating the risk of the feasible PCS parameters however, the additional (co)variation introduced by the weighting parameters has to be taken into account since the latter are functions of the same random vector. The choice for the joint loss function will be the (weighted) parameter vector MSE

$$l(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu}) = (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})' W (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}) \tag{2.5.1}$$

with the canonical weighting matrix being the inverse of the asymptotic variance of the OLS or MLE parameter vector $W = diag(\boldsymbol{\gamma})$. Thus W is proportional to the identity under homoskedasticity and equal group probabilities. The choice of W also renders the evaluation of the risk invariant to rotations of the parameter vector, such that PCS risk properties are preserved even if outcomes are not generated on the same scale across groups.

To assure existence of a criterion that properly approximates the risk, a trimmed expected scaled loss criterion is used with vanishing trimming boundaries as in Hansen (2016a). Alternatively one could impose additional moment assumptions along the sequences to assure uniform integrability of the scaled loss function. In the limit, the risk is determined by a function of random normal vectors and thus easier to compute through the distributional limit. Let $\hat{\boldsymbol{\mu}}_n$ be a sequence of estimators along $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. The asymptotic risk along the sequences of DGPs is then given by

$$\rho(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}) = \lim_{\zeta \to \infty} \liminf_{n \to \infty} E_{F_n}[\min\{nl(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}), \zeta\}]. \tag{2.5.2}$$

Note that under normality, this would collaps to the exact finite sample risk. The asymptotic risk of the OLS $\hat{\boldsymbol{\mu}}$ is then given by

$$\rho(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) = tr(WV_0).$$

The following theorem provides the asymptotic risk of the PCS estimator under estimated weights for close systems of locations.

Theorem 2.5.1 *Let $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. If $\hat{\boldsymbol{\omega}}$ is chosen according to (2.3.6), then*

$$\begin{aligned}\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) =& E\left[\frac{(\mathbf{Z}+\boldsymbol{\delta})'\Delta'\{M_2 - tr(\Delta' M_3 V_0) M_1 + 2M_3 V_0 \Delta' M_1\}\Delta(\mathbf{Z}+\boldsymbol{\delta})}{(tr(V_0^{-1}) + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_1 \Delta(\mathbf{Z}+\boldsymbol{\delta}))^2}\right] \\ &+ tr(WV_0) - 2tr(V_0^{-1}) tr(\Delta' M_3 V_0) E\left[\frac{1}{(tr(V_0^{-1}) + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_1 \Delta(\mathbf{Z}+\boldsymbol{\delta}))^2}\right]\end{aligned}$$

with $V_0 = diag(\boldsymbol{\gamma})^{-1}$, $M_1 = diag(\boldsymbol{\gamma}) \otimes diag(\boldsymbol{\gamma})$, $M_2 = diag(\boldsymbol{\gamma}) \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}'$ and $M_3 = diag(\boldsymbol{\gamma}) \otimes \boldsymbol{\gamma}$.

Thus, the asymptotic risk is given by the sum of the OLS risk, the expectation of the ratio of a quadratic form and a strictly positive random variable, and a strictly negative term. It depends on the limiting vector of the OLS, its asymptotic variance components and the unknown local parameter vector $\boldsymbol{\delta}$.

Theorem 2.5.1 allows us to establish sufficient conditions for a *strict uniform dominance* of the PCS estimator compared to the OLS. Uniform in the sense that for all bounded $\boldsymbol{\delta}$ vectors, the risk is strictly smaller than the risk of the OLS. It turns out that an easily interpretable sufficient condition is that the number of groups has to exceed three, i.e. we obtain the following corollary:

Corollary 2.5.1 *Let $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$ and $\hat{\boldsymbol{\omega}}$ be chosen according to (2.3.6). If $J \geq 4$, then*

$$\sup_{\boldsymbol{\delta} \in B} \quad \rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) < 0 \tag{2.5.3}$$

for any bounded $B \subset \mathbb{R}^{J^2}$.

Thus, a sufficient (but by no means necessary) condition for uniform dominance is a simple condition on the dimensionality of the mean vector. This is similar to classical James-Stein estimation that as a necessary condition requires at least a three-dimensional multivariate mean vector when shrinking to a fixed target for global risk reduction over the maximum-likelihood estimator (Stein, 1956). Here, the somewhat more flexible shrinkage target requires one additional dimension, i.e. at least four groups to assure a strictly smaller risk for any close system of locations. Consistency follows directly as a corollary. In a similar spirit, the PCS shrinks towards a restricted subspace, i.e. an estimator that equalizes the group locations under a generalized error term structure. The corresponding subspace has exactly dimensionality $l = 1$ thus the minimal condition for superior risk (Oman, 1982) is that $J \geq 3 + l$ which equals the sufficient condition from Corollary 2.5.1.

2.6 Monte Carlo Study

The following simulations compare the small sample behavior of the PCS estimator to potential alternatives over a large range of data generating processes that vary with respect to mean parameters and error variances across groups. We investigate the weighted parameter vector MSE under close systems for different local parameter values δ. The distant system behavior can be inferred for large values of δ. The following estimators are considered:

1. Ordinary least squares/frequency method (OLS),
2. pairwise cross-smoothing with plug-in smoothing parameters (PCS),
3. ridge regression estimator with the plug-in smoothing parameter (RR),
4. generalized ridge regression estimator with plug-in smoothing parameters (GRR),
5. nonparametric smoothing kernel with plug-in smoothing parameters (Kernel),
6. a selection/pretesting estimator based on Mallows C_p[14] (Mallows).

We study a setup with a moderate number of groups, i.e. $J = 4$, that are selected with equal likelihood. The three designs A, B and C are set such that mean vectors converge to the origin but vary in the degree of deviations from the origin

[14]We consider all possible submodels and choose the one with the lowest criterion value according to Mallows (1973). We also experimented with generalizations that are robust with respect to different error term structures. However, the classical C_p seems to dominate all adaptations in our simulations and thus results are omitted.

in finite samples. The mean vectors take following values $\boldsymbol{\mu}_A = (0, 0, 0, \delta/\sqrt{n})'$, $\boldsymbol{\mu}_B = (0, 0, -3\delta/\sqrt{n}, \delta/\sqrt{n})'$ and $\boldsymbol{\mu}_C = (0, 2\delta/\sqrt{n}, -3\delta/\sqrt{n}, \delta/\sqrt{n})'$ with δ varying over a positive grid starting at 0. For $\delta = 0$, we are in the case of identical means, i.e. a global average would be the most efficient estimator. Regarding the error term, we consider homoskedastic and heteroskedastic standardized log-normal distributions[15]. All results are based on 5000 simulations. For the plug-in weights we use the formulas from (2.3.6). In the homoskedastic designs, we change the variance estimator as if we knew in advance that the error is homoskedastic, i.e. all residuals are used to estimate a single variance parameter for all groups.

Weighted mean squared errors relative to OLS for $n = 400$ are reported in Figure 2.6.1. Results for other sample sizes follow the same patterns and are therefore omitted. First, note that GRR, RR and C_p estimator do not always outperform the OLS in the chosen setups. Only PCS and kernel estimators seem to uniformly dominate OLS. Depending on the value of δ one can get up to 30% improvement in the parameter vector MSE by using PCS over OLS. The largest benefits are obtained at lower values of δ as in these settings the shrinkage estimator can benefit from taking information from the other similar groups.

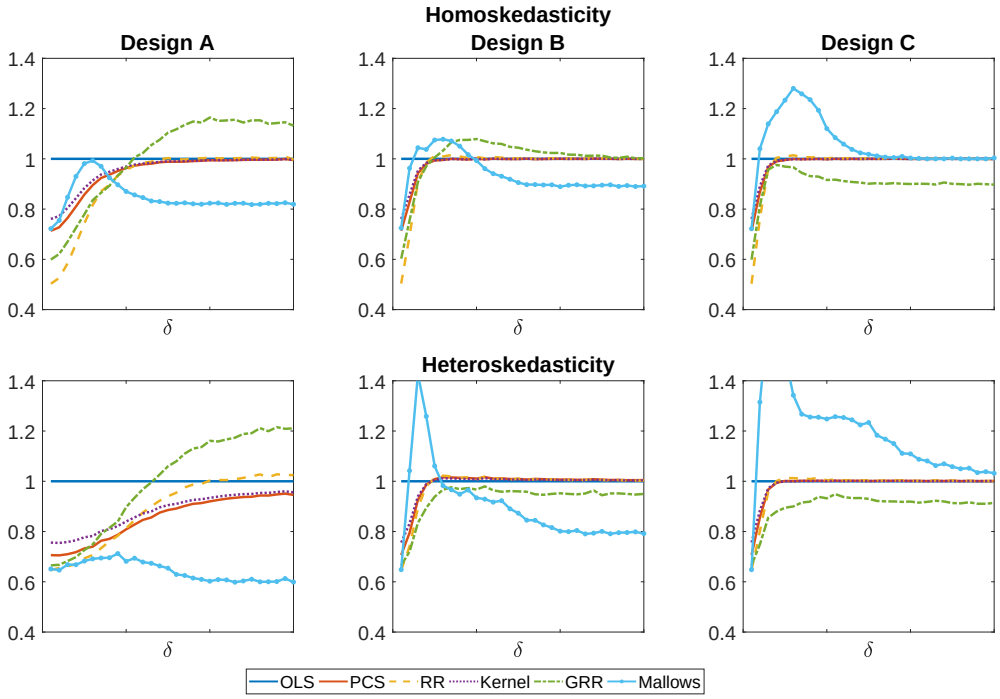
The kernel estimator follows virtually the same pattern as the PCS especially for larger δ parameters. For smaller δ values, however, kernel performs worse (up to 5 percentage points) due to its restricted flexibility when exploiting designs of almost equal means.

The RR generally performs well especially for small δ yielding MSE improvements up to 50% compared to OLS. In these designs, RR seems to benefit from the smaller amount of penalty parameters in comparison to PCS which has to deal with additional estimation noise. Nevertheless, RR loses its advantage in heterogeneous designs where the more flexible methods can further improve on the parameter vector MSE, i.e. under heteroskedasticity it even performs similar or slightly worse than OLS for moderate δ values in design A.

The GRR performs poorly in design A for large values of δ as it tends to shrink three groups to the rather distant leave the k-th group out targets in finite samples. Thus GRR estimation lacks robustness regarding the shrinkage target in asymmetric designs. It can perform substantially worse than ordinary least squares for a large range of δ values inflating the risk by more than 20%. Introducing a higher degree of symmetry around the origin in designs B and C helps GRR to shrink to correct targets and improves its performance. However, as these symmetries are usually unknown, we do not recommend the use of the plug-in GRR for shrinking categorical regressors in applied work.

[15] All results are robust with respect to the error distribution. Results for the normal distribution do not differ qualitatively and are therefore omitted.

Figure 2.6.1: Relative Weighted Parameter Vector Mean Squared Errors



The figure depicts the simulated relative weighted parameter vector mean squared errors for the estimators with plug-in risk-optimal shrinkage parameters under log-normally distributed errors for $n = 400$, equal selection probabilities, different parameter values δ, and different structures of the error variance with OLS as a benchmark. The parameter vectors are $\sqrt{n}\boldsymbol{\mu}_A = (0, 0, 0, \delta)'$, $\sqrt{n}\boldsymbol{\mu}_B = (0, 0, -3\delta, \delta)'$ and $\sqrt{n}\boldsymbol{\mu}_C = (0, 2\delta, -3\delta, \delta)'$. The group variances are $\boldsymbol{\sigma}^2_{hom} = (1, 1, 1, 1)'$ and $\boldsymbol{\sigma}^2_{het} = (1, 1, 1, 10)'$. Simulations are based on 5000 replications.

The pretest estimator based on Mallows C_p has its worst performance for moderate values of δ as in this range it is challenging for the model selection criterion to detect the optimal aggregation strategy. It often yields an underfitted model that introduces too much bias into the parameter estimates in line with the discussion in Section 2.4.1. This is particularly prominent in designs B and C with risk inflations of over 60% compared to OLS. With a higher degree of deviations from the origin (design C), the C_p estimator performs worse than OLS over a larger range of δ parameters. However, for extreme values of δ parameters, it can perform better or close to OLS if there are risk gains from aggregating identical groups (design A and B). In practice, these mean differences are usually unknown and thus using the model selection criterion can be detrimental to the estimation. The presence of data generating processes for which model selection criteria yield inferior risk is a well-known phenomenon in the literature on model selection and post-selection risk, see e.g. Leeb and Pötscher (2008) among many others.

PCS on the other hand is virtually never worse than OLS, i.e. it shows uniformly dominant behavior in line with our results in Section 2.5. However, PCS

is not always beating all the competitors over all the designs and δ values. For example, for small δ values GRR and RR are up to 20 percentage points superior profiting from an accurate shrinkage target in the design of almost equal means and less penalty parameters to estimate. For large values of δ, the pretest estimator can perform better than PCS as the risk optimal model can be obtained through strict aggregation. PCS and kernel estimator show similar behavior and both show robustness in terms of different designs. From the two, PCS slightly outperforms the kernel estimator for smaller values of δ.

Therefore, PCS seems to be a robust refinement over OLS for a wide range of DGPs as alternatives are either dominated by PCS (OLS and kernel) or are very design sensitive (GRR and C_p). For the first category, PCS seems to be particularly superior for small values of δ. For the latter category, there are always designs in which they perform (substantially) worse than OLS. Note that in general there is still room for further improvement since the large risk gains that can be obtained by the theoretically optimal PCS (see Section 2.2.1, Figure 2.2.1) cannot be reached by any method considered in the simulations designs.

2.7 Applications

2.7.1 Application I: A Fine is a Price

Gneezy and Rustichini (2000a) investigate the prediction of the deterrence hypothesis, i.e. that ceteris paribus introducing fines will decrease the likelihood of the associated action or behavior. They run a randomized control treatment study at ten day-care centers for young children in Haifa, Israel over a period of twenty weeks. It can be seen as a small panel data set with ten observations and twenty time periods. In period five, a fine is introduced for parents that come too late to pick up their children in six of these centers. They find that the fine increases the number of delayed parents and even after removal of the fine, the rate stayed at the same, higher level. The results have also been quoted in the literature on intrinsic and extrinsic motivation and crowding-out effects (Gneezy et al., 2011). Most of their major findings are summarized in a plot similar to the first subplot in Figure 2.7.1 which has been reused by e.g. Gneezy and Rustichini (2000b). In the variant used here, it depicts the share of late arrivals in both, treatment and control group over the duration of twenty weeks. Note that each point is an average over the subgroups of six and four data points in treatment and control group respectively which are basically predictions of a simple panel

data model[16]. In statistical terms, it contains estimates for the expected share of late arrivals conditional on time period and treatment status. Our method is well-suited for this application since by construction, there are small orthogonal groups that are determined by time and treatment status. We stabilize the estimates of the conditional means by using the plug-in PCS within treatment groups and time periods closely related to Gneezy and Rustichini (2000a), Table 2. Hence we smooth the averages within weeks 1-4, 5-8, 9-16 and 17-20 for both groups using the original means as first stage[17].

Figure 2.7.1: Mean Share of Late Arrivals, OLS and PCS estimates

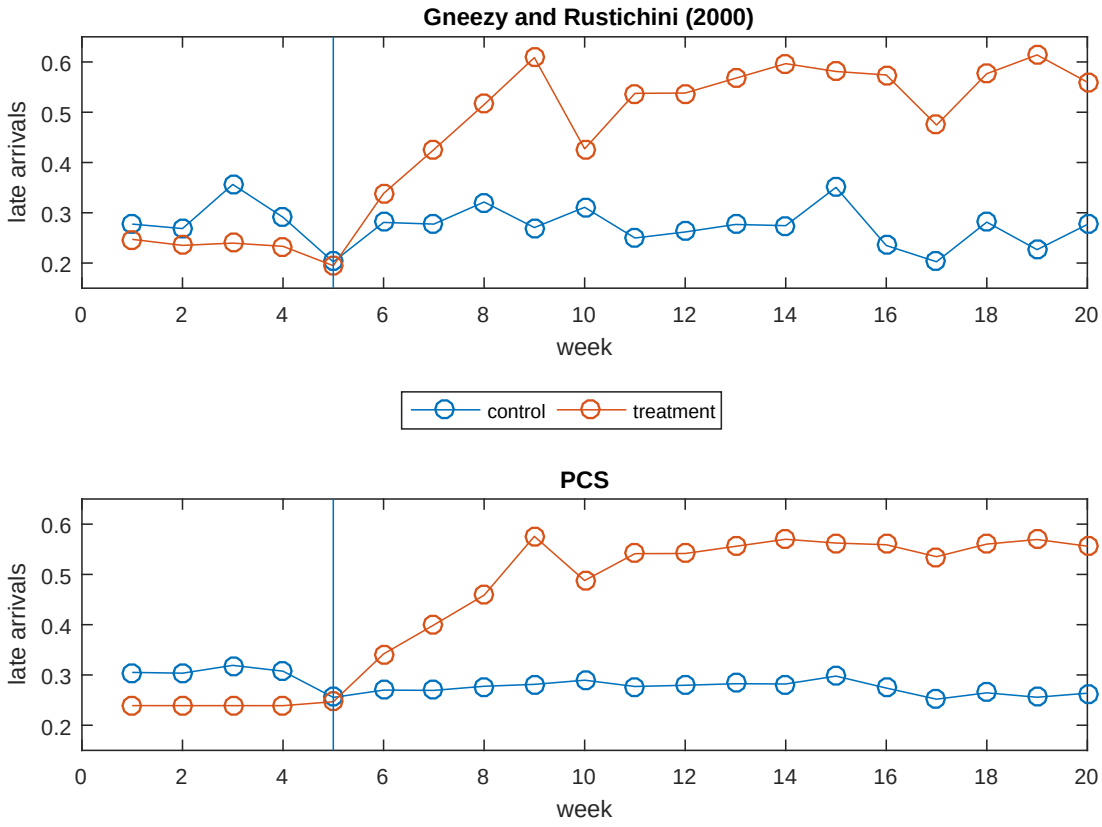


Figure 2.7.1 depicts OLS (Gneezy and Rustichini, 2000a) and PCS estimates for the conditional mean over time and treatment status. The major findings of the original visualization are confirmed. In fact, our estimates reveal the pattern much clearer since the PCS suggests a more stable share of the control group and a less fluctuating mean of the treatment group before and after the time of treatment.

[16]Note that if only time trend dummy variables are used, a pooled OLS, fixed effect and random effect models coincide.

[17]This is based on the prior characterization of the periods by Gneezy and Rustichini (2000a). Of course other smoothing strategies could be employed as well.

2.7.2 Application II: Minimum Wage Study

The Card and Krueger (1994) paper is a case study evaluating the effects of minimum wage increase on the employment of low-wage workers. They collected data from fast food chains in New Jersey and Pennsylvania in a telephone survey before and after a minimum wage increase in New Jersey from 4.25$ to 5.05$ in 1992. The dependent variable full-time employment equivalent is measured as the number of full-time workers plus 0.5 times the part-time workers.

The setup is well-suited for our method since there are four orthogonal groups by construction that are determined by state and time. The PCS estimator is applied to the difference-in-differences model on the original Card and Krueger (1994) data and for each fast food chain separately to account for potentially different time trends and heterogeneous effects on employment across chains. As mentioned in Card and Krueger (1994), KFC differs in its size, opening hours and type of food from the other chains and therefore might be a source of heterogeneity. The chain by chain analysis further reduces the observations per cell and thus benefits the application of PCS over OLS. An alternative strategy would be to also smooth across chains to further increase the possible number of shrinkage targets.

The OLS (Card and Krueger, 1994) and PCS results for pooled data and for each chain separately are reported in Table 2.7.1[18]. We find a positive significant change in employment for the pooled data. As further analysis shows, this result is driven by the significant positive employment change in Burger King. The other chains have no significant change in employment. Comparing the results across chains, KFC shows a different pattern from the other stores, as KFC is the only chain with a point estimate that is in line with the theory of increasing labor demand in a less labor costly environment, however the estimate is statistically insignificant.

All the estimated effects of the minimum wage on the employment are closer to zero for the PCS in comparison to the OLS. In the case of pooled data, Burger King, KFC and Roys, the differences between OLS and PCS are not as large. However in case of Wendys, the chain with smallest number of observations in the data set, the difference is more pronounced, showing the stabilizing property of PCS in such scenarios.

[18]Table A.1.1 in Appendix A.1.9 includes means, variances and number of observations for all subgroups.

Table 2.7.1: Mean and Difference-in-Differences Estimates

All Chains	OLS		PCS	
	NJ (t)	PEN (c)	NJ (t)	PEN (c)
B	20.44	23·33	20.53	22·87
A	21.03	21·17	21.01	21·12
DiD	2·75*		2·22*	
Burger King	OLS		PCS	
	NJ (t)	PEN (c)	NJ (t)	PEN (c)
B	22.16	29·42	22.25	29·06
A	23.63	26·22	23.63	26·06
DiD	4·67**		4·38**	
KFC	OLS		PCS	
	NJ (t)	PEN (c)	NJ (t)	PEN (c)
B	12.79	10·71	12.76	10·92
A	13.73	13·00	13.60	12·96
DiD	−1·35		−1·20	
Roys	OLS		PCS	
	NJ (t)	PEN (c)	NJ (t)	PEN (c)
B	23.14	19·74	22.99	19·80
A	21.73	15·81	21.68	16·12
DiD	2·52		2·37	
Wendys	OLS		PCS	
	NJ (t)	PEN (c)	NJ (t)	PEN (c)
B	22.08	24·12	22.43	23·46
A	23.40	22·10	23.10	22·44
DiD	3·35		1·69	

The table contains the mean estimates of the full-time employment equivalents. B = before the minimum wage increase, A = after the minimum wage increase, NJ = New Jersey, PEN = Pennsylvania, t = treated, c = control. DiD formula: $(\hat{\mu}_{NJ,A} - \hat{\mu}_{NJ,B}) - (\hat{\mu}_{PEN,A} - \hat{\mu}_{PEN,B})$. p-values (* < 0.1, ** < 0.05) are computed using conventional confidence bounds.

2.8 Concluding Remarks

Pairwise cross-smoothing provides a unifying framework to analyze and compare smoothing methods for categorical data that nests different approaches from the literature on (generalized) ridge regression, nonparametric smoothing kernels and model averaging. It penalizes L_2 differences between estimation parameters and first-stage estimates. The estimator can be easily implemented with standard software packages using the closed-form solutions derived in this paper. It has favorable risk properties compared to the ordinary least squares and other commonly used approaches. For future research, relaxing the assumption of a fixed number of groups, i.e. allowing for J to grow with the sample size with closeness restrictions that are related to sparsity in the sense of few different locations and alternative risk functions in the sense of (Hansen, 2016a) should be considered.

3 Detecting Structural Breaks using a Fusion Lasso Penalty

3.1 Introduction

In this paper, structural breaks are detected by adding a fusion penalty introduced by Land and Friedman (1997) to a linear regression model with time-varying parameters. The idea of the fusion penalty is to shrink differences of parameters consecutive in time to zero and at points at which these differences are non-zero structural breaks are detected. The optimal choice of shrinkage is an important issue. A non-optimal choice of shrinkage leads to underestimation or overestimation of the number of breaks and/or extremely biased estimates of the parameters. One possibility how to choose the optimal level of shrinkage is to apply an information criterion. Two criteria are chosen from the literature, the Extended Bayesian Information Criterion from Chen and Chen (2008) which is designed for shrinkage selection in high-dimensional models with polynomially increasing number of parameters and the Information Criterion introduced in Qian and Su (2016) for penalized linear regression model with time-varying parameters. The criterion in Chen and Chen (2008) is shown to be consistent for the regularization parameter and the criterion in Qian and Su (2016) is shown to detect the number of breaks consistently. The finite sample simulation study in this paper shows that they choose models in which the estimated parameters are strongly biased. Therefore, new criteria are introduced here which perform better regarding the bias and have a good performance regarding the detection of the correct number of breaks and their relative distance to the true positions.

The above mentioned regularization approach based on the fusion penalty is already implemented for detecting structural breaks in one-dimensional piecewise constant signals in Harchaoui and Lévy-Leduc (2010). Bleakley and Vert (2011) extend this approach to detect multiple structural breaks shared by a set of co-occurring one-dimensional piecewise constant signals. Chan et al. (2013) implement fusion penalty in structural break autoregressive (SBAR) models. Qian and Su (2016) consider a standard linear regression model and use a group lasso norm to penalize differences between coefficients to find the breaks, then they implement a post-estimation procedure to smooth out noisy breaks and at the end they estimate the coefficients of each regime by ordinary least squares (OLS). In comparison to Qian and Su (2016), the implementation in this paper will use an adaptive version of a fusion penalty to avoid the post-estimation smoothing

procedure.

In the simulation study, the regularization approach is compared to widely used tests of Bai and Perron (1998) who extend the supremum test of Andrews (1993) for a single break to detect multiple breaks. Their first test tests a null hypothesis of parameter stability against a presence of a fixed number of breaks. To relax this restriction of a fixed number of breaks, Bai and Perron (1998) also introduce a double maximum test, in which the null hypothesis of no break is tested against a presence of a fixed maximum number of breaks. Moreover, Bai and Perron (1998) propose a test with a null hypothesis of ℓ breaks against $\ell + 1$ breaks and based on this single test, they introduce a sequential method of estimating the number of structural breaks. Critical values for these tests are tabulated in Bai and Perron (2003). A disadvantage of these tests is that the range of potential structural breaks is restricted to get reliable results. This restriction may cause an inability to detect an early and/or a recent break point. The results of the simulations interestingly indicate that the regularization approach seems to detect also an absolute position of a break at an extreme position even though there are no theoretical guarantees for it. Otherwise, detection of the relative position of the break in the time series improves with the length of the time series as suggested by the theoretical results.

In comparison to other approaches to detect breaks, the potential advantages of the fusion penalty may be summarized as follows. The times of the breaks do not need to be known in advance. The maximum number of true breaks which can be detected grows with a logarithm of the time series length. This is less restrictive than in Bai and Perron (1998) who assume a fixed number of breaks at asymptotically distinct time points. By using an adaptive fusion penalty, the post-model selection estimation of the parameters could be avoided. Nevertheless, an alternative parameter estimation strategy could be to use positions of the breaks found by the fusion penalty as a starting point for an post-model selection estimation of the parameters by OLS.

The chapter is organized as follows. The Section 3.2 introduces the model, the estimation method and the information criteria for choosing the optimal amount of shrinkage. The Section 3.3 includes results of the simulation study which compares the regularization approach with tests from Bai and Perron (1998). The Section 3.4 describes the application. The Section 3.5 concludes.

3.2 Estimating Breaks by L_1-norm Regularization

3.2.1 Model

The following model is considered:

$$y_t = x_t'\beta_t + \varepsilon_t, \qquad t = 1, \dots, T, \tag{3.2.1}$$

where y_t is a centered dependent variable, x_t is a $p \times 1$ vector of standardized regressors, β_t is a $p \times 1$ vector of unknown coefficients allowed to vary across time and ε_t is an error term which is assumed to have a zero mean and at least a finite fourth moment. In a matrix notation, the model looks as follows:

$$\underset{T\times 1}{y} = \underset{T\times Tp}{X}\ \underset{Tp\times 1}{\beta} + \underset{T\times 1}{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_T \end{pmatrix} = \begin{pmatrix} x_1' & 0 & 0 & \cdots & 0 \\ 0 & x_2' & 0 & \cdots & 0 \\ 0 & 0 & x_3' & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & x_T' \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_T \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_T \end{pmatrix}.$$

Let us denote the variance-covariance matrix of the error term, $V[\varepsilon]$, as Σ.

In a presence of structural breaks, two types of structural breaks can occur in model (3.2.1): a break in the mean function, i.e. $\beta_t \neq \beta_{t+1}$ or a break in variance-covariance structure of the error terms. Let us assume for now that Σ is known. Without imposing any restrictions on the parameters, there are Tp parameters to estimate in (3.2.1). In this case, the number of parameters exceeds the number of observations, particularly $Tp \geq T$, i.e. (3.2.1) is a high-dimensional model. To reduce the dimensionality of the model, the assumption of stable parameters will be imposed. Then, in the case of m breaks the parameters satisfy the following restrictions:

$$\begin{aligned} \beta_{T_{j-1}} = \cdots = \beta_{T_j - 1} \equiv \alpha_j, \qquad & \text{for } j = 1, \dots, m, \\ \beta_{T_{j-1}} = \cdots = \beta_{T_j} \equiv \alpha_{m+1}, \qquad & \text{for } j = m+1. \end{aligned}$$

where $T_0 = 1$, $T_{m+1} = T$ and $T_1, \dots, T_m$ denote m break points. This assumption can be captured by an L_1-norm regularization of the differences of parameters, also called a fusion penalty, which will be added to the optimization problem.

In the case of an unknown Σ, it has to be estimated. Since the breaks in

the variance-covariance structure of the error terms might also be present, the variance-covariance parameters might be estimated with a fusion penalty or fused lasso introduced by Tibshirani et al. (2005) in the following way. In the first step, the (3.2.1) is estimated under homoscedasticity to obtain the residuals. In the case of heteroscedasticity, one can model squared residuals on a constant. The constant is allowed to vary in time and a fusion penalty is imposed to detect the changes in the variance. For the sake of completeness, the estimated model is:

$$e_t^2 = \sigma_t^2 + u_t, \qquad t = 1, \ldots, T, \tag{3.2.2}$$

where e_t^2 represents the squared residual from the estimated model (3.2.1), σ_t^2 is a time-varying variance to be estimated under the assumption that it is stable for some periods and u_t corrects for the model error. In this case, the method corresponds to the one in Harchaoui and Lévy-Leduc (2010).

In the case of autocorrelation, one can model an AR process with the obtained residuals and then use a fused lasso to detect the breaks and choose the appropriate lag of the AR process. If heteroscedasticity or autocorrelation is detected, the estimate of Σ can be used to transform the original model to get homoscedastic error terms and reestimate the model. In this case, the estimated model is:

$$e_t = \varphi_{t,1} e_{t-1} + \cdots + \varphi_{t,q} e_{t-q} + \nu_t, \qquad t = 1, \ldots, T, \tag{3.2.3}$$

where e_t represents the residual from the estimated model (3.2.1), $\varphi_{t,k}$'s are time-varying AR parameters to be estimated under the assumption that the AR process is stable for some periods and that we have the initial q observations before $t = 1$ at hand. ν_t is a zero mean error term.

Throughout the paper, the following notation is used. The superscript 0 denotes the true unknown parameters. Each period length is defined as $I_j^0 = T_j^0 - T_{j-1}^0$ for $j = 1, \ldots, m^0 + 1$ and the smallest period length as $I_{\min} = \min_{1 \le j \le m^0+1} |I_j^0|$. $J_{\min}$ and $J_{\max}$ denote the smallest and largest break in the parameters, i.e. $J_{\min} = \min_{1 \le j \le m^0+1} \|\alpha_{j+1}^0 - \alpha_j^0\|$ and $J_{\max} = \max_{1 \le j \le m^0+1} \|\alpha_{j+1}^0 - \alpha_j^0\|$.

3.2.2 L_1-norm Regularization of the Differences of Parameters

Adding a weighted L_1-norm constraint of the differences of parameters introduced by Land and Friedman (1997) to a least squares regression yields the

following minimization problem:

$$\hat{\beta}_s = \arg\min_{\beta} \quad (y - X\beta)'\Sigma^{-1}(y - X\beta) \quad \text{s.t.} \quad \sum_{t=2}^{T}\sum_{k=1}^{p} w_{t,k}|\beta_{t,k} - \beta_{t-1,k}| \leq s, \tag{3.2.4}$$

where $s \geq 0$ is a given positive constant, $\hat{\beta}_s$ is a $Tp \times 1$ vector stacking all the estimates and $w_{t,k}$ is a given positive weight allowing to penalize each difference differently. As it was already mentioned, the fusion penalty shrinks the differences of the coefficients to zero and therefore captures the stability of the parameters. The amount of shrinkage depends on the choice of s and weights $w_{t,k}$'s. In comparison to Qian and Su (2016), the penalty here allows every parameter to break separately of each other. The structure of the group lasso penalty in Qian and Su (2016) favors solutions in which all parameters break at the same time point.

The minimization problem (3.2.4) can be relaxed to the following penalized least squares problem:

$$\hat{\beta}_\lambda = \arg\min_{\beta} \quad (y - X\beta)'\Sigma^{-1}(y - X\beta) + \lambda\sum_{t=2}^{T}\sum_{k=1}^{p} w_{t,k}|\beta_{t,k} - \beta_{t-1,k}|, \tag{3.2.5}$$

where $\lambda \geq 0$ is a given tuning parameter, corresponding to a particular s in (3.2.4). The role of λ and $w_{t,k}$ in (3.2.5) can be summarized as follows. The smaller the λ and $w_{t,k}$, the bigger the differences between the parameters in time can be and potentially more breaks can be detected. The reversed relationship holds for high λ and high $w_{t,k}$.

The variance-covariance parameters can be estimated similarly. The parameters for heteroscedasticity:

$$\hat{\sigma}^2_\lambda = \arg\min_{\sigma^2} \quad (e^2 - I\sigma^2)'(e^2 - I\sigma^2) + \lambda\sum_{t=2}^{T} w_t|\sigma^2_t - \sigma^2_{t-1}|, \tag{3.2.6}$$

where e^2 is a $T \times 1$ vector containing all squared residuals, I is a $T \times T$ identity matrix and σ^2 is a $T \times 1$ vector of variances for each point in time. The rest of the parameters has the same role as mentioned in the previous optimization problem.

For the case of autocorrelation the fused lasso penalty is added to the optimization problem:

$$\hat{\varphi}_\lambda = \arg\min_{\varphi} \quad (e - E_q\varphi)'(e - E_q\varphi) + \lambda_1\sum_{t=1}^{T}\sum_{k=1}^{q} w^1_{t,k}|\varphi_{t,k}| + \lambda_2\sum_{t=2}^{T}\sum_{k=1}^{q} w^2_{t,k}|\varphi_{t,k} - \varphi_{t-1,k}|, \tag{3.2.7}$$

where e is a $T \times 1$ vector containing all residuals, E_q is a $T \times Tq$ matrix and φ is a

$Tq \times 1$ vector of all the AR parameters. The rest of the parameters has the same role as mentioned in the previous optimization problem. The only difference is that now there are two tuning parameters and two sets of weights. Further, the paper focuses on the estimation of the β parameters and leaves the case with an unknown variance as a potential extension.

Relaxing the problem (3.2.4) into (3.2.5) yields a convex function in β for a given λ and given $w_{t,k}$'s, however it is non-smooth. Regarding the algorithm used to solve a fusion penalty such as in (3.2.5), the Majorization-Minimization (MM) algorithm proposed by Yu et al. (2013) can be used. In each iteration of the algorithm, a majorization function, which is a differentiable approximation of (3.2.5), is constructed. The majorization function has to satisfy the following properties. It is always above the objective function and it goes through the optimal point found in the previous iteration. In each iteration, the majorization function represents a better and better approximation of the objective function. The algorithm stops when the change in the value of the objective function at the minimizer is small enough (10^{-10} or 10^{-6} is chosen in the simulation depending on the estimation step[1]) or if the maximum number of iterations is reached (10 000).

Further, the assumptions on the size of the break has to be introduced as detailed in Section 3.2.3. Unless the break is large enough, it cannot be detected by the proposed method. Below a certain threshold the break will be smoothed out in the second step. The number of breaks also cannot grow too fast with the sample size. The regressors have to have at least a finite fourth moment. Otherwise, the behavior of the regressors would mask the breaks.

3.2.3 Asymptotic Properties

To show asymptotic properties of the method regarding the consistent detection of the relative positions of the breaks given that the correct number of breaks is detected, the following assumptions are made.

Assumption 1 $m_0 = O(\log(T))$.

Assumption 2 $\sup_{t\geq 1} E\|x_t\|^{4q} < \infty$, $\sup_{t\geq 1} E|\varepsilon_t|^{4q} < \infty$ *for some* $q > 1$, $E(x_t\varepsilon_t) = 0$.

Assumption 3 $\exists \infty > \underline{c}_{xx} > 0$ *and* $\infty > \bar{c}_{xx} > 0$ *and a positive sequence* δ_T *decreasing to zero as* $T \to \infty$ *such that*

[1]Smaller values would improve the estimates but at a high computational cost. As it will be explained later, the estimates for the first step estimation have to be more precise, therefore 10^{-10} is used there. In the second step, 10^{-6} is used to gain computational speed.

$$\underline{c}_{xx} \leq \inf_{\substack{1\leq s<r\leq T+1 \\ r-s\geq\delta_T T}} \mu_{min}\left(\frac{1}{r-s}\sum_{t=s}^{r-1} E(x_t x_t')\right)$$
$$\leq \sup_{\substack{1\leq s<r\leq T+1 \\ r-s\geq\delta_T T}} \mu_{max}\left(\frac{1}{r-s}\sum_{t=s}^{r-1} E(x_t x_t')\right) \leq \bar{c}_{xx}.$$

where μ_{min} and μ_{max} are the smallest and largest eigenvalues.

Assumption 4 $\sup_{\substack{1\leq s<r\leq T+1 \\ r-s\geq\delta_T T}} \left(\frac{1}{\sqrt{r-s}}\sum_{t=s}^{r-1} x_t\varepsilon_t\right) = O_p((\log T)^3)$.

Assumption 5 *$T\delta_T \geq c_\nu T^{1/q}$ for some $c_\nu > 0$.*

Assumption 6 *$I_{min}/(T\delta_T) \to \infty$ as $T \to \infty$.*

Assumption 7 *$J_{max} = O(1)$, $T\delta_T J_{min}^2/(\log T)^6 \to \infty$ as $T \to \infty$.*

Assumption 8 *$\lambda = \lambda_T$ satisfies $\frac{\lambda\sqrt{p}\xi_T}{T\delta_T J_{min}} \to 0$ as $T \to \infty$ where ξ_T bounds the largest $w_{t,k}$.*

Assumption 1 captures the condition on m_0, the true number of breaks. The true number of breaks can grow with the length of the time series at a logarithmic rate. This is less restrictive than the assumption in Bai and Perron (1998) who are considering a setup with fixed number of breaks at asymptotically distinct time points. Assumptions 2 - 4 put restrictions on the true DGP. The Assumption 2 requires at least finite fourth moments of x_t and u_t. Assumptions 2 and 4 allow to use concentration inequalities. The Assumption 4 is not too restrictive as it is satisfied by several standard processes such as e.g. autoregressive moving average (ARMA) processes. Assumption 3 bounds the eigenvalues of $E(x_t x_t')$ and ensures the invertibility of the $X'X$ matrix with probability going to 1 for any closed interval $[s, r-1]$ of the minimum length $\delta_T T$ which grows to infinity as in Assumption 5. The sequence δ_T plays a role for the consistency of the relative break point positions T_j/T. Assumption 6 bounds the sequence from below, i.e. determines the slowest rate at which it converges to zero. Assumptions 7 and 8 bound the sequence from above, i.e. determine the fastest convergence to zero.

The following theorem shows consistency of $\{\hat{T}_j/T\}$ when $\hat{m} = m_0$.

Theorem 3.2.1 *Suppose that Assumptions A1-A8 hold. If $\hat{m} = m_0$, then*

$$P\left(\max_{1\leq j\leq m_0} |\hat{T}_j - T_j^0|/T \leq \delta_T\right) \to 1 \text{ as } T \to \infty.$$

The proof of the Theorem 3.2.1 is an augmented version of the proofs in Harchaoui and Lévy-Leduc (2010) and Qian and Su (2016) extended to the linear regression model estimated by the weighted fusion penalty. The proof is in Appendix B.2.1.

3.2.4 Estimation Procedure and Selection of the Tuning Parameter

The estimation procedure of problems such as (3.2.5) has to be done in two steps. In the first step, $w_{t,k}$'s need to be estimated. In the second step, the parameters in (3.2.5) are estimated while plugging $\hat{w}_{t,k}$'s from the first step for $w_{t,k}$'s. To ensure that the estimates in the second step are reliable and consistent, it is important to estimate $w_{t,k}$'s properly.

To get estimates for the weights $w_{t,k}$ in a high-dimensional model, β is first estimated from problem (3.2.5) in which all $w_{t,k} = 1$ and parameter λ is set optimally (let us denote this value as $\hat{\lambda}_1$). Then, the weights for the second step are defined as:

$$\hat{w}_{t,k} = \begin{cases} 1/|\hat{\beta}_{t,k,\hat{\lambda}_1} - \hat{\beta}_{t-1,k,\hat{\lambda}_1}| & \text{for } |\hat{\beta}_{t,k,\hat{\lambda}_1} - \hat{\beta}_{t-1,k,\hat{\lambda}_1}| \neq 0 \\ 1/\gamma & \text{for } |\hat{\beta}_{t,k,\hat{\lambda}_1} - \hat{\beta}_{t-1,k,\hat{\lambda}_1}| = 0 \end{cases}$$

where γ is set adequately high and still satisfying the Assumption 8. Then, the problem (3.2.5) is estimated again with $\hat{w}_{t,k}$ and new optimal λ.

The key question for estimating β from equations (3.2.5) is how to choose λ optimally. Assumption 8 provides an optimal rate at which it should converge to 0. In practice, information criteria are preferred to a cross-validation estimation of the tuning parameters because of lower computational costs. Information criterion always takes into an account the trade-off between fit and number of parameters. By choosing λ such that the information criterion is minimized, the best combination between the fit and the number of parameters is found. In the simulation study the following three criteria are used to choose λ from a grid of values evenly distributed between $[0.01, 0.5T]$ and their performance in finite samples is assessed. Alternatively, one could also take a grid of values which are logarithmically distributed as the fusion penalty estimates are more sensitive to smaller values of λ.

Extended BIC (EBIC)

The first information criterion is taken from Chen and Chen (2008). They study a problem of selecting a tuning parameter in penalized likelihood functions for high-dimensional models. In their setting, they allow parameters to grow polynomially with sample size, which is also the case in the model introduced in this paper since $Tp = O(T)$. They show that under normality of the error term and under an asymptotic identifiability condition, their Extended Bayesian Information Criterion ($EBIC$) chooses the true model consistently. The $EBIC$

takes the following form:

$$EBIC(\lambda) = T \log\left(\frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\hat{\beta}_{t,\lambda})^2\right) + (\log(T) + 2\log(Tp))\,|\hat{\beta}_\lambda|,$$

where $|\hat{\beta}_\lambda|$ represents the number of nonzero unique parameters in the model, i.e.

$$|\hat{\beta}_\lambda| = \sum_{k=1}^{p} 1(\hat{\beta}_{1,k,\lambda} \neq 0) + \sum_{t=2}^{T}\sum_{k=1}^{p} \mathbb{1}(\hat{\beta}_{t-1,k,\lambda} \neq \hat{\beta}_{t,k,\lambda} | \hat{\beta}_{t,k,\lambda} \neq 0),$$

where $\mathbb{1}(\cdot)$ is an indicator function taking a value 1, if the condition in the brackets is satisfied. $EBIC$ criterion will be used in both estimation steps to choose the optimal λ.

IC from Qian and Su (2016)

The second information criterion is taken from Qian and Su (2016). They detect multiple structural breaks with a group lasso norm in a linear model and choose the tuning parameter based on the following criterion:

$$IC_{QS}(\lambda) = \log\left(\frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\hat{\beta}_{t,\lambda})^2\right) + \rho_T|\hat{\beta}_\lambda|,$$

where they choose $\rho_T = 1/\sqrt{T}$. They show that under mild regularity conditions, the IC_{QS} is detecting the correct number of breaks. IC_{QS} criterion will be used in both estimation steps to choose λ.

Augmented IC from Qian and Su (2016)

The third way of choosing the tuning parameter is based on the observation that the two information criteria above are prone to underfit the models in small samples if used in the two step procedure described above, i.e. they detect often less breaks then there are in the true model.[2] The reason is that on the one hand, the parameter λ chosen as optimal by $EBIC$ or IC_{QS} in the first estimation step is closer to the truth in terms of the number of breaks (as shown e.g. in Qian and Su (2016)), however on the other hand for a cost of a high bias in the estimated parameters. This bias negatively influences the weights for the second estimation step. Therefore, an adjustment parameter is introduced into the criterion in the first estimation step. The role of the adjustment parameter is to choose λ such that less biased estimates of β are obtained, so that the second estimation step will be improved by more appropriate weights.

[2]See the results in the next section devoted to the simulation study.

The first augmented IC_{QS} is then given by:

$$IC_1(\lambda) = \log\left(\frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\hat{\beta}_{t,\lambda})^2\right) + \frac{1}{\sqrt{T}}(|\hat{\beta}_\lambda| - c),$$

where

$$c = \sum_{t=2}^{T}\sum_{k=1}^{p} 1\{|\hat{\beta}_{t,k,\lambda} - \hat{\beta}_{ad,k,\lambda}| < \delta(\max_{1\leq i\leq T}\hat{\beta}_{i,k,\lambda} - \min_{1\leq i\leq T}\hat{\beta}_{i,k,\lambda}) \mid \hat{\beta}_{t,k,\lambda} \neq 0, \hat{\beta}_{t,k,\lambda} \neq \hat{\beta}_{t-1,k,\lambda}\},$$

and ad represents the time at which parameter k added 1 to c the last time (for $t = 2$, $ad = 1$) and $\delta \in (0, 1)$ represents a percentage of a difference between the largest and the smallest estimated parameter β_k which determines a negligible break. In other words, c contains how many times the change in the parameter was not big enough to consider it as a break which is then deducted from the number of nonzero unique parameters in the model. By the inequality in the indicator function, c increases if the difference in the parameter between the current value and the value after the last break (contained in $\hat{\beta}_{ad,k,\lambda}$) is smaller than $100 \cdot \delta\%$ of a difference between the largest and the smallest value of the estimated parameter β_k over the whole period given that the parameter did not switch to zero and given that it is different from the parameter in the previous period. Introducing $\hat{\beta}_{ad,k,\lambda}$ is important to cover a case, in which the parameter changes slowly but monotonically over time. Then, these permanent small changes will be captured at least by a step function. If $|\hat{\beta}_{t,k,\lambda} - \hat{\beta}_{t-1,k,\lambda}|$ was used in c instead, then all the small breaks would be discarded without taking into account that their sum is already significant. The condition $\hat{\beta}_{t,k,\lambda} \neq \hat{\beta}_{t-1,k,\lambda}$ is important for not taking into account the same small break twice. The idea of capturing the parameter changes in a step function also affects the weights $w_{k,t}$. All the negligible breaks are smoothed out and only the (accumulated) significant ones are translated into the weights based on $IC_1(\lambda)$.

In the second step, the final model is selected according to a criterion similar to IC_{QS} only ρ_T is chosen differently but still within the consistency condition:

$$IC_2(\lambda) = \log\left(\frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\hat{\beta}_{t,\lambda})^2\right) + \frac{1}{\sqrt[3]{T^2}}|\hat{\beta}_\lambda|.$$

The adaptive fusing penalty should be able to suppress the insignificant breaks and leave only the most important ones. Therefore, a break is simply defined at a point t at which $\hat{\beta}_t \neq \hat{\beta}_{t-1}$ in the second estimation step.

3.3 Simulation Study

There are three different sample sizes in the simulation study: $T = \{50, 100, 200\}$. For each T, one thousand different samples were drawn ($S = 1000$). Optimal tuning parameters λ_1 and λ_2 are searched on a grid of 100 linearly spaced values from the interval $[0.01, T]$. Smoothing parameter δ in the IC_1 was set to three different values (0.025, 0.05 and 0.075) to see its effect on the estimation. The best results were obtained for $\delta = 0.075$ in models with one parameter and one break and for $\delta = 0.025$ in models with two parameters or no break, however the results over the chosen δ do not vary much.

The fusion penalty is compared to two standard tests for detecting breaks in parameters: Double Maximum test (DM) and/or the Sequential test (Seq) from Bai and Perron (1998) with the trimming parameter set to 0.15 at the 5% significance level. DM test tests a null hypothesis of no breaks against an alternative of unknown number of breaks given an upper bound of breaks set to 5. Sequential test tests stepwise a null hypothesis of ℓ breaks against an alternative of $\ell + 1$ breaks, starting with $\ell = 0$ until the null hypothesis is rejected.

All simulations are evaluated using the following criteria, whose formulas are given in Table 3.3.1. Firstly, three cases are distinguished based on the number of the detected breaks:

(1) Too few breaks were detected = "False Negatives" (FN),

(2) Correct number of breaks is detected = "True Positives" (TP),

(3) Too many breaks were detected = "False Positives" (FP).

Regarding these values, TP should be maximized and $FN + FP$ minimized.

Under a presence of break, the closeness of the estimated break to the true one plays a role. The criterion Q measures the average relative distance to the closest true break in the following fashion. In the case of no detected breaks, the estimate is fully penalized (i.e. attains value 1). In other cases, the average relative distance of the estimated breaks to the closest true break is computed. The true breaks which have no estimated counterpart (i.e. some other true breaks are closer to the estimated breaks) are fully penalized. Notice that elements in Q are constructed in such a way that the final result will be in a $[0, 1]$ interval, where 0 indicates a perfect estimate in all draws and 1 indicates that no breaks were detected in all draws. Methods with Q closer to 0 are evaluated as more precise.

In the case of $p > 1$, it is also important if the break was found in the correct parameters. This is captured in criterion P. The idea is to add 1 if

the estimated break is in exactly the same parameters as the closest true break. If the break is found in too many or too few parameters, only a corresponding fraction of correctly estimated parameters is added. As in the case of Q, P is also constructed such that the final result will be in a $[0, 1]$ interval. Values closer to 1 indicate better performance in detecting a break in the correct parameter.

Table 3.3.1: Evaluation Criteria

Criterion	Formula
FN	$\sum_{s=1}^{S} \mathbf{1}(\hat{m}_s < m_0)/S$
TP	$\sum_{s=1}^{S} \mathbf{1}(\hat{m}_s = m_0)/S$
FP	$\sum_{s=1}^{S} \mathbf{1}(\hat{m}_s > m_0)/S$
Q	$\frac{1}{S}\sum_{s=1}^{S}\{\mathbf{1}(\hat{m}_s = 0) + + \mathbf{1}(\hat{m}_s \neq 0)\frac{1}{m_0}\sum_{j=1}^{m_0}\left[\frac{1}{\lvert K_j\rvert}\sum_{k\in K_j}\frac{\lvert\hat{T}_k^s - T_j^0\rvert}{T}\Big\rvert_{K_j\neq\emptyset} + 1\mid_{K_j=\emptyset}\right]\}$
P	$\frac{1}{S}\sum_{s=1}^{S}\frac{1}{m_0}\sum_{j=1}^{m_0}\left[\frac{1}{\lvert K_j\rvert}\sum_{k\in K_j}\frac{\hat{q}_k^{s,c}}{\max(\hat{q}_k^s, q_j^0)}\Big\rvert_{K_j\neq\emptyset}\right]$
B	$\frac{1}{S}\sum_{s=1}^{S}\frac{1}{Tp}\sum_{t=1}^{T}\sum_{k=1}^{p}(\hat{\beta}_{t,k,\lambda}^s - \beta_{t,k}^0)^2$

Notation: $s \ldots$ s-th draw, $S \ldots$ total number of draws, $\hat{m}_s \ldots$ number of breaks estimated in draw s, $m_0 \ldots$ true number of breaks, $T_j^0 \ldots$ time point of the j-th true break, K_j is a set containing subindeces of all time points of the estimated breaks for which T_j^0 is the nearest time point, $\hat{T}_k^s \ldots$ time point of the k-th break estimated in the s-th draw, $\hat{q}_k^{s,c} \ldots$ number of parameters at which the k-th break was correctly detected (i.e. in comparison to the true parameter breaks at T_j^0) in the s-th draw, $\hat{q}_k^s \ldots$ number of all parameters at which the k-th break was detected in the s-th draw, $q_j^0 \ldots$ number of true parameter breaks at T_j^0, B is the average squared bias.

3.3.1 One Parameter - One Break - Middle

The data are simulated from the following model:

$$\begin{aligned} y_t &= 1 + 2x_t + \varepsilon_t, \qquad && t = 1, \ldots, T/2, \\ y_t &= 1 + 4x_t + \varepsilon_t, \qquad && t = T/2 + 1, \ldots, T, \end{aligned} \tag{3.3.1}$$

where x_t is an iid random variable from the standard normal distribution and ε_t is an iid Gaussian error term with zero mean and the standard deviation is either $\sigma_\varepsilon = \sqrt{5}$, i.e. the signal-to-noise ratio (SNR) is equal to 2 or $\sigma_\varepsilon = \sqrt{10}$, i.e. SNR = 1. In other words, a relatively large break in the middle of the series is present.

Tables B.3.25 - B.3.27 and Tables B.3.29 - B.3.31 (in the Appendix B.3)[3] contain rates of detected breaks in the first step (using $EBIC$, IC_{QS} and IC_1) and in both steps (based on several combinations of $EBIC$, IC_{QS}, IC_1 and IC_2)

[3] All the Tables and Figures are in the Appendix B.3.

for each pair of δ and SNR. The regularization techniques are not as strong in detecting exactly one break as the standard Sequential test. However, if we do not condition on finding exactly one break, Figures B.3.1 - B.3.6[4] show that the regularized estimates based on IC_1 and IC_2 are detecting the position of the break better with a slight overfitting as Tables B.3.27 and B.3.31 confirm. If we condition on finding exactly one break, Figures B.3.7 - B.3.12 confirm the results of Tables B.3.27 and B.3.31 that the standard test performs better. The exception is $T = 50$ and $T = 100$ for the lower SNR where the fusion penalty performs better than the Sequential test also if we condition on exactly one found break.

Regarding the bias of the estimates, an average squared bias is computed as

$$B = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{Tp}\sum_{t=1}^{T}\sum_{k=1}^{p}(\hat{\beta}^s_{t,k,\lambda} - \beta^s_{t,k,\lambda})^2,$$

where S denotes number of simulations and s a particular simulation. Tables B.3.17 - B.3.19 and Tables B.3.21 - B.3.23 reveal that the estimates in the first step based on the $EBIC$ or IC_{QS} are highly biased and therefore inappropriate as weights for the second step of the estimation in comparison to IC_1. The severe underfitting by using $EBIC$ or IC_{QS} in both steps indicates that $EBIC$ and IC_{QS} have a tendency to attain the minimum at a high λ and then lead to rather flat estimates. Tables B.3.17 - B.3.19 and Tables B.3.21 - B.3.23 show that even if IC_1 is used in the first step, using $EBIC$ or IC_{QS} in the second step still leads to many cases of underfitting.

The combination of IC_1 and IC_2 does not suffer from underfitting (see Tables B.3.25 - B.3.27 and Tables B.3.29 - B.3.31) but it has a tendency to overfit in larger samples. However, in the case of $T = 200$, the majority of the detected breaks is equal to 1 or 2, i.e. it is still very close to the true value. Moreover, the position of the break is also estimated close to the true value as mentioned above. It can be therefore concluded that IC_1 and IC_2 perform well, especially in smaller samples.

Based on the results in Tables B.3.25 - B.3.27 and Tables B.3.29 - B.3.31, a higher ratio of noise influences the standard test negatively only in smaller samples ($T = 50$ and $T = 100$). The influence of the higher noise on the two step regularization techniques affects the distribution of the breaks towards underfitting, i.e. in some cases the break is masked by the noise.

[4]Figures are reported only for $\delta = 0.075$.

3.3.2 One Parameter - One Break - End

A model with one relatively large break in one parameter at the end of the sample is introduced in this part. The data are simulated from the following model:

$$\begin{aligned} y_t &= 1 + 2x_t + \varepsilon_t, \qquad && t = 1, \dots, T-11, \\ y_t &= 1 + 4x_t + \varepsilon_t, \qquad && t = T-10, \dots, T, \end{aligned} \tag{3.3.2}$$

where x_t is an iid random variable from the standard normal distribution and ε_t is an iid Gaussian error term with zero mean and the standard deviation of the error term is once set so that SNR = 2: if $T = 50$: $\sigma_\varepsilon = \sqrt{3.2}$, if $T = 100$: $\sigma_\varepsilon = \sqrt{2.6}$, if $T = 200$: $\sigma_\varepsilon = \sqrt{2.3}$ and once it is set so that SNR = 1: if $T = 50$: $\sigma_\varepsilon = \sqrt{6.4}$, if $T = 100$: $\sigma_\varepsilon = \sqrt{5.2}$, if $T = 200$: $\sigma_\varepsilon = \sqrt{4.6}$.

Tables B.3.41 - B.3.43 and Tables B.3.45 - B.3.47 contain rates of detected breaks for each combination of δ, information criteria and SNR. For $T = 50$, the standard Sequential test performs in most of the cases better than the regularized estimators in finding exactly one break. For $T = 100$ and $T = 200$, regularized estimator based on IC_1 and IC_2 performs better than the standard test (especially for higher δ). If we do not condition on finding exactly one break, Figures B.3.13 - B.3.18[5] show that the regularized estimates based on IC_1 and IC_2 are detecting the position of the break better. Especially, when the break is so extreme that it is out of the bounds in which the standard test allows for a break, i.e. for $T = 100$ and $T = 200$. If we condition on finding exactly one break, Figures B.3.19 - B.3.24 confirm the results of Tables B.3.41 - B.3.43 and Tables B.3.45 - B.3.47 that the standard test performs better for $T = 50$ and that the standard test suffers from the bounding effect for $T = 100$ and $T = 200$.

Regarding the bias of the estimates, Tables B.3.33 - B.3.35 and Tables B.3.37 - B.3.39 show that the estimates in the first step based on the $EBIC$ or IC_{QS} are more biased then the ones based on the IC_1. The differences are not so big as before, since the break is set to $T - 10$, therefore smoothing out the estimates close to 2 do not add that much to the bias, however there is an improvement if IC_1 is used.

A higher ratio of noise influences the standard tests negatively only in smaller samples ($T = 50$) based on the results in Tables B.3.41 - B.3.43 and Tables B.3.45 - B.3.47. The influence of the higher noise on the two step regularization techniques affects the distribution of the breaks towards underfitting as in the previous subsection. This simulation exercise is about finding an absolute position of a break, i.e. the relative position changes with higher T. Even though, there

[5]Figures are reported only for $\delta = 0.075$.

is no theoretical result for consistency, it is interesting that the fusing penalty seems to perform well in such an extreme case.

3.3.3 One Parameter - No Break

The data are simulated from the following model:

$$y_t = 1 + 2x_t + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{3.3.3}$$

where x_t is an iid random variable from the standard normal distribution and ε_t is an iid Gaussian error term with zero and the standard deviation of the error term is either $\sigma_\varepsilon = \sqrt{2}$, i.e. the signal-to-noise ratio (SNR) is equal to 2 or $\sigma_\varepsilon = \sqrt{8}$, i.e. SNR = 0.5.

Tables B.3.9 - B.3.11 and B.3.13 - B.3.15 contain percentage rates of falsely detected breaks for each combination of δ, information criteria and SNR. As we can see, two step weighted fusion penalty based on applying in each step either $EBIC$ or IC_{QS} is performing the best. As it is already mentioned in the previous part, this result is mainly due to a general strong bias towards flat estimates of β based on $EBIC$ and IC_{QS}. As a result, the weights in the second step are big and push the weighted estimates to be even flatter (i.e. no break is estimated). The two step procedure based on IC_1 and IC_2 is supposed to solve the strong smoothing of the coefficient estimates. As a result in a case of no break, procedure based on IC_1 and IC_2 has higher false detection but still in a reasonable range. Tables B.3.1 - B.3.7 contain mean values of squared bias for the estimates of the model (3.3.3). In these tables, there are no big differences as in Tables B.3.17 - B.3.23 because all the procedures shrink towards the flat estimates correctly, therefore the bias is similar for all of them. Regarding the smoothing parameter δ, the best accuracy was achieved with $\delta = 0.025$ and the bias does not vary much over the δ.

The accuracy of the number of detected breaks is higher in the setup with a higher SNR for all the procedures. However, the regularization methods seems to be more negatively influenced by a higher noise in comparison to the standard tests. Overall, the regularization techniques perform better than the two standard tests in a case of no break.

3.3.4 Two Parameters - One Break - Middle

The data are simulated from the following model:

$$\begin{aligned} y_t &= 1 + 2x_{1t} + x_{2t} + \varepsilon_t, \qquad t = 1, \ldots, T/2, \\ y_t &= 1 + x_{1t} + x_{2t} + \varepsilon_t, \qquad t = T/2 + 1, \ldots, T, \end{aligned} \tag{3.3.4}$$

where x_t is an iid random variable from the standard normal distribution and ε_t is an iid Gaussian error term with zero mean and the standard deviation is either $\sigma_\varepsilon = \sqrt{6.4}$, i.e. SNR = 2 or $\sigma_\varepsilon = \sqrt{12.8}$, i.e. SNR = 1.

Tables B.3.57 - B.3.59 and Tables B.3.62 - B.3.64 contain rates of detected breaks for each combination of δ, information criteria and SNR. The regularization techniques are not as strong in detecting exactly one break as the standard Sequential test. The regularization techniques suffer from overfit again in terms of the number of detected breaks. Figures B.3.25 - B.3.30[6] show that the regularized estimates based on IC_1 and IC_2 are detecting the position of the break better even in the presence of a slight overfitting as Tables B.3.57 - B.3.59 and Tables B.3.62 - B.3.64 show. Regarding the detection of a break, the regularization method detects the break often corrcectly in the first parameter. Frequently it detects incorrectly a break in both parameters. However, it correctly almost never detects a break only in the second parameter as Figures B.3.25 - B.3.30 show.

Regarding the bias of the estimates, Tables B.3.49 - B.3.51 and Tables B.3.53 - B.3.55 reveal that the estimates in the first step based on the $EBIC$ or IC_{QS} are highly biased and therefore inappropriate as weights for the second step. The IC_1 improves the performance.

3.4 Application

The method is applied to detect breaks in the labor productivity in the US from 1955 to 2004. It is assumed that the US production follows a CES function:

$$Y_t = a(\varepsilon_t) \left[a_K(t) K_t^{\left(\frac{\sigma-1}{\sigma}\right)} + a_L(t) L_t^{\left(\frac{\sigma-1}{\sigma}\right)} \right]^{\left(\frac{\sigma}{\sigma-1}\right)},$$

where
Y_t = production in t,
K_t = capital input in t,
L_t = labor input in t,
$a_K(t)$ = capital augmenting technical progress,

[6] Figures are reported only for $\delta = 0.025$.

$a_L(t)$ = labor augmenting technical progress,
$a(\varepsilon_t)$ = productivity shock,
σ = elasticity of substitution, $\sigma \geq 0$.

Using the relationship for the marginal productivity of labor, the logarithm of labor productivity can be under the following assumptions

$$a(\varepsilon_t) = [\exp(\beta_0 + \varepsilon_t)]^{\frac{1}{1-\sigma}},$$
$$a_L(t) = [\exp(\alpha_L t)]^{\frac{\sigma-1}{\sigma}}$$

expressed as follows:

$$\ln\left(\frac{Y_t}{L_t}\right) = \beta_0 + \beta_1 \cdot t + \beta_2 \cdot \ln\left(\frac{w_t}{p_t}\right),$$

where w_t/p_t is a real wage in t and the coefficients in the linear model are related to the original parameters as:

$$\beta_1 = (1 - \sigma) * \alpha_L,$$
$$\beta_2 = \sigma.$$

The Figures 3.4.1-3.4.3 show the estimated results. Based on β_2, we can see that the elasticity of substitution is stable over the whole period and is equal to 0.93. The estimate satisfies the non-negativity constraint and suggests that capital and labor are almost perfect substitutes.

Since there is no break in β_2, coefficient β_1 captures directly the changes in labor productivity. On Figure 3.4.1, there are 6 breaks at the following times: 1972-3Q, 1975-1Q, 1979-2Q, 1984-3Q, 1988-2Q and 1992-1Q. The first break is close to a 1973 Oil crisis which negatively influenced the whole economy, as well as labor productivity. The price oil peaked in 1979 in the second Oil crisis. The third break representing 1979-2Q shows a rather big drop in β_2 which can be directly translated into a big drop in labor productivity caused by the situation on the oil market. After the mid-80s, the oil market became more stable and we can see that the productivity labor starts getting back to its pre-Oil crisis values. The pre-crisis state was reached in 1992-1Q.

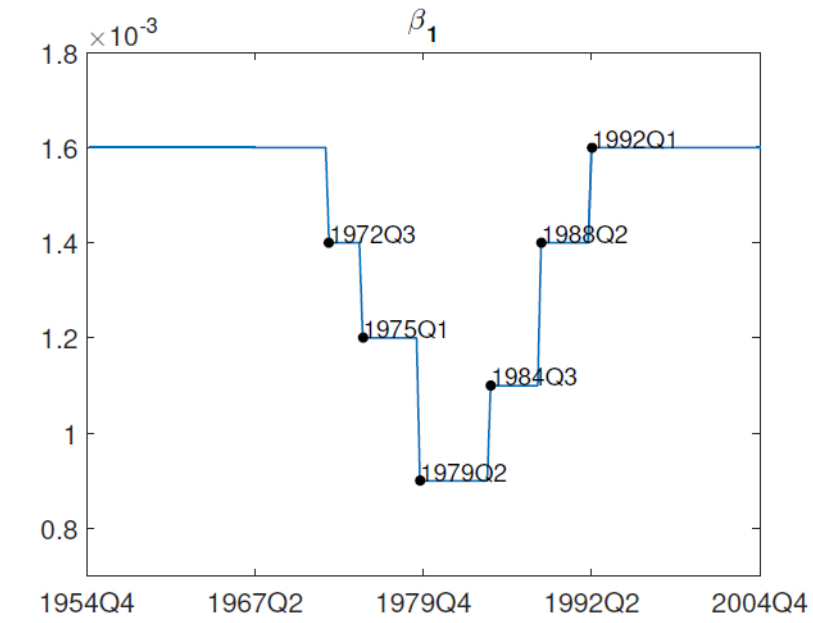


Figure 3.4.1: First Slope

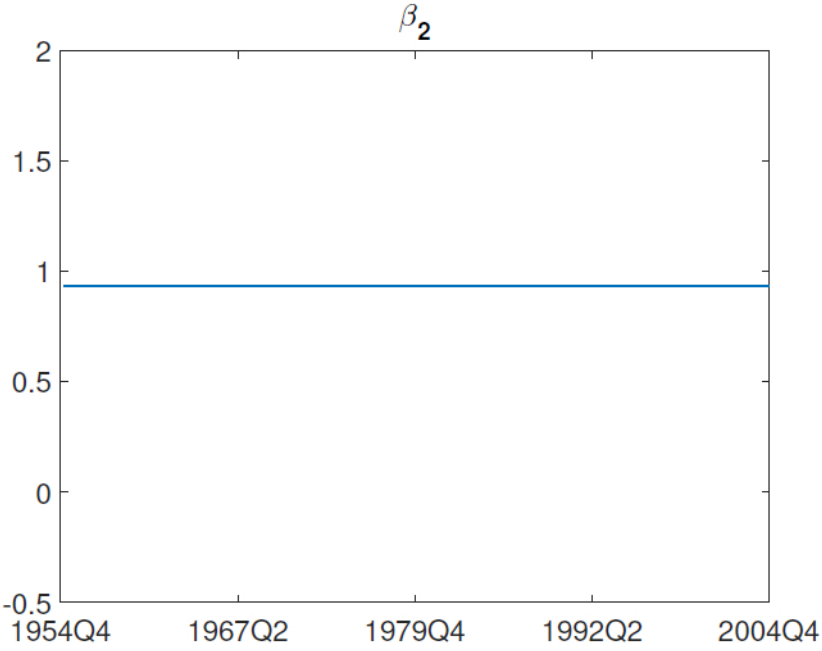


Figure 3.4.2: Second Slope

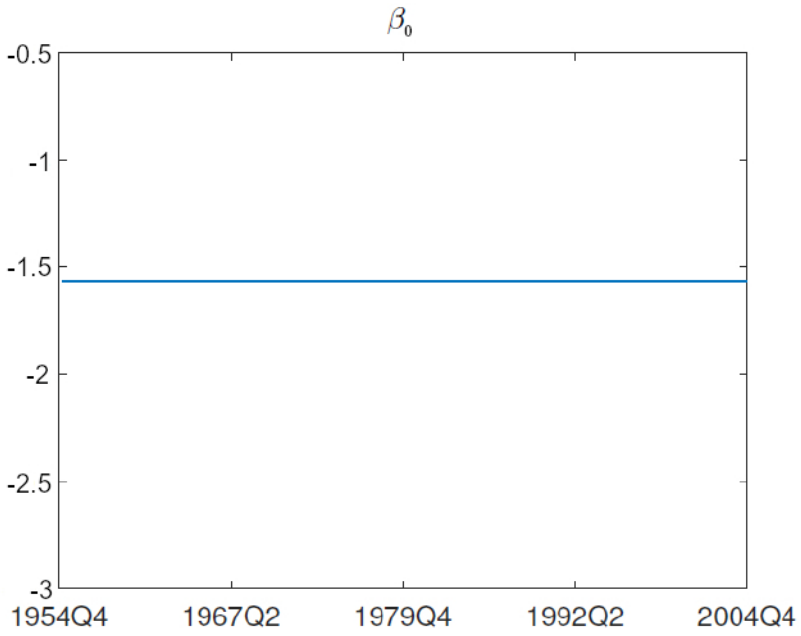


Figure 3.4.3: Intercept Term

3.5 Conclusion

In this paper, the weighted fusion penalty is applied to detect structural breaks in a linear regression model in a simple simulation study. The fusion penalty approach has a couple of advantages in comparison to standard tests which are widely used. It is not necessary to know in advance when structural breaks occurred. Moreover, the assumption of fixed number of breaks is not needed. Another advantage is that by the weighted fusion penalty, smoothing of falsely detected breaks is controlled within the method. No additional smoothing techniques are needed as in Qian and Su (2016).

The simulation results for samples with 50, 100 and 200 observations show that the weighted fusion penalty yields reasonable results, given the high-dimensional characteristic of the model. Information criteria IC_1 and IC_2 seem to perform well in small finite samples with low number of parameters. They reduce bias and serious underfitting of $EBIC$ and IC_{SQ} in the two step weighted estimation procedure. Comparing IC_1 and IC_2 to the standard tests in terms of estimating the exact correct number of breaks, the regularization approach is performing slightly worse. However, they detect the position of the break better. Lower SNR negatively influences the performance of all the techniques.

The method was applied for detecting structural breaks in the labor productivity in the US from 1955 to 2004. The breaks which were found are nicely interpretable. The drops in labor productivity were caused by 2 oil crises in 1973 and 1979. After stabilization of the oil market, the labor productivity got back to its pre-crisis value.

4 How well can Noncognitive Skills Predict Unemployment?: A Machine Learning Approach

4.1 Introduction

The role of noncognitive skills in explaining individual differences in educational attainment and labor market performance has been documented by labor economists and personality psychologists in numerous empirical and experimental studies. There is little doubt that beyond cognitive abilities, individual differences in noncognitive skills explain a large fraction of observed variation in educational attainment and labor market performance. Existing empirical evidence includes studies investigating the effects of noncognitive skills on earnings (Nyhus and Pons (2005) Mueller and Plug (2006)), job search (Uysal and Pohlmeier (2011), Viinikainen and Kokko (2012), Caliendo et al. (2014a)), occupational choice (John and Thomsen (2014)), self-employment (Caliendo et al. (2014b)) and educational attainment (Duckworth and Seligman (2005), Piatek and Pinger (2016)). Borghans et al. (2008) and Almlund et al. (2011) provide comprehensive overviews over the empirical findings from labor economics and personality psychology.

In this paper, we study the importance of noncognitive skills in predicting differences in individual unemployment by means of machine learning techniques. We introduce a multi-step procedure combining unsupervised and supervised machine learning techniques accomplishing: (1) dimensionality reduction of the survey items into noncognitive skill indices in two data-driven steps, (2) a target-oriented index construction of noncognitive skill measures (3) yielding an interpretable predictive model which can be (4) either statistically or economically tuned for classification. Contrary to the standard empirical studies, we follow a strictly predictive modeling approach to data analysis by focusing on the out-of-sample classification qualities of noncognitive skill measures for future unemployment. In the presence of a panoply of competing personality concepts, regularization is particularly attractive in terms of variable and model selection. Our predictive modeling approach contributes to a better understanding of the relevance of noncognitive skills for individual labor market performance for several reasons. First, variable selection is strictly based on pseudo-out-of-sample performance, i.e. the selected empirical models have a higher external validity.

Second, machine learning techniques can easily cope with the challenge of selecting the most relevant measures from data sources with a whopping number of items capturing several dimensions of noncognitive skills. Thus, they are not prone to in-sample overfitting, a problem that is likely to occur if many covariates are available. Third, machine learning techniques are particularly suitable for sparse modeling, i.e. they are able to select relevant variables and/or sets of variables among a large number of potential alternative specifications and provide final specifications which are easy to interpret. Finally, selecting a sparse model specification with superior predictive performance is attractive when it comes to rank competing predictive factors with respect to their relevance for the design of appropriate intervention strategies or manpower training programs.

Thus far, practical experience with machine learning techniques in the context of psychometric or econometric studies is rather limited. It is the aim of this study to investigate to what extent and how machine learning techniques can contribute to a better understanding of the predictive power of noncognitive skills for individual unemployment. In the center of our approach is the group lasso (Yuan and Lin, 2006). As an L_1-norm penalization strategy, "lassoing" is able to select relevant factors out of a large set of potentially relevant factors and eliminate factors of minor relevance in order to obtain a sparse and easy to interpret model. In addition to the simple lasso (Tibshirani, 1996), grouping leads to a further dimension reduction by selecting complete groups of variables for the model specification while suppressing the less relevant groups. In particular, we show, how lassoing and grouping can be used to construct target-oriented indices of noncognitive skills. These indices incorporate the most relevant information from a larger set of factors of noncognitive skills where the weights are determined by the predictive relevance for a given outcome variable. In this sense, our approach can be seen as an alternative to the Bayesian exploratory factor approach by Conti et al. (2014) that also produces low-dimensional aggregates from high-dimensional psychological measurements. In our empirical study based on the British Cohort Study (BCS), we construct target-oriented indices of noncognitive skills for unemployment based on 119 survey items. Noticeably, our approach is not limited to the empirical questions studied here, but has the potential to be a useful tool in similar settings where indices are constructed to reduce dimensionality of the estimation problem and where the focus of interest is external validity. In particular, our approach may have practical implications for pre-employment screening by providing valuable information to what extent a job candidate is likely to perform well in the job he is assigned for.

In this study, we propose an interpretable predictive modeling approach. Contrary to a purely predictive modeling approach optimized to find the best pre-

dictions based on a large set of covariates, our strategy is to select groups of variables which can serve as the basis for psychological intervention strategies or manpower training programs. For example, the individually identified groups of factors (e.g. lack of conscientiousness or lack of openness) driving the probability of becoming unemployed may call for an appropriate intervention strategy. Moreover, if the purpose of predictive modeling is an economically optimal design of intervention strategies, tuning the quality of model's predictions with respect to a conventional statistical prediction criterion (e.g. accuracy of classification or informedness) is suboptimal. Therefore we present, in addition to the selection of groups of factors, two approaches to select the threshold parameter in classification models. The first approach is a standard in the literature based on optimizing a statistically motivated criterion ignoring economic costs of the resulting classification. In the alternative second approach we replace the statistical criterion by an economic criterion to tune the threshold parameter so that the resulting classification leads to economic efficiency. Using unemployment classification as an example, with the economic approach the model is tuned to take the overall costs and benefits of an unemployment program into account and classify only those as unemployed so that the whole program is cost efficient.

The paper is organized as follows. In Section 4.2 we elaborate on the potential merits of predictive modeling to assess the impact of noncognitive skills on individual labor market performance. Moreover, we introduce our group lasso approach to select measures of noncognitive skills and compare it with traditional factor analytic approaches. In Section 4.3 we introduce the approach of intervention optimal classification. We describe the BCS sample and discuss further data issues in Section 4.4. The structure of BCS data allows us to set up an early warning system predicting unemployment over a horizon of 24 years using covariates measured at the age of 10. Our empirical findings are presented in Section 4.5, while Section 4.6 concludes and provides an outlook on future research.

4.2 A Machine Learning Approach to Select Skill Factors

In what follows, we present a general modeling strategy to select noncognitive skill factors to predict individual unemployment. Our empirical application focuses on the selection of factors, which are capable of predicting future unemployment best over a forecasting horizon of 24 years. We argue, that being able to identify the factors with high predictive ability may be helpful to optimize the design for the selection of individuals at risk into intervention strategies and

manpower training programs.

Explaining vs. Predicting

When modeling the impact of noncognitive skills for individual labor market performance, conventional statistical approaches in labor economics and personnel psychology almost exclusively follow an explanatory research strategy. Their goal is to identify and to quantitatively assess at best causal relationships with a high internal validity obtained by maximizing the in-sample explanatory power. In contrast, machine learning techniques generally follow a predictive modeling strategy focusing on high external validity and true out-of-sample predictive performance, i.e. the best model is the one that predicts or classifies the outcome variable best on data which have not been used for the estimation process. In the first place, we regard our empirical strategy, which is a combination of various machine learning techniques (unsupervised learning, group lasso, cross-validation etc.), as being complementary to the explanatory strategies of conventional empirical studies on observational data. Shmueli (2010) summarizes the scientific purpose and added value of predictive modeling in the following four main points:

i. Detection of complex relationships and patterns that are hard to hypothesize in large and rich data sets, especially given theories that exclude newly measurable concepts.

ii. Discovery of new measures as well as comparison of different operationalizations of constructs and different measurement instruments.

iii. Enhancement of existing explanatory models.

iv. Reality check of the strength of causal relationships.

As competing noncognitive skill factors are compared in terms of their predictive performance, our machine learning approach follows a predictive modeling strategy serving to some extent all four purposes listed above. Typically, empirical studies with many potential explanatory factors focusing on the in-sample fit suffer from the risk of over-fitting. Predictive modeling strategies naturally overcome this issue leading to a sparser model specification which separates relevant predictors from less relevant ones.

However, our modeling strategy is not a purely predictive one. It rather can be considered as a hybrid approach in the sense that not single, difficult to interpret variables (e.g. single survey items) are identified among a large set of variables as having a high predictive power like in a kitchen-sink regression. We rather follow a two-stage strategy where groups of survey items are identified in the first stage by means of cluster analysis unveiling the items that measure the

same noncognitive skill factors. In the latter stage, the grouped survey items are then studied with respect to their predictive performance using group-wise L_1-regularization to yield sparse model specifications, which are easy to interpret and reveal a high predictive quality.

Large-scale Skills Measures

Empirical studies based on large-scale observational data usually contain a large number of measures of cognitive and noncognitive skills. Typically, some type of dimension reduction technique is applied in order to reduce the dimensionality of the covariate space and to obtain interpretable empirical results. Predominantly, this is done ex-ante via preprocessing the data by principal component analysis (PCA) and related factor modeling strategies yielding a low-dimensional index representation. Alternatively, some type of dimension reduction is implicitly accomplished by focusing on certain noncognitive skills (e.g. Big Five, locus of control) and disregarding covariates reflecting more closely alternative (complementary or competing) concepts.

In case of the BCS with its rich information on noncognitive skill measures, the dimensionality of the predictive model grows along two dimensions: (i) number of survey items capturing different noncognitive skills and (ii) number of noncognitive skill indices built from the survey items. As mentioned above, instead of using all the survey items as covariates in the model, the dimensionality of the data set is first reduced by calculating an index from survey items capturing the same noncognitive skill factor. Either the researcher has a prior information on which survey items measure certain noncognitive skills or he applies a statistical dimension reduction technique like PCA or factor analysis to find the index structure in the data set. Apart from PCA, any other unsupervised machine learning technique can be applied that is able to unveil the grouping of the items. Therefore as a first step, we use clustering to collect survey items into groups representing particular noncognitive skills. This leads to the first dimensionality reduction along the dimension (i). The survey items in a group are then used later to calculate an index, which proxies a certain noncognitive skill factor.

In the second step, we plug the grouped survey items into the model and let group lasso to decide which of these groups are important for predicting individual unemployment and to estimate the group lasso index weights. More details on alternative approaches how to construct the weights are given in Section 4.2.2. The second step reduces the model dimensionality along the dimension (ii) by reducing the number of noncognitive skill indices in the model. We compare in Section 4.5 results of the group lasso estimates with more conventional ways of index construction for different model variants.

4.2.1 L_q-Regularization

In order to understand how the group lasso helps to select and estimate index weights, let us briefly introduce the general idea behind L_q-regularization first. Consider the following least squares minimization problem:

$$\min_{\beta} \|Y - X\beta\|_2^2 \qquad \text{s.t.} \quad \sum_{j=1}^{k} |\beta_j|^q \leq s, \tag{4.2.1}$$

where $\|Y - X\beta\|_2^2$ denotes sum of squared errors of a linear model, Y is an $N \times 1$ vector of the dependent variable, N represents the number of observations, X is an $N \times k$ matrix of covariates, β is an unknown $k \times 1$ parameter vector and s is a chosen non-negative constant for $q \geq 0$. When $q = 1$, the (4.2.1) is called LASSO (Least Absolute Shrinkage and Selection Operator) optimization problem introduced by Tibshirani (1996). The advantage of the L_1-regularization (i.e. $q = 1$, LASSO) over the L_2-regularization ($q = 2$, ridge) is that under the given restriction there is a high probability that some of the parameters will be set to zero at the minimum. The LASSO is therefore able to select relevant predictors yielding a sparse model specification among a large set of possible specifications. The parameter q can take any positive value. However, it can be shown that only for $0 \leq q \leq 1$ there is a non-zero probability to select covariates and to shrink the impact of less relevant covariates to zero (e.g. Hastie et al. (2015)).

The primal problem (4.2.1) can be reformulated in a Lagrangian representation:

$$\min_{\beta} \|Y - X\beta\|_2^2 + \lambda \sum_{j=1}^{k} |\beta_j|^q, \tag{4.2.2}$$

where $\sum_{j=1}^{k} |\beta_j|^q$ is the penalty function and λ the regularization (or penalty) parameter. For $0 \leq q \leq 1$, the choice of λ determines the sparsity of the solution, i.e. the size of λ determines how many elements in the estimated β are set to 0. The larger the λ is, the sparser is the specification obtained and the fewer covariates are included in the regression equation. However, applying the standard lasso as a selection tool to choose the most relevant survey items as proxies for noncognitive skills is somewhat opaque as some of the items in a questionnaire are likely to be highly correlated and in such a setting, the standard lasso technique randomly chooses one predictor from each group of highly correlated predictors not caring which one is selected, therefore making the model estimates hard to interpret (Zou and Hastie, 2005).

Group Lasso

The group lasso introduced by Yuan and Lin (2006) overcomes this problem and accounts for a given natural group structure in the data. It guarantees that all coefficients within a group become nonzero or zero simultaneously. Consider now the case where covariates (e.g. facets in the Big Five framework or a set of survey items measuring noncognitive skills) can be divided into J groups with k_j covariates in group j. The selection of groups is achieved by solving:[1]

$$\min_{\beta} \|Y - \sum_{j=1}^{J} X_j \beta_j \ - Z\delta\|_2^2 + \lambda \sum_{j=1}^{J} \sqrt{k_j} \|\beta_j\|_2, \tag{4.2.3}$$

where Y is an $N \times 1$ vector of the dependent variable, N represents the number of observations, X_j is an $N \times k_j$ matrix of covariates corresponding to the j-th group of skill factors, β_j is a $k_j \times 1$ parameter vector for group j. The method allows for additional covariates contained in Z with a corresponding parameter vector δ which is not subject to regularization, i.e. the covariates Z are always in the model.

Since $q = 0.5$ for every parameter vector β_j, the choice of the regularization parameter λ determines how many groups are selected. Typically, the optimal λ is chosen by K-fold cross-validation, see e.g. Hastie et al. (2015) for more details. The larger the λ, the more elements in the penalty term $\sum_{j=1}^{J} \sqrt{k_j} \|\beta_j\|_2 = \sum_{j=1}^{J} \sqrt{k_j} \sqrt{\beta_{j,1}^2 + \cdots + \beta_{j,k_j}^2}$ are forced to zero in order to minimize (4.2.3). In case of group lasso, the elements in the penalty term are Euclidean norms. A Euclidean norm is equal to zero, when all the components are zero. This means that the whole group is eliminated from the model and sparsity in groups is achieved. In addition, the penalty for a group j receives more weight, the larger the number of group items k_j is. This balances out the effect that bigger groups would be more likely to be selected without this correction. Therefore, the group lasso is a natural approach to select the best predicting group of variables related to a certain personality theory among a large set of competing personality theories.

Lassoing and group lassoing as briefly described above are not restricted to the estimation of linear regression models. They can easily be extended to the estimation of nonlinear models. For our application to individual unemployment, we use a regularized maximum likelihood logit approach where the least squares part of objective function in (4.2.2) is replaced by the negative log likelihood of the logit model. Under sparsity the group lasso is consistent in a sense that a difference between the estimated and true logit link function is bounded with high probability even when the number of predictors exceeds the number of ob-

[1] See Hastie et al. (2015) for different variants of the group lasso.

servations (Meier et al. (2008)).

4.2.2 Index Construction

In the following, we use the properties of L_q-norm regularization for the construction and selection of skill indices along with the conventional approaches. Assume that the grouping of the survey items with respect to different noncognitive skill factors is given. Detailed information about how ex-ante grouping can be achieved by clustering as used in this paper is provided in Section 4.4. For example, let $C_{i1}, C_{i2}, \ldots, C_{iK_C}$ be the responses of individual i on K_C different items related to conscientiousness and $E_{i1}, E_{i2}, \ldots, E_{iK_E}$ be the responses of individual i on K_E different items related to extraversion, then in general the indices for the two personality traits take the form of a weighted average across the corresponding items:

$$IC_i^C = \sum_{l=1}^{K_C} \omega_l^C C_{il}, \qquad IC_i^E = \sum_{l=1}^{K_E} \omega_l^E E_{il},$$

with $\sum_l \omega_l^C = \sum_l \omega_l^E = 1$. In the following, we consider three different strategies to determine the index weights ω_l^I, where I is a set collecting all available personality traits. In our example: $I = \{C, E\}$.

1. *Equal weights*: $\hat{\omega}_l^I = 1/K_I$
   This is the most simple and most often used way of constructing a skill index. It is not data driven and ignores that some items maybe superior to others in approximating the underlying skill factor.

2. *PCA based weights*: $\hat{\omega}_l^I = \pi_l^{PCA,I} / \sum_l \pi_l^{PCA,I}$.
   The weights are constructed by taking loadings of the first component as $\pi_l^{PCA,I}$'s. This index is data driven and assigns weights to items in a sense that the variance of the obtained scores is maximal.

3. *Group lasso based weights*: Here we use for illustrative purposes a linear model

$$Y_i = \underbrace{\beta_1^C C_{i1} + \beta_2^C C_{i2} \ldots \beta_K^C C_{iK_C}}_{\text{total impact of C on Y}} + \underbrace{\beta_1^E E_{i1} + \beta_2^E E_{i2} \ldots \beta_L^E E_{iK_E}}_{\text{total impact of E on Y}} + \cdots + OtherControls_i' \delta + \varepsilon_i,$$

   and estimate it by a group lasso. The index weights are constructed from the estimated regression coefficients: $\hat{\omega}_l^I = \frac{\hat{\beta}_l^I}{\sum_l \hat{\beta}_l^I}$ (see Appendix C.2). Note that the size of the weights for a certain item now depends on how strongly

this item can predict the outcome variable. The size of the weights is now data driven and context specific, i.e. the relative importance of the items can differ and depends on the outcome variable under consideration. This may be important for items from surveys which were not designed for a specific purpose and items that are only loosely connected to a specific theory of noncognitive skills. Moreover, note that the use of group lasso also allows to select the most relevant groups (indices) for the model, e.g. the group lasso might assign zero coefficients to all conscientiousness items, i.e. $\hat{\beta}_l^C = 0, \forall l$. In this case, all the $\hat{\omega}_l^C$ will get zero values and conscientiousness would be considered as irrelevant for predicting Y_i. In contrast to the other two methods of index construction, the group lasso has a selection property while the (i) equally weighted and (ii) the PCA based approaches always include the entire set of indices in the predictive model. The resulting index with context specific weights is called target-oriented as the outcome variable of interest determines the index weights.

4.2.3 The Predictive Model

As in the previous subsection, assume that the grouping of items is given. The goal is to find the best predictive model based on the following form:

$$Y_i = g_i(\theta) = g(\alpha + Indices_i' \, \gamma + OtherControls_i' \, \delta) + \varepsilon_i, \tag{4.2.4}$$

where Y_i is the individual labor market outcome to be predicted and $\theta = (\alpha, \gamma', \delta')'$ the parameter vector to be estimated. The functional form $g(\cdot)$ depends on the type of dependent variable to be predicted. In our empirical application Y_i is the binary indicator for unemployment, so that we simply choose $g(\cdot)$ to be the cdf of the logistic distribution. The different noncognitive skill factors are captured by the *Indices*. In order to study the predictive performance of noncognitive skills and to check how the different methods of index construction affect the quality of the predictions we compare the following 9 model variants:

A. Baseline model

   1. Model with control variables only (M_0): $\gamma = 0$

B. Restricted models with noncognitive skill factors without control variables: $\delta = 0$

   2. Model with equally weighted indices (*onlyEW*),
   3. Model with PCA indices (*onlyPCA*),
   4. Model with group lasso indices (*onlyGL*),

5. Model with post group lasso weighted indices (*post only GL*).

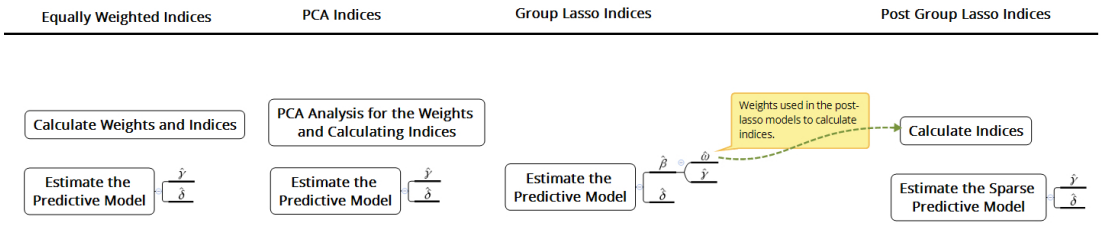
C. Unrestricted models with noncognitive skill factors and control variables: $\delta \neq 0$ and $\gamma \neq 0$

6. Model with noncognitive skills with equally weighted indices and controls (EW),
7. Model with noncognitive skills with PCA indices and controls (PCA),
8. Model with noncognitive skills with group lasso indices and controls (GL)[2],
9. Model with noncognitive skills with post group lasso indices and controls (*post GL*).

Model A.1 serves as the benchmark model. A comparison with models C.6-C.9 yields insights to what extent noncognitive skill factors contain any additional predictive power. Comparing models B.2-B.5 with models C.6-C.9 reveals how controls including cognitive skills improve the predictive performance given the set of noncognitive skill factors. Additional cross-comparisons can show whether cognitive and noncognitive skill overlap in predictive power for individual labor market outcomes.

Assuming that the grouping of the survey items for the index construction is known, Figure 4.2.1 represents the steps for the model estimation for each index type. Equally weighted indices can be calculated directly. For the PCA indices, one has to run the PCA analysis and calculate the indices using the first principal component for the weights. For the group lasso indices, a logit model with all survey items as covariates is estimated by group lasso imposing the known group structure. The target-oriented index weights can be calculated from the β's yielding context specific weights. The optimal λ is obtained from the 10-fold-cross-validation.

Figure 4.2.1: Model Estimation Step for the different Index Schemes



After the group lasso estimation, the researcher can decide whether to continue the analysis with the group lasso model using the survey items as covariates or use

[2]In our implementation, we regularized also the controls, i.e. speaking in terms of notation used in (3), the covariates Z were integrated into X_j's with a correspondingly augmented group structure.

the indices in a so-called post-lasso estimation.[3] The next step in the post-lasso estimation is to calculate the indices and re-estimate (4.2.4) with the selected indices and other controls by a standard maximum likelihood to get $\hat{\gamma}$ and $\hat{\delta}$. The potential advantage in this step is to get conditionally debiased prediction coefficients in a sparse model.

4.3 Intervention Optimal Classification

In order to use the model estimates for classification, a threshold parameter τ has to be chosen, which determines which class is assigned to each observation. For our binary logit model the classification rule is given by:

$$\hat{Y}_i(\tau) = \begin{cases} 1 & \text{if } g_i(\hat{\theta}) > \tau \\ 0 & \text{if } g_i(\hat{\theta}) \leq \tau \end{cases}. \tag{4.3.1}$$

Table 4.3.1 captures outcomes of a predictive exercise for any threshold parameter: (1) correctly predicted cases (TP = 'true positives' and TN = 'true negatives') and (2) misclassified cases (FP = 'false positives' and FN = 'false negatives'). In our empirical application correctly classified individuals are those who

Table 4.3.1: Confusion Matrix

	true value	
prediction	$Y = 1$	$Y = 0$
$\hat{Y}(\tau) = 1$	True Positive (TP)	False Positive (FP)
$\hat{Y}(\tau) = 0$	False Negative (FN)	True Negative (TN)
total	All positives in the data (P)	All negatives in the data (N)

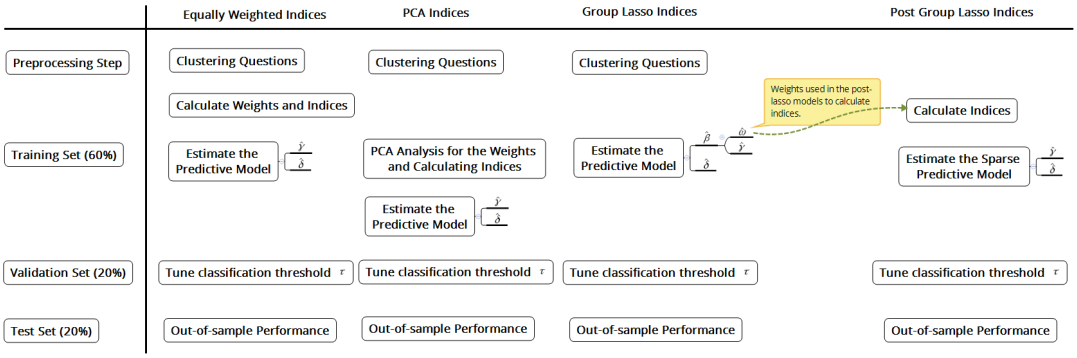
are unemployed and for whom the early warning system has predicted unemployment ('true positives') as well as individuals, who are employed and were not predicted to become unemployed ('true negatives'). An individual, who is falsely predicted as unemployed but will not experience unemployment belongs to the 'false positives'. Finally, the group of 'false negatives' consists of workers who are unemployed but were not detected to be at risk by the early warning system.

Following the goal to get a predictive model which performs well out-of-sample, the threshold parameter is tuned out-of-sample on a part of the data set which was not used to estimate the model coefficients θ. We refer to the data set used for estimation of θ as training set and the data set used for tuning τ

[3] It is possible to identify the index weights ω and index coefficients γ from β in the group lasso model. Regarding group lasso model predictions, it is not necessary to do so as using β with the survey items would yield the same predictions as calculating the index and using γ. Identifying ω is necessary for the post-lasso estimation.

as validation set. Having obtained the optimal τ, we then use it to perform the out-of-sample classifications for the test set data. The full procedure is captured in Figure 4.3.1.

Figure 4.3.1: Modeling Steps for the different Index Schemes



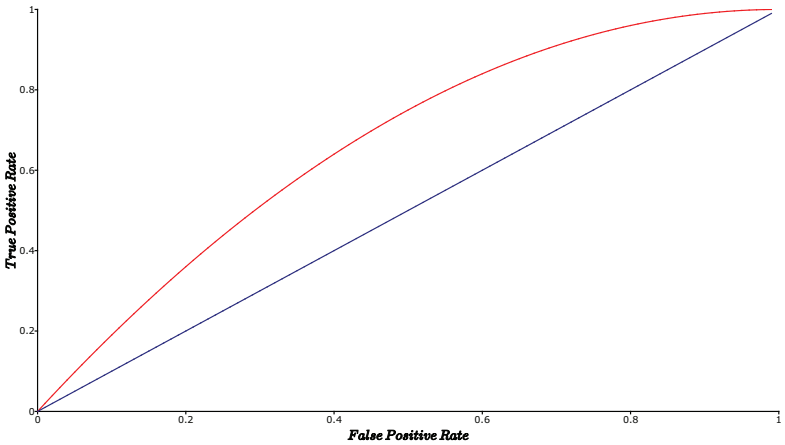
Statistically Optimal Intervention

From a statistical point of view, the optimal threshold parameter is chosen such that a certain statistical measure is optimized (e.g. accuracy or informedness). Conventional statistical measures to assess several threshold parameters are: (i) accuracy (share of correct predictions = (TP + TN)/(P+N)), (ii) sensitivity (true positive rate = $1 - \beta(\tau)$ = TP/P), (iii) specificity (true negative rate = $1 - \alpha(\tau)$ = TN/N) and (iv) positive predicted value (PPV, share of correctly predicted positives among all predicted positives = TP/(TP+FP)). Such statistical criteria may fail to deal with several factors. One of them is an unbalanced sample. By focusing only on accuracy in a setting such as unemployment, an "accurate" predictive model might predict everyone as employed yielding 89-90% level of accuracy given that unemployment rates are between 10 and 11%. Such a scenario would lead to high accuracy, low sensitivity and high specificity. Therefore, it is important to focus also on sensitivity and specificity. The optimal scenario is to have a model with both, high specificity and sensitivity. High specificity implies that a positive prediction can be used to ruling in unemployment. Together with a high PPV, this would be an indication of a model which is reliable to predict unemployment. Low PPV indicates many false positives relative to true positives, i.e. predicting someone as unemployed might be a false alarm with a high probability. A standard approach for fine tuning of the threshold parameter in a statistical sense is to use the information in the Receiver Operating Characteristics (ROC) curve, which captures the sensitivity and specificity trade-off (see Fawcett (2006) for an introduction).

Figure 4.3.2 illustrates the out-of-sample ROC curve for a classifier. The ROC

curve is used to assess statistical qualities of a classifier and to find the optimal threshold in a statistical sense. For our numerical example, the classification is effective, since the ROC curve lies above the 45-degree line representing the location of purely random sorting where the predicted outcomes are independent of the true outcomes. The area under the curve is $AUC = 0.66$ representing the probability that a randomly chosen unemployed person is classified as more likely to be unemployed than a randomly chosen employed person. The best threshold parameter τ in terms of Youden's index arises for a false positive rate, FPR $= \alpha(\tau_{opt}) = 0.5$. Geometrically, the best Youden's J-statistic maximizes the vertical distance of the ROC curve from the 45-degree line, i.e. the most informative model is chosen relatively to the random guess.

Figure 4.3.2: Receiver Operator Characteristics Curve



ROC curve (red) and 45-degree line (blue), AUC = 0.66, Youden's index = 0.25.

Economically Optimal Intervention

The statistical criteria do not capture the real price of misclassification. In case of predicting a low fraction of positives in the data, the typical trade-off is that by improving the number of true positives (high sensitivity), the number of false positives increases as well (low specificity). In an economic cost-benefit analysis this trade-off translates to the task of finding a point at which the marginal cost equals marginal benefit to determine how many people should enter a program in order to achieve economic efficiency. In the unemployment context, we are thinking of any kind of early stage interventions, such as a manpower training program (MTP) for young adults, who are at risk of becoming unemployed in later years. The assignment into treatment and non-treatment is determined by choosing the threshold parameter of classification such that expected per-capita costs of the intervention are minimized in the same fashion as suggested by Metz

(1978). The economically optimal threshold parameter could be very different from the statistically optimal one as it will be shown.

In our economic calculation, cases predicted as unemployed ('predicted positives') are assigned to an MTP and cases predicted as employed ('predicted negatives') are considered as in no need for an MTP. In this framework, the group of 'true negatives' does not generate any costs for the training program, while the 'true positives' receive an intervention with c as the treatment cost per capita. Moreover, assume that for the 'true positives' net unemployment costs per capita are φu, where u denotes unemployment benefits per capita and $0 < \varphi < 1$ is the reduced share of the unemployment benefit payments due to successful participation in the intervention. However, intervention cost also arises for an individual, who is falsely assigned to a training program but will not experience unemployment ('false positive'). Finally, for the group of 'false negatives', the unemployment benefits, u, have to be paid to full extent.

Table 4.3.2 reports the costs depending on the underlying (mis-)classification, where $Y = 1$ denotes a true unemployment status and $Y = 0$ employment. An individual is predicted to become unemployed $\hat{Y} = 1$, if the estimated predictor function $g_i(\hat{\theta})$ exceeds a given classification threshold τ.

Table 4.3.2: Confusion Matrix of Costs per Worker

	unemployment status	
predicted	$Y = 1$	$Y = 0$
$\hat{Y}(\tau) = 1$	$\varphi u + c$ (TP)	c (FP)
$\hat{Y}(\tau) = 0$	u (FN)	0 (TN)

Table 4.3.3 gives the classification probabilities in terms of true negative rate, $1-\alpha(\tau) = \Pr\left[\hat{Y}(\tau) = 0 \,|Y = 0\right]$, i.e. specificity, and true positive rate $1-\beta(\tau) = \Pr\left[\hat{Y}(\tau) = 1 \,|Y = 1\right]$, i.e. sensitivity, where $\pi = \Pr[Y = 1]$ is the probability of becoming unemployed (prevalence).

Table 4.3.3: Confusion Matrix of the Classification Probabilities

	unemployment status		
predicted	$Y = 1$	$Y = 0$	
$\hat{Y}(\tau) = 1$	$\pi(1 - \beta(\tau))$	$(1 - \pi)\alpha(\tau)$	$\pi(1 - \beta(\tau)) + (1 - \pi)\alpha(\tau)$
$\hat{Y}(\tau) = 0$	$\pi\beta(\tau)$	$(1 - \pi)(1 - \alpha(\tau))$	$\pi\beta(\tau) + (1 - \pi)(1 - \alpha(\tau))$
total	π	$1 - \pi$	1

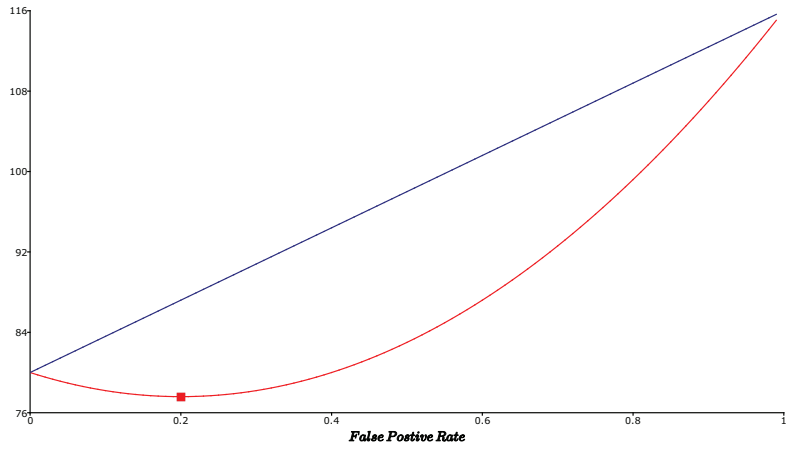
For a given classification scheme with threshold value τ, expected costs of assignment into an early treatment are given by:

$$\mathrm{E}\left[C(\tau)\right] = \Big(\pi(1 - \beta(\tau)) + (1 - \pi)\alpha(\tau)\Big)c + \Big(\pi(1 - \beta(\tau))\varphi + \pi\beta(\tau)\Big)u. \quad (4.3.2)$$

Thus, optimal classification is defined by the threshold value that minimizes expected costs $\tau_{opt} = \arg\min_{\tau} \quad \mathrm{E}\left[C(\tau)\right]$.

For our numerical example, expected assignment costs are depicted in Figure 4.3.3. The economically optimal threshold parameter is determined by the minimum of the expected assignment costs. In our example, the intervention optimal classification is obtained for a false positive rate of FPR $= \alpha(\tau_{opt}) = 0.2$ and a true positive rate of TPR $= 1 - \beta(\tau_{opt}) = 0.36$. In terms of these cost considerations, economically optimal assignment indicates that the requirements for assignment into treatment should be more restrictive than the statistically based optimal assignment, such that only 20 p.c. of the individuals not at risk of becoming unemployed will be assigned for the program (FPR $= 0.2$). Obviously the cost reduction results from not assigning too many individuals to the program who do not need it. The gain from such a policy would be that around 36 p.c. of those who will face unemployment in the future can profit from the intervention. Compared to the statistically optimal classification, the cost minimizing assignment strategy is more restrictive as for the Youden's index maximizing threshold parameter there are considerably more false positives and more true positives assigned to the program.

Figure 4.3.3: Expected Assignment Costs



Expected cost of assignment as a function of the false positive rate (red line). The blue line indicates the expected cost under random assignment for $\pi = 0.04$, $u = 2000$, $c = 100$ and $\varphi = 0.2$.

4.4 Data and Variable Construction

Our empirical study is based on data from the British Cohort Study (BCS). The BCS is a wide-ranging data set containing a rich variety of variables of the study members and their families regarding medical, physical, educational, social

and economic development as well as several measurements of noncognitive skills captured in 119 survey items. The collection of data began in 1970. Babies born in a particular week in 1970 were tracked during their childhood, youth and adult life roughly every four to five years. The longitudinal character of the data set enables us to analyze predictive power of early childhood environment and early cognitive and noncognitive skills measured at the age of 10 on adult labor market outcomes at the age of 34. Thus, our prediction horizon amounts to 24 years. After cleaning, imputing answers for the missing answers of the noncognitive skills[4] and matching the data for males from 1980 to the adult wave of 2004 our sample consists of 2749 observations.

The BCS has been used in several empirical studies on the link between personality traits and individual labour market outcomes. Notable examples are Prevoo and ter Weel (2015), who analyze the effect of early conscientiousness on a variety of adult outcomes, Blanden et al. (2007) for role of noncognitive skills for intergenerational mobility and Uysal (2015) for the causal effects of education on earnings.

Outcome Variable

The information on individual unemployment to be predicted with the variables measured at the age of 10 is taken from the 2004 wave of the BCS, i.e. when the study members were 34 years old. Similar information on unemployment is contained in the consecutive waves of 2008 and 2012. However, the attrition rate is around 18% relative to the 2004 wave. Due to this significant drop in the sample size, we refrained from predicting unemployment at even later stages of the life-cycle.

Moreover, we focus on male employees only, since for females at this particular stage of the life-cycle it is hard to distinguish between voluntary career breaks arising from family planning issues and involuntary ones. Predicting these frictional unemployment events due to voluntary career breaks is not in the focus of our study, in particular, if the goal is to optimize the predictive model with respect to the cost-minimal assignment into early intervention schemes. Employees who reported that they are: full-time or part-time employed and had a job in the last 4 years are coded as "Employed". Those who reported that they are currently unemployed and look for a job, receiving a job-seeker's allowance (JSA) or were unemployed at least once in the last 4 years are coded as "Unemployed". Thus, the outcome variable represents in this context a 4-year period prevalence

[4]If more than 10 answers were missing, the observations were discarded from the data set, i.e. at most 10 missing values were estimated for each study member. The R package Amelia was used for the imputation.

of unemployment for male at 34 years of age. Study members who reported they are self-employed, in full-time education, on a government scheme for training, sick, disabled, looking after the family, wholly retired or do not fit in any category are discarded from the sample. The unemployment rate of the sample is 9.79%.

Survey Items

There are several questionnaires in the BCS measuring behavior and noncognitive skills of the study members. There are three sources of measurements: questionnaires filled out by (1) the study members, (2) their mothers and (3) their teachers. Answers to these survey items are traditionally collected into indices (scales) representing particular noncognitive skills. We follow this approach. For a construction of the index, the answers have to be represented by points and have their corresponding index weights. The point representation of the answers is described in the Appendix C.1.1, Tables C.1.2-C.1.6.

At the age of 10, the study members were asked to complete two questionnaires: the Self-Esteem Scale (LAWSEQ) introduced by Lawrence (1973, 1978) and the Locus of Control Scale (CARALOC) based on Gammage (1975). Self-esteem is a concept capturing a self-evaluation of one's own worthiness. The Locus of Control Scale is supposed to capture how much one believes that he is in control of his life (Rotter, 1966). Higher scores represent an internalizer, a person who thinks that he has control over the outcomes in his life and that he can influence them by his own actions.

In the 1980 wave, mothers of the study members were asked to answer a set of survey items about the behavior of their 10 years old children. In total, there were 38 items which are listed in Table C.1.4. Based on these items, two well known instruments - the Rutter Behavior Scale (Rutter et al., 1970) and the Conners Rating Scale (Conners, 1969; Goyette et al., 1978) - are typically constructed in the literature to capture hyperactivity and anti-social behavior, see e.g. Conti et al. (2014) or Uysal (2015). Instead of applying the two measures mentioned above, we apply cluster analysis in a similar spirit as Butler et al. (1982) and Prevoo and ter Weel (2015) and go beyond the two standard measures yielding a more detailed structure of noncognitive skills from the mother's questionnaire.

Teachers of the study members also answered several survey items about the behavior of the children. In total, the teachers answered 53 items listed in Tables C.1.5 and C.1.6. As in the case of the parents, there are teacher's versions of the Rutter Behavior Scale (Rutter, 1967) and the Conners Rating Scale (Conners, 1969; Goyette et al., 1978). The cluster analysis was applied to identify the scales based on the teacher's questionnaire.

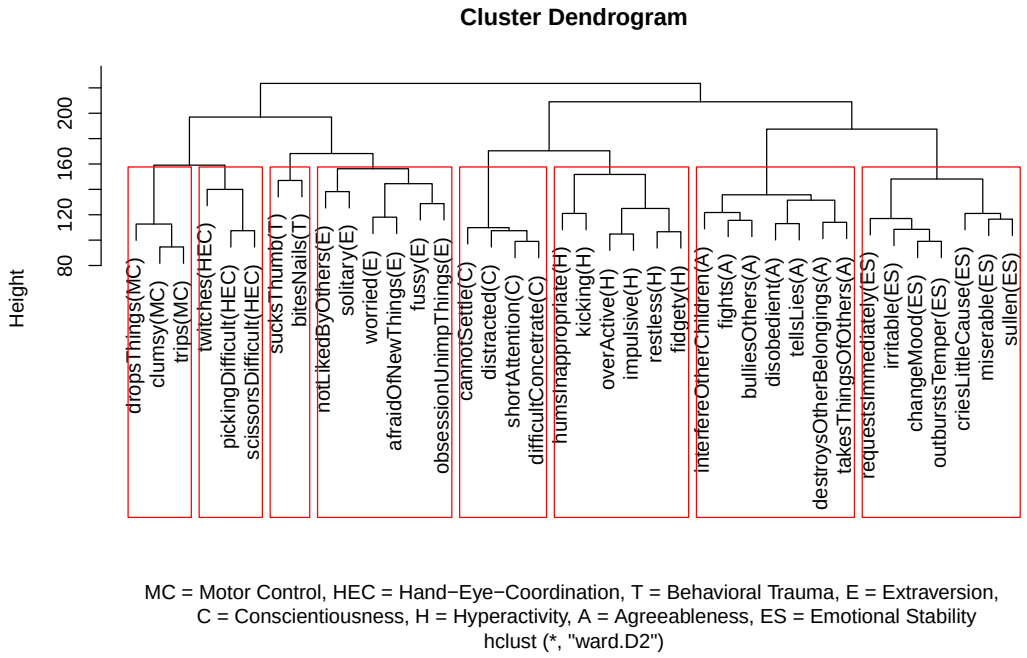
As controls, we use variables listed in Table C.1.1. Social class of the family

captures home environment and enters every model as 5 dummies (the first social class level is left out). Cognitive skills are captured in two "Ability" variables based on a Friendly Math Test and Edinburgh Reading Test.

Cluster Analysis

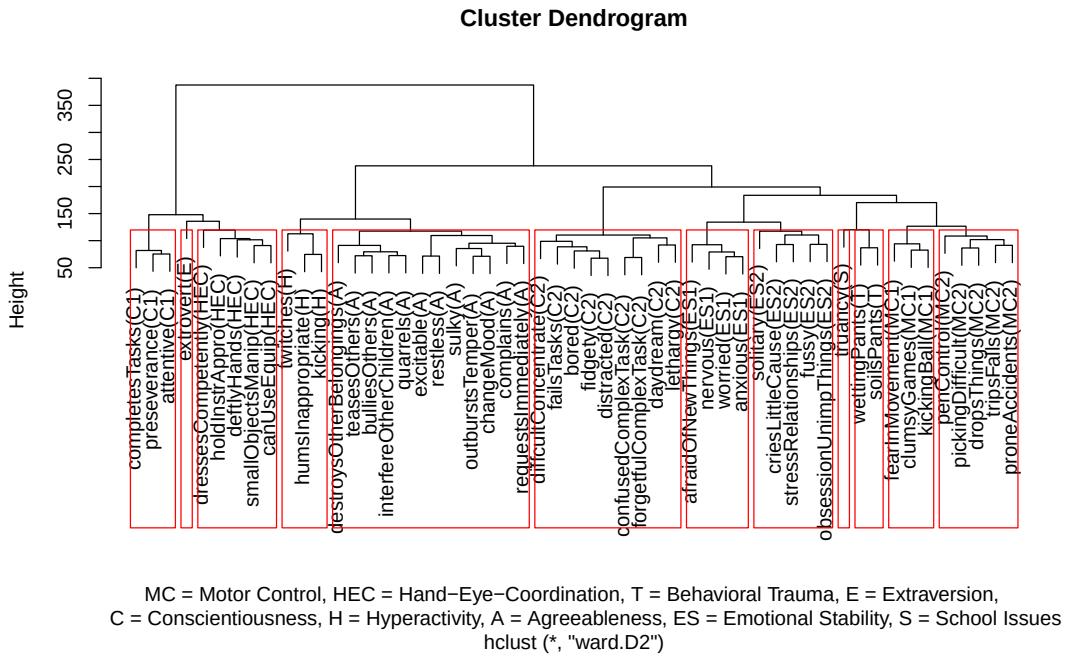
The results of the cluster analysis are captured in Figures 4.4.1 and 4.4.2 for the assessments by the mothers and the teachers, respectively. We use hierarchical clustering based on Ward's method, which starts with single item clusters and in each step decides which pair to merge such that the within-cluster variance minimally increases (Ward Jr, 1963). The outcome of the hierarchical clustering is a dendogram which plots the whole path of the merging steps. With the help of Figure 4.4.1, the procedure can be illustrated as follows. The algorithm starts with 38 clusters, i.e. each item is a cluster. In the next step, the algorithm finds 2 clusters which are the most similar to each other and merges them. In this case, it merges first `clumsy` and `trips` yielding 37 clusters. Then the algorithm finds again the two most similar clusters and merges them until there is only 1 cluster with all the items. The order of merging steps is represented in Figures 4.4.1 and 4.4.2 by the height level of the junctions. The lower the level, the earlier in the algorithm the clusters were merged.

Figure 4.4.1: Clusters of Mother-rated Items



The next step is how to choose the optimal level of clustering. For the hierarchical modeling, one of the recommended methods is to plot the number of

Figure 4.4.2: Clusters of Teacher-rated Items



clusters against a measure of similarity (Belsis et al., 2014). Based on this plot, one looks for a break (jump) which represents that the algorithm put together clusters which are too dissimilar and the merging should stop. In our case, we plot the number of clusters against the increase in the within sum of squares after merging, sometimes called as "merging cost" (Pragarauskaite and Dzemyda, 2012). When the merging cost is relatively too high, the merging should stop as the newly created cluster is too heterogeneous. The analysis of merging costs suggests 8 clusters for the mother-rated items since going from 8 to 7 clusters is relatively costly, see Figure C.1.1 in the Appendix C.1.2. For the teacher-rated items, a jump occurs at merging 12 to 11 clusters. The clusters are represented by red boxes in Figures 4.4.1 and 4.4.2 and got the following labels collected in Table 4.4.1.

In total, we have 22 indices for the analysis: (a) 8 clusters from the mother-rated items, (b) 12 clusters from the teacher-rated items, (c) self-esteem index and the locus of control index. The four personality traits coming from Big Five (E, C, A and ES) are very similar to the clusters obtained in Prevoo and ter Weel (2015) regarding the mother's survey items. Blanden et al. (2007) point out that the relevant variables in the BCS70 are rather close to the variables of Five Factor model. For a preliminary analysis, Tables C.1.7 and C.1.8 contain pairwise correlations between the 22 equally weighted indices separated for mothers and teachers respectively and two *Ability* indices. The results show that there is a

Table 4.4.1: Index Labels

Mother-rated items (8 clusters)	Teacher-rated items (12 clusters)
MC_M = Motor Control,	$MC1_T$ = Motor Control - fast movement,
HEC_M = Hand-Eye-Coordination,	$MC2_T$ = Motor Control - coordination,
T_M = Behavioral Trauma,	HEC_T = Hand-Eye-Coordination,
E_M = Extraversion,	T_T = Behavioral Trauma,
C_M = Conscientiousness,	E_T = Extraversion,
H_M = Hyperactivity,	$C1_T$ = Conscientiousness - (+) formulation,
A_M = Agreeableness,	$C2_T$ = Conscientiousness - (-) formulation,
ES_M = Emotional Stability.	H_T = Hyperactivity,
	A_T = Agreeableness,
	$ES1_T$ = Emotional Stability - inner,
	$ES2_T$ = Emotional Stability - outer,
	S_T = School Behavior.

Subscripts M an T indicate survey items from the mother's and teacher's questionnaire respectively.

slightly positive correlation between *Locus of Control* and both *Ability* indices indicating a potential that there might be a channel between noncognitive and cognitive skills.

By using information from both mother's and teacher's questionnaires, several noncognitive skills are measured from two different sources. This allows us to look into a question whether the source of information plays an important role for predictive purposes. In other words, who should be asked to evaluate the children in order to predict whether they would profit from an early MTP. Low levels of pairwise correlations between the personality traits measured in mother's and teacher's questionnaires in Table C.1.9 indicate that including the views of the parents and teachers is not leading to redundant measurements but to additional information for the model.

4.5 Empirical Results

4.5.1 Unemployment Classifications

In the following, we will evaluate the different model specifications and index constructions solely in terms of their statistical and economic predictive performance. We use the multi-step procedure described in Figure 4.3.1. The overall data set was split randomly into training, validation and test data sets with the shares close to 60%, 20% and 20%, see Table 4.5.1. We apply a specific splitting procedure which divides the three subsets such that the unemployment rate in each subset is close to the overall unemployment rate, i.e. in the first step, the training set is drawn such that the unemployment rate is close to 9.79% in

approximately 60% of the data. In the second step, the test set is drawn such that its unemployment rate is close to 9.79% in approximately 20% of the data. The validation set consists of the remaining observations not selected for the training and the test set. This yields an imprecise 50:50 split between validation and test set meanwhile assuring that the unemployment rate is as close to 9.79% as possible in each subset. Our splitting procedure guarantees that training set and validation set are representative in terms of the unemployment rate, so that choice of the tuning parameter τ selected on the observations of the validation set is based on a representative sample. After applying this splitting procedure, our test sample consists of 549 observations with an unemployment rate of 9.65%, for which we predict their employment status 24 years ahead.

Table 4.5.1: Training, Test and Validation Samples

	Training Set	Validation Set	Test Set	Total
Sample Size	1649	551	549	2749
Unemployment Rate (%)	9.76	9.98	9.65	9.79

Male workers at age 34, source: BCS wave 2004.

All models are estimated by maximum likelihood logit using the equally weighted index, the PCA index and the group lasso index. To get the group lasso index a penalized maximum likelihood method was implemented with λ chosen by a 10-fold cross validation based on a margin-based loss function which approximates the upper bound of misclassification (Lin, 2004). Table 4.5.2 contains various classification measures for our nine different model variants where τ was tuned based on Youden's statistics. The results show that models with high overall accuracy (over 80%) have a tendency to predict majority of cases as employed translating into low true positives (low sensitivity) and a high false negative rate. The results show the typical trade-off for binary outcome models, i.e. when the model detects many true positives, the number of false positives increases as well (making an unemployment program less efficient as many people take it unnecessarily) and when the model detects many true negatives, the number of false negatives increases (i.e. it misses many of truly unemployed people increasing the social cost of unemployment).

In terms of prediction quality, the models solely based on noncognitive skill factors are comparable with the richer models including in addition conventional regressors such as socio-economic controls (social class) and cognitive ability, i.e. even without conventional economic predictors the correct classification is rather high and improves only slightly for most of the classifiers. As a matter of fact, for the group lasso case including additional predictors turns out to be

Table 4.5.2: Out-of-Sample Classification Measures: Statistical Tuning

	M_0	onlyEW	onlyPCA	onlyGL	post only GL	EW	PCA	GL	post GL
Accuracy	0.65	0.53	0.51	0.65	0.81	0.61	0.60	0.62	**0.82**
Sensitivity	0.42	0.47	**0.49**	0.43	0.15	0.42	0.47	**0.49**	0.17
Specificity	0.68	0.54	0.52	0.67	0.88	0.64	0.61	0.64	**0.89**
PPV	0.12	0.10	0.10	0.12	0.12	0.11	0.11	0.13	**0.14**
Costs	204.98	203.94	201.47	200.69	249.34	208.61	197.39	**190.74**	243.96

Classification with τ chosen based on Youden's statistic evaluated on the validation set. The first four rows contain the standard classification measures. The last row contains the expected economic costs for the classifiers for $\varphi = 0.1$, $c = 100$ GBP, $u = 2800$ GBP (average annual JSA payments per JSA claimant) and $\pi = 0.0979$, the corresponding unemployment rate in the 2004 BCS sample. Bold numbers indicate the best values.

counterproductive in terms of accuracy. Comparing the predictive performance with respect to the type of index construction, the group lasso based index unambiguously dominates the two more standard procedures of index construction. The comparison of the models based on the area under the curve (AUC) described in Section 4.3 also underlines the statistical in- and out-of-sample dominance of the models based on the group lasso target-oriented indices as displayed in Table 4.5.3. The model based on conventional regressors (M_0) yields good results in terms of overall accuracy. However, in terms of sensitivity the majority of models including noncognitive skills yield a better result, i.e. they predict better the truly unemployed making noncognitive skills an important predictor for unemployment.

Table 4.5.3: AUC for all Models, Male, 34Y sample

	M_0	onlyEW	onlyPCA	onlyGL	post onlyGL	EW	PCA	GL	post GL
In-sample (Training Set)	0.574	0.658	0.655	0.739	0.756	0.667	0.664	0.736	**0.760**
Out-of-sample (Test Set)	0.533	0.528	0.533	0.567	0.563	0.535	0.538	**0.576**	0.568

AUC calculated for each model's ROC curve. The bold figures represent the best values.

For illustration purposes the last row of Table 4.5.2 contains the economic costs of unemployment implied by sorting based on a purely statistically trained program. For both groups of models, the models containing additional predictors and the models including noncognitive skill factors only, the two group lasso approaches reveal the largest potential to improve on the economic costs without optimizing the tuning parameter τ in this respect.

Table 4.5.4 contains various classification measures for our nine different model variants where τ was tuned regarding the expected assignment costs. It is not too surpring that prediction accuracy drops compared to the classifiers which are tuned w.r.t. to Youden's statistic. However, comparing the costs between the models with economically tuned threshold parameter τ in Table 4.5.4 and statistically tuned threshold parameter τ in Table 4.5.2, economic tuning yields a

considerable 33% reduction in economic costs on average. In the chosen illustrative setting, the cost of the program is relatively low ($c = 100$ GBP) in comparison to the JSA ($u = 2800$ GBP), therefore all the models predict many cases as becoming unemployed (high sensitivity and low specificity) as it is economically more efficient to send people to the training then to pay high unemployment benefits even though many participants actually did not need the training. As a consequence, the statistical measures will now crucially depend on the parameters of the expected costs function.

Table 4.5.4: Out-of-Sample Classification Measures: Economic Tuning

	M_0	onlyEW	onlyPCA	onlyGL	post only GL	EW	PCA	GL	post GL
Accuracy	0.12	**0.39**	0.35	0.13	0.12	0.24	0.26	0.15	0.11
Sensitivity	0.94	0.70	0.72	0.96	**0.98**	0.89	0.91	0.96	0.96
Specificity	0.03	**0.35**	0.31	0.04	0.02	0.17	0.19	0.06	0.02
PPV	0.09	0.10	0.10	0.10	0.10	0.10	**0.11**	0.10	0.10
Costs	138.26	167.06	166.59	132.52	**129.68**	138.75	132.46	131.06	134.15

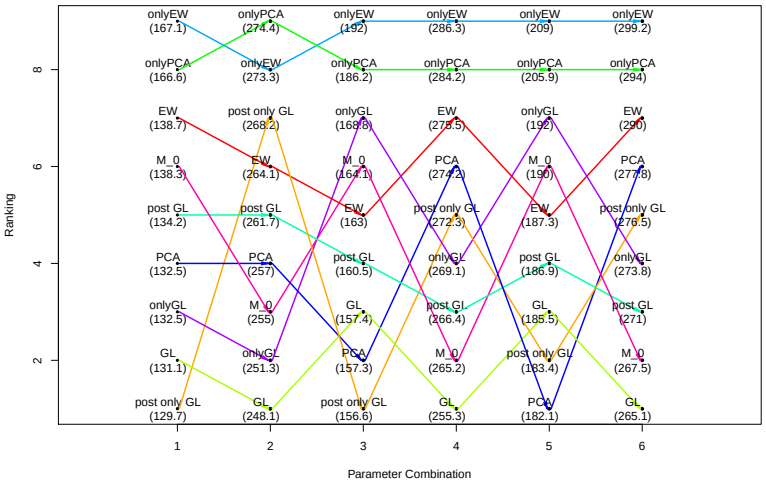
Classification based on cost-minimal τ evaluated on the validation set. The first four rows contain the standard classification measures. The last row contains the expected economic costs for the classifiers for $\varphi = 0.1$, $c = 100$ GBP, $u = 2800$ GBP (average annual JSA payments per JSA claimant) and $\pi = 0.0979$ the corresponding unemployment rate in the 2004 BCS sample. Bold numbers indicate the best values.

The group lasso based models show clear cost reductions as Table 4.5.4 suggests. To get a more robust picture, several configurations of the following parameters were run to analyze the potential of each method in order to reduce economic costs of employment training:

- expected reduced share of unemployment benefit payments: $\varphi = \{0.1, 0.2, 0.3\}$,
- per capita cost of intervention program: $c = \{100, 250\}$ GBP,
- average annual JSA payment per JSA claimant: $u = 2800$ GBP,
- unemployment rate in the 2004 BCS sample (incidence): $\pi = 0.0979$.

The parameter φ is set to three different values to illustrate the method. Alternatively, one could take values of the estimated treatment effects from the labor policy evaluation literature. As the literature leads to very different results depending on the program type and other covariates (Kluve, 2010), we restrain from this strategy at the moment. The values listed above yield 6 different combinations which were analyzed in terms of economic costs. Figure 4.5.1 captures the ranks over all parameter combinations for the 34 year old males. The group lasso model (GL) shows the lowest most stable expected costs. Table 4.5.5 confirms further the results of the graphical analysis. The median and mean ranks of the GL model are far ahead of the other methods underlying the potential of regularization in terms of yielding economically efficient predictive models over a range of parameter combinations.

Figure 4.5.1: Ranks - Economically Trained Models, Male 34Y Sample



Results for 6 parameter combinations. $\varphi = 0.1$ in combinations 1 and 2 $\varphi = 0.2$ in 3 and 4, $\varphi = 0.3$ in 5 and 6. $c = 100$ in 1, 3 and 5 and $c = 250$ in 2, 4 and 6. $u = 2800$ and $\pi = 0.0979$.

Table 4.5.5: Ranks for the Economic Costs, Economically Trained, Male, 34Y sample

	M_0	onlyEW	onlyPCA	onlyGL	post onlyGL	EW	PCA	GL	post GL
Mean	4.17	8.83	8.17	4.50	3.50	6.17	3.67	1.83	4.00
Median	4.50	9.00	8.00	4.00	3.50	6.50	3.50	1.50	4.00

Rank counts based on the economic costs across all parameter combinations.

4.5.2 Selected Skill Factors

In comparison to a pure machine learning approach or a kitchen-sink regression our group lasso approach yields interpretable prediction equations which may provide valuable information which skill factors are highly predictive and which may then be relevant for the design of appropriate intervention programs. Table 4.5.6 contains the estimates of M_0, *onlyPCA*, *onlyGL*, *PCA* and *GL* representing the main combinations of model covariate sets and types of weighting schemes. The models with equal indices were left out as their coefficient estimates are similar to PCA models. In the following, the main focus will be on the group lasso results as they perform economically well. The regularized models selected 16 indices out of 22 indicating that noncognitive skills play an important role for predictions and speaking against the models picking only one specific statistically significant noncognitive skill as in Prevoo and ter Weel (2015). Except of the extraversion (E_M) index, the signs of the covariates are the same in regularized and unregularized models. Interestingly, the regularized models do not include social background, *SocialClass*, in the model suggesting that this typically considered unemployment channel would not be helpful for prediction. In other words, so-

Table 4.5.6: Unemployment Equation Estimates, Sample 34Y

	M_0	onlyPCA	onlyGL	PCA	GL
Intercept	-1.9599^{***}	-2.3424^{***}	-2.2867	-2.0996^{***}	-2.2614
Noncognitive skills indices based on self-reported items					
SE		0.1098	0.0589	0.1083	0.0741
LC		-0.0738	-0.1274	-0.0309	-0.0774
Noncognitive skills indices based on mother's assessment					
MC_M		-0.0197	—	-0.0295	—
HEC_M		0.0398	—	0.0449	—
C_M		0.2772^{*}	0.1677	0.3100^{**}	0.1863
H_M		-0.0082	—	0.0018	—
A_M		-0.2390^{*}	-0.165	-0.2352^{*}	-0.1541
ES_M		-0.0750	-0.0765	-0.0757	-0.0762
TR_M		-0.0557	—	-0.0552	—
E_M		0.0218	-0.0022	0.0312	-0.0002
Noncognitive skills indices based on teacher's assessment					
$MC1_T$		0.2218	0.0399	0.2190	0.0291
$MC2_T$		0.1412	0.0436	0.1307	0.0423
HEC_T		-0.1405	-0.0447	-0.1203	-0.0358
$C1_T$		0.2708	—	0.2778	—
$C2_T$		-0.5155^{**}	-0.2458	-0.4618^{*}	-0.1742
H_T		0.0760	0.0494	0.0797	0.0503
A_T		-0.1677	-0.0783	-0.1794	-0.0824
$ES1_T$		0.0717	—	0.0888	—
$ES2_T$		-0.0063	-0.01	-0.0214	-0.0168
TR_T		0.0681	0.0246	0.0705	0.0237
E_T		-0.2642^{*}	-0.084	-0.2702^{*}	-0.0768
S_T		0.0421	0.0195	0.0413	0.0182
Control variables					
SocialClass10Y2	-0.3082			-0.3211	—
SocialClass10Y3	-0.2814			-0.2682	—
SocialClass10Y4	-0.2756			-0.1942	—
SocialClass10Y5	-0.1398			-0.0954	—
SocialClass10Y6	-0.1458			0.0734	—
AbilityMath10Y	-0.2640^{*}			-0.2586	-0.1362
AbilityRead10Y	-0.0112			0.0868	—

ML logit estimates of the unemployment equation. Dependent variable: employed = 0, unemployed = 1. Models M_0, *onlyPCA* and *PCA* are estimated by standard ML. The stars represent the following p-values: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Models *onlyGL* and *GL* are estimated by the regularized ML. The regularization parameter λ is chosen by the 10-fold cross-validation. The index coefficients are identified by the sum up constraint $\sum_j \hat{\omega}_j^I = 1$. See Appendix C.2 for more details. Currently there is no standardized significance test for regularized coefficients allowing to assess the statistical significance. Subscripts of the NCS indices indicate whether it is built based on mother-rated questions (M) or teacher-rated questions (T). — means that lasso excluded the variable from the model.

cial background has no predictive discriminatory effect for unemployment as well as the reading ability, *AbilityRead*. The typical predictor for the unemployment, mathematical ability (*AbilityMath*), was picked up by the group lasso with a neg-

ative effect on the predicted probability of unemployment similar to Feinstein (2000).

Looking into the Big Five indices, *Agreeableness* decreases the probability of the 4-year prevalence of being unemployed as it would be expected from people who get along well with others. The same logic holds for *Emotional Stability* and *Extraversion* which are in general highly appreciated noncognitive skills in the society and therefore expected to have a positive effect on employment. Going beyond Big Five, *Hand Eye Coordination (HEC)* measured at the age of 10 predicts lower unemployment probability. This could mean that a child with better developed fine physical coordination is at a higher cognitive level which is then indirectly measured through the *HEC* index. A survey devoted to studies which analyze eye and reaching movements as measures of real-time cognitive processing can be found in Spivey et al. (2008). Conversely, *Motor Control* increases the probability of unemployment. By a thorough look at the survey items, the motor control indices could be proxies of competitiveness and self-confidence as indirectly measured by focusing on the coordination during games and what impression a child leaves when performing a focused movement. Piek et al. (2006) study an impact of gross and fine motor skills in children and adolescents on their perceived self-worth concluding that poor motor control negatively impacts self-perception. In this light, both *Motor Control* indices could be then predictors for a higher probability of the 4-year prevalence of being unemployed as more self-confident people are not afraid of short-term unemployment with an outlook of getting a position they want as it would be predicted by the Expectancy-Value Theory which is concisely outlined in Vansteenkiste et al. (2005). This would also justify the logic behind the positive prediction coefficient for *Self-Esteem*. Children with higher level of *Hyperactivity* and signs of *Trauma* have a higher predicted probability of being unemployed as they could have problems with concentration leading to difficulties on the labor market as descriptively analyzed in a phone survey by Biederman and Faraone (2006). The last predictor is related to *School Issues*, in particular to truancy. A truant child has a higher probability of being unemployed as it is probably a first sign of a potential dropout leading to a lower level of education with its consequences for the labor market.

4.6 Conclusion

This paper takes a closer look on the predictive power of non-cognitive skills for unemployment by means of machine learning techniques. This existence of considerable predictive power of noncognitive skills has been claimed in many empirical studies which led to the claim that public policies should pay more

attention to programs to enhance these skills (Heckman and Kautz (2012)). This paper is trying to provide further evidence on this hypothesis based on real out-of-sample forecast and applying several novel steps in the modeling procedure.

The group lasso approach proposed here accounts for several desirable features. It guarantees a parsimonious selection of factors and avoids over-fitting in the presence of data sets containing a large number of skill factors. Moreover lassoing helps to construct target-oriented skill indices and select only those indices which are most relevant in predicting a certain outcome variable. We show that our group lasso approach cannot be persistently outperformed by standard approaches using equally weighted indices or PCA based indices, especially in terms of economic costs. In comparison to other black box classification techniques, our predictive model is easily interpretable and can contribute to a discussion about important predictors of unemployment and help to design selection processes at an early stage of life cycle.

We also investigated two different perspectives for the optimal classification: statistical and economic one. Meanwhile statistical tuning of classification models is the dominant approach in the machine learning literature, ignoring economic costs in applied economic research can lead to inefficient policy recommendations in the larger cost and benefit perspective. The idea of focusing on economic tuning of classification models instead of the statistical tuning can be easily extended to other classification techniques such as support vector machines or decision trees yielding potentially new interesting results.

Our empirical findings are based on data from the British Cohort Study (BCS) containing life cycle information of individuals born in 1970. Admittedly, expecting non-cognitive skill factors which were surveyed at the age of 10 to be predictors of individual labor market outcomes 24 years later is very ambitious. Recognizing the fact that the survey items were not even designed for a purpose of individual unemployment prediction, the predictive quality of the factors is astonishing and leading to reduction of economic costs if the group lasso indices are used.

The BCS is a very rich, but also very specific data source in terms of its design and the definition of non-cognitive skill factors. Therefore, in future work our findings should be confronted with results based on different data sources with alternative definitions of the skill factors and different forecasting horizons. Moreover, the choice of the shrinkage parameter of the group lasso is rather conventionally chosen by means of 10-fold cross-validation. Here we see room for further improvement, since cross-validation is known to yield rather unstable estimates of the optimal shrinkage parameter. Future work could consider stability selection strategy based on subsampling to create more stable solutions as proposed by Meinshausen and Bühlmann (2010).

5 Conclusion

The thesis introduces two generalizations of existing regularization methods based on L_1- and L_2-norm regularization. One focuses on reduction of point estimation risk of conditional means in data sets with a clear group structure by exploiting a flexible L_2-norm regularization scheme. The other one detects structural breaks in a model with time-varying coefficients by using an L_1-norm regularization. The last empirical contribution develops a new index building scheme for noncognitive skill indices based on maximization of the predictive power of the indices for classification problems combining a machine learning and regularization approaches and adds a way of economic tuning of the model parameters which is not widely used in the literature. The method yields an interpretable predictive economic model and its potential applicability is illustrated on classifying of unemployed.

As partly illustrated in the three chapters of the thesis, the methodological contributions together with more detailed data sets open a way for the regularization and machine learning techniques to answer relevant economic questions, reveal heterogeneity in data and improve the predictive or explanatory power of the models depending on the problem and the adequate method. The recent contributions about the general value added of the regularization and machine learning techniques in Fan et al. (2011), Varian (2014), Mullainathan and Spiess (2017) and Athey (2018) for economic research seem to predict that these methods will become standard tools in econometrics.

Zusammenfassung

Die Dissertation führte zwei Verallgemeinerungen bestehender Regularisierungsmethoden basierend auf der L_1- und L_2-Norm Regularisierung ein. Eine konzentriert sich auf die Verringerung des Punktschätzungsrisikos von bedingten Mittelwerten in Datensätzen mit einer klaren Gruppenstruktur durch Nutzung eines flexiblen L_2-Norm Regulationsschemas. Die andere erkennt strukturelle Brüche in einem Modell mit zeitlich variierenden Koeffizienten mithilfe einer L_1-Norm Regularisierung. Der letzte empirische Beitrag entwickelt ein neues Indexbildungsschema für nichtkognitive Fähigkeitsindizes, das auf der Maximierung der Vorhersagekraft der Indizes für Klassifikationsprobleme basiert, wobei maschinelle Lern- und Regularisierungsansätze kombiniert werden, und fügt einen Weg zur wirtschaftlichen Abstimmung der Modellparameter hinzu, der nicht weit verbreitet in der Literatur ist. Die Methode liefert ein interpretierbares prädiktives Modell, und ihre mögliche Anwendbarkeit wird bei der Klassifikation von Arbeitslosen veranschaulicht.

6 Author's Contributions

Chapter	**Co-Author (CA)**	**CA's Status**	**Own - CA's Contribution** (in %)
2	Phillip Heiler	PhD Student	50-50
3	-	-	100-0
4	Winfried Pohlmeier	1st Supervisor	60-40

References

Aitchison, J. and Aitken, C. G. (1976). Multivariate Binary Discrimination by the Kernel Method. Biometrika, 63(3):413–420.

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. In 2nd International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973, pages 267–281.

Almlund, M., Duckworth, A. L., Heckman, J., and Kautz, T. (2011). Chapter 1 - personality psychology and economics. In Eric A. Hanushek, S. M. and Woessmann, L., editors, Handbook of The Economics of Education, volume 4 of Handbook of the Economics of Education, pages 1 – 181. Elsevier.

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

Athey, S. (2018). The Impact of Machine Learning on Economics. University of Chicago Press.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. Ann. Statist., 47(2):1148–1178.

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

Bai, J. and Perron, P. (2003). Critical values for multiple structural change tests. Econometrics Journal, 6:72–78.

Belsis, P., Koutoumanos, A., and Sgouropoulou, C. (2014). Pburc: a patterns-based, unsupervised requirements clustering framework for distributed agile software development. Requirements engineering, 19(2):213–225.

Bertrand, M., Crépon, B., Marguerie, A., and Premand, P. (2017). Contemporaneous and Post-Program Impacts of a Public Works Program: Evidence from Côte d'Ivoire.

Biederman, J. and Faraone, S. V. (2006). The effects of attention-deficit/hyperactivity disorder on employment and household income. Medscape General Medicine, 8(3):12.

Blanden, J., Gregg, P., and Macmillan, L. (2007). Accounting for intergenerational income persistence: Non-cognitive skills, ability and education. Economic Journal, 117:C43 – C60.

Bleakley, K. and Vert, J.-P. (2011). The group fused Lasso for multiple change-point detection. ArXiv e-prints.

Borghans, L., Duckworth, A. L., Heckman, J. J., and Ter Weel, B. (2008). The Economics and Psychology of Personality Traits. Journal Human Resources, 43(4):972–1059.

Brodie, J., Daubechies, I., De Mol, C., Giannone, D., and Loris, I. (2009). Sparse and stable markowitz portfolios. Proceedings of the National Academy of Sciences, 106(30):12267–12272.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model Selection: An Integral Part of Inference. Biometrics, 53(2):603–618.

Burnham, K. P. and Anderson, D. (2003). Model Selection and Multi-Model Inference. Springer New York.

Butler, N., Haslum, M., Barker, W., and Morris, A. (1982). Child health and education study: First report to the department of education and science on the 10-year follow-up. Bristol: Department of Child Health, University of Bristol.

Caliendo, M., Cobb-Clark, D. A., and Uhlendorff, A. (2014a). Locus of control and job search strategies. Review of Economics and Statistics, 97(1):88–103.

Caliendo, M., Fossen, F., and Kritikos, A. S. (2014b). Personality characteristics and the decisions to become and stay self-employed. Small Business Economics, 42(4):787–814.

Card, D. and Krueger, A. B. (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. The American Economic Review, 84(4):772–793.

Chan, N. H., Yau, C. Y., and Zhang, R.-M. (2013). Group lasso for structural break time series. Journal of the American Statistical Association, In Press, Accepted Manuscript(ja):–.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. Biometrika, 95(3):759–771.

Cheng, X., Liao, Z., and Shi, R. (2015). Uniform Asymptotic Risk of Averaging GMM Estimator Robust to Misspecification. PIER Working Paper 15-017.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. American Economic Review, 107(5):261–65.

Chernozhukov, V., Hansen, C., and Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. American Economic Review, 105(5):486–90.

Claeskens, G., Hjort, N. L., et al. (2008). Model Selection and Model Averaging. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Cohen, A. (1966). All Admissible Linear Estimates of the Mean Vector. The Annals of Mathematical Statistics, pages 458–463.

Conners, C. K. (1969). A teacher rating scale for use in drug studies with children. American journal of Psychiatry, 126(6):884–888.

Conti, G., Frühwirth-Schnatter, S., Heckman, J. J., and Piatek, R. (2014). Bayesian exploratory factor analysis. Journal of Econometrics, 183:31 – 57.

Duckworth, A. L. and Seligman, M. E. (2005). Self-discipline outdoes iq in predicting academic performance of adolescents. Psychological Science, 16:939–944.

Fan, J., Lv, J., and Qi, L. (2011). Sparse high-dimensional models in economics. Annual Review of Economics, 3(1):291–317.

Fawcett, T. (2006). An introduction to roc analysis. Pattern Recognition Letters, 27(8):861 – 874. ROC Analysis in Pattern Recognition.

Feinstein, L. (2000). The relative economic importance of academic, psychological and behavioural attributes developed on childhood. CEPDP (443), Centre for Economic Performance, London School of Economics and Political Science.

Fienberg, S. E. and Holland, P. W. (1973). Simultaneous Estimation of Multinomial Cell Probabilities. Journal of the American Statistical Association, 68(343):683–691.

Gammage, P. (1975). Socialisation, schooling and locus of control. PhD thesis, Bristol University, Bristol, England.

Gneezy, U., Meier, S., and Rey-Biel, P. (2011). When and Why Incentives (Don't) Work to Modify Behavior. The Journal of Economic Perspectives, 25(4):191–209.

Gneezy, U. and Rustichini, A. (2000a). A Fine is a Price. The Journal of Legal Studies, 29:1–17.

Gneezy, U. and Rustichini, A. (2000b). Pay Enough or Don't Pay at All. The Quarterly Journal of Economics, 115(3):791–810.

Gould, N. I. M. (1985). On Practical Conditions for the Existence and Uniqueness of Solutions to the General Equality Quadratic Programming Problem. Mathematical Programming, 32(1):90–99.

Goyette, C. H., Conners, C. K., and Ulrich, R. F. (1978). Normative data on revised conners parent and teacher rating scales. Journal of Abnormal Child Psychology, 6(2):221–236.

Hall, P. (1981). On Nonparametric Multivariate Binary Discrimination. Biometrika, 68(1):287–294.

Hall, P., Li, Q., and Racine, J. S. (2007). Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors. The Review of Economics and Statistics, 89(4):784–789.

Hall, P., Racine, J., and Li, Q. (2004). Cross-Validation and the Estimation of Conditional Probability Densities. Journal of the American Statistical Association, 99(468).

Hansen, B. E. (2007). Least Squares Model Averaging. Econometrica, 75(4):1175–1189.

Hansen, B. E. (2014). Model Averaging, Asymptotic Risk, and Regressor Groups. Quantitative Economics, 5(3):495–530.

Hansen, B. E. (2016a). Efficient Shrinkage in Parametric Models. Journal of Econometrics, 190(1):115–132.

Hansen, B. E. (2016b). The Risk of James–Stein and Lasso Shrinkage. Econometric Reviews, 35(8-10):1456–1470.

Hansen, B. E. and Racine, J. S. (2012). Jackknife Model Averaging. Journal of Econometrics, 167(1):38–46.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. Journal of the American Statistical Association, 105(492):1480–1493.

Hastie, T., Tibshirani, R., and Wainright, M. (2015). Statistical Learning with Sparisty The Lasso and Generalizations. Monographs on Statistics and Probability. CRC Press.

Heckman, J. J. and Kautz, T. (2012). Hard evidence on soft skills. Labour Economics, 19(4):451 – 464.

Hjort, N. L. and Claeskens, G. (2003). Frequentist Model Average Estimators. Journal of the American Statistical Association, 98(464):879–899.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12(1):55–67.

John, K. and Thomsen, S. L. (2014). Heterogeneous returns to personality: the role of occupational choice. Empirical Economics, 47(2):553–592.

Kluve, J. (2010). The effectiveness of european active labor market programs. Labour Economics, 17(6):904 – 918.

Knaus, M., Lechner, M., and Strittmatter, A. (2017). Heterogeneous employment effects of job search programmes: A machine learning approach.

Land, S. R. and Friedman, J. H. (1997). Variable fusion: A new adaptive signal regression method. Technical report, Technical Report 656, Department of Statistics, Carnegie Mellon University Pittsburgh.

Lawrence, D. (1973). Improved Reading through Counselling. Ward Lock Educational: London.

Lawrence, D. (1978). Counselling students with reading difficulties: a handbook for tutors and organizers. Good Reading: London.

Leeb, H. and Pötscher, B. M. (2008). Sparse estimators and the oracle property, or the return of hodges' estimator. Journal of Econometrics, 142(1):201–211.

Li, K.-C. (1987). Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set. The Annals of Statistics, 15(3):958–975.

Li, Q. and Racine, J. (2003). Nonparametric Estimation of Distributions with Categorical and Continuous Data. Journal of Multivariate Analysis, 86(2):266–292.

Liang, H., Zou, G., Wan, A. T., and Zhang, X. (2011). Optimal Weight Choice for Frequentist Model Average Estimators. Journal of the American Statistical Association, 106(495):1053–1066.

Lin, Y. (2004). A note on margin-based loss functions in classification. Statistics & Probability Letters, 68(1):73 – 82.

Liu, C.-A. (2015). Distribution Theory of the Least Squares Averaging Estimator. Journal of Econometrics, 186(1):142–159.

Mallows, C. L. (1973). Some Comments on C_p. Technometrics, 15(4):661–675.

Meier, L., van de Geer, S., and Bühlmann, P. (2008). The group lasso for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 70(1):53 – 71.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4):417–473.

Merlevède, F., Peligrad, M., and Rio, E. (2009). Bernstein inequality and moderate deviations under strong mixing conditions, volume Volume 5 of Collections, pages 273–292. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Metz, C. E. (1978). Basic principles of roc analysis. Seminars in Nuclear Medicine, 8(4):283 – 298.

Mueller, G. and Plug, E. (2006). Estimating the effect of personality on male-female earnings. Industrial and Labor Relations Review, 60:3–22.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. Journal of Economic Perspectives, 31(2):87–106.

Nyhus, E. and Pons, E. (2005). The effects of personality on earnings. Journal of Economic Psychology, 26:363–384.

Oman, S. D. (1982). Shrinking Towards Subspaces in Multiple Linear Regression. Technometrics, 24(4):307–311.

Ouyang, D., Li, Q., and Racine, J. S. (2009). Nonparametric Estimation of Regression Functions with Discrete Regressors. Econometric Theory, 25(01):1–42.

Piatek, R. and Pinger, P. (2016). Maintaining (locus of) control? data combination for the identification and inference of factor structure models. Journal of Applied Econometrics, 31:734–755.

Piek, J. P., Baynam, G. B., and Barrett, N. C. (2006). The relationship between fine and gross motor ability, self-perceptions and self-worth in children and adolescents. Human Movement Science, 25(1):65 – 75. Approaches to Sensory-Motor Development in Infants and Children.

Pragarauskaite, J. and Dzemyda, G. (2012). Visual decisions in the analysis of customers online shopping behavior. Nonlinear Analysis: Modelling and Control, 17(3):355–368.

Prevoo, T. and ter Weel, B. (2015). The importance of early conscientiousness for socio-economic outcomes: evidence from the british cohort study. Oxford Economic Papers, 67:918–948.

Qian, J. and Su, L. (2016). Shrinkage estimation of regression models with multiple structural changes. Econometric Theory, 32(6):1376–1433.

Rotter, J. (1966). Generalized expectancies for internal versus external control of reinforcement. Psychological Monographs, 80(609):1–28.

Rutter, M. (1967). A children's behaviour questionnaire for completion by teachers: preliminary findings. Journal of child Psychology and Psychiatry, 8(1):1–11.

Rutter, M., Tizard, J., and Whitmore, K. (1970). Education, health and behaviour. Longman.

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6(2):461–464.

Shmueli, G. (2010). To explain or to predict? Statististical Science, 25(3):289–310.

Simonoff, J. S. (1995). Smoothing Categorical Data. Journal of Statistical Planning and Inference, 47(1-2):41–69.

Simonoff, J. S. (1996). Smoothing Methods in Statistics. Springer Series in Statistics. Springer New York.

Spivey, M., Richardson, D., and Dale, R. (2008). Oxford handbook of human action, chapter 12: The movement of eye and hand as a window into language and cognition, pages 225–249. New York: Oxford University Press.

Stein, C. M. (1956). Efficient Nonparametric Testing and Estimation. In Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. The Regents of the University of California.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58:267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(1):91–108.

Titterington, D. and Bowman, A. (1985). A Comparative Study of Smoothing Procedures for Ordered Categorical Data. Journal of Statistical Computation and Simulation, 21(3-4):291–312.

Tutz, G. and Oelker, M.-R. (2017). Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. International Statistical Review, 85(2):204–227.

Uysal, S. D. (2015). Doubly robust estimation of causal effects with multivalued treatments: An application to the returns to schooling. Journal of Applied Econometrics, 30(5):763–786.

Uysal, S. D. and Pohlmeier, W. (2011). Unemployment duration and personality. Journal of Economic Psychology, 32:980–992.

Vansteenkiste, M., Lens, W., De Witte, H., and Feather, N. T. (2005). Understanding unemployed people's job search behaviour, unemployment experience and well-being: A comparison of expectancy-value theory and self-determination theory. British Journal of Social Psychology, 44(2):269–287.

Varian, H. R. (2014). Big data: New tricks for econometrics. Journal of Economic Perspectives, 28(2):3–28.

Viinikainen, J. and Kokko, K. (2012). Personality traits and unemployment: Evidence from longitudinal data. Journal of Economic Psychology, 33(6):1204 – 1222.

Ward Jr, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301):236–244.

Yu, D., Won, J.-H., Lee, T., Lim, J., and Yoon, S. (2013). High-dimensional Fused Lasso Regression using Majorization-Minimization and Parallel Processing. ArXiv e-prints.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 68(1):49–67.

Zhang, X., Liang, H., et al. (2011). Focused Information Criterion and Model Averaging for Generalized Additive Partial Linear Models. The Annals of Statistics, 39(1):174–200.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 67(2):301–320.

A Appendices for Shrinkage for Categorical Regressors

A.1 Regular Appendix

A.1.1 FOC and SOC Conditions for (2.2.2)

Let $S_{\mathbf{\Lambda}}(\boldsymbol{\mu})$ denote the objective function in (2.2.2) where $\mathbf{\Lambda} = (\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1J}, \lambda_{21}, \ldots, \lambda_{JJ})$ and $\lambda_{kk} = 0$ for all $k = \{1, \ldots, J\}$. Note that:

$$\frac{\partial S_{\mathbf{\Lambda}}(\boldsymbol{\mu})}{\partial \mu_k} = -2\sum_{i=1}^{n}(Y_i - \mathbf{D}_i'\boldsymbol{\mu})D_{ik} + 2\sum_{j=1}^{J}\lambda_{kj}(\mu_k - \hat{\mu}_j) \tag{A.1.1}$$

$$\frac{\partial^2 S_{\mathbf{\Lambda}}(\boldsymbol{\mu})}{\partial \mu_k^2} = 2n_k + 2\sum_{j=1}^{J}\lambda_{kj} \tag{A.1.2}$$

$$\frac{\partial^2 S_{\mathbf{\Lambda}}(\boldsymbol{\mu})}{\partial \mu_k \partial \mu_l} = 0 \qquad l \neq k \tag{A.1.3}$$

Setting (A.1.1) equal to zero to solve for $\hat{\mu}_k^{PCS}$ and rearranging the terms yields

$$\sum_{i=1}^{n} Y_i D_{ik} + \sum_{j=1}^{J}\lambda_{kj}\hat{\mu}_j = \hat{\mu}_k^{PCS}(\sum_{j=1}^{J}\lambda_{kj} + n_k).$$

The estimate $\hat{\mu}_k^{PCS}$ exists if and only if $\sum_{j=1}^{J}\lambda_{kj} \neq -n_k$.

The matrix of second derivatives of $S_{\mathbf{\Lambda}}(\boldsymbol{\mu})$ is a diagonal matrix that leads to a strictly convex penalty if and only if

$$\sum_{j=1}^{J}\lambda_{kj} > -n_k \text{ for all } k \in \{1, \ldots, J\}.$$

An estimator defined as the solution to (2.2.2) exists and is a unique global minimizer if and only if $\sum_{j=1}^{J}\lambda_{kj} > -n_k$ for all $k \in \{1, \ldots, J\}$.

A.1.2 First Order Approximation of the Modified Least Squares

$$\hat{\mu}_k = \frac{\sum_{i=1}^{n} D_{ik}Y_i}{\sum_{i=1}^{n} D_{ik} + \mathbb{1}(\sum_{i=1}^{n} D_{ik} = 0)}$$

By definition

$$\hat{\mu}_k - \mu_k = \frac{\sum_{i=1}^n D_{ik}(Y_i - \mu_k)}{\sum_{i=1}^n D_{ik} + \mathbb{1}(\sum_{i=1}^n D_{ik} = 0)} - \mu_k \frac{\mathbb{1}(\sum_{i=1}^n D_{ik} = 0)}{\sum_{i=1}^n D_{ik} + \mathbb{1}(\sum_{i=1}^n D_{ik} = 0)}.$$

Using iid and $p_k = E[D_{ik}] > 0$, the WULLN and CLT imply

$$n^{-1} \sum_{i=1}^n D_{ik} = p_k + o_p(1)$$
$$n^{-1/2} \sum_{i=1}^n D_{ik}(Y_i - \mu_k) \xrightarrow{d} \mathcal{N}(0, \sigma_k^2 p_k).$$

Because $p_k > 0$ and sample means and the indicator are bounded a.s., we have that

$$n^{-1/2} \mu_k \mathbb{1}(\sum_{i=1}^n D_{ik} = 0) = o_p(1)$$

and

$$n^{-1/2} \frac{\sum_{i=1}^n D_{ik}(Y_i - \mu_k)}{n^{-1} \sum_{i=1}^n D_{ik} + n^{-1} \mathbb{1}(\sum_{i=1}^n D_{ik} = 0)} = n^{-1/2} \frac{\sum_{i=1}^n D_{ik}(Y_i - \mu_k)}{p_k} + o_p(1)$$

which together imply that

$$\hat{\mu}_k - \mu_k = \frac{1}{n p_k} \sum_{i=1}^n D_{ik}(Y_i - \mu_k) + o_p(n^{-1/2}).$$

A.1.3 Proof of Proposition 2.3.1

The bias of the PCS using the first order approximation is given by

$$\begin{aligned} E[\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k)] - \mu_k &= \sum_{j=1}^J \omega_{kj}(E[\hat{\mu}_j] - \mu_k) \\ &= \sum_{j=1}^J \omega_{kj}(\mu_j - \mu_k). \end{aligned}$$

The variance of the approximated PCS is given by

$$V[\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k)] = \sum_{j=1}^J \omega_{kj}^2 V[\hat{\mu}_j]$$

as $E[\hat{\mu}_j] - \mu_j \approx 0$, observations being independent, and cell means being approximately uncorrelated, i.e.

$$E[(\hat{\mu}_k - \mu_k)(\hat{\mu}_l - \mu_l)] \approx \frac{1}{n^2 p_k p_l} \sum_{i=1}^n \sum_{j=1}^n E[D_{ik}(Y_i - \mu_k)] E[D_{jl}(Y_j - \mu_l)]$$

$$= 0$$

for all $k \neq l$. The variances are given by

$$\begin{aligned} V[\hat{\mu}_j] &\approx E\left[\left(\frac{1}{np_j}\sum_{i=1}^{n} D_{ij}(Y_i - \mu_j)\right)^2\right] \\ &= \frac{1}{np_j^2}E[D_{ij}(Y_i - \mathbf{D}_i'\boldsymbol{\mu})^2] \\ &= \frac{1}{np_j^2}E[D_{ij}\varepsilon_i^2] \\ &= \frac{1}{np_j}E[\varepsilon_i^2|D_{ij} = 1] \\ &= \frac{\sigma_j^2}{np_j}. \end{aligned}$$

where second step is due to independent observations. The MSE in (2.3.4) then follows by the usual bias-variance decomposition.

A.1.4 Proof of Theorem 2.3.1

The problem of the constrained minimization of (2.3.4) can be rewritten as the following Lagrangian:

$$\begin{aligned} \min_{\boldsymbol{\omega}_k, \alpha_k} \mathcal{L}(\boldsymbol{\omega}_k, \alpha_k) &= \min_{\boldsymbol{\omega}_k, \alpha_k} \boldsymbol{\omega}_k' H_k \boldsymbol{\omega}_k + 2\alpha_k(1 - \boldsymbol{\iota}_J'\boldsymbol{\omega}_k), \\ \text{with } \boldsymbol{\omega}_k &= (\omega_{k1}, \omega_{k2}, \ldots, \omega_{kJ})', \\ H_k &= \Delta_k \boldsymbol{\mu}\boldsymbol{\mu}'\Delta_k' + diag(\boldsymbol{\gamma})^{-1}/n, \\ \alpha_k &= \text{ Lagrange multiplier.} \end{aligned}$$

The FOCs are given by

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{\omega}_k, \alpha_k)}{\partial \boldsymbol{\omega}_k} &= 2H_k\boldsymbol{\omega}_k - 2\alpha_k\boldsymbol{\iota}_J = 0 \\ \frac{\partial \mathcal{L}(\boldsymbol{\omega}_k, \alpha_k)}{\partial \alpha_k} &= 2(1 - \boldsymbol{\iota}_J'\boldsymbol{\omega}_k) \quad = 0 \end{aligned}$$

The solution of setting the FOC to zero gives optimal values:

$$\begin{aligned} \alpha_k^* &= [\boldsymbol{\iota}_J'(H_k'H_k)^{-1}H_k'\boldsymbol{\iota}_J]^{-1} \\ \boldsymbol{\omega}_k^* &= [\boldsymbol{\iota}_J'(H_k'H_k)^{-1}H_k'\boldsymbol{\iota}_J]^{-1}(H_k'H_k)^{-1}H_k'\boldsymbol{\iota}_J \end{aligned}$$

The expression for ω_{kj}^* can be inferred from the j-th entry of $\boldsymbol{\omega}_k^*$. For uniqueness conditions consider Extended Appendix A.2.3.

A.1.5 WMSE Optimal and Plugin Smoothing Parameters for Kernel and (generalized) Ridge Regression

For all methods we choose the weighted MSE criterion with $W = V_0^{-1}$ being the inverse of the MLE/least squares variance-covariance matrix. The first-stage estimate is chosen to be the (modified) ordinary least squares. Note that for a given PCS estimator (or restricted version thereof) $\hat{\boldsymbol{\mu}}(\boldsymbol{\Lambda})$ with smoothing parameter vector $\boldsymbol{\Lambda}$, the criterion can be written as:

$$E[(\hat{\boldsymbol{\mu}}(\boldsymbol{\Lambda}) - \boldsymbol{\mu})' W (\hat{\boldsymbol{\mu}}(\boldsymbol{\Lambda}) - \boldsymbol{\mu})] = \sum_{k=1}^{J} \gamma_k E[(\hat{\mu}_k(\boldsymbol{\Lambda}) - \mu_k)^2].$$

Kernel Smoothing

The implicit kernel constraints yield the estimator of the form

$$\hat{\mu}_k^{Kernel}(\boldsymbol{\Lambda}) = \frac{n_k \bar{Y}_k}{n_k + \sum_{l \neq k} \lambda_l} + \frac{\sum_{j \neq k} \lambda_j \hat{\mu}_j}{n_k + \sum_{l \neq k} \lambda_l}. \tag{A.1.4}$$

or equivalently by using the WLLN and $\bar{Y}_k = \hat{\mu}_k$

$$\hat{\mu}_k^{Kernel}(\boldsymbol{\Lambda}) - \mu_k = \frac{n p_k (\hat{\mu}_k - \mu_k) + \sum_{j \neq k} \lambda_j (\hat{\mu}_j - \mu_k)}{n p_k + \sum_{l \neq k} \lambda_l} + o_p(1) \tag{A.1.5}$$

and thus

$$\begin{aligned}(\hat{\mu}_k^{Kernel}(\boldsymbol{\Lambda}) - \mu_k)^2 \approx & \frac{(n p_k)^2 (\hat{\mu}_k - \mu_k)^2 + 2 n p_k (\hat{\mu}_k - \mu_k) \sum_{j \neq k} \lambda_j (\hat{\mu}_j - \mu_k)}{(n p_k + \sum_{l \neq k} \lambda_l)^2} \\ & + \frac{\sum_{j \neq k} \sum_{m \neq k, j} \lambda_j \lambda_m (\hat{\mu}_j - \mu_k)(\hat{\mu}_m - \mu_k) + \sum_{j \neq k} \lambda_j^2 (\hat{\mu}_j - \mu_k)^2}{(n p_k + \sum_{l \neq k} \lambda_l)^2}.\end{aligned} \tag{A.1.6}$$

Since the group means are uncorrelated, the expected risk for group k is given by

$$\begin{aligned}E[(\hat{\mu}_k^{Kernel}(\boldsymbol{\Lambda}) - \mu_k)^2] \approx & \frac{(n p_k)^2 \frac{1}{\gamma_k n} + \sum_{j \neq k} \sum_{m \neq k, j} \lambda_j \lambda_m (\mu_j - \mu_k)(\mu_m - \mu_k)}{(n p_k + \sum_{l \neq k} \lambda_l)^2} \\ & + \frac{\sum_{j \neq k} \lambda_j^2 [\frac{1}{\gamma_j n} + (\mu_j - \mu_k)^2]}{(n p_k + \sum_{l \neq k} \lambda_l)^2}\end{aligned} \tag{A.1.7}$$

which yields the overall weighted MSE

$$\sum_{k=1}^{J} \gamma_k E[(\hat{\mu}_k^{Kernel}(\boldsymbol{\Lambda}) - \mu_k)^2] \approx \sum_{k=1}^{J} \left[\frac{n p_k}{n p_k + \sum_{l \neq k} \lambda_l} \right]^2 \frac{1}{n} + \frac{\sum_{k=1}^{J} \sum_{j \neq k} \lambda_j^2 \frac{\gamma_k}{\gamma_j} \frac{1}{n}}{(n p_k + \sum_{l \neq k} \lambda_l)^2}$$

$$+ \frac{\sum_{k=1}^{J} \sum_{j \neq k} \sum_{m \neq k} \gamma_k \lambda_j \lambda_m (\mu_j - \mu_k)(\mu_m - \mu_k)}{(np_k + \sum_{l \neq k} \lambda_l)^2}. \tag{A.1.8}$$

The plug-in estimator can be obtained by replacing the expression for p_k, γ_k and μ_k for $k = 1 \ldots, J$ by the corresponding estimates as for the PCS and optimizing with respect to $\mathbf{\Lambda}$ using numerical optimization. Similar to kernel estimation for continuous data, there is no closed-form solution for the general case.

Ridge Regression

For the ridge regression (RR), the restrictions imposed on the smoothing parameters are $\lambda_{kj} = \lambda$. This yields an estimator:

$$\hat{\mu}_k^{RR}(\lambda) = \frac{n_k}{n_k + (J-1)\lambda} \bar{Y}_k + \frac{(J-1)\lambda}{n_k + (J-1)\lambda} \sum_{j \neq k} \frac{\hat{\mu}_j}{J-1}.$$

Under the choice of $\hat{\mu}_j = \bar{Y}_j$, the weighted MSE of the leading term of a first-order approximation of the RR estimator takes form[1]:

$$\begin{aligned} MSE(\hat{\boldsymbol{\mu}}^{RR}(\lambda)) &= E\left[(\hat{\boldsymbol{\mu}}^{RR}(\lambda) - \boldsymbol{\mu})' W (\hat{\boldsymbol{\mu}}^{RR}(\lambda) - \boldsymbol{\mu})\right] = E\left[\sum_{k=1}^{J} \frac{p_k}{\sigma_k^2} (\hat{\mu}_k^{RR}(\lambda) - \mu_k)^2\right] \\ &\approx \sum_{k=1}^{J} \frac{p_k}{\sigma_k^2} \left(\left(\frac{(J-1)\lambda}{np_k + (J-1)\lambda} \right)^2 \left[\left(\sum_{j \neq k} \frac{\mu_j}{(J-1)} - \mu_k \right)^2 + \sum_{j \neq k} \frac{\sigma_j^2}{(J-1)^2 np_j} \right] \right. \\ &\qquad \left. + \left(\frac{np_k}{np_k + (J-1)\lambda} \right)^2 \frac{\sigma_k^2}{np_k} \right). \end{aligned}$$

The FOC for the weighted MSE optimal parameter λ takes form:

$$\begin{aligned} \sum_{k=1}^{J} \frac{p_k}{\sigma_k^2} \left(\frac{\lambda(J-1)np_k}{(np_k + (J-1)\lambda)^3} \left[\left(\sum_{j \neq k} \frac{\mu_j}{(J-1)} - \mu_k \right)^2 + \sum_{j \neq k} \frac{\sigma_j^2}{(J-1)^2 np_j} \right] \right. \\ \left. - \frac{np_k^2}{(np_k + (J-1)\lambda)^3} \frac{\sigma_k^2}{np_k} \right) \stackrel{!}{=} 0 \end{aligned}$$

The sum notation of the FOC reveals that the smoothing parameter λ is non-trivially intertwined across the reference groups. This implies that a closed-form solution exists only in special cases, e.g. for a balanced design when $p_k = 1/J$ for all k or for a design with 2 groups. In general, the FOC is a polynomial equation

[1] For more details about the first-order approximation please refer to Section 2.3.

of any order between 1 and $3(J-1)+1$. This means that already for some designs with more than two groups, one has to solve a polynomial equation of order larger than 4. According to the Abel-Ruffini theorem, there is no guarantee that a solution in radicals exists for polynomial equations of order five and higher with arbitrary coefficients. In these cases, one has to solve the FOC numerically and find the global minimum.

Generalized Ridge Regression

For the generalized ridge regression (GRR), the restrictions imposed on the smoothing parameters are $\lambda_{kj} = \lambda_k$. This yields an estimator:

$$\hat{\mu}_k^{GRR}(\lambda_k) = \frac{n_k}{n_k + (J-1)\lambda_k}\bar{Y}_k + \frac{(J-1)\lambda_k}{n_k + (J-1)\lambda_k}\sum_{j\neq k}\frac{\hat{\mu}_j}{J-1},$$

which can be rewritten in the following weighted form:

$$\hat{\mu}_k^{GRR}(\omega_k) = (1-\omega_k)\bar{Y}_k + \omega_k\sum_{j\neq k}\frac{\hat{\mu}_j}{J-1}.$$

The GRR estimator depends on the weights within its own reference category k. Therefore, optimization of the parameter vector MSE can be done group by group and is invariant to any MSE weighting. Under the choice of $\hat{\mu}_j = \bar{Y}_j$, the MSE of the leading term of a first-order approximation of the GRR estimator takes form[2]:

$$MSE(\hat{\mu}_k^{GRR}(\omega_k)) \approx \omega_k^2\left[\sum_{j\neq k}\frac{\mu_j}{(J-1)} - \mu_k\right]^2 + (1-\omega_k)^2\frac{\sigma_k^2}{np_k} + \omega_k^2\sum_{j\neq k}\frac{\sigma_j^2}{(J-1)^2 np_j}.$$

Optimal solution for ω_k is:

$$\omega_k^* = \frac{\frac{\sigma_k^2}{np_k}}{\frac{\sigma_k^2}{np_k} + \sum_{j\neq k}\frac{\sigma_j^2}{(J-1)^2 np_j} + \left[\sum_{j\neq k}\frac{\mu_j}{J-1} - \mu_k\right]^2}.$$

A.1.6 Proof of Lemma 2.4.1

Proof: $\omega_{kj}^f = O_p(n^{-1})$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0)$:

$$n\omega_{kj}^f = \frac{\lambda_{kj}}{n_k/n + \sum_{l\neq k}\lambda_{kl}/n} \xrightarrow{p} \frac{\lambda_{kj}}{p_k} = O(1)$$

[2]For more details about the first-order approximation please refer to Section 2.3.

by WULLN for n_k/n, continuous mapping and assuming λ_{kj} fixed. w_{kk}^f follows by definition.

Proof: $\omega_{kj}^* \to \bar{w}_{kj} = \frac{\gamma_j(1+\boldsymbol{\delta}'\Delta_k' diag(\gamma)\Delta_j\boldsymbol{\delta})}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}\boldsymbol{\delta}'\Delta' M_1 \Delta\boldsymbol{\delta}}$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. Use the closed-form in (2.3.5) and continuity together with $\sqrt{n}\Delta_k\boldsymbol{\mu} \to \Delta_k\boldsymbol{\delta}$ and $\sqrt{n}\Delta\boldsymbol{\mu} \to \Delta\boldsymbol{\delta}$.

Proof: $\omega_{kj}^* \to \bar{w}_{kj} = \gamma_j \frac{\boldsymbol{\mu}'\Delta_k' diag(\gamma)\Delta_j\boldsymbol{\mu}}{\frac{1}{2}\boldsymbol{\mu}'\Delta' M_1 \Delta\boldsymbol{\mu}}$ if $\{F_n\} \in S(\infty, V_0)$. Follows from dividing by n and taking simple limits, i.e.

$$\omega_{kj}^* = \frac{\gamma_j(1/n + \boldsymbol{\mu}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\mu})}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J/n + \frac{1}{2}\boldsymbol{\mu}'\Delta' M_1\Delta\boldsymbol{\mu}} \to 2\gamma_j \frac{\boldsymbol{\mu}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\mu}}{\boldsymbol{\mu}'\Delta' M_1\Delta\boldsymbol{\mu}}$$

which exists as $\{F_n\} \in S(\infty, V_0)$.

Proof: $\hat{\omega}_{kj} \xrightarrow{d} w_{kj}^a = \frac{\gamma_j(1+(\mathbf{Z}+\boldsymbol{\delta})'\Delta_k diag(\gamma)\Delta_j(\mathbf{Z}+\boldsymbol{\delta}))}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_1 \Delta(\mathbf{Z}+\boldsymbol{\delta})}$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. Take $\hat{\omega}_{kj}$ according to (2.3.6). Note that $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$ and thus $\hat{M}_1 \xrightarrow{p} M_1$. Additionally $\sqrt{n}(\hat{\mu}_k - \hat{\mu}_j) = \sqrt{n}(\hat{\mu}_k - \mu_k) - \sqrt{n}(\hat{\mu}_j - \mu_j) + \sqrt{n}(\mu_k - \mu_j) \xrightarrow{d} Z_k - Z_j + \delta_k - \delta_j$ since $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. Similarly $\sqrt{n}\Delta_k\hat{\boldsymbol{\mu}} \xrightarrow{d} \Delta_k(\mathbf{Z} + \boldsymbol{\delta})$ and $\sqrt{n}\Delta\hat{\boldsymbol{\mu}} \xrightarrow{d} \Delta(\mathbf{Z} + \boldsymbol{\delta})$. The rest follows from continuity of $\hat{\omega}_{kj}$.

Proof: $\hat{\omega}_{kj} \xrightarrow{p} \bar{w}_{kj}$ if $\{F_n\} \in S(\infty, V_0)$. Note that $\hat{\boldsymbol{\mu}} \xrightarrow{p} \boldsymbol{\mu}$, $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$ and thus $\hat{M}_1 \xrightarrow{p} M_1$. Thus by continuous mapping

$$\hat{\omega}_{kj} = \frac{\hat{\gamma}_j(1/n + \hat{\boldsymbol{\mu}}'\Delta_k' diag(\hat{\boldsymbol{\gamma}})\Delta_j\hat{\boldsymbol{\mu}})}{\hat{\boldsymbol{\gamma}}'\boldsymbol{\iota}_J/n + \frac{1}{2}\hat{\boldsymbol{\mu}}'\Delta'\hat{M}_1\Delta\hat{\boldsymbol{\mu}}} \xrightarrow{p} 2\gamma_j \frac{\boldsymbol{\mu}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\mu}}{\boldsymbol{\mu}'\Delta' M_1\Delta\boldsymbol{\mu}}$$

which exists as $\{F_n\} \in S(\infty, V_0)$.

A.1.7 Proof of Theorem 2.4.1

Proof: $\sqrt{n}(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k^f) - \mu_k - B_{1k}(\boldsymbol{\omega}_k^f)) \xrightarrow{d} Z_k \sim \mathcal{N}\left(0, \frac{\sigma_k^2}{p_k}\right)$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0)$. By definition of the PCS and using fixed weights we have

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k^f) - \mu_k) = \sqrt{n}\sum_{j=1}^J \omega_{kj}^f(\hat{\mu}_j - \mu_j) + \sqrt{n}\sum_{j=1}^J \omega_{kj}^f(\mu_j - \mu_k)$$

Using Lemma 2.4.1 together with $\sqrt{n}(\hat{\mu}_j - \mu_j) = O_p(1)$ for all $\{F_n\}$ we have that

$$\begin{aligned}\sqrt{n}(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k^f) - \mu_k - \sum_{j=1}^J \omega_{kj}^f(\mu_j - \mu_k)) &= \sqrt{n}(\hat{\mu}_k - \mu_k) + o_p(1) \\ &\xrightarrow{d} Z_k\end{aligned}$$

Proof: $\sqrt{n}(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k^*)-\mu_k-B_{2k}(\boldsymbol{\omega}_k^*)) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=1}^J \bar{\omega}_{kj}^2 \frac{\sigma_j^2}{p_j}\right)$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0)$. Using the definition from the PCS, the CLT for $\sqrt{n}(\hat{\mu}_j - \mu_j)$ together with Lemma 2.4.1 yields

$$\begin{aligned}\sqrt{n}(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k^*) - \mu_k - \sum_{j=1}^J \omega_{kj}^*(\mu_j - \mu_k)) &= \sqrt{n}\sum_{j=1}^J \omega_{kj}^*(\hat{\mu}_j - \mu_j) \\ &= \sqrt{n}\sum_{j=1}^J \bar{\omega}_{kj}(\hat{\mu}_j - \mu_j) + o(1) \\ &\xrightarrow{d} \sum_{j=1}^J \bar{\omega}_{kj} Z_j\end{aligned}$$

with the final quantity being distributed $\mathcal{N}(0, \sum_{j=1}^J \bar{\omega}_{kj}^2 \sigma_j^2/p_j)$ since Z_j, Z_k are asymptotically independent for all $j \neq k$ due to the orthogonality of the groups.

Proof: $\sqrt{n}(\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}_k)-\mu_k) \xrightarrow{d} \sum_{j=1}^J \omega_{kj}^a Z_j + \sum_{j=1}^J \omega_{kj}^a(\delta_j - \delta_k)$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. Rewriting the PCS in the usual manner yields

$$\begin{aligned}\sqrt{n}(\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}_k) - \mu_k) &= \sqrt{n}\sum_{j=1}^J \hat{\omega}_{kj}(\hat{\mu}_j - \mu_j) + \sqrt{n}\sum_{j=1}^J \hat{\omega}_{kj}(\mu_j - \mu_k) \\ &\xrightarrow{d} \sum_{j=1}^J \omega_{kj}^a Z_j + \sum_{j=1}^J \omega_{kj}^a(\delta_j - \delta_k)\end{aligned}$$

where convergence in distribution follows from joint convergence of the $\hat{\omega}_{kj}$'s and $\sqrt{n}(\hat{\mu}_j - \mu_j)$'s as they are continuous functions of the same random normal vector and using the distributional Lemma for the weights for $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$ and $\sqrt{n}(\mu_j - \mu_k) \to \delta_k - \delta_j$ by definition of sequences in $S(\boldsymbol{\delta}, V_0)$.

Proof: $\sqrt{n}(\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}_k) - \mu_k - B_{3k}(\bar{\boldsymbol{\omega}}_k)) \xrightarrow{d} Z_k \sim \mathcal{N}\left(0, \frac{\sigma_k^2}{p_k}\right)$ if $\{F_n\} \in S(\infty, V_0)$. By Lemma 2.4.1, $\hat{\omega}_{kj} \xrightarrow{p} \bar{\omega}_{kj}$ as $\{F_n\} \in S(\infty, V_0)$. Rewriting the PCS yields

$$\begin{aligned}\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}_k) - \mu_k &= \sum_{j=1}^J \hat{\omega}_{kj}(\hat{\mu}_j - \mu_j + \mu_j - \mu_k) \\ &= \sum_{j=1}^J \hat{\omega}_{kj}(\hat{\mu}_j - \mu_j) + \sum_{j=1}^J \bar{\omega}_{kj}(\mu_j - \mu_k) + \sum_{j=1}^J (\hat{\omega}_{kj} - \bar{\omega}_{kj})(\mu_j - \mu_k)\end{aligned}$$

or equivalently

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}_k) - \mu_k - \sum_{j \neq k} \bar{\omega}_{kj}(\mu_j - \mu_k)) = \sum_{j=1}^J (\hat{\omega}_{kj} - \bar{\omega}_{kj})\sqrt{n}(\hat{\mu}_j - \mu_j)$$

$$+\sum_{j=1}^{J}\bar{\omega}_{kj}\sqrt{n}(\hat{\mu}_j-\mu_j)+\sum_{j\neq k}\sqrt{n}(\hat{\omega}_{kj}-\bar{\omega}_{kj})(\mu_j-\mu_k)$$
$$=\sum_{j=1}^{J}\bar{\omega}_{kj}\sqrt{n}(\hat{\mu}_j-\mu_j)+\sum_{j\neq k}\sqrt{n}(\hat{\omega}_{kj}-\bar{\omega}_{kj})(\mu_j-\mu_k)+o_p(1).$$

The right hand side is asymptotically normal as the components are stabilizing transformations of continuous functions of the same random normal vector. In terms of its asymptotic variance, one can either show the equivalence to Z_k using the delta method or simpler by Theorem 2.5.1. It implies that as $||\Delta\boldsymbol{\delta}||_\infty \to \infty$, the PCS risk is converging to the OLS. Since both estimators are asymptotically normal, the asymptotic variances have to coincide.

A.1.8 Proof of Theorem 2.5.1 and Corollary 2.5.1

Proof: Let $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. The plugin weights are given by

$$\hat{\omega}_{kj}=\frac{\hat{\gamma}_j+n\sum_{m=1}^{J}(\hat{\mu}_k-\hat{\mu}_m)(\hat{\mu}_j-\hat{\mu}_m)\hat{\gamma}_j\hat{\gamma}_m}{\sum_{l=1}^{J}\hat{\gamma}_l+0.5n\sum_{l=1}^{J}\sum_{m=1}^{J}(\hat{\mu}_l-\hat{\mu}_m)^2\hat{\gamma}_l\hat{\gamma}_m}.$$

which by Lemma 2.4.1 converge in distribution, i.e.

$$\hat{\omega}_{kj}\xrightarrow{d}w_{kj}^a=\frac{\gamma_j+\sum_{m=1}^{J}(Z_k-Z_m+\delta_k-\delta_m)(Z_j-Z_m+\delta_j-\delta_m)\gamma_j\gamma_m}{d_0}$$

with $d_0=\boldsymbol{\gamma}'\boldsymbol{\iota}_J+\frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_1\Delta(\mathbf{Z}+\boldsymbol{\delta})$. By Theorem 2.4.1, the distributional limit for the PCS under $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$ is given by

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}_k)-\mu_k)\xrightarrow{d}\sum_{j=1}^{J}\omega_{kj}^a Z_j+\sum_{j=1}^{J}\omega_{kj}^a(\delta_j-\delta_k)=\sum_{j=1}^{J}\omega_{kj}^a(Z_j-Z_k+\delta_j-\delta_k)+Z_k\equiv\psi_k$$

since $\sum_{j=1}^{J}\omega_{kj}^a=1$ for all k. By Lemma 1 of Hansen (2016a), the asymptotic weighted MSE criterion then yields

$$\begin{aligned}\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}),\boldsymbol{\mu})&=\sum_{k=1}^{J}\gamma_k E[\psi_k^2]\\&=E[\sum_{k=1}^{J}\sum_{j=1}^{J}\sum_{l=1}^{J}\gamma_k\gamma_j\gamma_l(Z_k-Z_j+\delta_k-\delta_j)(Z_k-Z_l+\delta_k-\delta_l)/d_0^2]\\&-2E[\sum_{k=1}^{J}\sum_{j=1}^{J}\gamma_k\gamma_j(Z_k-Z_j+\delta_k-\delta_j)Z_k/d_0]+E[\sum_{k=1}^{J}\gamma_k Z_k^2]\\&\equiv E[A]-2E[B]+\rho(\hat{\boldsymbol{\mu}},\boldsymbol{\mu})\end{aligned}$$

with

$$\begin{aligned} A &= (\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_2 \Delta(\mathbf{Z}+\boldsymbol{\delta})/d_0^2 \\ M_2 &= diag(\boldsymbol{\gamma}) \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}' \\ B &= (\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_3 \mathbf{Z}/d_0 \\ M_3 &= diag(\boldsymbol{\gamma}) \otimes \boldsymbol{\gamma} \end{aligned}$$

To further simplify $E[B]$ we use a multivariate version of Stein's Lemma given by Lemma 2 in Hansen (2016a) which yields

$$E[B] = E[\eta(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_3 \mathbf{Z}] = E\left[tr\left(\frac{\partial}{\partial \mathbf{x}}\eta(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_3 V_0\right)\right]$$

with $\eta(\mathbf{x}) = \mathbf{x}/(\boldsymbol{\gamma}'\boldsymbol{\iota}_J + 0.5\mathbf{x}'\Delta' M_1 \Delta \mathbf{x})$ and derivative

$$\frac{\partial}{\partial \mathbf{x}}\eta(\mathbf{x})' = \frac{1}{d_0} I_J - \frac{\Delta' M_1 \Delta}{d_0^2}\mathbf{x}\mathbf{x}'$$

and hence

$$\begin{aligned} E[B] &= tr(\Delta' M_3 V_0) E\left[\frac{1}{d_0}\right] - E\left[\frac{tr(\Delta' M_1 \Delta(\mathbf{Z}+\boldsymbol{\delta})(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_3 V_0)}{d_0^2}\right] \\ &= \boldsymbol{\gamma}'\boldsymbol{\iota}_J tr(\Delta' M_3 V_0) E\left[\frac{1}{d_0^2}\right] + \frac{1}{2} tr(\Delta' M_3 V_0) E\left[\frac{(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_1 \Delta(\mathbf{Z}+\boldsymbol{\delta})}{d_0^2}\right] \\ &\quad - E\left[\frac{(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_3 V_0 \Delta' M_1 \Delta(\mathbf{Z}+\boldsymbol{\delta})}{d_0^2}\right]. \end{aligned}$$

Since $\boldsymbol{\gamma}'\boldsymbol{\iota}_J = tr(V_0^{-1})$, plugging in and bringing the terms together with $E[A]$ yields the following asymptotic risk:

$$\begin{aligned} \rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) =& \rho(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) + E\left[\frac{(\mathbf{Z}+\boldsymbol{\delta})'\Delta' C \Delta(\mathbf{Z}+\boldsymbol{\delta})}{(tr(V_0^{-1}) + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_1 \Delta(\mathbf{Z}+\boldsymbol{\delta}))^2}\right] \\ &- 2tr(V_0^{-1}) tr(\Delta' M_3 V_0) E\left[\frac{1}{(tr(V_0^{-1}) + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta' M_1 \Delta(\mathbf{Z}+\boldsymbol{\delta}))^2}\right] \end{aligned}$$

with $C = M_2 - tr(\Delta' M_3 V_0) M_1 + 2M_3 V_0 \Delta' M_1$. The corollary then follows from C being negative semidefinite if $J > 3$. We show that for $J > 3$, $-C$ is positive semidefinite. Some algebra yields the following characterization for $-C$:

$$\begin{aligned} [left = -C_{ij} = \{] &\gamma_i(2J-7)\sum_{l\neq i}\gamma_l \sum_{m=1}^{J}\gamma_m \qquad \text{if } i = j \\ &-\gamma_i\gamma_j(2J-7)\sum_{l=1}^{J}\gamma_l \qquad \text{if } i \neq j. \end{aligned}$$

Due to positivity of the γ_j's, the diagonal elements of $-C$ are strictly positive if $(2J-7) > 3$. Thus, a sufficient condition for $-C$ being positive semidefinite is (absolute) diagonal dominance, i.e. $|-C_{ii}| \geq \sum_{j \neq i} |-C_{ij}|$ which yields

$$\gamma_i(2J-7)\sum_{l \neq i}\gamma_l \sum_{m=1}^{J}\gamma_m \geq \sum_{j \neq i}\gamma_i\gamma_j(2J-7)\sum_{l=1}^{J}\gamma_j \quad \Leftrightarrow 0 \geq 0$$

which proofs the sufficiency.

A.1.9 Supplementary Material for Section 2.7.2

Table A.1.1: Summary Statistics of the Card and Krueger (1994) Data

Chain	NJ (treated)						PEN (control)					
	Before			After			Before			After		
	$\hat{\mu}$	σ^2	n	$\hat{\mu}$	σ^2	n	$\hat{\mu}$	σ^2	n	$\hat{\mu}$	σ^2	n
All chains	20.44	82.92	321	21.03	86.36	319	23.33	140.57	77	21.17	68.5	77
Burger King	22.16	61.95	131	23.63	70.63	131	29.42	182.81	33	26.22	50.31	35
KFC	12.79	21.83	67	13.73	39.60	68	10.71	7.83	12	13.00	11.59	12
Roys	23.14	109.36	81	21.73	89.30	78	19.74	32.96	17	15.81	43.89	17
Wendys	22.08	79.99	42	23.40	96.64	42	24.12	61.20	15	22.10	39.35	13

A.2 Extended Appendix

Note: Some of the derivations are based on an earlier formulation of the paper and rely on the following definition for the MSE optimal smoothing parameters λ^*_{kj}:

$$\lambda^*_{kj} = \frac{\sigma^2_k np_j \big/ \sigma^2_j}{a_{kj} - \sum_{l\neq k} \frac{\sigma^2_k p_l a_{kj}}{\sigma^2_l p_k a_{kl}}} \quad \text{for all } k \neq j, \quad \lambda^*_{kk} = 0 \tag{A.2.1}$$

where $a_{kj} = \left(1 + \frac{\sigma^2_k/np_k}{1+b_{kj}} \sum_{l\neq k} \frac{1+b_{kl}}{\sigma^2_l/np_l} + \frac{\mu_{kj}}{1+b_{kj}} \sum_{l\neq k} \frac{\mu_{kl}}{\sigma^2_l/np_l}\right)$ and $b_{kj} = \sum_{m\neq k} \frac{\mu_{km}\mu_{jm}}{\sigma^2_m/np_m}$ and $\mu_{kj} = \mu_k - \mu_j$.

A.2.1 Mixed Data

The original framework is rather restrictive beyond applications that form orthogonal groups by construction. In this section we provide a multi-level approach for applying the PCS in additive linear models with mixed categorical and continuous regressors. Imagine an $n \times k$-dimensional continuous regressor matrix $X = (\mathbf{X}(1), \ldots, \mathbf{X}(k))$ and an additive linear model:

$$\mathbf{Y} = X\boldsymbol{\beta} + D\boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

with $\mathbf{Y} = (Y_1, \ldots, Y_n)'$, D being an $n \times J$ matrix where $\mathbf{D}'_i$ is the i-th row, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_J)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$, $E[\boldsymbol{\varepsilon}|X, D] = \mathbf{0}$ and $V[\boldsymbol{\varepsilon}|D] = \Sigma_D$. We propose to use the PCS in a two-step procedure to estimate both the location parameters as well as the parameters of the continuous regressors. It works by partialling out the expectations conditional on the set of orthogonal dummies. Note that

$$\mathbf{Y} - E[\mathbf{Y}|D] = (X - E[X|D])\boldsymbol{\beta} + \boldsymbol{\varepsilon} - E[\boldsymbol{\varepsilon}|D].$$

Replacing the conditional expectations by PCS estimators, that only rely on the orthogonal data, yields the following estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}^{PCS} = [(X - \hat{E}[X|D])'(X - \hat{E}[X|D])]^{-1}(X - \hat{E}[X|D])'(\mathbf{Y} - \hat{E}[\mathbf{Y}|D])$$

with

$$\begin{aligned} \hat{E}[X|D] &= (\hat{E}[\mathbf{X}(1)|D], \ldots, \hat{E}[\mathbf{X}(k)|D]) \\ \hat{E}[\mathbf{X}(j)|D] &= D(D'D + U'W_jU)^{-1}(I + U'W_jV(D'D)^{-1})D'\mathbf{X}(j) \end{aligned}$$

$$\hat{E}[\mathbf{Y}|D] = D(D'D + U'W_yU)^{-1}(I + U'W_yV(D'D)^{-1})D'\mathbf{Y}$$

where $U = (I_J \otimes \boldsymbol{\iota}_J)$, $V = (\boldsymbol{\iota}_J \otimes I_J)$ and W_j, W_y being the diagonal matrix of PCS smoothing parameters for the regression model of $\mathbf{X}(j)$ on D and $\mathbf{Y}$ on D respectively. Note that the step for estimating slope parameter $\boldsymbol{\beta}$ could be combined with additional means for regularization such as e.g. LASSO or ridge estimation in the presence of e.g. high-dimensional or multicollinear regressors. To obtain an estimator for $\boldsymbol{\mu}$ one can subtract the continuous component and use a projection, i.e.

$$\hat{\boldsymbol{\mu}}^{PCS,a} = (D'D)^{-1}D'(\mathbf{Y} - X\hat{\boldsymbol{\beta}}^{PCS})$$

One can show that both estimators are root-n consistent. The two-stage approach is computationally very efficient and does not require numerical optimization since the closed-form of the PCS can be used directly in the first stage. Note however, that optimality can now only be achieved with respect to the risk of the first stages. A better approach would be to use the residuals $\mathbf{Y}^* := (\mathbf{Y} - X\hat{\boldsymbol{\beta}}^{PCS})$ in another PCS step, i.e. estimate the model[3]

$$\mathbf{Y}^* = D\boldsymbol{\mu} + \mathbf{u}$$

which yields

$$\hat{\boldsymbol{\mu}}^{PCS,b} = (D'D + U'W_{y^*}U)^{-1}(I + U'W_{y^*}V(D'D)^{-1})D'\mathbf{Y}^*$$

where $\mathbf{u} = X(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{PCS}) + \boldsymbol{\varepsilon}$ and W_{y^*} being the diagonal matrix of PCS smoothing parameters for the regression of $\mathbf{Y}^*$ on D. Note that due to $\mathbf{u}$ depending on the first stage estimator, MSE optimal smoothing parameters are not exact anymore. Simulations reveal that the regularization positively affects the parameter risk for both the continuous and the discrete part. In fact, the improvements on the continuous part are non negligible. As for the exclusively discrete regressor case, the parameter risk of the discrete part seems to dominate the simple OLS, in particular in the presence of small and/or very volatile groups. The gains are more pronounced if D and X are correlated.

[3] A potential drawback of this approach is its violation of Neyman orthogonality due to construction of the second stage that differs from a standard partialling out procedure which uses residuals on both sides of the equation.

A.2.2 One-to-one Correspondence between λ_{kj} and ω_{kj}

By using the following relationship $n_k + \sum_{j \neq k} \lambda_{kj} = n_k/\omega_{kk}$, the smoothing parameters can be expressed as

$$\lambda_{kj} = \frac{n_k \omega_{kj}}{\omega_{kk}} \qquad j \neq k. \tag{A.2.2}$$

Now we want to check if there is a one-to-one correspondence between λ_{kj} and ω_{kj}. To get the same λ_{kj}'s from two different sets $\boldsymbol{\omega}_k^{(1)}$ and $\boldsymbol{\omega}_k^{(2)}$, where $\boldsymbol{\omega}_k^{(1)} = \left(\omega_{k1}^{(1)}, \ldots, \omega_{kJ}^{(1)}\right)'$ and $\boldsymbol{\omega}_k^{(2)} = \left(\omega_{k1}^{(2)}, \ldots, \omega_{kJ}^{(2)}\right)'$, the following ratios have to hold:

$$\frac{\omega_{kj}^{(1)}}{\omega_{kk}^{(1)}} = \frac{\omega_{kj}^{(2)}}{\omega_{kk}^{(2)}} \qquad \forall (k,j), j \neq k. \tag{A.2.3}$$

Without a loss of generality, we can express the last elements of $\boldsymbol{\omega}_k^{(1)}$ and $\boldsymbol{\omega}_k^{(2)}$ as follows:

$$\omega_{kJ}^{(1)} = 1 - \sum_{j \neq J} \omega_{kj}^{(1)}, \tag{A.2.4}$$

$$\omega_{kJ}^{(2)} = 1 - \sum_{j \neq J} \omega_{kj}^{(2)}. \tag{A.2.5}$$

By construction, $\frac{\omega_{kj}^{(1)}}{\omega_{kj}^{(2)}} = \frac{\omega_{kk}^{(1)}}{\omega_{kk}^{(2)}} = c_k, \forall (k,j), j \neq k$, where c_k is any non-zero constant. Thus,

$$c_k = \frac{\omega_{kJ}^{(1)}}{\omega_{kJ}^{(2)}} = \frac{1 - \sum_{j \neq J} \omega_{kj}^{(1)}}{1 - \sum_{j \neq J} \omega_{kj}^{(2)}} = \frac{1 - c_k \sum_{j \neq J} \omega_{kj}^{(2)}}{1 - \sum_{j \neq J} \omega_{kj}^{(2)}}.$$

For this equality to hold, $c_k = 1$. This implies that there are no two different sets of $\boldsymbol{\omega}_k$ which would give us the same λ_{kj}'s and vice versa.

A.2.3 Uniqueness of the MSE Optimal Regularization Parameters

The problem of the constrained minimization of (2.3.4) can be rewritten as the following Lagrangian:

$$\min_{\boldsymbol{\omega}_k, \alpha_k} \mathcal{L}(\boldsymbol{\omega}_k, \alpha_k) = \min_{\boldsymbol{\omega}_k, \alpha_k} \boldsymbol{\omega}_k' H_k \boldsymbol{\omega}_k + 2\alpha_k (1 - \boldsymbol{\iota}_J' \boldsymbol{\omega}_k), \tag{A.2.6}$$

$$\text{with } \boldsymbol{\omega}_k = (\omega_{k1}, \omega_{k2}, \ldots, \omega_{kJ})',$$

$$H_k = \Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k' + diag(\boldsymbol{\gamma})^{-1}/n,$$

$$\alpha_k = \text{ Lagrange multiplier.}$$

The FOCs are given by

$$\begin{aligned}\frac{\partial \mathcal{L}(\boldsymbol{\omega}_k, \alpha_k)}{\partial \boldsymbol{\omega}_k} &= 2H_k\boldsymbol{\omega}_k - 2\alpha_k\boldsymbol{\iota}_J = 0\\ \frac{\partial \mathcal{L}(\boldsymbol{\omega}_k, \alpha_k)}{\partial \alpha_k} &= 2(1 - \boldsymbol{\iota}_J'\boldsymbol{\omega}_k) \quad = 0\end{aligned}$$

The solution of setting the FOC to zero gives optimal values:

$$\begin{aligned}\alpha_k^* &= [\boldsymbol{\iota}_J'(H_k'H_k)^{-1}H_k'\boldsymbol{\iota}_J]^{-1}\\ \boldsymbol{\omega}_k^* &= [\boldsymbol{\iota}_J'(H_k'H_k)^{-1}H_k'\boldsymbol{\iota}_J]^{-1}(H_k'H_k)^{-1}H_k'\boldsymbol{\iota}_J\end{aligned}$$

The expression for ω_{kj}^* can be inferred from the j-th entry of $\boldsymbol{\omega}_k^*$.

To investigate if the $\boldsymbol{\omega}_k^*$ is a unique global minimizer of (A.2.6), we first rewrite the optimization problem (A.2.6) into a form used in null-space methods to solve equality quadratic problems, see e.g. Gould (1985). The idea behind the null-space methods is to reduce the dimensionality of the optimization problem by exploiting the constraints and obtain the original solution as a combination of the optimal solution from the reduced space and a corresponding vector stemming from the constraint. By choosing a $J \times (J-1)$ matrix Z such that $\boldsymbol{\iota}_J'Z = \mathbf{0}$ and $\text{rank}(\boldsymbol{\iota}_J \vdots Z) = J$, solving the following null-space method problem yields the same solution as minimizing (2.3.4) or its equivalent (A.2.6):

$$\min_{\boldsymbol{\omega}_{Z,k} \in \mathbb{R}^{J-1}} \boldsymbol{\omega}_{Z,k}'Z'H_kZ\boldsymbol{\omega}_{Z,k} + \boldsymbol{\omega}_{Z,k}'Z'H_k\boldsymbol{\iota}_J\omega_{\iota,k} \tag{A.2.7}$$

$$\text{where } \boldsymbol{\iota}_J'\boldsymbol{\iota}_J\omega_{\iota,k} = 1 \tag{A.2.8}$$

and then $\boldsymbol{\omega}_k^* = Z\boldsymbol{\omega}_{Z,k}^* + \boldsymbol{\iota}_J\omega_{\iota,k}$ and $\alpha_k^* = (\boldsymbol{\iota}_J'\boldsymbol{\iota}_J)^{-1}\boldsymbol{\iota}_J'H_k\boldsymbol{\omega}_k^*$. Note that (A.2.8) just determines the value of $\omega_{\iota,k}$ and (A.2.7) is in fact an unconstrained problem.

Case 1 - Finite n: The advantage of rewriting the problem into the form used in null-space methods is the possibility to deduce whether the problem has a unique solution. For completeness, theorem quoted below from (Gould, 1985, Theorem 1.1(i)) enables to assess the uniqueness of the solution.

Theorem A.2.1 *Suppose (A.2.7) is as given with $\boldsymbol{\iota}_J'$ of full row rank and Z is constructed so that $\boldsymbol{\iota}_J'Z = \mathbf{0}'$ and rank$(\boldsymbol{\iota}_J \vdots Z) = J$. Then (A.2.7) has a strong minimizer if and only if $Z'H_kZ$ is positive definite.*

It is easy to see that $\boldsymbol{\iota}_J'$ has a full row rank of one. We choose

$$Z = \begin{pmatrix} -1 & 0 & \cdots & 0 \\ 1 & -1 & \ddots & \vdots \\ 0 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & 1 \end{pmatrix}$$

such that $\boldsymbol{\iota}_J' Z = \mathbf{0}'$ and $\text{rank}(\boldsymbol{\iota}_J \vdots Z) = J$. Note that Z has a full column rank $= J - 1$. This implies that if H_k is positive definite, then $Z' H_k Z$ is also positive definite.

We know that $H_k = \Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k' + diag(\boldsymbol{\gamma})^{-1}/n$. For a finite n, $diag(\boldsymbol{\gamma})^{-1}/n$ is a diagonal matrix with positive elements and thus positive definite. Since $\Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k'$ gives a matrix which is rank deficient, it can happen that for some non-zero vector $\mathbf{x}$ we get that $\mathbf{x}' \Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k' \mathbf{x} = 0$ but it cannot be negative as illustrated below:

$$\mathbf{x}' \Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k' \mathbf{x} = (\boldsymbol{\mu}' \Delta_k' \mathbf{x})' \boldsymbol{\mu}' \Delta_k' \mathbf{x} = a^2 \geq 0 \quad \text{for all } \mathbf{x} \neq \mathbf{0}.$$

Therefore, $\Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k'$ is a positive semi-definite matrix. A sum of a positive definite and positive semi-definite matrix gives a positive definite matrix:

$$\mathbf{x}' H_k \mathbf{x} = \mathbf{x}' (\Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k' + diag(\boldsymbol{\gamma})^{-1}/n)\mathbf{x} = \underbrace{\mathbf{x}' \Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k' \mathbf{x}}_{\geq 0} + \underbrace{\mathbf{x}' diag(\boldsymbol{\gamma})^{-1}/n \mathbf{x}}_{>0} > 0$$

for all $\mathbf{x} \neq \mathbf{0}$.

This means that H_k is a positive definite matrix and that $\boldsymbol{\omega}_k^*$ is a unique minimizer of (2.3.4).

Case 2: $n \to \infty$:

Subcase 2.1 - asymptotically close group location system, $\sqrt{n}\Delta_k \boldsymbol{\mu} \to \Delta_k \boldsymbol{\delta} \in \mathbb{R}^J$: In this case, squared bias and variance converge to zero. The variance vanishes to zero at rate n dominating (2.3.4). In order to assess the uniqueness of the solution in this case, we have to look at the positive definiteness of nH_k to take care of both convergence rates:

$$\mathbf{x}' n H_k \mathbf{x} = n \mathbf{x}' \underbrace{\Delta_k \boldsymbol{\mu}\boldsymbol{\mu}' \Delta_k'}_{=O(n^{-1})} \mathbf{x} + \mathbf{x}' diag(\boldsymbol{\gamma})^{-1} \mathbf{x} = \underbrace{O(1)}_{\geq 0} + \underbrace{\mathbf{x}' diag(\boldsymbol{\gamma})^{-1} \mathbf{x}}_{>0} > 0$$

for all $\mathbf{x} \neq \mathbf{0}$.

As mentioned before, $\mathbf{x}'\Delta_k\boldsymbol{\mu}\boldsymbol{\mu}'\Delta_k'\mathbf{x}$ cannot be negative, thus the $O(1)$ element is non-negative. Matrix $diag(\boldsymbol{\gamma})^{-1}$ is a diagonal matrix with positive elements and thus positive definite. This means that nH_k is a positive definite matrix and according to Theorem A.2.1, $\boldsymbol{\omega}_k^*$ is a unique minimizer of (2.3.4).

Subcase 2.2 - asymptotically distant group location system $\sup_{k,j}\sqrt{n}|\mu_k-\mu_j|\to\infty$*:* In this case, the variance in (2.3.4) asymptotically vanishes to zero at rate n meanwhile the squared bias dominates the MSE. In order to assess the uniqueness of the solution in this case, we have to look again at the positive definiteness of H_k. To do that properly, we assess the positive (semi)definiteness of $\Delta_k\boldsymbol{\mu}\boldsymbol{\mu}'\Delta_k'$ and $diag(\boldsymbol{\gamma})^{-1}/n$ separately to take proper care of their different convergence rates.

Denoting $\sup_{k,j}\sqrt{n}|\mu_k-\mu_j| = O(f(n))$, we get the following convergence rate corrected form to analyse positive (semi)definiteness of $\Delta_k\boldsymbol{\mu}\boldsymbol{\mu}'\Delta_k'$

$$\frac{n}{f(n)^2}\mathbf{x}'\underbrace{\Delta_k\boldsymbol{\mu}\boldsymbol{\mu}'\Delta_k'}_{\substack{=O(f(n)^2n^{-1}),\\ \alpha>0}}\mathbf{x} = O(1) \geq 0 \quad \text{for all } \mathbf{x}\neq\mathbf{0}.$$

As mentioned before, $\mathbf{x}'\Delta_k\boldsymbol{\mu}\boldsymbol{\mu}'\Delta_k'\mathbf{x}$ cannot be negative, thus the $O(1)$ element is non-negative and $\Delta_k\boldsymbol{\mu}\boldsymbol{\mu}'\Delta_k'$ is positive semidefinite.

After correcting for the convergence rate of the $diag(\boldsymbol{\gamma})^{-1}/n$ we get

$$\mathbf{x}'diag(\boldsymbol{\gamma})^{-1}\mathbf{x} > 0 \quad \text{for all } \mathbf{x}\neq\mathbf{0}.$$

Matrix $diag(\boldsymbol{\gamma})^{-1}$ is a diagonal matrix with positive elements and thus positive definite. This means that H_k is a sum of a positive semidefinite and positive definite matrix yielding a positive definite matrix. According to Theorem A.2.1, $\boldsymbol{\omega}_k^*$ is then a unique minimizer of (2.3.4). There is a one-to-one correspondence between $\boldsymbol{\omega}_k^*$ and $\boldsymbol{\Lambda}_k^*$, see Section A.2.2. Thus, these results translate to $\boldsymbol{\Lambda}_k^*$.

A.2.4 Proof $\sum_{j=1}^J \lambda_{kj}^* > -n_k$

Since $\omega_{kj}^* = \lambda_{kj}^*/(n_k+\sum_{j=1}^J\lambda_{kj}^*)$, if $\text{sign}\left(\lambda_{kj}^*\right) = \text{sign}\left(\omega_{kj}^*\right)$ then $\sum_{j=1}^J\lambda_{kj}^* > -n_k$ holds for all $j\neq k$. Assuming the reverse yields

$$\begin{aligned}&\text{sign}\left(\lambda_{kj}^*\right) \neq \text{sign}\left(\omega_{kj}^*\right)\\ \Leftrightarrow \text{sign}&\left(\frac{\sigma_k^2 np_j\big/\sigma_j^2}{a_{kj}-\sum_{l\neq k}\frac{\sigma_k^2p_la_{kj}}{\sigma_l^2p_ka_{kl}}}\right) \neq \text{sign}\left(\frac{\sigma_k^2p_j\big/\sigma_j^2p_k}{a_{kj}}\right)\end{aligned}$$

$$\Leftrightarrow \text{sign}\left(1 - \sum_{l \neq k} \frac{\sigma_k^2 p_l}{\sigma_l^2 p_k a_{kl}}\right) \neq 1$$

$$\Leftrightarrow 1 < \sum_{l \neq k} \frac{\sigma_k^2 p_l}{\sigma_l^2 p_k a_{kl}}, \tag{A.2.9}$$

where $a_{kl} = \left(1 + \frac{\sigma_k^2/np_k}{1+b_{kl}} \sum_{q \neq k} \frac{1+b_{kq}}{\sigma_q^2/np_q} + \frac{\mu_{kl}}{1+b_{kl}} \sum_{q \neq k} \frac{\mu_{kq}}{\sigma_q^2/np_q}\right)$ and $b_{kl} = \sum_{m \neq k} \frac{\mu_{km}\mu_{lm}}{\sigma_m^2/np_m}$. In the first step we plug in results from (A.2.1). Evaluating the individual terms in the sum of (A.2.9) yields

$$\frac{p_l \sigma_k^2}{p_k a_{kl} \sigma_l^2} = \frac{\frac{p_l}{\sigma_l^2} + \sum_{m \neq k} \frac{\mu_{km}\mu_{lm} n p_m p_l}{\sigma_m^2 \sigma_l^2}}{\sum_{q=1}^J \frac{p_q}{\sigma_q^2} + \sum_{q=1}^J \sum_{m \neq k} \frac{\mu_{km}\mu_{qm} n p_m p_q}{\sigma_q^2 \sigma_m^2}}. \tag{A.2.10}$$

Summing over all $l \neq k$ yields

$$\sum_{l \neq k} \frac{p_l \sigma_k^2}{p_k a_{kl} \sigma_l^2} = \frac{\sum_{l \neq k} \frac{p_l}{\sigma_l^2} + \sum_{l \neq k} \sum_{m \neq k} \frac{\mu_{km}\mu_{lm} n p_m p_l}{\sigma_m^2 \sigma_l^2}}{\sum_{q=1}^J \frac{p_q}{\sigma_q^2} + \sum_{q=1}^J \sum_{m \neq k} \frac{\mu_{km}\mu_{qm} n p_m p_q}{\sigma_q^2 \sigma_m^2}}. \tag{A.2.11}$$

Since $\frac{p_k}{\sigma_k^2} + \sum_{m \neq k} \frac{\mu_{km}^2 p_m p_k}{\sigma_k^2 \sigma_m^2} > 0$, (A.2.11) is smaller than 1. Thus by contradiction $\sum_{j=1}^J \lambda_{kj}^* > -n_k$ where $\lambda_{kk}^* = 0$.

A.2.5 A Wald Test for Equality of Means under Local Parameterization

Consider a Wald statistics for equality between all possible pairs in the system, i.e. the H_0 is that $\mu_k = \mu_j$ for all j, k. Again, we assume knowledge of the variances. Let $Z_{kj,n}$ be random variables that converge in distribution to a standard normal. The test statistics can be written as follows:

$$\begin{aligned} W_n &= n \sum_k \sum_{j>k} \frac{(\hat{\mu}_k - \hat{\mu}_j)^2}{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}} \\ &= \sum_k \sum_{j>k} \left(\frac{n(\mu_k - \mu_j)^2}{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}} + 2 \frac{\sqrt{n}(\mu_k - \mu_j)}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} Z_{kj,n} + (Z_{kj,n})^2 \right). \end{aligned}$$

Using the local parameterization it follows that

$$W_n(F_n) \xrightarrow{d} \sum_k \sum_{j>k} \left(\frac{(\delta_k - \delta_j)^2}{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}} + 2 \frac{\delta_k - \delta_j}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} \mathcal{N}_{kj} + \mathcal{X}_{kj}^2 \right) \quad \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0),$$

$$P(|W_n(F_n) > c|) \to 1 \text{ for all } c > 0 \quad \text{if } \{F_n\} \in S(\infty, V_0)$$

where $\mathcal{N}_{kj}$ are standard normal random variables and $\mathcal{X}^2_{kj}$ are Chi-squared random variables with one degree of freedom.

Hence under distant systems, the first term goes to infinity as the Wald statistics does under the alternative of at least one single different mean. If the local parameters are all zero, the statistics is classical $\mathcal{X}^2$ with $J(J-1)/2$ degrees of freedom. Under any other close system however, the asymptotic distribution is a mixture between a chi square and mean zero normal plus a strictly positive constant. Hence depending on the norm of the pairwise differences, the distribution of the test statistics can yield very different critical values compared to a standard Wald testing procedure. As in the case of two groups, one can expect that for moderate sizes of the local parameters, a test will not reject the equality leading to larger type-II errors. Therefore, the local asymptotic framework allows for a better representation of the finite sample behavior, in particular when test statistics are *moderate*.

A.2.6 Effective Degrees of Freedom

Let D be the $n \times J$ dimensional matrix of stacked transposed $\mathbf{D}_i$'s and $\mathbf{Y}$ the vector of outcomes. Many estimators for $\boldsymbol{\mu}$ have a corresponding linear map that maps $\mathbf{Y}$ to its predictions $\hat{\mathbf{Y}}$. In particular, these estimators determine an $n \times n$-dimensional matrix Π such that $\Pi\mathbf{Y} = \hat{\mathbf{Y}}$. In the case of the standard projection in the first stage one obtains that $\Pi = D(D'D)^{-1}D'$. The complexity of the linear map or of the estimator can be described by the effective degrees of freedom, i.e. the sum of the eigenvalues which can be computed as the trace over the linear operator Π. Consider the following examples that illustrate directly why L_2 penalization can be beneficial since it reduces the sum of the eigenvalues of the linear map that is its trace. For the OLS one obtains that

$$tr(D(D'D)^{-1}D') = J.$$

Without loss of generality, assume now we have some prior belief on why regularization of the group means towards zero should be beneficial. A simple ridge estimator with shrinkage parameter $\kappa > 0$ yields a corresponding matrix with effective degrees of freedom

$$tr(D(D'D + \kappa I_J)^{-1}D') = \sum_{j=1}^{J} \frac{n_j}{n_j + \kappa} < J \quad \text{for all} \quad \kappa > 0.$$

The ridge estimator basically pushes the eigenvalues towards zero and for a nonorthogonal design reduces the impact of large covariances between different

regressors. In the case of orthogonalized groups it limits the impact of each category specific observation by moving it towards zero. In a generalized ridge setup, other shrinkage targets such as the global average are feasible as well, i.e. the zero target does not affect the general conclusion about the complexity. Regularization lowers the effective degrees of freedom and therefore potentially reduces estimation noise. This illustrates why the effective degrees of freedom are often used as description of the dimensionality of the parameter space, i.e. the complexity of the statistical model.

The complexity of both, optimal as well as estimated PCS is non-standard, i.e. we obtain the following results for the effective degrees of freedom:

Theorem A.2.2 *Let* $\Pi(\mathbf{\Lambda}^*)$ *and* $\Pi(\hat{\mathbf{\Lambda}})$ *denote the linear operator based on the MSE optimal smoothing parameters and the plug-in estimator respectively. It follows that*

$$
\begin{aligned}
&tr(\Pi(\mathbf{\Lambda}^*)) \in [1,2] && \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0) \\
&tr(\Pi(\mathbf{\Lambda}^*)) \searrow 1 && \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0), \Delta\boldsymbol{\delta} = \mathbf{0} \\
&tr(\Pi(\mathbf{\Lambda}^*)) \to t,\ t \in (1,2) && \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0), \Delta\boldsymbol{\delta} \neq \mathbf{0} \\
&tr(\Pi(\mathbf{\Lambda}^*)) \nearrow 2 && \text{if } \{F_n\} \in S(\infty, V_0)
\end{aligned}
$$

and

$$
\begin{aligned}
&tr(\Pi(\hat{\mathbf{\Lambda}})) \in [1,2] \quad a.s. && \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0) \\
&tr(\Pi(\hat{\mathbf{\Lambda}})) = 2 + O_p(n^{-1/2}) && \text{if } \{F_n\} \in S(\infty, V_0)
\end{aligned}
$$

with $\nearrow$ *and* $\searrow$ *denoting convergence from below and from above.*

First we show the bounds for the trace. Note that $tr(\Pi(\mathbf{\Lambda}^*)) = \sum_{k=1}^{J} \omega_{kk}^*$. This can be rewritten as

$$
\omega_{kk}^* = \frac{\left(1 + n \sum_{m=1}^{J} \mu_{km}^2 \frac{p_m}{\sigma_m^2}\right)}{\frac{\sigma_k^2}{p_k}\left(\sum_{l=1}^{J} \frac{p_l}{\sigma_l^2} + n \sum_{l=1}^{J} \frac{p_l}{\sigma_l^2} \sum_{m=1}^{J} \mu_{km}\mu_{lm} \frac{p_m}{\sigma_m^2}\right)}.
$$

Note that

$$
\sum_{l=1}^{J} \frac{p_l}{\sigma_l^2} \sum_{m=1}^{J} \mu_{km}\mu_{lm} \frac{p_m}{\sigma_m^2} = \sum_{l=1}^{J} \sum_{m>l} \mu_{lm}^2 \frac{p_l p_m}{\sigma_l^2 \sigma_m^2} = \frac{1}{2} \sum_{l=1}^{J} \sum_{m=1}^{J} \mu_{lm}^2 \frac{p_l p_m}{\sigma_l^2 \sigma_m^2}
$$

and hence the denominator of w_{kk}^* is independent of k which yields

$$\sum_{k=1}^{J} \omega_{kk}^* = \sum_{k=1}^{J} \frac{\frac{p_k}{\sigma_k^2} + n \sum_{m \neq k} \mu_{km}^2 \frac{p_k p_m}{\sigma_k^2 \sigma_m^2})}{\sum_{l=1}^{J} \left(\frac{p_l}{\sigma_l^2} + \frac{n}{2} \sum_{m \neq l} \mu_{lm}^2 \frac{p_l p_m}{\sigma_l^2 \sigma_m^2} \right)}$$

which behaves like $(1+nx)/(1+nx/2)$ with $x \geq 0$ and hence is bounded between one and two. This derivation is independent from whether true or estimated means are used. Now we derive the convergence results for the optimal PCS under close and distant systems.

If $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$ with $\Delta\boldsymbol{\delta} = 0$ then $nx \to 0$, i.e. we have that

$$\omega_{kk}^* = \frac{\frac{p_k}{\sigma_k^2}}{\sum_{l=1}^{J} \frac{p_l}{\sigma_l^2}} + o(1) \Rightarrow \sum_{k=1}^{J} \omega_{kk}^* = 1 + o(1)$$

where convergence from above in n follows from the form.

If $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$ with $\Delta\boldsymbol{\delta} \neq \mathbf{0}$ then $nx \to \tilde{x} \in \mathbb{R}^+ \backslash \{0\}$ and thus

$$\sum_{k=1}^{J} \omega_{kk}^* = \frac{1 + \tilde{x}}{1 + \tilde{x}/2} + o(1)$$

with $(1 + \tilde{x})/(1 + \tilde{x}/2) \in (1, 2)$.

If $\{F_n\} \in S(\infty, V_0)$, then dividing numerator and denominator by n yields

$$\sum_{k=1}^{J} \omega_{kk}^* = \sum_{k=1}^{J} \frac{\frac{p_k}{\sigma_k^2} \sum_{m=1}^{J} \mu_{km}^2 \frac{p_m}{\sigma_m^2}}{\frac{1}{2} \sum_{l=1}^{J} \sum_{m=1}^{J} \mu_{lm}^2 \frac{p_l p_m}{\sigma_l^2 \sigma_m^2}} + O(n^{-1}) = 2 + O(n^{-1}).$$

where convergence from below follows from the form. For the estimated smoothing parameters however the equality only holds for probability limits and hence we have

$$\begin{aligned} \hat{\omega}_{kk} &= \frac{\left(1 + \sum_{m=1}^{J} \hat{\mu}_{km}^2 \frac{n_m}{\hat{\sigma}_m^2} \right)}{\frac{\hat{\sigma}_k^2}{n_k} \left(\sum_{l=1}^{J} \frac{n_l}{\hat{\sigma}_l^2} + \sum_{l=1}^{J} \frac{n_l}{\hat{\sigma}_l^2} \sum_{m=1}^{J} \hat{\mu}_{km} \hat{\mu}_{lm} \frac{n_m}{\hat{\sigma}_m^2} \right)} \\ &= \frac{\left(1 + n \sum_{m=1}^{J} \mu_{km}^2 \frac{p_m}{\sigma_m^2} \right)}{\frac{\sigma_k^2}{p_k} \left(\sum_{l=1}^{J} \frac{p_l}{\sigma_l^2} + n \sum_{l=1}^{J} \frac{p_l}{\sigma_l^2} \sum_{m=1}^{J} \mu_{km} \mu_{lm} \frac{p_m}{\sigma_m^2} \right)} + O_p(n^{-1/2}) \end{aligned}$$

which is equivalent to the expression for the optimal weights plus remainder. Hence the result from above for the sum follows by adjusting the approximation order from $O(n^{-1})$ to $O_p(n^{-1/2})$.

The effective degrees of freedom for both optimal and estimated PCS are bounded between one and two by construction. For the optimal PCS, the sum of the eigenvalues converges to one if the local parameters are all zero and to a fixed number between one and two if the local parameters are nonzero but finite. In the case of distant systems, it converges to the upper bound of two. Note, however, that the parameter t is not necessarily monotonic in the norm of the local parameter vector[4]. For the estimated smoothing parameters under distant systems, convergence in probability as in Lemma 2.4.1 assures that the trace converges to two in probability as well. However, under close systems convergence is not achieved.

As a corollary for the optimal PCS, if there are at least two different groups, the sum of the eigenvalues of the linear operator will converge to two as the sample size increases. If there is only one location, i.e. the global mean, it converges to one. Independently of the total number of different groups under the true DGP, the effective degrees of freedom will always be between one and two. This seems to imply that, for large samples, any system described by a finite number of locations should optimally (in a MSE sense) be modeled by at most two effective parameters. We are not aware of any comparable result in the literature. It implies that the eigenvalue conditions by Cohen (1966) for admissibility are met, however under data dependent smoothing parameters his framework does not exactly coincide with the PCS.

A.2.7 Finite-Sample Properties of Optimal Smoothing Parameters

This section contains a qualitative discussion of the small sample behavior of MSE optimal smoothing parameters with four groups, i.e. $J = 4$. Let group 1 be the reference category, i.e. the group whose mean is shrunk to the means of the target groups. We report the MSE optimal smoothing weights[5] ω_{11}, ω_{12}, ω_{13} and ω_{14} as functions of (1) number of observations: n_1, n_2 and n_3, (2) error variances: σ_1^2, σ_2^2 and σ_3^2 and (3) differences in group means[6]: $\Delta\mu_{12}$, $\Delta\mu_{23}$ and $\Delta\mu_{13}$ since the smoothing parameters depend only on differences in means, not on mean levels. For simplicity, effects are reported for shifting single group means which can be directly translated into mean differences.

[4]This statement is true for all the L_p norms, including the sup norm used for the sequence characterization in Definition 2.

[5]All ω's in the following text should have a star superscript which is left out unless necessary for readability.

[6]Note that the MSE optimal smoothing parameters depend on cell probabilities, not observations. The analysis is one-to-one if $n_k = np_k$.

For each DGP, a group mean vector, $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)'$, error variance vector, $\boldsymbol{\sigma^2} = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2)'$ and number of observations vector, $\mathbf{n} = (n_1, n_2, n_3, n_4)'$, have to be specified. There are three mean designs - equal means design (A), unequal means design (B) with three equal means and one distant mean and unequal means design (C) with 3 almost equal means and one distant mean. In addition, each design is combined with homoskedastic or heteroskedastic error variances and small or large number of observations. The upper part of the Table A.2.1 contains all parameter values. The lower part contains the ranges in which parameters are shifted ceteris paribus.

Table A.2.1: Design Values

n_1	n_2	n_3	n_4	σ_1^2	σ_2^2	σ_3^2	σ_4^2		μ_1	μ_2	μ_3	μ_4
Small Design				**Homosked.**					**Equal Means**			
5	5	5	5	1	1	1	1	(A)	0	0	0	0
Large Design				**Heterosked.**					**Non-Equal Means**			
100	100	100	100	3	5	1.5	1	(B)	0	0	0	100
								(C)	0	100	2	0
Ranges of Inputs in $\omega_{11}(\cdot)$, $\omega_{12}(\cdot)$, $\omega_{13}(\cdot)$, and $\omega_{14}(\cdot)$												
No. of obs., n_j				**Variances, σ_j^2**					**Means, μ_j**			
[2,100]				[1,10]					[-400,400]			

Note that for design (A) the effects of inputs are monotonic. The effects under the designs (B) and (C) are not monotonic and therefore illustrated graphically. For the sake of brevity, only the most relevant results are mentioned.

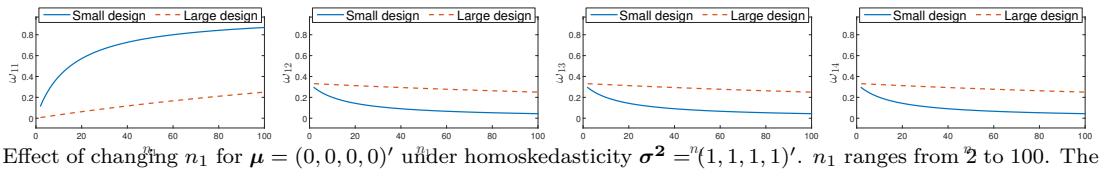
Equal Means Design (A) In design (A), $\boldsymbol{\mu} = (0, 0, 0, 0)'$, we analyze the effect of number of observations under both types of error variances and the effect of variance in a small and large sample design. Note that the MSE optimal smoothing parameters under equal means simplify to

$$\omega_{kj} = \frac{n_j}{\sigma_j^2} \Big/ \sum_{l=1}^{J} \frac{n_l}{\sigma_l^2}. \tag{A.2.12}$$

Effects of changing number of observations: Keeping the other parameters constant, ω_{kj} increases in n_j and decreases in n_m where $m \neq j$. In other words, the target group with more observations is relatively more informative for the other groups with equal means and therefore gets a relatively higher smoothing weight, see Figure A.2.1.

In smaller samples, additional observations play a more important role for ω_{kj} and help to smooth the mean of base category relatively more towards the largest group. In larger samples, additional observations are not as important since the whole system is already stabilized towards large groups.

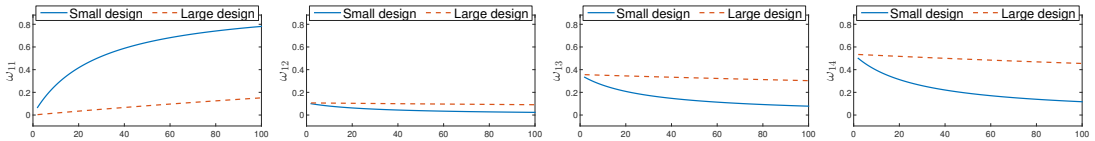
Figure A.2.1: Effect of n_1 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Equal Means (A), Homoskedasticity



Effect of changing n_1 for $\boldsymbol{\mu} = (0, 0, 0, 0)'$ under homoskedasticity $\boldsymbol{\sigma^2} = (1, 1, 1, 1)'$. n_1 ranges from 2 to 100. The other groups contain 5 observations (small design) or 100 observations (large design).

Under heteroskedasticity, target groups with larger variances have lower smoothing weights, see Figure A.2.2. In a small sample design under equal means, the number of observations in a group is more important to get a higher smoothing weight in comparison to a large design in which the variance is the more important factor (compare ω_{11} and ω_{14} in Figure A.2.2). This implies that a small variance can be exploited well only when the number of observations in the group is reasonably large.

Figure A.2.2: Effect of n_1 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Equal Means (A), Heteroskedasticity



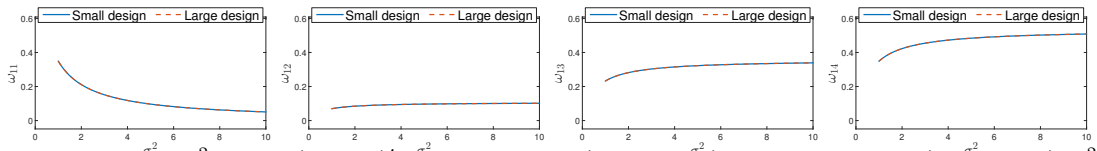
Effect of changing n_1 for $\boldsymbol{\mu} = (0, 0, 0, 0)'$ under heteroskedasticity, $\boldsymbol{\sigma^2} = (3, 5, 1.5, 1)'$. n_1 ranges from 2 to 100. The other groups contain 5 observations (small design) or 100 observations (large design).

Effects of changing error variance: In general, ω_{kk} is decreasing in its own variance σ_k^2 while ω_{kj} increases in σ_q^2 where $q \neq j$ and decreases in σ_j^2. Intuitively, groups with lower variance can provide more precise mean estimates and therefore the smoothing towards them is relatively stronger, see Figure A.2.3. Since group 4 has always the lowest variance, it has the largest smoothing weight.

At high variance levels, groups with large variances get almost zero weight and thus all smoothing weights stay relatively stable. Note that the results are independent of the sample size because the number of observations in all groups are scaled by the same constant and by (A.2.12) the values coincide.

Unequal Means Design (B) Design (B), $\boldsymbol{\mu} = (0, 0, 0, 100)'$, represents a situation of three equal means and one distant mean. We report the effect of an increase in the number of observations and of changing mean differences. In addition, the effect of changes in variance is presented for the heteroskedastic design.

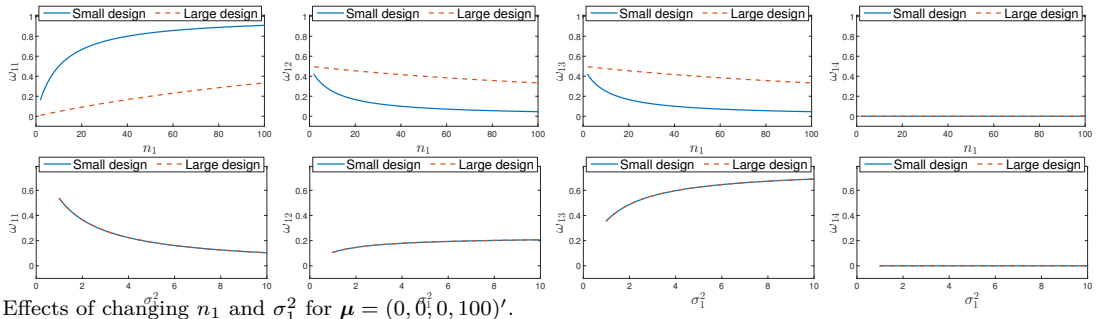
Figure A.2.3: Effect of σ_1^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Equal Means (A), Heteroskedasticity



Effect of changing σ_1^2 for $\boldsymbol{\mu} = (0, 0, 0, 0)'$ for 5 observations (small design) or 100 observations (large design). σ_1^2 ranges from 1 to 10. The variances are given by $\boldsymbol{\sigma^2} = (\sigma_1^2, 5, 1.5, 1)'$.

Effect of a distant mean The large distance between μ_1 and μ_4 decreases the smoothing of group 1 towards the distant mean considerably, see Figure A.2.4. As group 4 gets almost zero weight, the changes in n_1, n_2, n_3, σ_1^2, σ_2^2 and σ_3^2 have an effect mainly on the three equal mean groups as in the design (A).

Figure A.2.4: Effect of n_1 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means (B), Homoskedasticity and Effect of σ_1^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means (B), Heteroskedasticity



Effects of changing n_1 and σ_1^2 for $\boldsymbol{\mu} = (0, 0, 0, 100)'$.
1st Row: Effect of changing n_1 under homoskedasticity, $\boldsymbol{\sigma^2} = (1, 1, 1, 1)'$. n_1 ranges from 2 to 100. The other groups contain 5 observations (small design) or 100 observations (large design).
2nd Row: Effect of changing σ_1^2 for 5 observations (small design) or 100 observations (large design). σ_1^2 ranges from 1 to 10. The variances are given by $\boldsymbol{\sigma^2} = (\sigma_1^2, 5, 1.5, 1)'$.

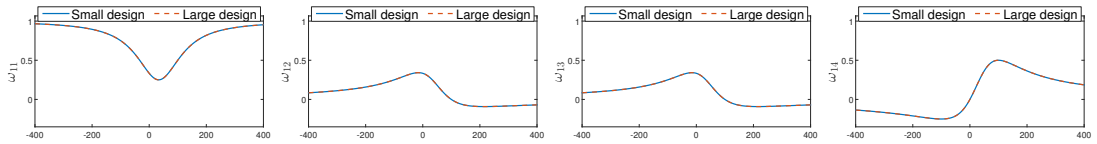
Effects of changing a group mean The impact of shifting a mean depends on: (1) which group mean changes, (2) the size of the shift and (3) the presence of another close group mean. Under heteroskedasticity, relatively larger variances lead to flatter curves and lower smoothing levels. Relatively smaller variances lead to more amplified curves and higher smoothing levels but the qualitative results described below do not change.

The main effects are illustrated on shifting μ_1, see Figure A.2.5. The effects are no longer monotonic: Shifting μ_1 far away from all the other group means puts most weight into ω_{11}, as there is no other close group mean to which it would be sensible to smooth. Note that ω_{12}, ω_{13} and ω_{14} are close to zero or even slightly negative for extreme values of μ_1. The smoothing weight ω_{11} takes its lowest values in an interval, in which there are other close means towards which

it pays off to smooth. Meanwhile, the other smoothing weights have their peaks in this interval.

The asymmetric decrease of the curves of ω_{12} and ω_{13} around 0 is caused by the absence or presence of other close means with respect to μ_1. Shifting μ_1 from 0 towards negative values causes a milder decrease, since there is no other close mean to compete with for groups 2 and 3. Shifting μ_1 from 0 towards positive values causes a steeper decrease, since in this direction there is a mean of group 4 to which it is also sensible to smooth. Similar logic can be used to explain the asymmetry around the peak of ω_{14}.

Figure A.2.5: Effect of shifting μ_1 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means (B), Homoscedasticity



Effects of shifting μ_1 under homoskedasticity, $\boldsymbol{\sigma^2} = (1, 1, 1, 1)'$, for 5 observations (small design) or 100 observations (large design). μ_1 ranges from -400 to 400. The group means are given by $\boldsymbol{\mu} = (\mu_1, 0, 0, 100)'$.

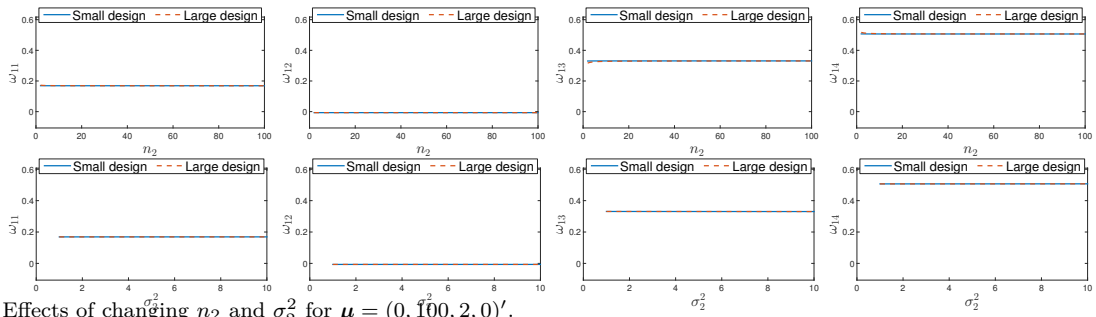
Unequal Means Design (C) Design (C), $\boldsymbol{\mu} = (0, 100, 2, 0)'$, represents a situation of three almost equal means (2 of them are equal to each other) and one distant mean. We report the effects under heteroskedasticity for small and large samples.

Effect of a distant mean In this design, the distant mean group 2 gets an almost zero or even slightly negative smoothing weight. Moreover, changes in n_2 or σ_2^2 have almost no effect on all the smoothing weights because $\Delta\mu_{21}$ is too large and therefore group 2 has close to zero weight regardless of how many observations it has and how small the variance is, see Figure A.2.6. Leaving group 2 aside, the effects of n_1, n_3, σ_1^2 and σ_3^2 are very similar to the effects in the design (B), since μ_3 is close enough to μ_1 and μ_4, i.e. they are almost as equal.

Effects of changing a group mean The effect of a mean shift depends on: (1) which group mean changes, (2) size of the shift, (3) presence of another close group mean and (4) error variances.

The effect of shifting μ_1 is comparable to shifting μ_1 in design (B). The only change is that in design (C), group 2 represents now the "distant group mean" and group 4 is now in the set of "equal group means". Shifting μ_2 follows a same logic as shifting μ_1 in design (B) only on a smaller scale, since there is no

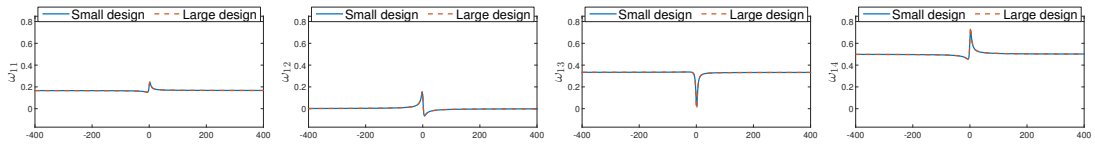
Figure A.2.6: Effect of n_2 and σ_2^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means (C), Heteroskedasticity



Effects of changing n_2 and σ_2^2 for $\boldsymbol{\mu} = (0, 100, 2, 0)'$.
1st Row: Effect of changing n_2 under heteroskedasticity, $\boldsymbol{\sigma^2} = (3, 5, 1.5, 1)'$. n_2 ranges from 2 to 100. The other groups contain 5 observations (small design) or 100 observations (large design).
2nd Row: Effect of changing σ_2^2 for 5 observations (small design) or 100 observations (large design). σ_2^2 ranges from 1 to 10. The variances are given by $\boldsymbol{\sigma^2} = (3, \sigma_2^2, 1.5, 1)'$.

large stabilizing mean, see Figure A.2.7. Lower error variances lead to a higher smoothing and vice versa.

Figure A.2.7: Effect of shifting μ_2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means (C), Heteroscedasticity



Effects of shifting μ_2 under heteroscedasticity, $\boldsymbol{\sigma^2} = (3, 5, 1.5, 1)'$, for 5 observations (small design) or 100 observations (large design). μ_2 ranges from -400 to 400. The group means are given by $\boldsymbol{\mu} = (0, \mu_2, 2, 0)'$.

If the means are very close to each other, the MSE optimal smoothing parameters change their values sharply in a narrow interval covering the close distance between the means. This behavior could potentially cause problems for any estimator of the optimal smoothing parameters that is subject to small sample variation. Finite sample deviations from the true parameters might yield smoothing parameters far away from the optimal ones leading to unfavorable aggregation. Once there is one distant mean, the smoothing parameters stabilize.

B Appendix for Detecting Structural Breaks using a Fusion Lasso Penalty

B.1 Additional Lemmas

The lemmas are adjusted versions of the lemmas in Qian and Su (2016) adopted to the adaptive fusion penalty.

Lemma B.1.1 *Consider the optimization problem (3.2.5) with a known* Σ *and without loss of generality* $\Sigma = I$*, then the following holds for the solution* $\{\hat{\beta}_t, t = 1, 2, \ldots, T\}$ *and* $\{\hat{\theta}_t, t = 1, 2, \ldots, T : \hat{\theta}_1 = \hat{\beta}_1,\ \hat{\theta}_t = \hat{\beta}_t - \hat{\beta}_{t-1}$ *for* $t \geq 2\}$*, i.e.* $\hat{\beta}_r = \sum_{s=1}^{r} \hat{\theta}_s$ *and* $k \in \{1, \ldots, p\}$*:*

(i-a) $\sum_{r=\hat{T}_j}^{T} x_{r,k}(y_r - x_r' \sum_{s=1}^{r} \hat{\theta}_s) = \frac{\lambda}{2} w_{\hat{T}_j,k} \operatorname{sgn}\left[\hat{\theta}_{\hat{T}_j,k}\right]$ *for* $j = 1, \ldots, \hat{m}$ *and* $\hat{\theta}_{t,k} \neq 0$,

(i-b) $|\sum_{r=\hat{T}_j}^{T} x_{r,k}(y_r - x_r' \sum_{s=1}^{r} \hat{\theta}_s)| \leq \frac{\lambda}{2} w_{\hat{T}_j,k}$ *for* $j = 1, \ldots, \hat{m}$ *and* $\hat{\theta}_{t,k} = 0$,

(ii) $\left\|\sum_{r=1}^{T} x_r(y_r - x_r'\hat{\beta}_r)\right\| = 0$ *for* $t = 1$,

(iii) $\left\|\sum_{r=t}^{T} x_r(y_r - x_r'\hat{\beta}_r)\right\| \leq \frac{\lambda}{2} \|diag(W_t)\|$ *for* $t = 2, \ldots, T$,

where W_t *is a diagonal* $p \times p$ *matrix with the weights* $w_{t,k}$*,* $t = \{2, \ldots, T\}$ *and* $k \in \{1, \ldots, p\}$*, on the diagonal and* $diag(\cdot)$ *returns the main matrix diagonal as a vector.*

Proof: To minimize (3.2.5), it simplifies the analysis when β_t's are replaced by θ_t's as defined in the lemma. The θ's are just the differences between the β coefficients and θ_1 is then equal to β_1. In this case, the following first order conditions have to hold for $\hat{\theta}$'s based on the subdifferential calculus:

$$-2\sum_{r=1}^{T}\left(y_r - x_r'\sum_{s=1}^{r}\hat{\theta}_s\right)x_r = 0 \qquad \text{if } t = 1, \tag{B.1.1}$$

$$-2\sum_{r=t}^{T}\left(y_r - x_r'\sum_{s=1}^{r}\hat{\theta}_s\right)x_r + \lambda W_t e_t = 0 \qquad \text{if } t = 2, \ldots, T, \tag{B.1.2}$$

where W_t is a diagonal $p \times p$ matrix having the weights $w_{t,k}$, $k \in \{1, \ldots, p\}$, on the diagonal and e_t is a $p \times 1$ vector with the following elements for $k \in \{1, \ldots, p\}$:

$$e_{t,k} = \operatorname{sgn}\left[\hat{\theta}_{t,k}\right] \qquad \text{if } \hat{\theta}_{t,k} \neq 0,$$

$$|e_{t,k}| \leq 1 \qquad \text{if } \hat{\theta}_{t,k} = 0.$$

From here it is straightforward to derive (i-a) and (i-b) for individual coefficients k at $\hat{T}_j$:

$$\sum_{r=\hat{T}_j}^{T} x_r(y_r - x_r' \sum_{s=1}^{r} \hat{\theta}_s) = \frac{\lambda}{2} W_{\hat{T}_j} e_{\hat{T}_j} \text{ for } j = 1, \ldots, \hat{m}.$$

The result follows from the definition of $e_{t,k}$. From (B.1.1) and $\hat{\beta}_r = \sum_{s=1}^{r} \hat{\theta}_s$, we get (ii). As $\|W_t e_t\| \leq \|diag(W_t)\|$, then:

$$\left\| \sum_{r=t}^{T} x_r(y_r - x_r' \hat{\beta}_r) \right\| \leq \frac{\lambda}{2} \|diag(W_t)\| \text{ for } t = 2, \ldots, T.$$

□

The following lemma is based on the proof of Theorem 2 in Merlevède et al. (2009).

Lemma B.1.2 *Let $\{\xi_t, t = 1, 2, \ldots\}$ be a zero-mean strong mixing process with the mixing coefficients $\alpha(\tau) \leq c_\alpha \rho^\tau$ for some $c_\alpha > 0$ and $\rho \in (0, 1)$. If $\sup_{1 \leq t \leq T} |\xi_t| \leq M_T$, then there exists a constant C_0 depending on c_α and ρ such that for any $T \geq 2$ and $\varepsilon > 0$,*

$$P\left(\left| \sum_{t=1}^{T} \xi_t \right| > \varepsilon \right) \leq \exp\left(- \frac{C_0 \varepsilon^2}{\nu_0^2 T + M_T^2 + \varepsilon M_T (\log T)^2} \right),$$

where $\nu_0^2 = \sup_{t \geq 1} \left[var(\xi_t) + 2 \sum_{s=t+1}^{\infty} |cov(\xi_t, \xi_s)| \right]$.

Proof (Merlevède et al., 2009, Theorem 2): In the mentioned source, the inequality is proved for $\alpha_\tau \leq \exp(-2c\tau)$ for some $c > 0$. If $c_\alpha = 1$, we can take the $\rho = \exp(-2c)$ and apply the theorem to get (i). □

Lemma B.1.3 *Let Assumptions 2, 3 and 5 hold and let $\nu_T = T\delta_T$. Under these assumptions:*

(i) $\sup_{\substack{1 \leq s < r \leq T+1 \\ r-s \geq \nu_T}} \mu_{max}\left(\frac{1}{r-s} \sum_{t=s}^{r-1} x_t x_t' \right) \leq \bar{c}_{xx} + o_p(1)$,

(ii) $\inf_{\substack{1 \leq s < r \leq T+1 \\ r-s \geq \nu_T}} \mu_{min}\left(\frac{1}{r-s} \sum_{t=s}^{r-1} x_t x_t' \right) \geq \underline{c}_{xx} + o_p(1)$.

Proof: (i)

Using the Weyl's inequality for eigenvalues of a perturbed Hermitian matrix, another important inequality $\mu_{\max}(A) \leq \|A\|_F$ for any matrix A where $\|\cdot\|_F$ denotes Frobenius norm and Assumption 3:

$$\mu_{\max}\left(\frac{1}{r-s} \sum_{t=s}^{r-1} x_t x_t' \right) = \mu_{\max}\left(\frac{1}{r-s} \sum_{t=s}^{r-1} E(x_t x_t') + \frac{1}{r-s} \sum_{t=s}^{r-1} (x_t x_t' - E(x_t x_t')) \right)$$

$$
\begin{aligned}
&\leq \mu_{\max}\left(\frac{1}{r-s}\sum_{t=s}^{r-1}E(x_tx_t')\right)+\mu_{\max}\left(\frac{1}{r-s}\sum_{t=s}^{r-1}(x_tx_t'-E(x_tx_t'))\right)\\
&\leq \mu_{\max}\left(\frac{1}{r-s}\sum_{t=s}^{r-1}E(x_tx_t')\right)+\left\|\frac{1}{r-s}\sum_{t=s}^{r-1}(x_tx_t'-E(x_tx_t'))\right\|_F\\
&\leq \bar{c}_{xx}+\left\|\frac{1}{r-s}\sum_{t=s}^{r-1}(x_tx_t'-E(x_tx_t'))\right\|_F.
\end{aligned}
$$

Now, it needs to be shown that $\left\|\frac{1}{r-s}\sum_{t=s}^{r-1}(x_tx_t'-E(x_tx_t'))\right\|_F=o_p(1)$. This can be done by proving that $\max_{\substack{1\leq s<r\leq T+1\\ r-s\geq\nu_T}}\left\|\frac{1}{r-s}\sum_{t=s}^{r-1}(x_tx_t'-E(x_tx_t'))\right\|_F=o_p(1)$.

The Assumption 5 implies $\nu_T\geq c_\nu T^{1/q}$. Let $\eta_T=T^{1/(2q)}$ and ι_{up} be an arbitrary $p\times 1$ unit vector such that $\|\iota_{up}\|=1$ for $u=1,2$. Let $\zeta_t\equiv\iota_{1p}'[x_tx_t'-E(x_tx_t')]\iota_{2p}$, $\zeta_{1t}\equiv\iota_{1p}'[x_tx_t'\mathbb{1}_t-E(x_tx_t'\mathbb{1}_t)]\iota_{2p}$ and $\zeta_{2t}\equiv\iota_{1p}'[x_tx_t'(1-\mathbb{1}_t)-E(x_tx_t'(1-\mathbb{1}_t))]\iota_{2p}$ where $\mathbb{1}_t\equiv\mathbb{1}(\|x_t\|^2\leq\eta_T)$. Under these definitions, $\zeta_t=\zeta_{1t}+\zeta_{2t}$. By Boole inequality and Lemma B.1.2, for a sufficiently large positive C:

$$
\begin{aligned}
&P\left(\sup_{\substack{1\leq s<r\leq T+1\\ r-s\geq\nu_T}}\left|\frac{1}{\sqrt{r-s}}\sum_{t=s}^{r-1}\zeta_{1t}\right|\geq C(\log T)^3\right)\\
&\leq T^2\sup_{\substack{1\leq s<r\leq T+1\\ r-s\geq\nu_T}}P\left(\left|\sum_{t=s}^{r-1}\zeta_{1t}\right|\geq C\sqrt{r-s}(\log T)^3\right)\\
&\leq T^2\sup_{\substack{1\leq s<r\leq T+1\\ r-s\geq\nu_T}}\exp\left(-\frac{C_0C^2(r-s)(\log T)^6}{\nu_0^2(r-s)+4\eta_T^2+2C\sqrt{r-s}(\log T)^3\eta_T[\log(r-s)]^2}\right)\\
&\leq T^2\exp\left(-\frac{C_0C^2\nu_T(\log T)^6}{\nu_0^2\nu_T+4\eta_T^2+2C\sqrt{\nu_T}(\log T)^3\eta_T[\log\nu_T]^2}\right)\\
&\leq \exp\left(-\frac{C_0C^2\nu_T(\log T)^6}{\nu_0^2\nu_T+4\eta_T^2+2C\sqrt{\nu_T}(\log T)^3\eta_T[\log\nu_T]^2}+2\log T\right)\\
&\to 0 \text{ as } T\to\infty
\end{aligned}
$$

The second but last inequality stems from a careful analysis of the exponential function. Analyzing the first derivative reveals that the function is decreasing in $r-s$ for a fixed T, therefore it achieves its maximum at ν_T, the lower bound of $r-s$. Further analysis of the limiting behavior yields the convergence to zero with $T\to\infty$.

As the next step, by Assumption 2, Boole and Markov inequalities and Lebesgue's dominated convergence theorem:

$$
P\left(\sup_{\substack{1\leq s<r\leq T+1\\ r-s\geq\nu_T}}\left|\frac{1}{\sqrt{r-s}}\sum_{t=s}^{r-1}\zeta_{2t}\right|\geq C(\log T)^3\right)\leq P\left(\max_{1\leq t\leq T}\|x_t\|^2\geq\eta_T\right)
$$

$$\leq T \max_{1\leq t\leq T} P\left(\|x_t\|^2 \geq \eta_T\right) \leq \frac{T}{\eta_T^{2q}} \max_{1\leq t\leq T} E\left[\|x_t\|^{4q} \mathbb{1}\{\|x_t\|^2 \geq \eta_T\}\right] \to 0 \text{ as } T \to \infty,$$

where the first inequality stems from the definition of ζ_{2t} which is non-zero only when $\|x_t\|^2 \geq \eta_T$, i.e. the supremum event as a more restrictive event can be bounded from above by $P\left(\max_{1\leq t\leq T} \|x_t\|^2 \geq \eta_T\right)$. Then, the Boole, Markov and Lebesgue's dominated convergence theorem are applied using the Assumption 2 on $\|x_t\|^{4q}$.

Given that the unit vectors ι_{1p} and ι_{2p} are arbitrary, it follows that

$$\max_{\substack{1\leq s<r\leq T+1 \\ r-s\geq \nu_T}} \left\| \frac{1}{r-s} \sum_{t=s}^{r-1} [x_t x_t' - E(x_t x_t')] \right\|_F = O_p(\nu_T^{-1/2} (\log T)^3) = o_p(1).$$

Then, (i) follows. The proof of (ii) follows analogical steps. □

B.2 Proofs

B.2.1 Proof of Theorem 3.2.1

Define $A_{T,j} = \{|\hat{T}_j - T_j^0| > T\delta_T\}$ and $C_T = \{\max_{1\leq l\leq m^0} |\hat{T}_l - T_l^0| < I_{min}/2\}$. Event C_T captures that the largest time distance between an estimated break point and the true break point is at most a half of the shortest period.

Since

$$P\left(\max_{1\leq j\leq m^0} |\hat{T}_j - T_j^0| > T\delta_T\right) \leq \sum_{j=1}^{m^0} P(A_{T,j}),$$

the consistency will be proven by showing (i) $\sum_{j=1}^{m^0} P(A_{T,j} \cap C_T) \to 0$ and (ii) $\sum_{j=1}^{m^0} P(A_{T,j} \cap C_T^C) \to 0$ where the superscript C denotes the set complement.

Proving (i): It will be shown that $\sum_{j=1}^{m^0} P(A_{T,j}^+ \cap C_T) \to 0$ and $\sum_{j=1}^{m^0} P(A_{T,j}^- \cap C_T) \to 0$ where $A_{T,j}^+ = \{\hat{T}_j - T_j^0 > T\delta_T\}$ and $A_{T,j}^- = \{T_j^0 - \hat{T}_j > T\delta_T\}$. First, focus on $A_{T,j}^-$ case, i.e. the estimated break is before the true break on the timeline, $\hat{T}_j < T_j^0$. The other case is proven analogously.

By definition of C_T, $T_{j-1}^0 < \hat{T}_j < T_{j+1}^0$ for all $j \in \{1, \ldots, m_0\}$. In other words, C_T describes an event when the j-th estimated break is between the true $(j-1)$-th and $(j+1)$-th break point.

By plugging the model $y_t = x_t'\beta_t^0 + \varepsilon_t$ into Lemma B.1.1 (iii), it holds for the estimated break points and true break points:

$$\left\| -\sum_{r=\hat{T}_j}^{T} x_r x_r' (\hat{\beta}_r - \beta_r^0) + \sum_{r=\hat{T}_j}^{T} x_r \varepsilon_r \right\| \leq \frac{\lambda}{2} \left\| diag(W_{\hat{T}_j}) \right\|,$$

$$\left\| -\sum_{r=T_j^0}^{T} x_r x_r'(\hat{\beta}_r - \beta_r^0) + \sum_{r=T_j^0}^{T} x_r \varepsilon_r \right\| \leq \frac{\lambda}{2} \left\| diag(W_{T_j^0}) \right\|.$$

By triangle inequality, the fact that $\hat{\beta}_r = \hat{\alpha}_{j+1}$ and $\beta_r^0 = \alpha_j^0$ for $r \in [\hat{T}_j, T_j^0 - 1]$ and by using another triangle inequality, the following result can be derived, where $c_{W_j} = \left\| diag(W_{\hat{T}_j}) \right\| + \left\| diag(W_{T_j^0}) \right\|$:

$$\begin{aligned}
\lambda c_{W_j}/2 &\geq \left\| -\sum_{r=\hat{T}_j}^{T} x_r x_r'(\hat{\beta}_r - \beta_r^0) + \sum_{r=\hat{T}_j}^{T} x_r \varepsilon_r + \sum_{r=T_j^0}^{T} x_r x_r'(\hat{\beta}_r - \beta_r^0) - \sum_{r=T_j^0}^{T} x_r \varepsilon_r \right\| \\
&= \left\| -\sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\hat{\beta}_r - \beta_r^0) + \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\| \\
&= \left\| -\sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\hat{\alpha}_{j+1} - \alpha_j^0) + \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\| \\
&\geq \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\hat{\alpha}_{j+1} - \alpha_j^0) \right\| - \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\| \\
&= \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\hat{\alpha}_{j+1} - \alpha_{j+1}^0 + \alpha_{j+1}^0 - \alpha_j^0) \right\| - \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\| \\
&\geq \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\alpha_{j+1}^0 - \alpha_j^0) \right\| - \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\hat{\alpha}_{j+1} - \alpha_{j+1}^0) \right\| - \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\| \\
\lambda &\geq \frac{1}{c_{W_j}/2} \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\alpha_{j+1}^0 - \alpha_j^0) \right\| - \frac{1}{c_{W_j}/2} \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\hat{\alpha}_{j+1} - \alpha_{j+1}^0) \right\| \\
&\quad - \frac{1}{c_{W_j}/2} \left\| \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\| \\
&\equiv R_{j,1} - R_{j,2} - R_{j,3}. \qquad \text{(B.2.1)}
\end{aligned}$$

Note that $c_{W_j} \leq 2\sqrt{p}\xi_T$. Define $\bar{R}_j(\lambda) = \{\lambda \geq \frac{1}{3}R_{j,1}\} \cup \{R_{j,2} \geq \frac{1}{3}R_{j,1}\} \cup \{R_{j,3} \geq \frac{1}{3}R_{j,1}\}$. The event $\bar{R}_j$ is equivalent to $\lambda \geq R_{j,1} - R_{j,2} - R_{j,3}$, i.e. $P(\bar{R}_j(\lambda)) = 1$. Then,

$$\begin{aligned}
P(A_{T,j}^- \cap C_T) &\leq P(A_{T,j}^- \cap C_T \cap \{\lambda \geq \frac{1}{3}R_{j,1}\}) \\
&\quad + P(A_{T,j}^- \cap C_T \cap \{R_{j,2} \geq \frac{1}{3}R_{j,1}\}) \\
&\quad + P(A_{T,j}^- \cap C_T \cap \{R_{j,3} \geq \frac{1}{3}R_{j,1}\}) \\
&\equiv AC_{j,1} + AC_{j,2} + AC_{j,3}.
\end{aligned}$$

Using that $\|Au\| = [tr(uu'A'A)]^{1/2} \geq \mu_{\min}(A'A)^{1/2}\|u\|$, where A is a matrix and u is a vector:

$$\begin{aligned}\sum_{j=1}^{m_0} AC_{j,1} &\leq \sum_{j=1}^{m_0} P(A_{T,j}^{-} \cap \{\lambda \geq \frac{1}{3}R_{j,1}\}) \\ &= \sum_{j=1}^{m_0} P\left(\frac{1}{c_{W_j}(T_j^0 - \hat{T}_j)}\left\|\sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\alpha_{j+1}^0 - \alpha_j^0)\right\| \leq \frac{3\lambda}{2(T_j^0 - \hat{T}_j)}; T_j^0 - \hat{T}_j > \delta_T T\right) \\ &\leq \sum_{j=1}^{m_0} P\left(c_{1T,j} \leq \frac{3\lambda\sqrt{p}\xi_T}{J_{\min}\delta_T T}; T_j^0 - \hat{T}_j > \delta_T T\right) \to 0,\end{aligned}$$

where the convergence to zero comes from $c_{1T,j} \equiv \mu_{\min}\left(\frac{1}{T_j^0 - \hat{T}_j}\sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'\right) \geq \underline{c}_{xx}/2 > 0$ with probability going to 1 by Lemma B.1.3 and the Assumption 8 says that $\frac{3\lambda\sqrt{p}\xi_T}{J_{\min}\delta_T T} \to 0$. This is a contradiction for $T \to \infty$ and therefore the probability converges to zero.

Next, the $AC_{j,2}$ will be bounded.

$$\begin{aligned}AC_{j,2} &= P(A_{T,j}^{-} \cap C_T \cap \{R_{j,2} \geq \frac{1}{3}R_{j,1}\}) \\ &\leq P(A_{T,j}^{-} \cap C_T \cap \left\{\bar{c}_{1T,j}\left\|\hat{\alpha}_{j+1} - \alpha_{j+1}^0\right\| \geq \frac{1}{3}c_{1T,j}\left\|\alpha_{j+1}^0 - \alpha_j^0\right\|\right\}),\end{aligned}$$

where $\bar{c}_{1T,j} \equiv \mu_{\max}\left(\frac{1}{T_j^0 - \hat{T}_j}\sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'\right) \leq 2\bar{c}_{xx}$ with probability going to 1 by Lemma B.1.3. Given $A_{T,j}^{-}$ and C_T, $\hat{\beta}_t = \hat{\alpha}_{j+1}$ for $t \in [T_j^0, (T_j^0 + T_{j+1}^0)/2 - 1]$. Using Lemma B.1.1 (iii) and following similar steps as to get (B.2.1):

$$\lambda c_{W_j}/2 \geq \left\|\sum_{r=T_j^0}^{(T_j^0+T_{j+1}^0)/2-1} x_r x_r'(\hat{\alpha}_{j+1} - \alpha_{j+1}^0)\right\| - \left\|\sum_{r=T_j^0}^{(T_j^0+T_{j+1}^0)/2-1} x_r \varepsilon_r\right\|,$$

where $c_{W_j} = \left\|diag(W_{T_j^0})\right\| + \left\|diag(W_{(T_j^0+T_{j+1}^0)/2-1})\right\|$. Conditional on C_T:

$$\left\|\hat{\alpha}_{j+1} - \alpha_{j+1}^0\right\| \leq (c_{2T,j})^{-1}\left[\frac{2\lambda\sqrt{p}\xi_T}{I_{\min}} + \left\|\frac{2}{T_{j+1}^0 - T_j^0}\sum_{r=T_j^0}^{(T_j^0+T_{j+1}^0)/2-1} x_r u_r\right\|\right],$$

where $c_{2T,j} \equiv \mu_{\min}\left(\frac{2}{T_{j+1}^0 - T_j^0}\sum_{r=T_j^0}^{(T_j^0+T_{j+1}^0)/2-1} x_r x_r'\right) \geq \underline{c}_{xx}/2 > 0$ with probability going to 1. Therefore:

$$\begin{aligned}&\sum_{j=1}^{m_0} P\left(\{\left\|\hat{\alpha}_{j+1} - \alpha_{j+1}^0\right\| \geq \frac{1}{3}(\bar{c}_{1T,j})^{-1}c_{1T,j}\left\|\alpha_{j+1}^0 - \alpha_j^0\right\|\} \cap C_T\right) \\ &\quad \leq \sum_{j=1}^{m_0} P\left(\frac{2\lambda\sqrt{p}\xi_T}{I_{\min}} \geq \frac{1}{3}(\bar{c}_{1T,j})^{-1}c_{1T,j}c_{2T,j}\left\|\alpha_{j+1}^0 - \alpha_j^0\right\|\right)\end{aligned}$$

$$+\sum_{j=1}^{m_0} P\left(\left\|\frac{2}{T_{j+1}^0-T_j^0}\sum_{r=T_j^0}^{(T_j^0+T_{j+1}^0)/2-1} x_r u_r\right\| \geq \frac{1}{3}(\bar{c}_{1T,j})^{-1} c_{1T,j} c_{2T,j} \left\|\alpha_{j+1}^0-\alpha_j^0\right\|\right).$$

The first term converges to zero since $\lambda\sqrt{p}\xi_T/(I_{\min}J_{\min}) \to 0$ under Assumptions 6 and 8. The second term is bounded from above by

$$P\left(\left\|\frac{2}{T_{j+1}^0-T_j^0}\sum_{r=T_j^0}^{(T_j^0+T_{j+1}^0)/2-1} x_r u_r\right\| \geq \bar{c}_{xx}\underline{c}_{xx}^2 \left\|\alpha_{j+1}^0-\alpha_j^0\right\|/24\right) \to 0.$$

The convergence to zero is a consequence of Assumptions 4, 6 and 7 as one can rearrange the elements:

$$P\left(\frac{1}{J_{\min}T^{1/2}\delta^{1/2}}\left\|\frac{\sqrt{2}}{\sqrt{T_{j+1}^0-T_j^0}}\sum_{r=T_j^0}^{(T_j^0+T_{j+1}^0)/2-1} x_r u_r\right\| \geq \frac{\bar{c}_{xx}\underline{c}_{xx}^2 I_{\min}^{1/2}}{24\sqrt{2}T^{1/2}\delta^{1/2}}\right) \to 0.$$

where the left side is $o\left(\frac{(logT)^3}{J_{\min}T^{1/2}\delta_T^{1/2}}\right)$ and the right hand side diverges as $\frac{I_{\min}^{1/2}}{T^{1/2}\delta^{1/2}} \to \infty$. Therefore, the $\sum_{j=1}^{m_0} AC_{j,2} \to 0$.

Remember that $c_{1T,j} \geq \underline{c}_{xx}/2 > 0$ with probability going to 1. Then,

$$\begin{aligned}
\sum_{j=1}^{m_0} AC_{j,3} &\leq P\left(A_{T,j}^- \cap \{R_{j,3} \geq \frac{1}{3}R_{j,1}\}\right) \\
&= P\left(A_{T,j}^- \cap \{\left\|\frac{1}{T_j^0-\hat{T}_j}\sum_{r=\hat{T}_j}^{T_j^0-1} x_r\varepsilon_r\right\| \geq \frac{1}{3}\left\|\frac{1}{T_j^0-\hat{T}_j}\sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\alpha_{j+1}^0-\alpha_j^0)\right\|\}\right) \\
&\leq \sum_{j=1}^{m_0} P\left(A_{T,j}^- \cap \{\left\|\frac{1}{\sqrt{T_j^0-\hat{T}_j}}\sum_{r=\hat{T}_j}^{T_j^0-1} x_r\varepsilon_r\right\| \geq \frac{1}{3}c_{1T,j}J_{\min}I_{\min}^{1/2}\}\right) \to 0,
\end{aligned}$$

where the convergence to zero follows from the same rearrangement as in the previous case based on Assumptions 4, 6 and 7. The part $\sum_{j=1}^{m^0} P(A_{T,j} \cap C_T) \to 0$ is proven.

Proving (ii): Similarly as before, it will be shown $\sum_{j=1}^{m^0} P(A_{T,j}^+ \cap C_T^C) \to 0$ and $\sum_{j=1}^{m^0} P(A_{T,j}^- \cap C_T^C) \to 0$. As the other result can be proven analogously, the $\sum_{j=1}^{m^0} P(A_{T,j}^- \cap C_T^C) \to 0$ is shown below. Define:

$$\begin{aligned}
D_T^{(l)} &\equiv \left\{\exists j \in \{1,\ldots,m_0\}, \hat{T}_j \leq T_{j-1}^0\right\} \cap C_T^C, \\
D_T^{(m)} &\equiv \left\{\forall j \in \{1,\ldots,m_0\}, T_{j-1}^0 < \hat{T}_j \leq T_{j+1}^0\right\} \cap C_T^C, \\
D_T^{(r)} &\equiv \left\{\exists j \in \{1,\ldots,m_0\}, \hat{T}_j \geq T_{j+1}^0\right\} \cap C_T^C.
\end{aligned}$$

From there, $\sum_{j=1}^{m^0} P(A_{T,j}^- \cap C_T^C) = \sum_{j=1}^{m^0} P(A_{T,j}^- \cap D_T^{(l)}) + \sum_{j=1}^{m^0} P(A_{T,j}^- \cap D_T^{(m)}) +$

$\sum_{j=1}^{m^0} P(A_{T,j}^- \cap D_T^{(r)})$.

First, $\sum_{j=1}^{m^0} P(A_{T,j}^- \cap D_T^{(m)})$ will be analyzed.

$$\begin{aligned}
P(A_{T,j}^- \cap D_T^{(m)}) &= P\left(A_{T,j}^- \cap \left\{\hat{T}_{j+1} - T_j^0 \geq \frac{1}{2} I_{\min}\right\} \cap D_T^{(m)}\right) \\
&\quad + P\left(A_{T,j}^- \cap \left\{\hat{T}_{j+1} - T_j^0 < \frac{1}{2} I_{\min}\right\} \cap D_T^{(m)}\right) \\
&\leq P\left(A_{T,j}^- \cap \left\{\hat{T}_{j+1} - T_j^0 \geq \frac{1}{2} I_{\min}\right\} \cap D_T^{(m)}\right) \\
&\quad + P\left(A_{T,j}^- \cap \left\{T_{j+1}^0 - \hat{T}_{j+1}^0 \geq \frac{1}{2} I_{\min}\right\} \cap D_T^{(m)}\right),
\end{aligned}$$

where the inequality is based on the fact that $0 \leq \hat{T}_{j+1} - T_j^0 \leq I_{\min}/2$ implies $T_{j+1}^0 - \hat{T}_{j+1}^0 = (T_{j+1}^0 - T_j^0) - (\hat{T}_{j+1}^0 - T_j^0) \geq I_{\min} - I_{\min}/2 = I_{\min}/2$. It will be helpful to see the following set hierarchy:

$$\begin{aligned}
&\left\{A_{T,j}^- \cap \left\{T_{j+1}^0 - \hat{T}_{j+1}^0 \geq \frac{1}{2} I_{\min}\right\} \cap D_T^{(m)}\right\} \subset \\
&\qquad \cup_{k=j+1}^{m^0-1} \left(\left\{T_k^0 - \hat{T}_k \geq \frac{1}{2} I_{\min}\right\} \cap \left\{\hat{T}_{k+1} - T_k^0 \geq \frac{1}{2} I_{\min}\right\} \cap D_T^{(m)}\right).
\end{aligned}$$

Then,

$$\begin{aligned}
&\sum_{j=1}^{m^0} P\left(A_{T,j}^- \cap D_T^{(m)}\right) \\
&\quad \leq P\left(A_{T,j}^- \cap \left\{\hat{T}_{j+1} - T_j^0 \geq \frac{1}{2} I_{\min}\right\} \cap D_T^{(m)}\right) \\
&\qquad + \sum_{j=1}^{m^0} \sum_{k=j+1}^{m^0-1} P\left(\left\{T_k^0 - \hat{T}_k \geq \frac{1}{2} I_{\min}\right\} \cap \left\{\hat{T}_{k+1} - T_k^0 \geq \frac{1}{2} I_{\min}\right\} \cap D_T^{(m)}\right).
\end{aligned} \tag{B.2.2}$$

To find the bounds, the following steps have to be made. In a similar fashion as in the previous part using Lemma B.1.1 (iii) for two points T_j^0 and $\hat{T}_j$:

$$\begin{aligned}
\frac{\lambda\sqrt{p}\xi_T}{T_j^0 - \hat{T}_j} &\geq \frac{1}{T_j^0 - \hat{T}_j} \left\| - \sum_{r=\hat{T}_j}^{T_j^0-1} x_r x_r'(\hat{\alpha}_{j+1} - \alpha_j^0) + \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\| \\
&\geq c_{1T,j} \left\|\hat{\alpha}_{j+1} - \alpha_j^0\right\| - \left\| \frac{1}{T_j^0 - \hat{T}_j} \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\|
\end{aligned}$$

Thus,

$$\left\|\hat{\alpha}_{j+1} - \alpha_j^0\right\| \leq (c_{1T,j})^{-1} \left[\frac{\lambda\sqrt{p}\xi_T}{T_j^0 - \hat{T}_j} + \left\| \frac{1}{T_j^0 - \hat{T}_j} \sum_{r=\hat{T}_j}^{T_j^0-1} x_r \varepsilon_r \right\| \right]. \tag{B.2.3}$$

Analogically, using Lemma B.1.1 (iii) for two points T_j^0 and $\hat{T}_{j+1}$:

$$\begin{aligned}\frac{\lambda\sqrt{p}\xi_T}{\hat{T}_{j+1}-T_j^0} &\geq \frac{1}{\hat{T}_{j+1}-T_j^0}\left\|-\sum_{r=T_j^0}^{\hat{T}_{j+1}-1} x_r x_r'(\hat{\alpha}_{j+1}-\alpha_{j+1}^0)+\sum_{r=T_j^0}^{\hat{T}_{j+1}-1} x_r\varepsilon_r\right\| \\ &\geq c_{3T,j}\left\|\hat{\alpha}_{j+1}-\alpha_{j+1}^0\right\| - \left\|\frac{1}{\hat{T}_{j+1}-T_j^0}\sum_{r=T_j^0}^{\hat{T}_{j+1}-1} x_r\varepsilon_r\right\|,\end{aligned}$$

where $c_{3T,j} \equiv \mu_{\min}\left(\frac{1}{\hat{T}_{j+1}-T_j^0}\sum_{r=T_j^0}^{\hat{T}_{j+1}-1} x_r x_r'\right) \geq \underline{c}_{xx}/2 > 0$ with probability going to 1. Thus,

$$\left\|\hat{\alpha}_{j+1}-\alpha_{j+1}^0\right\| \leq (c_{3T,j})^{-1}\left[\frac{\lambda\sqrt{p}\xi_T}{\hat{T}_{j+1}-T_j^0}+\left\|\frac{1}{\hat{T}_{j+1}-T_j^0}\sum_{r=T_j^0}^{\hat{T}_{j+1}-1} x_r\varepsilon_r\right\|\right]. \tag{B.2.4}$$

Define

$$\begin{aligned}E_{T,j} \equiv \Bigg\{&\left\|\alpha_{j+1}^0-\alpha_j^0\right\| \leq \lambda\sqrt{p}\xi_T\left(\frac{1}{T_j^0-\hat{T}_j}(c_{1T,j})^{-1}+\frac{1}{\hat{T}_{j+1}-T_j^0}(c_{3T,j})^{-1}\right) \\ &(c_{1T,j})^{-1}\left\|\frac{1}{T_j^0-\hat{T}_j}\sum_{r=\hat{T}_j}^{T_j^0-1} x_r\varepsilon_r\right\|+(c_{3T,j})^{-1}\left\|\frac{1}{\hat{T}_{j+1}-T_j^0}\sum_{r=T_j^0}^{\hat{T}_{j+1}-1} x_r\varepsilon_r\right\|\Bigg\}.\end{aligned} \tag{B.2.5}$$

Applying the triangle inequality on (B.2.3) and (B.2.4), yields that $E_{T,j}$ has probability one. Then,

$$\begin{aligned}&\sum_{j=1}^{m^0} P\left(A_{T,j}^- \cap \left\{\hat{T}_{j+1}-T_j^0 \geq \frac{1}{2}I_{\min}\right\} \cap D_T^{(m)}\right) \\ &\quad= \sum_{j=1}^{m^0} P\left(E_{T,j} \cap A_{T,j}^- \cap \left\{\hat{T}_{j+1}-T_j^0 \geq \frac{1}{2}I_{\min}\right\} \cap D_T^{(m)}\right) \\ &\quad\leq \sum_{j=1}^{m^0} P\left(E_{T,j} \cap \left\{T_j^0-\hat{T}_j > T\delta_T\right\} \cap \left\{\hat{T}_{j+1}-T_j^0 \geq \frac{1}{2}I_{\min}\right\}\right) \\ &\quad\leq \sum_{j=1}^{m^0} P\left(\frac{\lambda\sqrt{p}\xi_T}{T\delta_T}(c_{1T,j})^{-1}+\frac{2\lambda\sqrt{p}\xi_T}{I_{\min}}(c_{3T,j})^{-1} \geq \left\|\alpha_{j+1}^0-\alpha_j^0\right\|/3\right) \\ &\qquad+ P\left(\left\{(c_{1T,j})^{-1}\left\|\frac{1}{T_j^0-\hat{T}_j}\sum_{r=\hat{T}_j}^{T_j^0-1} x_r\varepsilon_r\right\| \geq \left\|\alpha_{j+1}^0-\alpha_j^0\right\|/3\right\} \cap \left\{T_j^0-\hat{T}_j > T\delta_T\right\}\right) \\ &\qquad+ P\left(\left\{(c_{3T,j})^{-1}\left\|\frac{1}{\hat{T}_{j+1}-T_j^0}\sum_{r=T_j^0}^{\hat{T}_{j+1}-1} x_r\varepsilon_r\right\|\right.\right.\end{aligned}$$

$$\geq \left\|\alpha^0_{j+1} - \alpha^0_j\right\|/3\Big\} \cap \Big\{\hat{T}_{j+1} - T^0_j \geq \frac{1}{2} I_{\min}\Big\}\Big) \tag{B.2.6}$$

The first term in (B.2.6) goes asymptotically to zero as $\frac{\lambda\sqrt{p}\xi_T}{J_{\min} T\delta_T} \to 0$ and $\frac{T\delta_T \lambda\sqrt{p}\xi_T}{I_{\min} J_{\min} T\delta_T} \to 0$ by Assumptions 6 and 8. The middle term converges to zero by applying Assumptions 4 and 7. The last term is also converging to zero by imposing Assumptions 4, 6 and 7. The proof is similar to the proof for the second term in (B.2.2), i.e. $\sum_{j=1}^{m^0} P(A^-_{T,j} \cap D^{(m)}_T) \to 0$.

Now, $\sum_{j=1}^{m^0} P(A^-_{T,j} \cap D^{(l)}_T)$ will be analyzed.

$$P(A^-_{T,j} \cap D^{(l)}_T) \leq P(D^{(l)}_T) \leq \sum_{r=1}^{m^0} 2^{r-1} P\left(\max\left\{l \in \left\{1,\ldots,m^0\right\} : \hat{T}_l \leq T^0_{l-1}\right\} = r\right),$$

where the probability of an event $P(D^{(l)}_T)$ is bounded by the sum of probabilities of all the individual events as follows. If the largest index which satisfies $\hat{T}_l \leq T^0_{l-1}$ is r, then the probability of this event bounds all the events in which any s-tuple of indices from $\{1,\ldots,r\}$ where $s \leq r$ satisfies the inequality $\hat{T}_l \leq T^0_{l-1}$. In total, there are 2^{r-1} of such s-tuples. The event $\max\{l \in \{1,\ldots,m^0\} : \hat{T}_l \leq T^0_{l-1}\} = r$ implies $\hat{T}_r \leq T^0_{r-1}$ and $\hat{T}_{l+1} \leq T^0_l$ for $l = r,\ldots,m^0$, and $\max\left\{l \in \{1,\ldots,m^0\} : \hat{T}_l \leq T^0_{l-1}\right\} = r \subset \cup_{k=r}^{m^0-1}(\left\{T^0_k - \hat{T}_k \geq I_{\min}/2\right\} \cap \{\hat{T}^0_{k+1} - T^0_k \geq I_{\min}/2\})$ by the geometry of the restrictions. Then,

$$\begin{aligned}
&\sum_{j=1}^{m^0} P\left(A^-_{T,j} \cap D^{(l)}_T\right) \\
&\quad \leq m^0 \sum_{r=1}^{m^0-1} 2^{r-1} \sum_{k=r}^{m^0-1} P\left(\left\{T^0_k - \hat{T}_k \geq I_{\min}/2\right\} \cap \left\{\hat{T}^0_{k+1} - T^0_k \geq I_{\min}/2\right\}\right) \\
&\qquad + m^0 2^{m^0-1} P\left(T^0_{m^0} - \hat{T}_{m^0} \geq I_{\min}/2\right)
\end{aligned} \tag{B.2.7}$$

Using (B.2.5) for $j = m^0$ says that probability of the event of E_{T,m^0} is 1. Then,

$$\begin{aligned}
&m^0 2^{m^0-1} P\left(T^0_{m^0} - \hat{T}_{m^0} \geq I_{\min}/2\right) \\
&\quad = m^0 2^{m^0-1} P\left(E_{T,m^0} \cap T^0_{m^0} - \hat{T}_{m^0} \geq I_{\min}/2\right) \\
&\quad \leq m^0 2^{m^0-1} P\left(\frac{\lambda\sqrt{p}\xi_T}{T\delta_T}(c_{1T,m^0})^{-1} + \frac{2\lambda\sqrt{p}\xi_T}{I_{\min}}(c_{3T,m^0})^{-1} \geq \left\|\alpha^0_{m^0+1} - \alpha^0_{m^0}\right\|/3\right) \\
&\qquad + m^0 2^{m^0-1} P\left(\left\{(c_{1T,m^0})^{-1}\left\|\frac{1}{T^0_{m^0} - \hat{T}_{m^0}} \sum_{r=\hat{T}_{m^0}}^{T^0_{m^0}-1} x_r \varepsilon_r\right\|\right. \right. \\
&\qquad\qquad \left.\left. \geq \left\|\alpha^0_{m^0+1} - \alpha^0_{m^0}\right\|/3\right\}, T^0_{m^0} - \hat{T}_{m^0} \geq I_{\min}/2\right)
\end{aligned}$$

$$+ m^0 2^{m^0-1} P\left(\left\{(c_{3T,m^0})^{-1}\left\|\frac{1}{T-T^0_{m^0}}\sum_{r=T^0_{m^0}}^{T} x_r\varepsilon_r\right\| \geq \left\|\alpha^0_{m^0+1}-\alpha^0_{m^0}\right\|/3\right\}\right)$$
$$\to 0,$$

in a similar fashion as for (B.2.6) using $m^0 2^{m^0-1} = O(T\log T)$ and the concentration inequality in Lemma B.1.2. The rate for the number of breaks comes from this concentration result as the $\log(T\log T)$ can be squeezed to the exponential without inflating the tail probability.

The first term in (B.2.7) can be bounded by using (B.2.5) for $j = k$, the Assumption 1, and the concentration inequality in Lemma B.1.2, i.e.

$$\begin{aligned}
&m^0\sum_{r=1}^{m^0-1} 2^{r-1}\sum_{k=r}^{m^0-1} P\left(\left\{T^0_k-\hat{T}_k\geq I_{\min}/2\right\}\cap\left\{\hat{T}^0_{k+1}-T^0_k\geq I_{\min}/2\right\}\right)\\
&\quad\leq m^0 2^{m^0-1}\sum_{k=1}^{m^0-1} P\left(E_{T,k}\cap\left\{T^0_k-\hat{T}_k\geq I_{\min}/2\right\}\cap\left\{\hat{T}^0_{k+1}-T^0_k\geq I_{\min}/2\right\}\right)\\
&\quad\leq m^0 2^{m^0-1}\sum_{k=1}^{m^0-1} P\left(\frac{\lambda\sqrt{p}\xi_T}{T\delta_T}(c_{1T,k})^{-1}+\frac{2\lambda\sqrt{p}\xi_T}{I_{\min}}(c_{3T,k})^{-1}\geq\left\|\alpha^0_{k+1}-\alpha^0_k\right\|/3\right)\\
&\qquad+ m^0 2^{m^0-1}\sum_{k=1}^{m^0-1} P\left(\left\{(c_{1T,k})^{-1}\left\|\frac{1}{T^0_k-\hat{T}_k}\sum_{s=\hat{T}_k}^{T^0_k-1} x_s\varepsilon_s\right\|\right.\right.\\
&\qquad\qquad\left.\left.\geq\left\|\alpha^0_{k+1}-\alpha^0_k\right\|/3\right\}\cap\left\{T^0_k-\hat{T}_k>I_{\min}/2\right\}\right)\\
&\qquad+ m^0 2^{m^0-1}\sum_{k=1}^{m^0-1} P\left(\left\{(c_{3T,k})^{-1}\left\|\frac{1}{\hat{T}_{k+1}-T^0_k}\sum_{s=T^0_k}^{\hat{T}_{k+1}-1} x_s\varepsilon_s\right\|\right.\right.\\
&\qquad\qquad\left.\left.\geq\left\|\alpha^0_{k+1}-\alpha^0_k\right\|/3\right\}\cap\left\{\hat{T}_{k+1}-T^0_k\geq\frac{1}{2}I_{\min}\right\}\right)\to 0.
\end{aligned}$$

This yields the final result $\sum_{j=1}^{m^0} P(A^-_{T,j}\cap D^{(l)}_T)\to 0$. The last part $\sum_{j=1}^{m^0} P(A^-_{T,j}\cap D^{(r)}_T)\to 0$ follows analogously. □

B.3 Tables and Figures

B.3.1 One Parameter - No Break - Average Squared Bias

Table B.3.1: Average Squared Bias, No Break, 1000 draws, SNR=2, $\delta = 0.025$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.044	0.045	0.092	0.044	0.047	0.137	0.069	0.089	0.107
100	0.020	0.020	0.046	0.020	0.020	0.043	0.023	0.025	0.039
200	0.011	0.011	0.030	0.011	0.011	0.014	0.011	0.011	0.020

Table B.3.2: Average Squared Bias, No Break, 1000 draws, SNR=2, $\delta = 0.050$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.044	0.045	0.090	0.044	0.047	0.142	0.071	0.091	0.107
100	0.020	0.020	0.041	0.020	0.020	0.042	0.024	0.025	0.039
200	0.011	0.011	0.025	0.011	0.011	0.014	0.011	0.011	0.020

Table B.3.3: Average Squared Bias, No Break, 1000 draws, SNR=2, $\delta = 0.075$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.044	0.045	0.089	0.044	0.047	0.140	0.075	0.096	0.107
100	0.020	0.020	0.040	0.020	0.020	0.044	0.024	0.025	0.039
200	0.011	0.011	0.024	0.011	0.011	0.014	0.011	0.011	0.020

Table B.3.4: Best δ for the Bias, No Break, 1000 draws, SNR=2

	1 step	2 step		
T	IC_1	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.025	0.025	0.025
100	0.025	0.025	0.025	0.025
200	0.025	0.025	0.025	0.025

Table B.3.5: Average Squared Bias, No Break, 1000 draws, SNR=1, $\delta = 0.025$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.104	0.105	0.176	0.103	0.107	0.321	0.196	0.239	0.205
100	0.040	0.040	0.087	0.041	0.041	0.106	0.055	0.062	0.082
200	0.020	0.020	0.053	0.020	0.020	0.034	0.022	0.022	0.037

Table B.3.6: Average Squared Bias, No Break, 1000 draws, SNR=1, $\delta = 0.050$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.104	0.105	0.171	0.103	0.107	0.310	0.205	0.241	0.205
100	0.040	0.040	0.079	0.041	0.041	0.102	0.057	0.062	0.082
200	0.020	0.020	0.044	0.020	0.020	0.034	0.023	0.023	0.037

Table B.3.7: Average Squared Bias, No Break, 1000 draws, SNR=1, $\delta = 0.075$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.104	0.105	0.170	0.103	0.107	0.309	0.216	0.247	0.205
100	0.040	0.040	0.076	0.041	0.041	0.104	0.058	0.065	0.082
200	0.020	0.020	0.042	0.020	0.020	0.035	0.023	0.023	0.037

Table B.3.8: Best δ for the Bias, No Break, 1000 draws, SNR=1

	1 step	2 step		
T	IC_1	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.075	0.075	0.025	0.025
100	0.075	0.050	0.025	0.025
200	0.075	0.025	0.025	0.025

B.3.2 One Parameter - No Break - Rates of Falsely Detected Breaks

Table B.3.9: Rates of False Positives (FP), 1000 draws, SNR=2, $\delta = 0.025$

	2 step					Bai Perron	
T	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	DM	Seq
50	0.8	1.0	37.8	19.8	22.5	53.2	22.4
100	0.0	0.0	25.0	11.7	12.0	32.6	13.6
200	0.2	0.2	10.7	5.5	5.5	18.3	11.4

Table B.3.10: Rates of False Positives (FP), 1000 draws, SNR=2, $\delta = 0.050$

	2 step					Bai Perron	
T	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	DM	Seq
50	0.8	1.0	38.7	20.1	22.6	53.2	22.4
100	0.0	0.0	24.9	10.7	10.9	32.6	13.6
200	0.2	0.2	10.6	5.1	5.1	18.3	11.4

Table B.3.11: Rates of False Positives (FP), 1000 draws, SNR=2, $\delta = 0.075$

	2 step					Bai Perron	
T	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	DM	Seq
50	0.8	1.0	38.3	19.5	21.7	53.2	22.4
100	0.0	0.0	26.0	10.3	10.8	32.6	13.6
200	0.2	0.2	10.8	4.6	4.6	18.3	11.4

Table B.3.12: Best δ for the FP, No Break, 1000 draws, SNR=2

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.075	0.075
100	0.050	0.075	0.075
200	0.050	0.075	0.075

Table B.3.13: Rates of False Positives (FP), 1000 draws, SNR=1, $\delta = 0.025$

	2 step					Bai Perron	
T	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	DM	Seq
50	4.2	4.5	48.4	40.0	40.6	46.7	21.4
100	0.4	0.4	33.5	23.1	23.3	30.5	12.6
200	0.0	0.0	19.1	12.0	12.0	16.6	9.9

Table B.3.14: Rates of False Positives (FP), 1000 draws, SNR=1, $\delta = 0.050$

	2 step					Bai Perron	
T	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	DM	Seq
50	4.2	4.5	46.1	37.0	37.6	46.7	21.4
100	0.4	0.4	31.0	20.0	20.0	30.5	12.6
200	0.0	0.0	17.7	10.1	10.1	16.6	9.9

Table B.3.15: Rates of False Positives (FP), 1000 draws, SNR=1, $\delta = 0.075$

	2 step					Bai Perron	
T	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	DM	Seq
50	4.2	4.5	46.0	36.8	37.3	46.7	21.4
100	0.4	0.4	30.9	18.9	19.1	30.5	12.6
200	0.0	0.0	17.2	9.4	9.4	16.6	9.9

Table B.3.16: Best δ for the FP, No Break, 1000 draws, SNR=1

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.075	0.075	0.075
100	0.075	0.075	0.075
200	0.075	0.075	0.075

B.3.3 One Parameter - One Break - Middle - Average Squared Bias

Table B.3.17: Average Squared Bias, Break - Middle, 1000 draws, SNR=2, $\delta = 0.025$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.709	0.792	0.479	0.699	0.756	0.566	0.630	0.606	0.797
100	0.412	0.421	0.275	0.445	0.417	0.281	0.375	0.335	0.356
200	0.235	0.235	0.151	0.266	0.252	0.145	0.200	0.191	0.147

Table B.3.18: Average Squared Bias, Break - Middle, 1000 draws, SNR=2, $\delta = 0.050$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.709	0.792	0.501	0.699	0.756	0.566	0.629	0.605	0.797
100	0.412	0.421	0.304	0.445	0.417	0.275	0.358	0.321	0.356
200	0.235	0.235	0.171	0.266	0.252	0.140	0.178	0.171	0.147

Table B.3.19: Average Squared Bias, Break - Middle, 1000 draws, SNR=2, $\delta = 0.075$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.709	0.792	0.539	0.699	0.756	0.630	0.655	0.659	0.797
100	0.412	0.421	0.315	0.445	0.417	0.271	0.344	0.313	0.356
200	0.235	0.235	0.185	0.266	0.252	0.136	0.165	0.155	0.147

Table B.3.20: Best δ for the Bias, Break - Middle, 1000 draws, SNR=2

	1 step	2 step		
T	IC_1	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.050	0.050	0.050
100	0.025	0.075	0.075	0.075
200	0.025	0.075	0.075	0.075

Table B.3.21: Average Squared Bias, Break - Middle, 1000 draws, SNR=1, $\delta = 0.025$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	1.005	1.464	0.699	1.043	1.420	0.975	0.962	0.964	1.260
100	0.624	0.642	0.414	0.653	0.647	0.495	0.578	0.559	0.708
200	0.403	0.403	0.247	0.445	0.430	0.290	0.399	0.371	0.319

Table B.3.22: Average Squared Bias, Break - Middle, 1000 draws, SNR=1, $\delta = 0.050$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	1.005	1.464	0.762	1.043	1.420	1.050	1.005	1.049	1.260
100	0.624	0.642	0.439	0.653	0.647	0.485	0.562	0.544	0.708
200	0.403	0.403	0.270	0.445	0.430	0.274	0.366	0.342	0.319

Table B.3.23: Average Squared Bias, Break - Middle, 1000 draws, SNR=1, $\delta = 0.075$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	1.005	1.464	0.869	1.043	1.420	1.268	1.223	1.277	1.260
100	0.624	0.642	0.450	0.653	0.647	0.479	0.554	0.538	0.708
200	0.403	0.403	0.285	0.445	0.430	0.273	0.356	0.334	0.319

Table B.3.24: Best δ for the Bias, Break - Middle, 1000 draws, SNR=1

	1 step	2 step		
T	IC_1	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.025	0.025	0.025
100	0.025	0.075	0.075	0.075
200	0.025	0.075	0.075	0.075

B.3.4 One Parameter - One Break - Middle - Rates of Detected Breaks

Table B.3.25: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=2, $\delta = 0.025$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	23.8	22.3	4.1	5.3	5.3	25.8
	1	17.0	15.7	41.9	45.2	44.0	63.2
	2	26.7	24.0	38.0	36.2	37.0	10.6
	3	17.4	17.2	12.3	10.2	10.3	0.4
	4	6.6	7.5	2.7	2.5	2.6	0.0
	5	4.0	5.0	0.7	0.4	0.6	0.0
	6+	4.5	8.3	0.3	0.2	0.2	0.0
	Q	0.34	0.33	0.13	0.14	0.14	0.33
100	0	12.9	12.2	1.3	1.8	1.8	4.8
	1	19.5	18.2	32.1	35.7	34.7	85.8
	2	28.3	27.3	40.1	42.0	41.6	9.3
	3	18.2	18.6	19.1	15.9	16.6	0.1
	4	10.8	11.7	6.0	3.8	4.3	0.0
	5	4.0	4.2	0.8	0.4	0.6	0.0
	6+	6.3	7.8	0.6	0.4	0.4	0.0
	Q	0.22	0.21	0.07	0.08	0.08	0.11
200	0	6.4	6.1	0.0	0.1	0.1	0.0
	1	16.4	15.8	19.7	23.2	22.9	94.7
	2	28.8	28.3	36.8	39.5	39.2	5.3
	3	23.2	23.5	25.4	24.9	24.7	0.0
	4	12.3	12.0	12.2	9.7	9.9	0.0
	5	5.2	5.9	4.5	2.2	2.7	0.0
	6+	7.7	8.4	1.4	0.4	0.5	0.0
	Q	0.14	0.14	0.05	0.05	0.05	0.03

Q ... Average Relative Distance from the True Break.

Table B.3.26: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=2, $\delta = 0.050$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	23.8	22.3	4.3	6.7	6.4	25.8
	1	17.0	15.7	48.3	50.3	49.4	63.2
	2	26.7	24.0	34.8	32.1	32.7	10.6
	3	17.4	17.2	9.8	8.6	9.1	0.4
	4	6.6	7.5	2.0	1.7	1.8	0.0
	5	4.0	5.0	0.5	0.4	0.4	0.0
	6+	4.5	8.3	0.3	0.2	0.2	0.0
	Q	0.34	0.33	0.12	0.15	0.14	0.33
100	0	12.9	12.2	1.5	2.9	2.8	4.8
	1	19.5	18.2	40.3	42.7	42.0	85.8
	2	28.3	27.3	39.8	40.1	40.1	9.3
	3	18.2	18.6	15.0	11.5	12.3	0.1
	4	10.8	11.7	2.7	2.3	2.3	0.0
	5	4.0	4.2	0.5	0.5	0.3	0.0
	6+	6.3	7.8	0.2	0.0	0.2	0.0
	Q	0.22	0.21	0.07	0.08	0.08	0.11
200	0	6.4	6.1	0.0	0.2	0.2	0.0
	1	16.4	15.8	27.0	30.0	29.5	94.7
	2	28.8	28.3	40.9	42.3	42.2	5.3
	3	23.2	23.5	22.0	20.4	20.4	0.0
	4	12.3	12.0	8.4	5.8	6.4	0.0
	5	5.2	5.9	1.0	0.9	0.9	0.0
	6+	7.7	8.4	0.7	0.4	0.4	0.0
	Q	0.14	0.14	0.04	0.04	0.04	0.03

Q ... Average Relative Distance from the True Break.

Table B.3.27: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=2, $\delta = 0.075$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	23.8	22.3	4.5	7.6	7.2	25.8
	1	17.0	15.7	51.1	51.3	50.7	63.2
	2	26.7	24.0	33.3	30.7	31.4	10.6
	3	17.4	17.2	7.9	7.6	7.6	0.4
	4	6.6	7.5	2.0	1.7	2.0	0.0
	5	4.0	5.0	0.5	0.6	0.6	0.0
	6+	4.5	8.3	0.7	0.5	0.5	0.0
	Q	0.34	0.33	0.13	0.16	0.15	0.33
100	0	12.9	12.2	1.3	3.4	3.1	4.8
	1	19.5	18.2	44.6	46.4	46.0	85.8
	2	28.3	27.3	39.9	39.0	39.4	9.3
	3	18.2	18.6	11.9	9.8	10.0	0.1
	4	10.8	11.7	1.9	1.3	1.3	0.0
	5	4.0	4.2	0.4	0.1	0.2	0.0
	6+	6.3	7.8	0.0	0.0	0.0	0.0
	Q	0.22	0.21	0.06	0.08	0.08	0.11
200	0	6.4	6.1	0.0	0.0	0.0	0.0
	1	16.4	15.8	33.8	35.7	35.2	94.7
	2	28.8	28.3	41.2	42.6	42.2	5.3
	3	23.2	23.5	18.4	17.2	17.6	0.0
	4	12.3	12.0	5.8	3.9	4.3	0.0
	5	5.2	5.9	0.6	0.4	0.5	0.0
	6+	7.7	8.4	0.2	0.2	0.2	0.0
	Q	0.14	0.14	0.04	0.03	0.04	0.03

Q ... Average Relative Distance from the True Break.

Table B.3.28: Best δ for the Q, Break - Middle , 1000 draws, SNR=2

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.050	0.025	0.025
100	0.075	0.025	0.025
200	0.075	0.075	0.075

Table B.3.29: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=1, $\delta = 0.025$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	23.4	22.3	4.8	4.9	4.9	46.4
	1	22.4	21.1	50.6	52.3	51.8	47.8
	2	21.1	19.9	34.7	33.9	34.1	5.7
	3	14.4	13.3	7.8	7.1	7.3	0.1
	4	7.1	6.6	1.6	1.3	1.4	0.0
	5	4.7	4.6	0.3	0.3	0.3	0.0
	6+	6.9	12.2	0.2	0.2	0.2	0.0
	Q	0.34	0.34	0.15	0.15	0.15	0.53
100	0	18.4	17.5	1.8	1.9	1.9	24.0
	1	19.8	19.8	37.1	41.4	40.5	70.5
	2	23.6	23.5	38.9	39.0	39.1	5.5
	3	16.9	16.5	16.5	14.6	15.2	0.0
	4	8.9	8.0	4.9	2.7	2.8	0.0
	5	4.2	4.8	0.6	0.4	0.5	0.0
	6+	8.2	9.9	0.2	0.0	0.0	0.0
	Q	0.28	0.28	0.10	0.10	0.10	0.31
200	0	11.9	11.7	0.6	0.6	0.6	2.5
	1	14.8	14.4	21.7	25.2	24.9	93.7
	2	29.1	29.1	39.3	44.0	42.9	3.8
	3	18.5	18.2	25.2	21.6	22.4	0.0
	4	10.6	11.0	9.4	6.8	7.1	0.0
	5	6.4	6.4	2.9	1.3	1.6	0.0
	6+	8.7	9.2	0.9	0.5	0.5	0.0
	Q	0.21	0.21	0.08	0.07	0.07	0.08

Q ... Average Relative Distance from the True Break.

Table B.3.30: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=1, $\delta = 0.050$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	23.4	22.3	6.1	7.2	7.2	46.4
	1	22.4	21.1	57.4	58.6	58.0	47.8
	2	21.1	19.9	29.7	28.2	28.6	5.7
	3	14.4	13.3	5.9	5.2	5.3	0.1
	4	7.1	6.6	0.3	0.2	0.3	0.0
	5	4.7	4.6	0.3	0.3	0.3	0.0
	6+	6.9	12.2	0.3	0.3	0.3	0.0
	Q	0.34	0.34	0.16	0.17	0.17	0.53
100	0	18.4	17.5	2.3	2.9	2.9	24.0
	1	19.8	19.8	46.6	49.5	48.9	70.5
	2	23.6	23.5	35.9	35.2	35.1	5.5
	3	16.9	16.5	12.6	10.7	11.0	0.0
	4	8.9	8.0	2.0	1.4	1.8	0.0
	5	4.2	4.8	0.6	0.3	0.3	0.0
	6+	8.2	9.9	0.0	0.0	0.0	0.0
	Q	0.28	0.28	0.10	0.10	0.10	0.31
200	0	11.9	11.7	0.7	1.0	1.0	2.5
	1	14.8	14.4	29.6	33.7	33.3	93.7
	2	29.1	29.1	43.6	44.0	43.4	3.8
	3	18.5	18.2	19.1	15.9	16.8	0.0
	4	10.6	11.0	5.6	4.4	4.4	0.0
	5	6.4	6.4	1.3	0.9	1.0	0.0
	6+	8.7	9.2	0.1	0.1	0.1	0.0
	Q	0.21	0.21	0.06	0.07	0.07	0.08

Q ... Average Relative Distance from the True Break.

Table B.3.31: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=1, $\delta = 0.075$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	23.4	22.3	6.6	7.9	7.9	46.4
	1	22.4	21.1	60.0	60.4	60.0	47.8
	2	21.1	19.9	26.9	25.5	25.8	5.7
	3	14.4	13.3	4.7	4.4	4.5	0.1
	4	7.1	6.6	1.1	1.1	1.1	0.0
	5	4.7	4.6	0.1	0.1	0.1	0.0
	6+	6.9	12.2	0.6	0.6	0.6	0.0
	Q	0.34	0.34	0.17	0.18	0.18	0.53
100	0	18.4	17.5	2.4	3.0	3.0	24.0
	1	19.8	19.8	53.6	57.2	56.1	70.5
	2	23.6	23.5	32.7	30.6	31.0	5.5
	3	16.9	16.5	10.1	8.5	9.1	0.0
	4	8.9	8.0	1.0	0.6	0.7	0.0
	5	4.2	4.8	0.2	0.1	0.1	0.0
	6+	8.2	9.9	0.0	0.0	0.0	0.0
	Q	0.28	0.28	0.10	0.10	0.10	0.31
200	0	11.9	11.7	1.1	1.5	1.5	2.5
	1	14.8	14.4	36.0	39.9	39.1	93.7
	2	29.1	29.1	43.3	41.8	42.3	3.8
	3	18.5	18.2	16.3	13.9	14.3	0.0
	4	10.6	11.0	2.9	2.6	2.5	0.0
	5	6.4	6.4	0.4	0.3	0.3	0.0
	6+	8.7	9.2	0.0	0.0	0.0	0.0
	Q	0.21	0.21	0.06	0.07	0.07	0.08

Q ... Average Relative Distance from the True Break.

Table B.3.32: Best δ for the Q, Break - Middle, 1000 draws, SNR=1

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.025	0.025
100	0.075	0.025	0.025
200	0.075	0.050	0.050

B.3.5 One Parameter - One Break - Middle - All

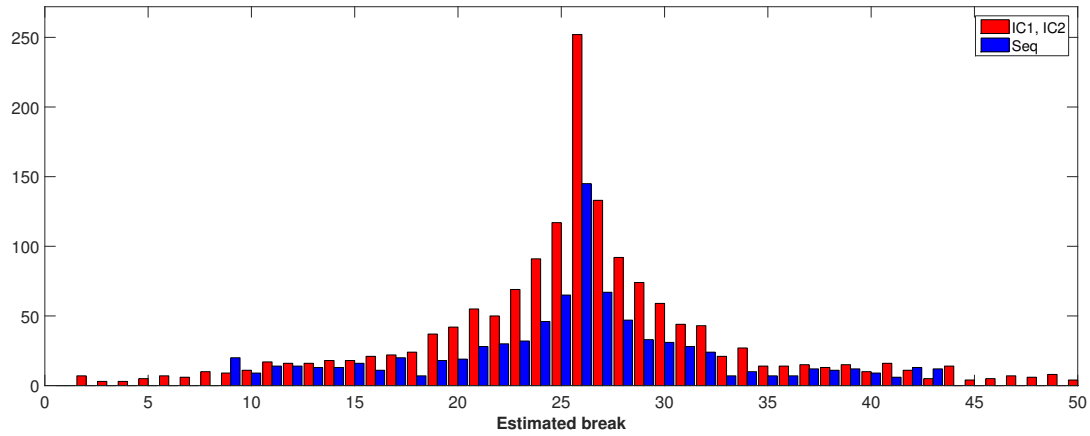


Figure B.3.1: Positions of the Estimated Breaks, $T = 50$, SNR $= 2$, $\delta = 0.075$, Break in the Middle

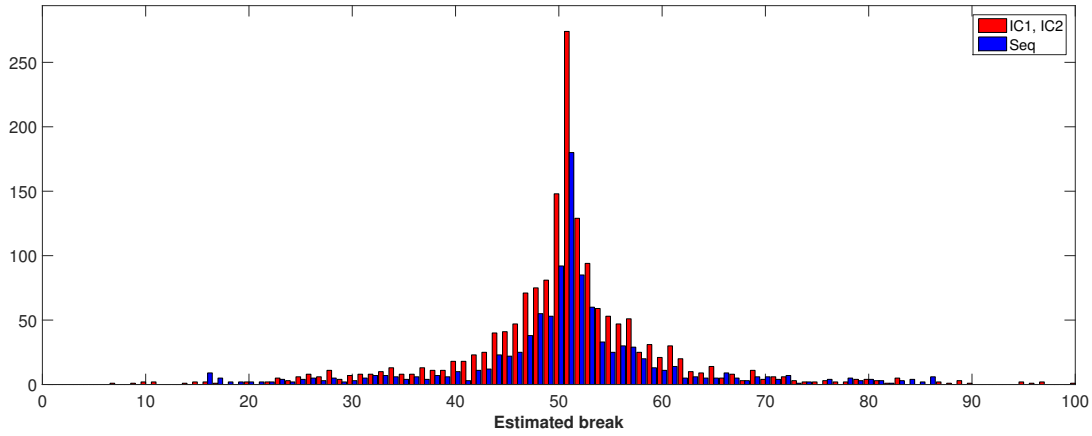


Figure B.3.2: Positions of the Estimated Breaks, $T = 100$, SNR $= 2$, $\delta = 0.075$, Break in the Middle

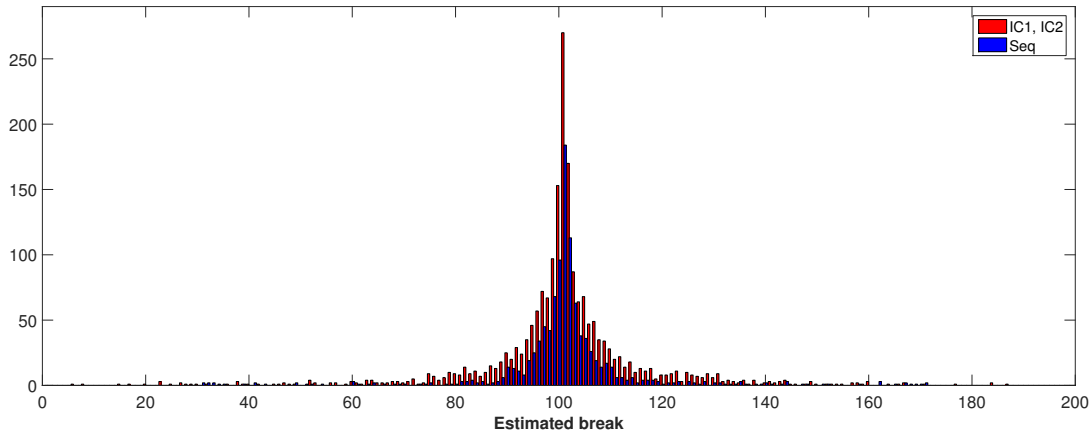


Figure B.3.3: Positions of the Estimated Breaks, $T = 200$, SNR $= 2$, $\delta = 0.075$, Break in the Middle

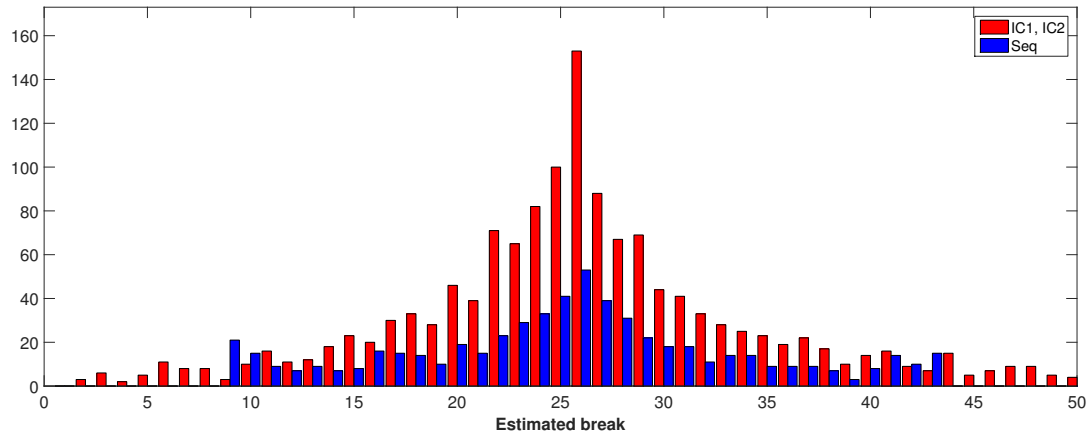


Figure B.3.4: Positions of the Estimated Breaks, $T = 50$, SNR $= 1$, $\delta = 0.075$, Break in the Middle

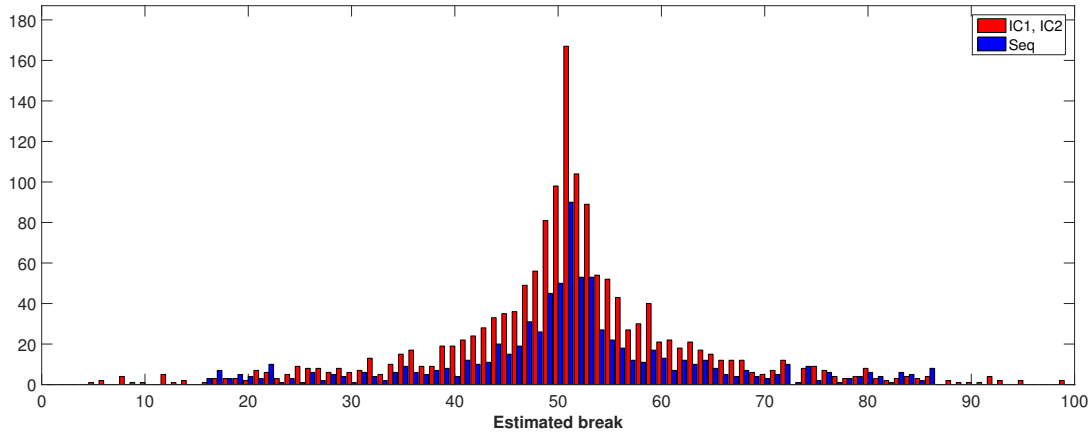


Figure B.3.5: Positions of the Estimated Breaks, $T = 100$, SNR $= 1$, $\delta = 0.075$, Break in the Middle

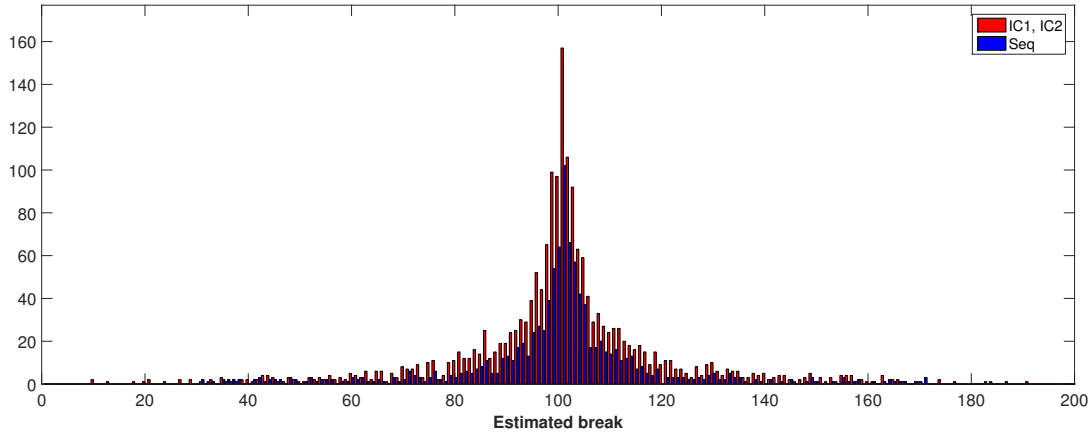


Figure B.3.6: Positions of the Estimated Breaks, $T = 200$, SNR $= 1$, $\delta = 0.075$, Break in the Middle

B.3.6 One Parameter - One Break - Middle - One Found

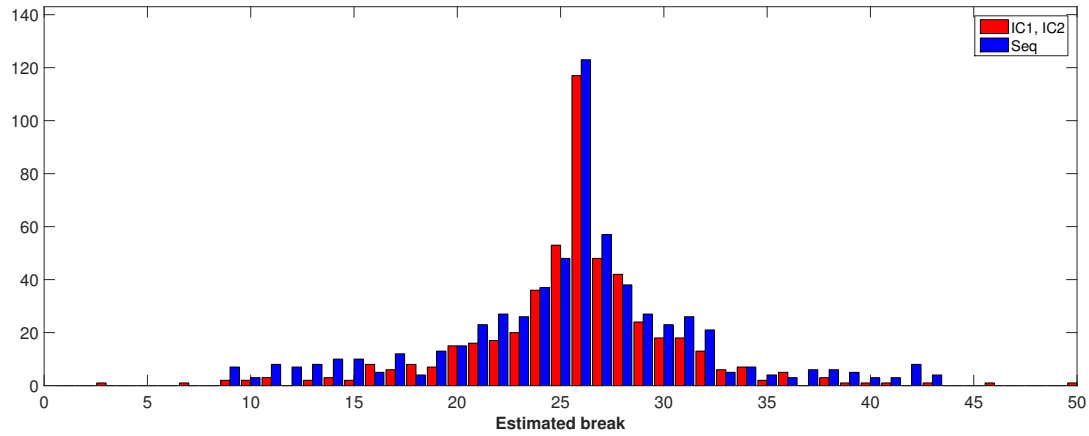


Figure B.3.7: Positions of the Estimated Breaks given One Break Found, $T = 50$, SNR $= 2$, $\delta = 0.075$, Break in the Middle

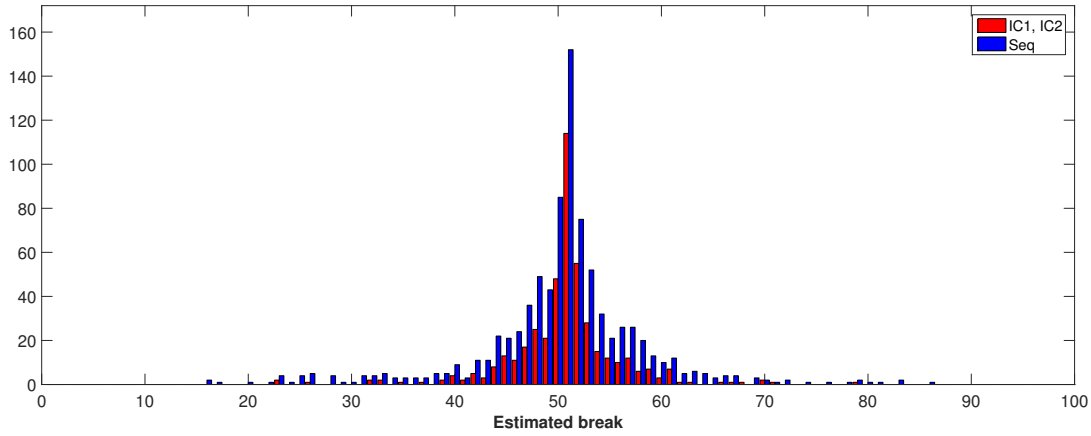


Figure B.3.8: Positions of the Estimated Breaks given One Break Found, $T = 100$, SNR $= 2$, $\delta = 0.075$, Break in the Middle

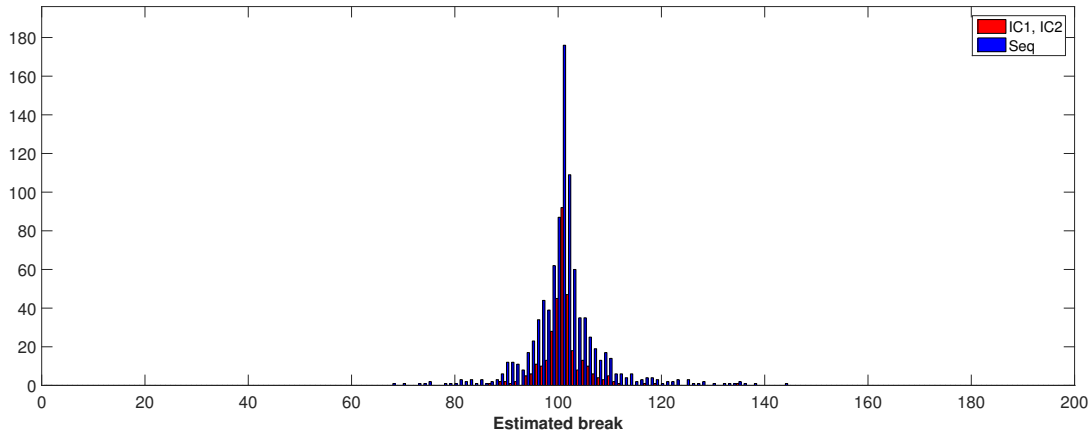


Figure B.3.9: Positions of the Estimated Breaks given One Break Found, $T = 200$, SNR $= 2$, $\delta = 0.075$, Break in the Middle

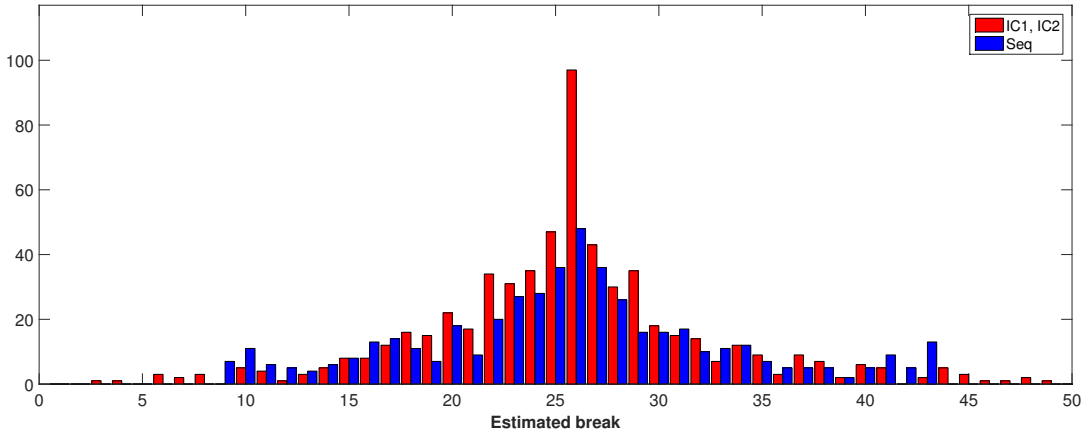


Figure B.3.10: Positions of the Estimated Breaks given One Break Found, $T = 50$, SNR $= 1$, $\delta = 0.075$, Break in the Middle

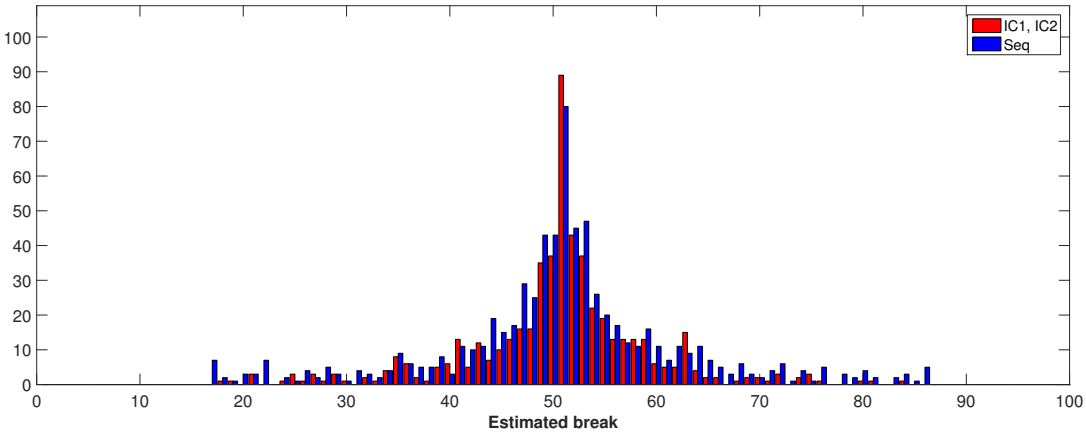


Figure B.3.11: Positions of the Estimated Breaks given One Break Found, $T = 100$, SNR $= 1$, $\delta = 0.075$, Break in the Middle

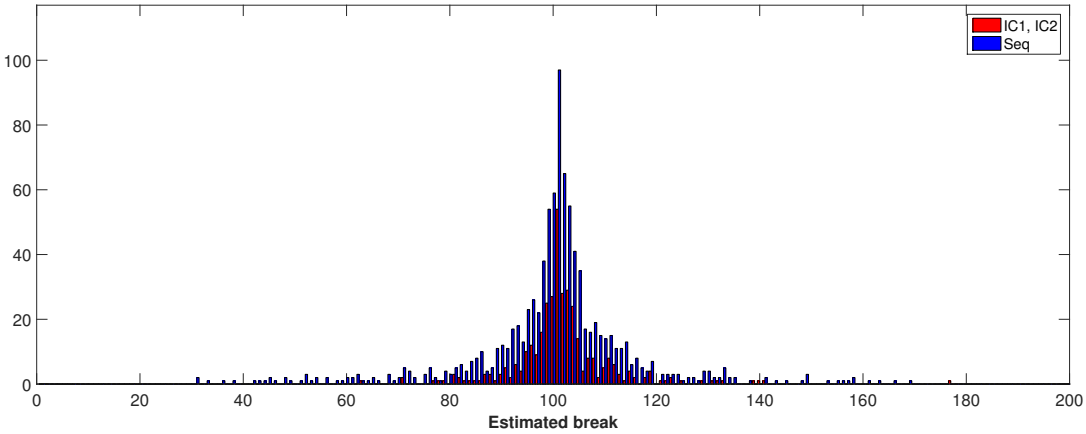


Figure B.3.12: Positions of the Estimated Breaks given One Break Found, $T = 200$, SNR $= 1$, $\delta = 0.075$, Break in the Middle

B.3.7 One Parameter - One Break - End - Average Squared Bias

Table B.3.33: Average Squared Bias, Break - End, 1000 draws, SNR=2, $\delta = 0.025$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.573	0.623	0.363	0.576	0.573	0.435	0.488	0.462	0.493
100	0.369	0.369	0.199	0.371	0.368	0.224	0.281	0.260	0.359
200	0.202	0.202	0.111	0.202	0.202	0.127	0.165	0.160	0.208

Table B.3.34: Average Squared Bias, Break - End, 1000 draws, SNR=2, $\delta = 0.050$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.573	0.623	0.380	0.576	0.573	0.438	0.494	0.473	0.493
100	0.369	0.369	0.210	0.371	0.368	0.228	0.283	0.265	0.359
200	0.202	0.202	0.116	0.202	0.202	0.127	0.165	0.160	0.208

Table B.3.35: Average Squared Bias, Break - End, 1000 draws, SNR=2, $\delta = 0.075$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.573	0.623	0.392	0.576	0.573	0.455	0.507	0.488	0.493
100	0.369	0.369	0.214	0.371	0.368	0.229	0.281	0.263	0.359
200	0.202	0.202	0.119	0.202	0.202	0.129	0.165	0.159	0.208

Table B.3.36: Best δ for the Bias, Break - End, 1000 draws, SNR=2

	1 step	2 step		
T	IC_1	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.025	0.025	0.025
100	0.025	0.025	0.075	0.025
200	0.025	0.025	0.075	0.075

Table B.3.37: Average Squared Bias, Break - End, 1000 draws, SNR=1, $\delta = 0.025$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.764	0.940	0.552	0.773	0.954	0.783	0.763	0.775	0.813
100	0.411	0.411	0.281	0.404	0.405	0.360	0.370	0.366	0.444
200	0.217	0.217	0.165	0.217	0.217	0.192	0.202	0.202	0.233

Table B.3.38: Average Squared Bias, Break - End, 1000 draws, SNR=1, $\delta = 0.050$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.764	0.940	0.564	0.773	0.954	0.782	0.778	0.788	0.813
100	0.411	0.411	0.289	0.404	0.405	0.365	0.379	0.377	0.444
200	0.217	0.217	0.169	0.217	0.217	0.192	0.202	0.201	0.233

Table B.3.39: Average Squared Bias, Break - End, 1000 draws, SNR=1, $\delta = 0.075$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.764	0.940	0.579	0.773	0.954	0.795	0.802	0.803	0.813
100	0.411	0.411	0.296	0.404	0.405	0.374	0.381	0.384	0.444
200	0.217	0.217	0.170	0.217	0.217	0.194	0.202	0.201	0.233

Table B.3.40: Best δ for the Bias, Break - End, 1000 draws, SNR=1

	1 step	2 step		
T	IC_1	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.050	0.025	0.025
100	0.025	0.025	0.025	0.025
200	0.025	0.025	0.075	0.075

B.3.8 One Parameter - One Break - End - Rates of Detected Breaks

Table B.3.41: Percentage Rates of detecting 0-5 breaks and Distance, Break - End, 1000 draws, SNR=2, $\delta = 0.025$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	53.4	50.7	9.6	17.2	16.2	27.5
	1	11.8	11.4	42.2	43.9	40.6	65.6
	2	16.0	14.0	31.0	27.7	29.2	6.6
	3	9.5	10.3	13.4	9.4	11.5	0.3
	4	4.2	4.6	3.1	1.4	2.0	0.0
	5	2.2	2.9	0.7	0.4	0.5	0.0
	6+	2.9	6.1	0.0	0.0	0.0	0.0
	Q	0.61	0.59	0.21	0.27	0.26	0.34
100	0	89.2	88.7	17.5	36.4	32.7	56.4
	1	1.8	1.6	29.2	29.7	29.4	41.6
	2	4.3	3.8	31.3	25.4	26.0	2.0
	3	2.9	3.2	14.8	7.0	9.5	0.0
	4	1.1	1.1	5.2	1.2	1.9	0.0
	5	0.3	0.6	1.9	0.3	0.5	0.0
	6+	0.4	1.0	0.1	0.0	0.0	0.0
	Q	0.91	0.90	0.29	0.44	0.41	0.63
200	0	100.0	100.0	36.6	61.3	58.6	83.6
	1	0.0	0.0	20.7	20.9	22.0	15.7
	2	0.0	0.0	22.9	13.7	14.4	0.7
	3	0.0	0.0	12.4	3.2	3.7	0.0
	4	0.0	0.0	5.4	0.7	1.1	0.0
	5	0.0	0.0	1.8	0.2	0.2	0.0
	6+	0.0	0.0	0.2	0.0	0.0	0.0
	Q	1.00	1.00	0.45	0.66	0.63	0.88

Q ... Average Relative Distance from the True Break.

Table B.3.42: Percentage Rates of detecting 0-5 breaks and Distance, Break - End, 1000 draws, SNR=2, $\delta = 0.050$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	53.4	50.7	10.3	20.4	18.4	27.5
	1	11.8	11.4	49.2	46.3	45.1	65.6
	2	16.0	14.0	29.6	25.7	27.6	6.6
	3	9.5	10.3	8.6	6.2	7.1	0.3
	4	4.2	4.6	1.8	1.0	1.3	0.0
	5	2.2	2.9	0.5	0.4	0.5	0.0
	6+	2.9	6.1	0.0	0.0	0.0	0.0
	Q	0.61	0.59	0.21	0.29	0.28	0.34
100	0	89.2	88.7	17.7	39.2	34.4	56.4
	1	1.8	1.6	34.7	31.2	32.3	41.6
	2	4.3	3.8	30.2	22.8	23.7	2.0
	3	2.9	3.2	12.3	5.6	7.5	0.0
	4	1.1	1.1	4.2	1.1	1.8	0.0
	5	0.3	0.6	0.7	0.0	0.2	0.0
	6+	0.4	1.0	0.2	0.1	0.1	0.0
	Q	0.91	0.90	0.29	0.46	0.42	0.63
200	0	100.0	100.0	35.4	60.8	57.9	83.6
	1	0.0	0.0	23.2	23.3	24.0	15.7
	2	0.0	0.0	25.1	13.2	14.3	0.7
	3	0.0	0.0	10.6	2.2	2.9	0.0
	4	0.0	0.0	4.5	0.4	0.7	0.0
	5	0.0	0.0	1.1	0.1	0.2	0.0
	6+	0.0	0.0	0.1	0.0	0.0	0.0
	Q	1.00	1.00	0.44	0.65	0.62	0.88

Q ... Average Relative Distance from the True Break.

Table B.3.43: Percentage Rates of detecting 0-5 breaks and Distance, Break - End, 1000 draws, SNR=2, $\delta = 0.075$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	53.4	50.7	10.8	21.4	19.0	27.5
	1	11.8	11.4	50.2	46.0	45.6	65.6
	2	16.0	14.0	28.4	25.1	26.8	6.6
	3	9.5	10.3	8.1	5.5	6.3	0.3
	4	4.2	4.6	1.9	1.4	1.7	0.0
	5	2.2	2.9	0.3	0.4	0.3	0.0
	6+	2.9	6.1	0.3	0.2	0.3	0.0
	Q	0.61	0.59	0.22	0.30	0.28	0.34
100	0	89.2	88.7	16.6	39.2	33.6	56.4
	1	1.8	1.6	39.8	31.5	34.3	41.6
	2	4.3	3.8	29.5	23.1	23.2	2.0
	3	2.9	3.2	11.2	5.1	7.3	0.0
	4	1.1	1.1	2.4	0.9	1.3	0.0
	5	0.3	0.6	0.5	0.2	0.3	0.0
	6+	0.4	1.0	0.0	0.0	0.0	0.0
	Q	0.91	0.90	0.28	0.46	0.41	0.63
200	0	100.0	100.0	33.5	60.6	57.2	83.6
	1	0.0	0.0	27.2	24.8	25.9	15.7
	2	0.0	0.0	24.9	11.4	12.6	0.7
	3	0.0	0.0	10.3	2.7	3.5	0.0
	4	0.0	0.0	3.3	0.4	0.7	0.0
	5	0.0	0.0	0.7	0.1	0.1	0.0
	6+	0.0	0.0	0.1	0.0	0.0	0.0
	Q	1.00	1.00	0.42	0.64	0.61	0.88

Q ... Average Relative Distance from the True Break.

Table B.3.44: Best δ for the Q, Break - End, 1000 draws, SNR=2

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.025	0.025
100	0.075	0.025	0.025
200	0.075	0.075	0.075

Table B.3.45: Percentage Rates of detecting 0-5 breaks and Distance, Break - End, 1000 draws, SNR=1, $\delta = 0.025$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	52.4	50.0	14.3	17.2	17.2	45.6
	1	11.1	11.1	44.5	46.4	45.4	47.0
	2	14.5	14.0	28.6	27.2	27.4	6.9
	3	10.5	9.6	9.5	7.4	7.8	0.4
	4	5.8	5.7	2.5	1.6	2.0	0.1
	5	2.0	2.6	0.3	0.2	0.1	0.0
	6+	3.7	7.0	0.3	0.0	0.1	0.0
	Q	0.62	0.60	0.29	0.31	0.31	0.53
100	0	86.7	85.4	23.6	33.1	32.8	65.8
	1	1.8	2.3	32.5	36.8	35.2	31.9
	2	4.2	4.4	26.8	24.2	24.5	2.2
	3	3.7	3.7	13.4	5.3	6.8	0.1
	4	1.2	1.9	2.9	0.5	0.6	0.0
	5	0.8	0.6	0.8	0.1	0.1	0.0
	6+	1.6	1.7	0.0	0.0	0.0	0.0
	Q	0.90	0.89	0.38	0.44	0.44	0.72
200	0	99.4	99.4	46.8	59.0	58.7	85.3
	1	0.2	0.2	21.5	24.4	24.3	14.1
	2	0.2	0.1	17.7	13.1	13.4	0.6
	3	0.0	0.1	9.9	3.0	3.1	0.0
	4	0.2	0.2	3.1	0.4	0.4	0.0
	5	0.0	0.0	0.8	0.1	0.1	0.0
	6+	0.0	0.0	0.2	0.0	0.0	0.0
	Q	1.00	1.00	0.58	0.67	0.67	0.90

Q ... Average Relative Distance from the True Break.

Table B.3.46: Percentage Rates of detecting 0-5 breaks and Distance, Break - End, 1000 draws, SNR=1, $\delta = 0.050$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	52.4	50.0	16.0	20.8	20.7	45.6
	1	11.1	11.1	49.7	49.1	48.0	47.0
	2	14.5	14.0	24.9	22.8	23.9	6.9
	3	10.5	9.6	7.4	5.8	5.6	0.4
	4	5.8	5.7	1.9	1.4	1.7	0.1
	5	2.0	2.6	0.1	0.1	0.1	0.0
	6+	3.7	7.0	0.0	0.0	0.0	0.0
	Q	0.62	0.60	0.30	0.34	0.34	0.53
100	0	86.7	85.4	24.3	37.5	36.7	65.8
	1	1.8	2.3	38.2	36.2	34.9	31.9
	2	4.2	4.4	26.0	22.1	23.4	2.2
	3	3.7	3.7	9.2	3.6	4.4	0.1
	4	1.2	1.9	1.6	0.4	0.4	0.0
	5	0.8	0.6	0.7	0.2	0.2	0.0
	6+	1.6	1.7	0.0	0.0	0.0	0.0
	Q	0.90	0.89	0.38	0.48	0.47	0.72
200	0	99.4	99.4	47.2	63.0	62.8	85.3
	1	0.2	0.2	25.4	23.4	22.8	14.1
	2	0.2	0.1	18.5	12.4	12.8	0.6
	3	0.0	0.1	7.4	1.2	1.6	0.0
	4	0.2	0.2	1.3	0.0	0.0	0.0
	5	0.0	0.0	0.2	0.0	0.0	0.0
	6+	0.0	0.0	0.0	0.0	0.0	0.0
	Q	1.00	1.00	0.57	0.69	0.69	0.90

Q ... Average Relative Distance from the True Break.

Table B.3.47: Percentage Rates of detecting 0-5 breaks and Distance, Break - End, 1000 draws, SNR=1, $\delta = 0.075$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	52.4	50.0	17.1	21.5	21.4	45.6
	1	11.1	11.1	52.1	51.4	50.2	47.0
	2	14.5	14.0	23.2	20.9	22.0	6.9
	3	10.5	9.6	5.5	4.5	4.6	0.4
	4	5.8	5.7	1.6	1.3	1.3	0.1
	5	2.0	2.6	0.4	0.3	0.4	0.0
	6+	3.7	7.0	0.1	0.1	0.1	0.0
	Q	0.62	0.60	0.31	0.34	0.34	0.53
100	0	86.7	85.4	25.8	39.8	38.9	65.8
	1	1.8	2.3	41.8	37.3	36.9	31.9
	2	4.2	4.4	24.7	19.4	20.2	2.2
	3	3.7	3.7	6.0	3.1	3.4	0.1
	4	1.2	1.9	1.3	0.2	0.3	0.0
	5	0.8	0.6	0.4	0.2	0.3	0.0
	6+	1.6	1.7	0.0	0.0	0.0	0.0
	Q	0.90	0.89	0.39	0.50	0.49	0.72
200	0	99.4	99.4	46.8	64.8	64.4	85.3
	1	0.2	0.2	27.9	23.0	22.9	14.1
	2	0.2	0.1	18.7	11.0	11.4	0.6
	3	0.0	0.1	6.0	1.0	1.1	0.0
	4	0.2	0.2	0.4	0.2	0.2	0.0
	5	0.0	0.0	0.1	0.0	0.0	0.0
	6+	0.0	0.0	0.1	0.0	0.0	0.0
	Q	1.00	1.00	0.57	0.71	0.71	0.90

Q ... Average Relative Distance from the True Break.

Table B.3.48: Best δ for the Q, Break - End, 1000 draws, SNR=1

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.025	0.025	0.025
100	0.050	0.025	0.025
200	0.075	0.025	0.025

B.3.9 One Parameter - One Break - End - All

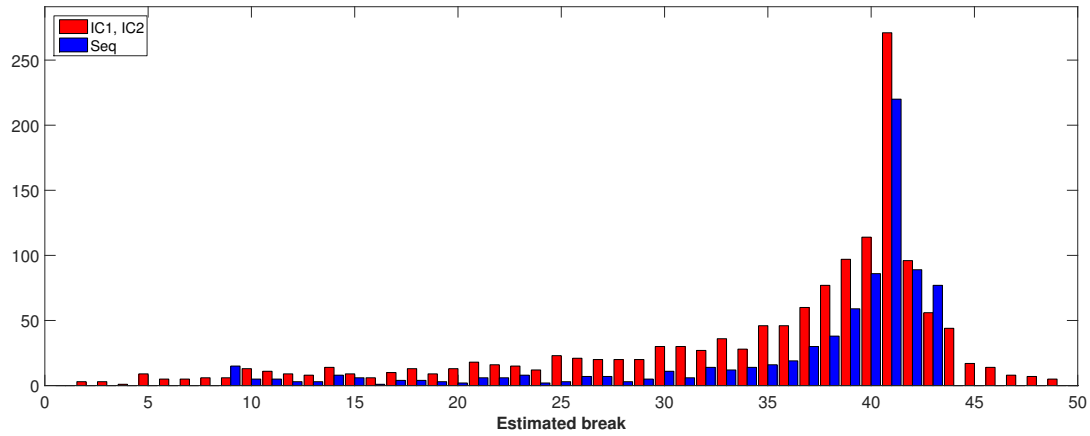


Figure B.3.13: Positions of the Estimated Breaks, $T = 50$, SNR $= 2$, $\delta = 0.075$, Break at the End

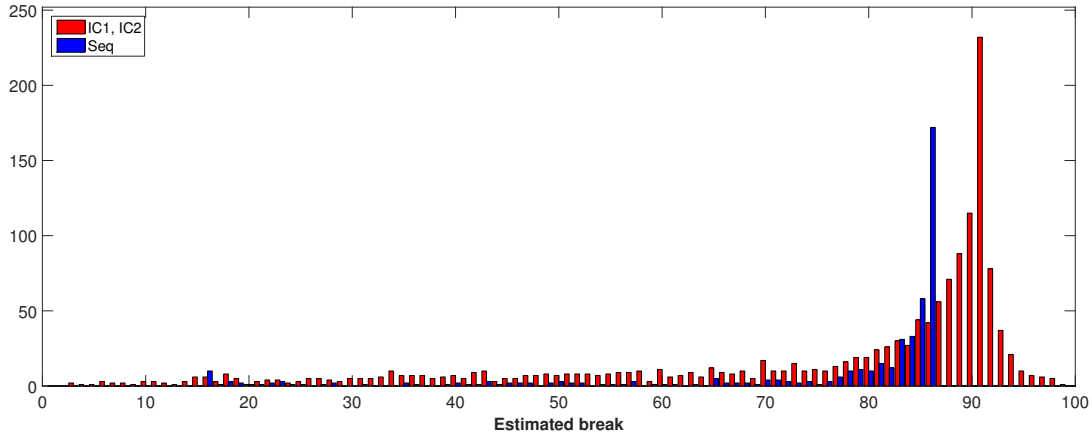


Figure B.3.14: Positions of the Estimated Breaks, $T = 100$, SNR $= 2$, $\delta = 0.075$, Break at the End

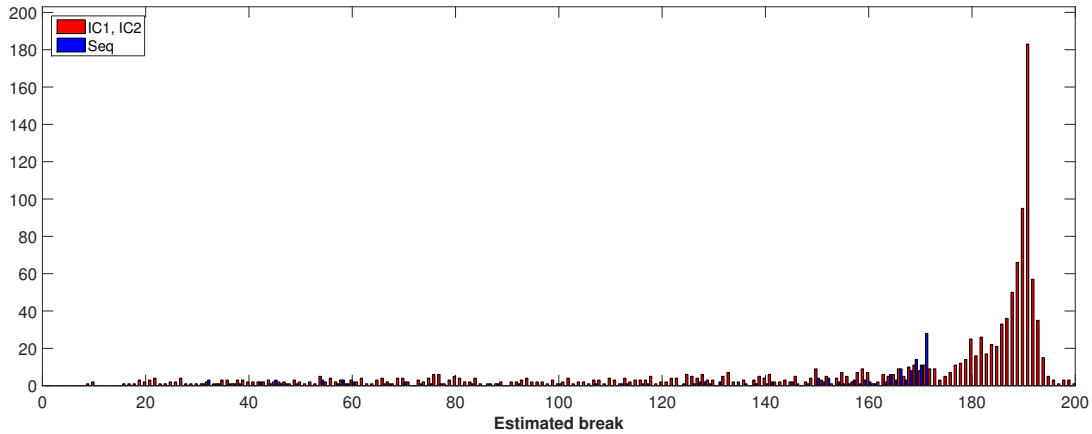


Figure B.3.15: Positions of the Estimated Breaks, $T = 200$, SNR $= 2$, $\delta = 0.075$, Break at the End

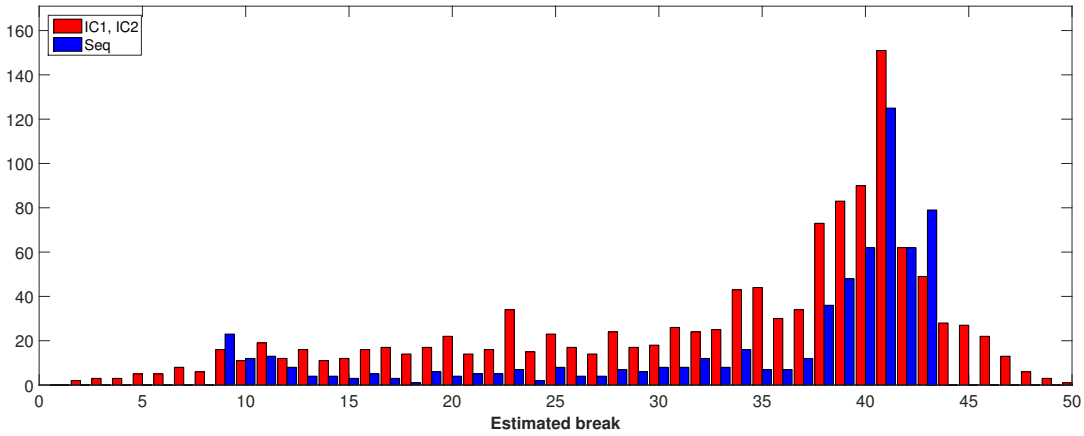


Figure B.3.16: Positions of the Estimated Breaks, $T = 50$, SNR $= 1$, $\delta = 0.075$, Break at the End

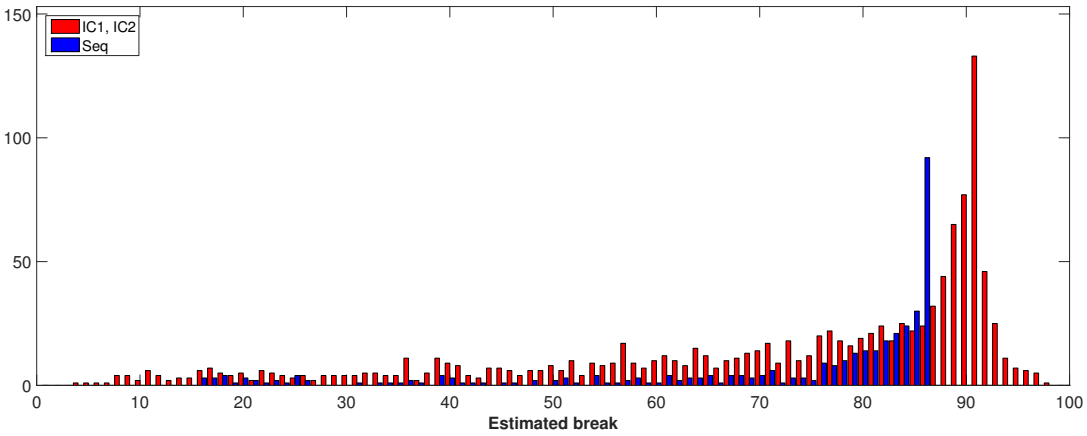


Figure B.3.17: Positions of the Estimated Breaks, $T = 100$, SNR $= 1$, $\delta = 0.075$, Break at the End

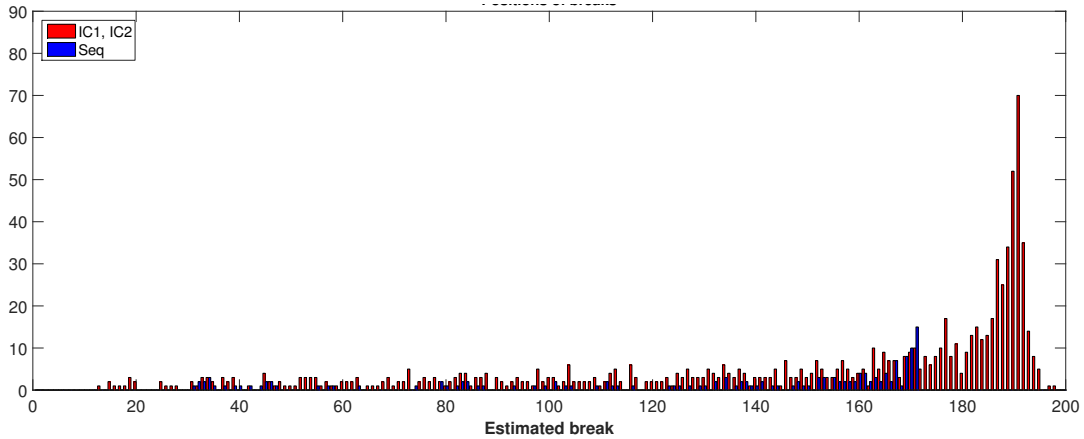


Figure B.3.18: Positions of the Estimated Breaks, $T = 200$, SNR $= 1$, $\delta = 0.075$, Break at the End

B.3.10 One Parameter - One Break - End - One Found

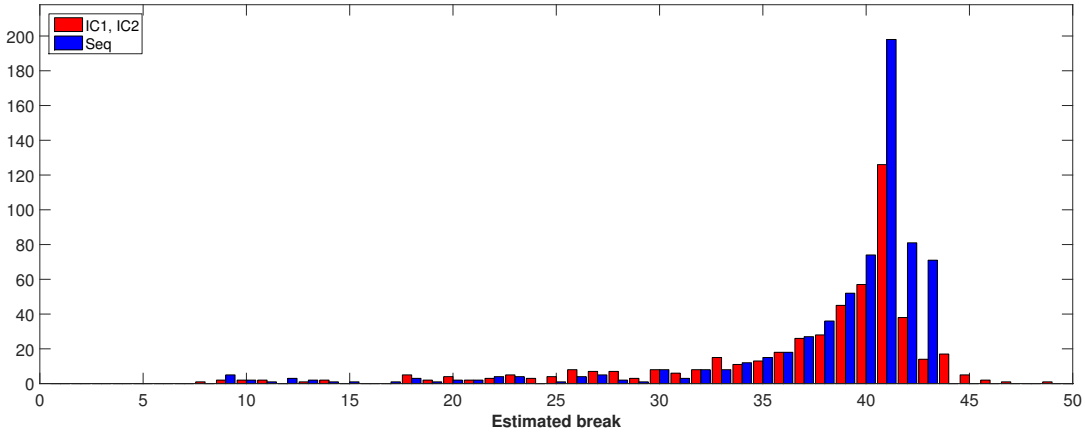


Figure B.3.19: Positions of the Estimated Breaks given One Break Found, $T = 50$, SNR $= 2$, $\delta = 0.075$, Break at the End

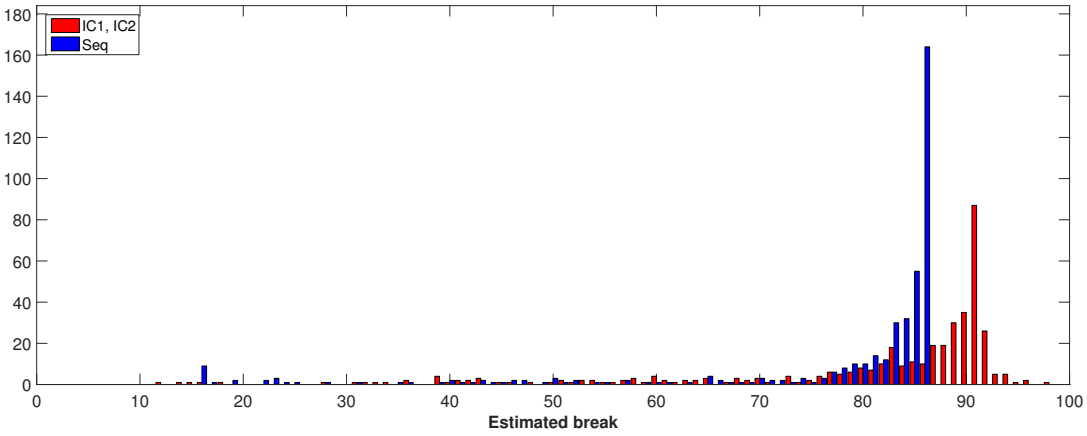


Figure B.3.20: Positions of the Estimated Breaks given One Break Found, $T = 100$, SNR $= 2$, $\delta = 0.075$, Break at the End

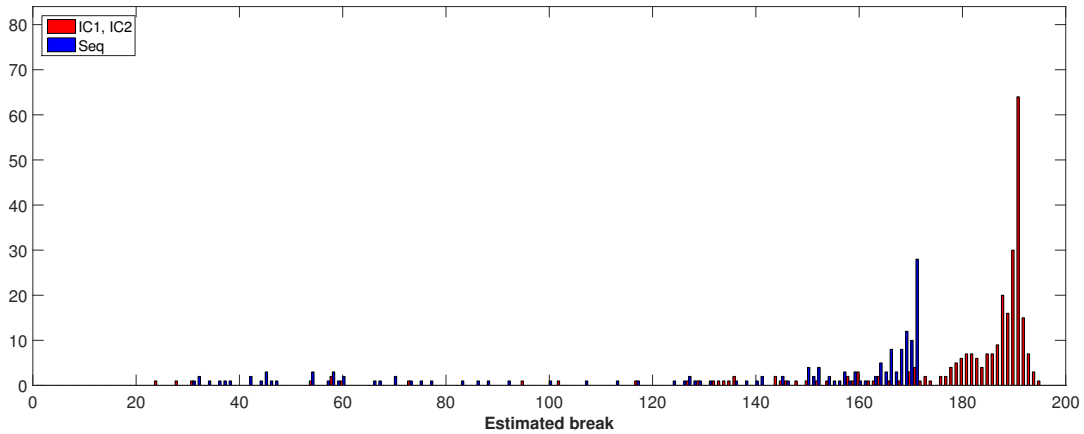


Figure B.3.21: Positions of the Estimated Breaks given One Break Found, $T = 200$, SNR $= 2$, $\delta = 0.075$, Break at the End

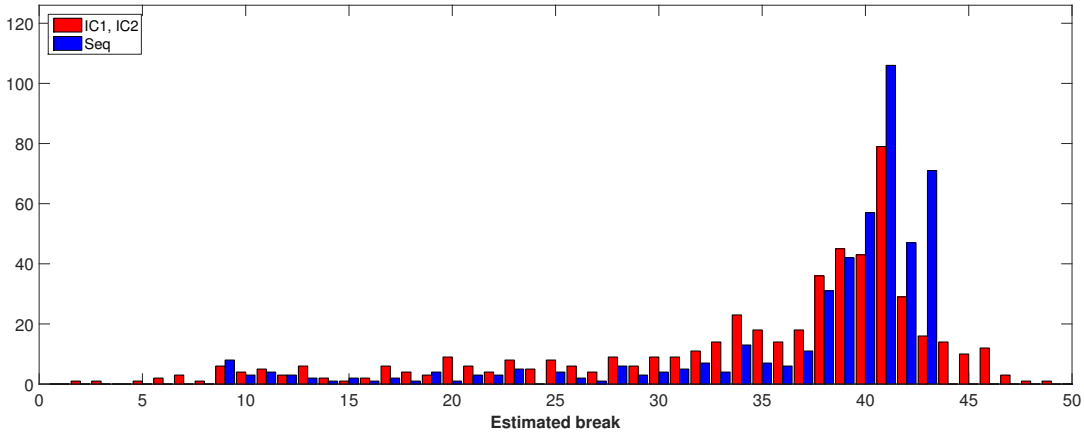


Figure B.3.22: Positions of the Estimated Breaks given One Break Found, $T = 50$, SNR $= 1$, $\delta = 0.075$, Break at the End

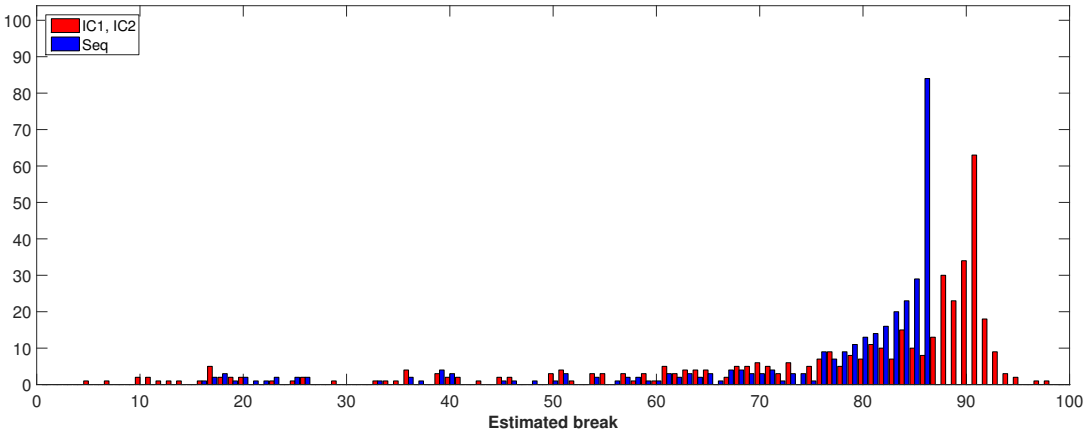


Figure B.3.23: Positions of the Estimated Breaks given One Break Found, $T = 100$, SNR $= 1$, $\delta = 0.075$, Break at the End

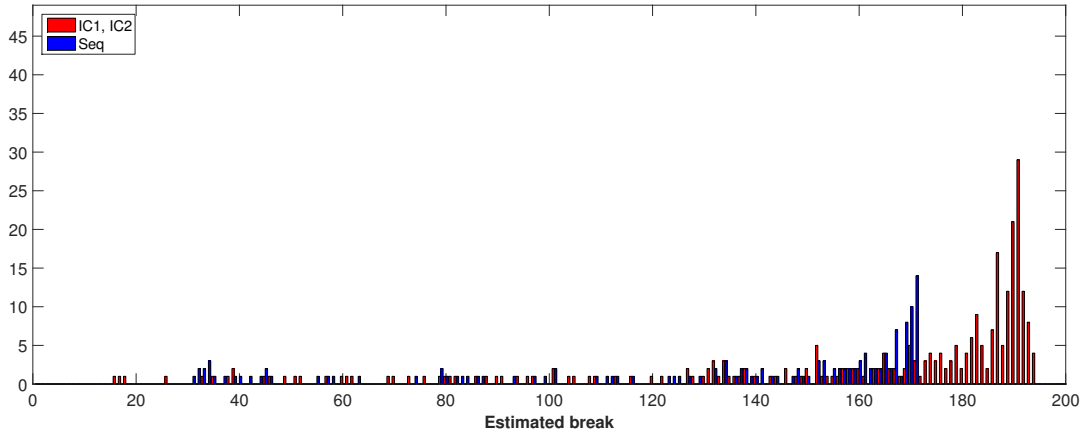


Figure B.3.24: Positions of the Estimated Breaks given One Break Found, $T = 200$, SNR $= 1$, $\delta = 0.075$, Break at the End

B.3.11 Two Parameters - One Break - Middle - Average Squared Bias

Table B.3.49: Average Squared Bias, Break - Middle, 1000 draws, SNR=2, $\delta = 0.025$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.571	0.875	0.432	0.581	0.574	0.675	0.546	0.547	0.763
100	0.424	0.411	0.233	0.379	0.391	0.328	0.326	0.328	0.376
200	0.378	0.348	0.131	0.332	0.320	0.153	0.213	0.236	0.161

Table B.3.50: Average Squared Bias, Break - Middle, 1000 draws, SNR=2, $\delta = 0.050$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.571	0.875	0.431	0.581	0.574	0.662	0.540	0.540	0.763
100	0.424	0.411	0.227	0.379	0.391	0.321	0.318	0.320	0.376
200	0.378	0.348	0.125	0.332	0.320	0.152	0.201	0.226	0.161

Table B.3.51: Average Squared Bias, Break - Middle, 1000 draws, SNR=2, $\delta = 0.075$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.571	0.875	0.441	0.581	0.574	0.681	0.547	0.549	0.763
100	0.424	0.411	0.226	0.379	0.391	0.312	0.316	0.317	0.376
200	0.378	0.348	0.124	0.332	0.320	0.152	0.196	0.221	0.161

Table B.3.52: Best δ for the Bias, Break - Middle, 1000 draws, SNR=2

	1 step	2 step		
T	IC_1	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.050	0.050	0.050	0.050
100	0.075	0.075	0.075	0.075
200	0.075	0.050	0.075	0.075

Table B.3.53: Average Squared Bias, Break - Middle, 1000 draws, SNR=1, $\delta = 0.025$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.824	2.501	0.750	0.958	1.710	1.370	1.050	1.052	1.299
100	0.540	0.536	0.408	0.553	0.544	0.660	0.523	0.517	0.691
200	0.429	0.406	0.230	0.412	0.411	0.325	0.335	0.337	0.342

Table B.3.54: Average Squared Bias, Break - Middle, 1000 draws, SNR=1, $\delta = 0.050$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.824	2.501	0.742	0.958	1.710	1.349	1.064	1.064	1.299
100	0.540	0.536	0.399	0.553	0.544	0.640	0.508	0.508	0.691
200	0.429	0.406	0.221	0.412	0.411	0.314	0.331	0.335	0.342

Table B.3.55: Average Squared Bias, Break - Middle, 1000 draws, SNR=1, $\delta = 0.075$

	1 step			2 step					BP
T	$EBIC$	QS	IC_1	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0.824	2.501	0.945	0.958	1.710	1.651	1.366	1.386	1.299
100	0.540	0.536	0.399	0.553	0.544	0.629	0.510	0.507	0.691
200	0.429	0.406	0.220	0.412	0.411	0.312	0.327	0.329	0.342

Table B.3.56: Best δ for the Bias, Break - Middle, 1000 draws, SNR=1

	1 step	2 step		
T	IC_1	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.050	0.050	0.025	0.025
100	0.075	0.075	0.050	0.075
200	0.075	0.075	0.075	0.075

B.3.12 Two Parameters - One Break - Middle - Rates of Detected Breaks

Table B.3.57: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=2, $\delta = 0.025$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	41.5	35.3	2.1	2.4	2.4	35.3
	1	22.6	21.0	7.1	13.2	13.0	46.3
	2	16.7	18.6	28.3	35.3	35.5	16.5
	3	8.3	7.9	29.7	28.9	29.0	1.8
	4	4.8	4.3	18.1	13.4	13.2	0.1
	5	2.7	2.0	7.5	4.4	4.6	0.0
	6+	3.4	10.9	7.2	2.4	2.3	0.0
	Q	0.48	0.43	0.17	0.17	0.17	0.44
	P	0.47	0.51	0.57	0.63	0.63	0.32
100	0	40.7	41.0	1.1	1.1	1.1	13.8
	1	24.8	23.8	6.6	11.5	11.4	72.9
	2	18.0	19.8	22.5	31.6	31.6	12.6
	3	7.6	8.2	25.2	29.8	30.8	0.7
	4	4.2	3.4	19.1	17.4	17.5	0.0
	5	2.6	1.1	13.4	6.4	5.6	0.0
	6+	2.1	2.7	12.1	2.2	2.0	0.0
	Q	0.45	0.45	0.14	0.13	0.13	0.21
	P	0.52	0.54	0.63	0.72	0.72	0.43
200	0	46.9	41.7	0.1	0.2	0.2	0.9
	1	19.5	24.0	5.2	8.9	8.8	89.9
	2	18.2	21.4	21.3	31.8	33.6	8.8
	3	8.9	8.7	26.0	31.3	32.2	0.4
	4	3.5	2.9	21.0	16.4	15.9	0.0
	5	1.6	0.6	13.8	7.6	6.8	0.0
	6+	1.4	0.7	12.6	3.8	2.5	0.0
	Q	0.49	0.44	0.12	0.10	0.10	0.05
	P	0.48	0.55	0.69	0.77	0.79	0.50

Q ... Average Relative Distance from the True Break.

Table B.3.58: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=2, $\delta = 0.050$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	41.5	35.3	2.8	3.2	3.2	35.3
	1	22.6	21.0	7.0	13.1	13.0	46.3
	2	16.7	18.6	31.3	38.7	38.8	16.5
	3	8.3	7.9	31.4	29.3	29.1	1.8
	4	4.8	4.3	16.3	10.5	10.6	0.1
	5	2.7	2.0	6.9	3.8	3.9	0.0
	6+	3.4	10.9	4.3	1.4	1.4	0.0
	Q	0.48	0.43	0.17	0.17	0.17	0.44
	P	0.47	0.51	0.57	0.63	0.63	0.32
100	0	40.7	41.0	1.2	1.3	1.3	13.8
	1	24.8	23.8	7.2	12.2	12.1	72.9
	2	18.0	19.8	27.7	37.0	37.4	12.6
	3	7.6	8.2	24.4	28.5	29.1	0.7
	4	4.2	3.4	20.0	13.8	14.4	0.0
	5	2.6	1.1	11.2	5.0	4.1	0.0
	6+	2.1	2.7	8.3	2.2	1.6	0.0
	Q	0.45	0.45	0.14	0.13	0.13	0.21
	P	0.52	0.54	0.63	0.72	0.73	0.43
200	0	46.9	41.7	0.2	0.4	0.4	0.9
	1	19.5	24.0	8.3	12.1	12.6	89.9
	2	18.2	21.4	21.1	32.2	33.3	8.8
	3	8.9	8.7	28.0	29.9	30.9	0.4
	4	3.5	2.9	20.5	16.7	16.4	0.0
	5	1.6	0.6	13.1	6.0	4.5	0.0
	6+	1.4	0.7	8.8	2.7	1.9	0.0
	Q	0.49	0.44	0.11	0.10	0.09	0.05
	P	0.48	0.55	0.71	0.79	0.81	0.50

Q . . . Average Relative Distance from the True Break.

Table B.3.59: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=2, $\delta = 0.075$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	41.5	35.3	3.0	3.4	3.4	35.3
	1	22.6	21.0	8.2	14.5	14.4	46.3
	2	16.7	18.6	33.2	37.5	37.6	16.5
	3	8.3	7.9	31.3	29.8	29.9	1.8
	4	4.8	4.3	15.3	10.7	10.5	0.1
	5	2.7	2.0	5.2	2.4	2.5	0.0
	6+	3.4	10.9	3.8	1.7	1.7	0.0
	Q	0.48	0.43	0.17	0.17	0.17	0.44
	P	0.47	0.51	0.57	0.63	0.63	0.32
100	0	40.7	41.0	1.2	1.3	1.3	13.8
	1	24.8	23.8	7.4	13.6	13.8	72.9
	2	18.0	19.8	27.4	36.1	36.1	12.6
	3	7.6	8.2	30.0	30.2	31.8	0.7
	4	4.2	3.4	18.5	13.6	12.8	0.0
	5	2.6	1.1	9.0	4.2	3.4	0.0
	6+	2.1	2.7	6.5	1.0	0.8	0.0
	Q	0.45	0.45	0.13	0.12	0.12	0.21
	P	0.52	0.54	0.63	0.73	0.73	0.43
200	0	46.9	41.7	0.2	0.3	0.3	0.9
	1	19.5	24.0	9.3	12.7	13.0	89.9
	2	18.2	21.4	23.5	35.5	37.2	8.8
	3	8.9	8.7	27.2	28.0	29.7	0.4
	4	3.5	2.9	20.8	16.8	15.0	0.0
	5	1.6	0.6	12.1	5.3	4.1	0.0
	6+	1.4	0.7	6.9	1.4	0.7	0.0
	Q	0.49	0.44	0.10	0.09	0.09	0.05
	P	0.48	0.55	0.70	0.79	0.81	0.50

Q ... Average Relative Distance from the True Break.

Table B.3.60: Best δ for the Q, Break - Middle, 1000 draws, SNR=2

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.050	0.025	0.025
100	0.075	0.025	0.025
200	0.075	0.025	0.025

Table B.3.61: Best δ for the P, Break - Middle, 1000 draws, SNR=2

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.050	0.025	0.025
100	0.075	0.025	0.025
200	0.075	0.025	0.025

Table B.3.62: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=1, $\delta = 0.025$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	31.7	26.4	1.5	1.5	1.5	50.6
	1	23.4	19.7	5.8	8.4	8.3	36.2
	2	18.8	17.0	28.7	32.7	32.2	11.4
	3	8.6	7.6	32.1	32.1	32.6	1.7
	4	6.5	4.4	17.7	15.6	15.6	0.1
	5	3.0	3.0	7.6	6.2	6.2	0.0
	6+	8.0	21.9	6.6	3.5	3.6	0.0
	Q	0.41	0.38	0.18	0.18	0.18	0.59
	P	0.48	0.48	0.53	0.55	0.55	0.25
100	0	45.8	42.4	1.2	1.2	1.2	44.3
	1	21.5	22.5	7.7	10.3	10.4	47.0
	2	14.3	16.3	25.6	32.9	32.9	8.0
	3	7.8	8.7	27.3	30.3	30.3	0.7
	4	4.1	4.2	20.2	16.0	16.0	0.0
	5	2.2	1.9	10.5	6.0	6.1	0.0
	6+	4.3	4.0	7.5	3.3	3.1	0.0
	Q	0.51	0.48	0.17	0.16	0.16	0.50
	P	0.45	0.48	0.59	0.64	0.64	0.28
200	0	48.4	44.9	0.4	0.4	0.4	14.2
	1	19.4	20.6	5.1	7.1	7.1	79.6
	2	15.9	18.0	21.5	29.9	30.5	6.1
	3	8.4	8.9	30.1	34.6	34.4	0.1
	4	4.2	4.1	18.2	16.2	16.7	0.0
	5	1.8	1.9	13.9	8.2	7.8	0.0
	6+	1.9	1.6	10.8	3.6	3.1	0.0
	Q	0.52	0.49	0.14	0.13	0.13	0.20
	P	0.45	0.50	0.66	0.72	0.72	0.43

Q ... Average Relative Distance from the True Break.

Table B.3.63: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=1, $\delta = 0.050$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	31.7	26.4	2.0	2.0	2.0	50.6
	1	23.4	19.7	7.7	10.4	10.3	36.2
	2	18.8	17.0	34.6	38.7	38.6	11.4
	3	8.6	7.6	29.5	28.1	28.1	1.7
	4	6.5	4.4	15.6	13.6	13.8	0.1
	5	3.0	3.0	5.8	4.9	4.9	0.0
	6+	8.0	21.9	4.8	2.3	2.3	0.0
	Q	0.41	0.38	0.18	0.18	0.18	0.59
	P	0.48	0.48	0.53	0.55	0.55	0.25
100	0	45.8	42.4	1.6	1.6	1.6	44.3
	1	21.5	22.5	9.4	14.6	14.5	47.0
	2	14.3	16.3	28.8	35.2	35.1	8.0
	3	7.8	8.7	29.9	28.6	28.5	0.7
	4	4.1	4.2	17.3	13.5	13.9	0.0
	5	2.2	1.9	7.3	4.6	4.5	0.0
	6+	4.3	4.0	5.7	1.9	1.9	0.0
	Q	0.51	0.48	0.16	0.16	0.16	0.50
	P	0.45	0.48	0.59	0.65	0.65	0.28
200	0	48.4	44.9	0.9	0.9	0.9	14.2
	1	19.4	20.6	7.7	10.7	10.7	79.6
	2	15.9	18.0	25.3	32.8	33.4	6.1
	3	8.4	8.9	29.4	31.3	31.8	0.1
	4	4.2	4.1	18.6	17.6	17.2	0.0
	5	1.8	1.9	10.6	4.9	4.7	0.0
	6+	1.9	1.6	7.5	1.8	1.3	0.0
	Q	0.52	0.49	0.13	0.13	0.13	0.20
	P	0.45	0.50	0.67	0.74	0.74	0.43

Q . . . Average Relative Distance from the True Break.

Table B.3.64: Percentage Rates of detecting 0-5 breaks and Distance, Break - Middle, 1000 draws, SNR=1, $\delta = 0.075$

		2 step					BP
T	Criterion	$EBIC$ $EBIC$	QS QS	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS	Seq
50	0	31.7	26.4	1.6	1.7	1.7	50.6
	1	23.4	19.7	8.7	11.7	11.7	36.2
	2	18.8	17.0	37.0	40.1	39.9	11.4
	3	8.6	7.6	29.2	27.6	27.8	1.7
	4	6.5	4.4	13.3	10.6	10.9	0.1
	5	3.0	3.0	5.0	3.9	3.8	0.0
	6+	8.0	21.9	5.2	4.4	4.2	0.0
	Q	0.41	0.38	0.18	0.18	0.18	0.59
	P	0.48	0.48	0.54	0.56	0.56	0.25
100	0	45.8	42.4	2.4	2.6	2.6	44.3
	1	21.5	22.5	8.8	13.2	13.2	47.0
	2	14.3	16.3	32.5	38.0	37.9	8.0
	3	7.8	8.7	32.8	30.2	30.6	0.7
	4	4.1	4.2	13.2	10.7	10.3	0.0
	5	2.2	1.9	6.5	4.0	4.1	0.0
	6+	4.3	4.0	3.8	1.3	1.3	0.0
	Q	0.51	0.48	0.17	0.17	0.16	0.50
	P	0.45	0.48	0.59	0.64	0.64	0.28
200	0	48.4	44.9	1.2	1.2	1.2	14.2
	1	19.4	20.6	8.0	12.0	12.0	79.6
	2	15.9	18.0	25.6	34.0	34.8	6.1
	3	8.4	8.9	30.7	30.4	30.7	0.1
	4	4.2	4.1	19.4	16.3	15.7	0.0
	5	1.8	1.9	9.9	4.9	4.5	0.0
	6+	1.9	1.6	5.2	1.2	1.1	0.0
	Q	0.52	0.49	0.13	0.12	0.12	0.20
	P	0.45	0.50	0.68	0.75	0.75	0.43

Q ... Average Relative Distance from the True Break.

Table B.3.65: Best δ for the Q, Break - Middle, 1000 draws, SNR=1

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.050	0.025	0.025
100	0.075	0.025	0.025
200	0.075	0.025	0.025

Table B.3.66: Best δ for the P, Break - Middle, 1000 draws, SNR=1

	2 step		
T	IC_1 IC_2	IC_1 $EBIC$	IC_1 QS
50	0.050	0.025	0.025
100	0.075	0.025	0.025
200	0.075	0.025	0.025

B.3.13 Two Parameters - One Break - Middle - All

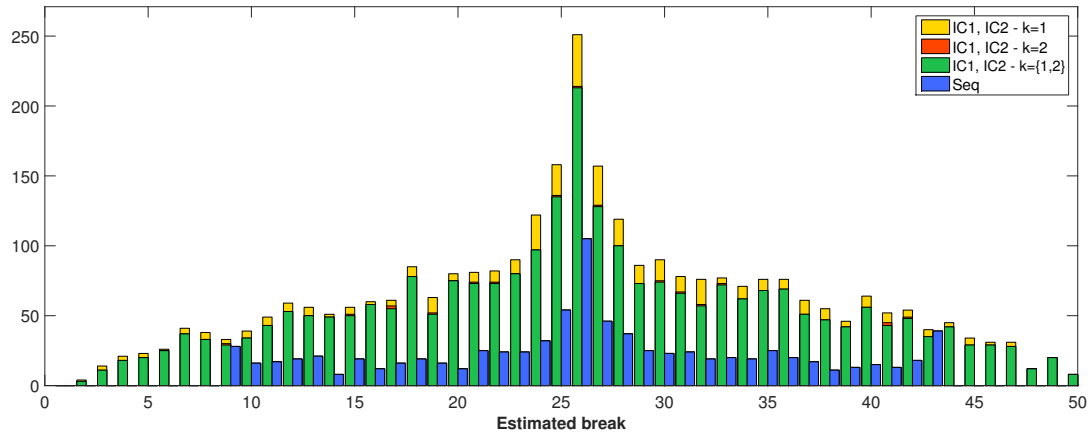


Figure B.3.25: Positions of the Estimated Breaks, $T = 50$, SNR $= 2$, $\delta = 0.025$, Break in the Middle

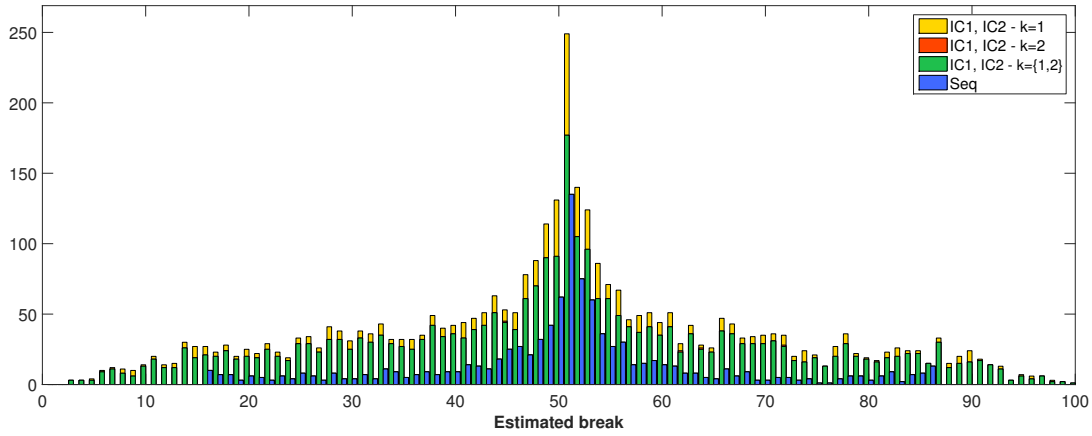


Figure B.3.26: Positions of the Estimated Breaks, $T = 100$, SNR $= 2$, $\delta = 0.025$, Break in the Middle

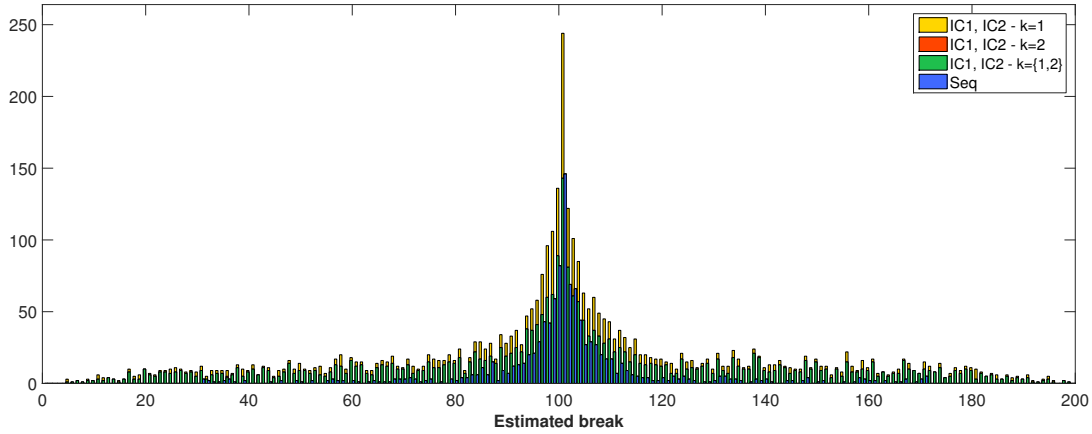


Figure B.3.27: Positions of the Estimated Breaks, $T = 200$, SNR $= 2$, $\delta = 0.025$, Break in the Middle

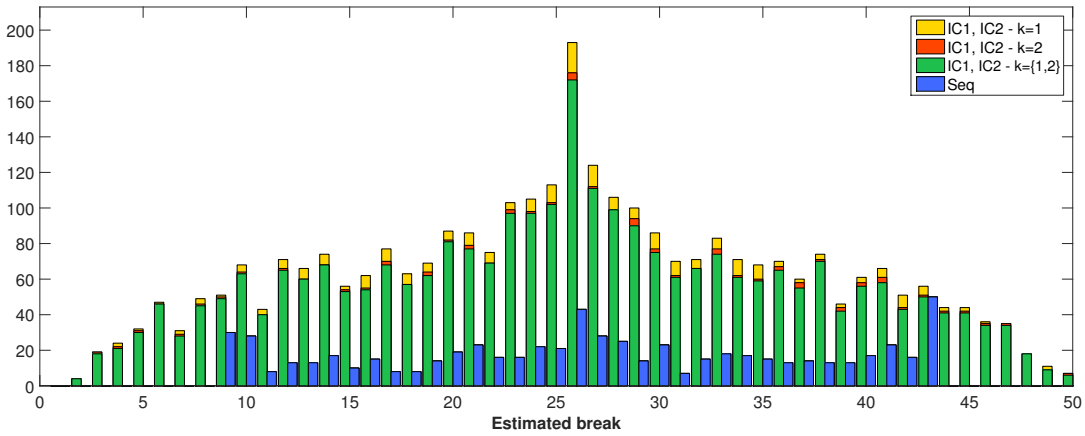


Figure B.3.28: Positions of the Estimated Breaks, $T = 50$, SNR $= 1$, $\delta = 0.025$, Break in the Middle

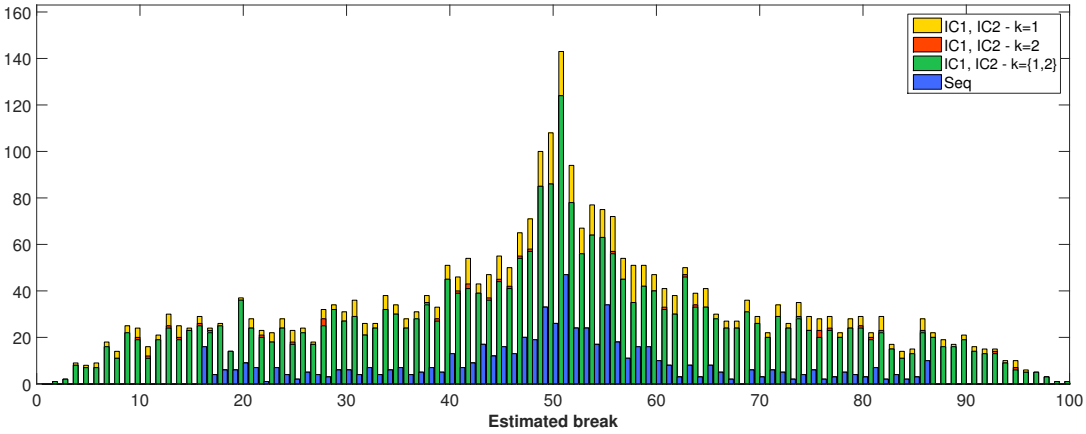


Figure B.3.29: Positions of the Estimated Breaks, $T = 100$, SNR $= 1$, $\delta = 0.025$, Break in the Middle

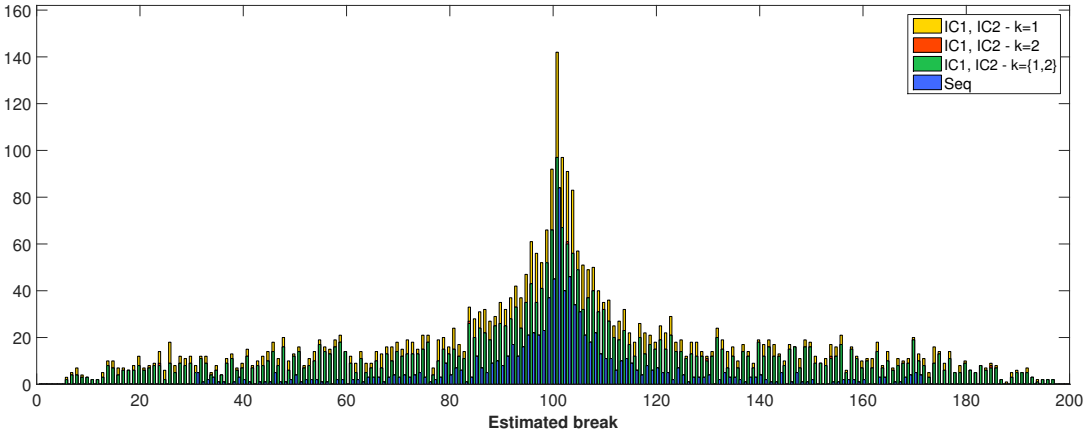


Figure B.3.30: Positions of the Estimated Breaks, $T = 200$, SNR $= 1$, $\delta = 0.025$, Break in the Middle

C Appendices for How well can Noncognitive Skills Predict Unemployment?: A Machine Learning Approach

C.1 Regular Appendix

C.1.1 Tables

Variable Definitions, Summary Statistics

Table C.1.1: Description Control Variables

Variable	Description
SocialClass10Y•	social class of the family at the age of 10, the better occupation of the parents taken. 6 categories: • = 6 if professional, = 5 if managerial-technical, = 4 if skilled non-manual, = 3 if skilled manual, = 2 if partly skilled, = 1 if unskilled
AbilityMath10Y	standardized score of a "Friendly Math Test" taken at the age of 10
AbilityRead10Y	standardized score of an "Edinburgh Reading Test" taken at the age of 10

Table C.1.2: Non-cognitive Skills Measures: Self-Esteem

Self-Esteem Scale Questions	
1	Do you think that your parents like to hear about your ideas?
2	Do you often feel lonely at school?
3	Do other children often break friends or fall out with you?
4	Do you think that other children often say nasty things about you?
5	When you have to say things in front of teachers, do you usually feel shy?
6	Do you often feel sad because you have nobody to play with at school?
7	Are there lots of things about yourself you would like to change?
8	When you have to say things in front of other children, do you usually feel foolish?
9	When you want to tell a teacher something, do you usually feel foolish?
10	Do you often have to find new friends because your old friends are playing with somebody else?
11	Do you usually feel foolish when you talk to your parents?
12	Do other people often think that you tell lies?

Answers coded as: "Yes" = 1 point, "No" = 2 points and "I don't know" = 1.5 point, except Question 1 for which the points for "Yes" and "No" are switched. Higher score indicates higher self-esteem.

Table C.1.3: Non-cognitive Skills Measures: Locus of Control

Locus of Control Scale Questions	
1	Do you feel that most of the time it's not worth trying hard because things never turn out right anyway?
2	Do you feel that wishing can make good things happen?
3	Are people good to you no matter how you act towards them?
4	Do you usually feel that it's almost useless to try in school because most children are cleverer than you?
5	Is a high mark just a matter of "luck" for you?
6	Are tests lust a lot of guess work for you?
7	Are you often blamed for things which just aren't your fault?
8	Are you the kind of person who believes that planning ahead makes things turn out better?
9	When bad things happen to you, is it usually someone else's fault?
10	When someone is very angry with you, is it impossible to make him your friend again?
11	When nice things happen to you is it only good luck?
12	Do you feel sad when it's time to leave school each day?
13	When you get into an argument is it usually the other person's fault?
14	Are you surprised when your teacher says you've done well?
15	Do you usually get low marks, even when you study hard?
16	Do you think studying for tests is a waste of time?

Answers coded as: "Yes" = 1 point, "No" = 2 points and "I don't know" = 1.5 point, except Question 8 for which the points for "Yes" and "No" are switched. Higher score indicates an internalizer.

Table C.1.4: Non-cognitive Skills Measures: Mother-rated Items

Motor Control (MC) Items	
1	Child drops things which are being carried
2	Child is noticeably clumsy
3	Child trips or falls easily into objects or other people
Hand Eye Coordination (HEC) Items	
4	Child has twitches mannerisms or tics of the face and body
5	Child has difficulty picking up small objects
6	Child has difficulty in using scissors
Conscientiousness (C) Items	
7	Child cannot settle to anything for more than a few moments
8	Child is inattentive, easily distracted
9	Child fails to finish things he/she starts, short attention span
10	Child has difficulty concentrating on any particular task though may return to it
Hyperactivity (H) Items - recoded as (100 - answer) to capture hyperactivity	
11	Child is squirmy or fidgety
12	Child is very restless, i.e. running often, jumping up and down
13	Child shows restless or over-active behavior
14	Child hums or makes other odd noises at inappropriate times
15	Child is given to rhythmic tapping or kicking
Agreeableness (A) Items	
16	Child frequently fights with other children
17	Child bullies other children
18	Child interferes with the activity of others
19	Child is often disobedient
20	Child often tells lies
21	Child often destroys own or others' belongings
22	Child sometimes takes things belonging to others
Emotional Stability (ES) Items	
23	Child cries for little cause
24	Child often appears miserable, unhappy
25	Child is sullen or sulky
26	Child is irritable
27	Child changes mood quickly and drastically
28	Child displays outbursts of temper, explosive or unpredictable behavior
29	Child's requests must be met immediately, easily frustrated
30	Child is impulsive, excitable
Behavioral Trauma (TR) Items - recoded as (100 - answer) to capture behavioral trauma	
31	Child frequently sucks thumb or fingers
32	Child frequently bites nails or fingers
Extraversion (E) Items	
33	Child is not much liked by other children
34	Child tends to do things on his own
35	Child is fussy or over particular
36	Child is often worried
37	Child tends to be fearful or afraid of new things or new situations
38	Child becomes obsessional about unimportant things

Answers coded on a scale from 0 to 100 where 0 = "certainly" and 100 = "does not apply".

Table C.1.5: Non-cognitive Skills Measures: Teacher-rated Items

Motor Control 1 (MC1) Items - recoded as (48 - answer) to capture motor control	
1	Child is noticeably clumsy in formal or informal games
2	Child finds it difficult to kick a ball forward
3	Child is fearful in movements, requires much encouragement to move faster
Motor Control 2 (MC2) Items - recoded as (48 - answer) to capture motor control	
4	Child trips or falls easily or bumps into objects or other children
5	Child shows difficulty when picking up small objects
6	Child drops things which are being carried
7	Child shows inadequate control when handlling a pencil or paint brush
8	Child experiences classroom or playground accidents
Hand Eye Coordination (HEC) Items	
9	Child works deftly with his or her hands
10	Child dresses and undresses competently (e.g. for P.E)
11	Child manipulates small objects easily with his/her hands
12	Child can use scissors and similar manipulative equipment competently
13	Child holds writing and drawing instruments appropriately
Conscientiousness 1 (C1) Items	
14	Child shows preservance, persists with difficult or routine work
15	Child pays attention to what is being explained in class
16	Child completes tasks which are started
Conscientiousness 2 (C2) Items - recoded as (48 - answer) to capture conscientiousness	
17	Child is given to daydreaming
18	Child has difficulty concentrating on any particular task though may return to it
19	Child becomes bored during class
20	Child becomes confused or hesitant when given a complex task
21	Child is squirmy or fidgety
22	Child is easily distracted
23	Child is forgetful when given a complex task
24	Child shows lethargic and listless behaviour
25	Child fails to finish things he/she starts
Hyperactivity (H) Items	
26	Child hums or makes other odd noises at inappropriate times
27	Child is given to rhythmic tapping or kicking during class
28	Child has twitches mannerisms or tics of the face and body
Agreeableness (A) Items - recoded as (48 - answer) to capture agreeableness	
29	Child complains about things
30	Child displays outbursts of temper, explosive or unpredictable behaviour
31	Child teases other children to excess
32	Child interferes with the activity of other children
33	Child changes mood quickly and drastically
34	Child is excitable, impulsive
35	Child shows restless or over-active behaviour
36	Child quarrels with other children
37	Child destroys own or other children's belongings
38	Child bullies other children
39	Child is sullen or sulky
40	Child's requests must be met immediately, easily frustrated

Answers coded on a scale from 1 to 47 where 1 = "not at all" and 47 = "a great deal".

Table C.1.6: Non-cognitive Skills Measures: Teacher-rated Items - cont.

Emotional Stability 1 (ES1) Items - recoded as (48 - answer) to capture emotional stability	
41	Child is fearful or afraid of new things or situations
42	Child behaves nervously
43	Child is worried and anxious about many things
44	Child is anxious (opposite: unworried)
Emotional Stability 2 (ES2) Items - recoded as (48 - answer) to capture emotional stability	
45	Child cries for little cause
46	Child is fussy or over particular
47	In relations with others child appears to be miserable, unhappy, tearful or distressed
48	Child becomes obsessional about unimportant things
49	Child tends to do things on his own
Behavioral Trauma (TR) Items	
50	Child has problems with wetting pants during class
51	Child has problems of soiling pants during class
Extraversion (E) Items	
52	Child is an extrovert (opposite: introvert)
School Issues (S) Items	
53	Child truants from school

Answers coded on a scale from 1 to 47 where 1 = “not at all” and 47 = “a great deal”.

Table C.1.7: Correlation Matrix for Equally Weighted Indexes (Mother's Questionnaire) based on the Hierarchical Clustering for Male - 34Y

	MC_M	HEC_M	C_M	H_M	A_M	ES_M	TR_M	E_M	SE	LC	A-M
HEC_M	0.46										
C_M	0.35	0.28									
H_M	-0.36	-0.26	-0.56								
A_M	0.34	0.35	0.47	-0.48							
ES_M	0.32	0.29	0.42	-0.52	0.59						
TR_M	-0.17	-0.16	-0.14	0.15	-0.20	-0.16					
E_M	0.29	0.30	0.26	-0.33	0.30	0.51	-0.15				
SE	0.09	0.03	0.21	-0.13	0.14	0.14	-0.06	0.08			
LC	0.09	0.08	0.25	-0.13	0.19	0.14	-0.07	0.09	0.41		
A-M	0.09	0.10	0.34	-0.16	0.21	0.16	-0.06	0.08	0.26	0.46	
A-R	0.09	0.09	0.32	-0.14	0.20	0.15	-0.05	0.06	0.25	0.48	0.75

A-M = Ability Math, A-R = Ability Read, subscript M indicates questions in the Mother's

Questionnaire.

Table C.1.8: Correlation Matrix for Equally Weighted Indeces (Teacher's Questionnaire) based on Hierarchical Clustering for Male - 34Y

	$C1_T$	$C2_T$	$ES1_T$	$ES2_T$	TR_T	A_T	$MC1_T$	$MC2_T$	HEC_T	S_T	E_T	H_T	SE	LC	A-M
$C2_T$	0.83														
$ES1_T$	0.32	0.47													
$ES2_T$	0.28	0.43	0.62												
TR_T	-0.06	-0.09	-0.08	-0.17											
A_T	0.48	0.57	0.26	0.47	-0.11										
$MC1_T$	0.29	0.41	0.42	0.53	-0.15	0.25									
$MC2_T$	0.42	0.54	0.37	0.49	-0.22	0.49	0.60								
HEC_T	0.46	0.44	0.30	0.30	-0.17	0.27	0.45	0.59							
S_T	-0.16	-0.20	-0.12	-0.20	0.36	-0.25	-0.16	-0.25	-0.18						
E_T	0.07	0.11	0.46	0.41	-0.05	-0.19	0.35	0.11	0.18	-0.03					
H_T	-0.41	-0.50	-0.27	-0.39	0.16	-0.63	-0.30	-0.55	-0.32	0.27	0.07				
SE	0.24	0.26	0.15	0.20	-0.03	0.21	0.14	0.17	0.16	-0.07	0.08	-0.16			
LC	0.35	0.37	0.19	0.17	-0.01	0.19	0.13	0.20	0.20	-0.09	0.09	-0.15	0.41		
A-M	0.47	0.50	0.25	0.14	-0.04	0.22	0.16	0.25	0.28	-0.11	0.08	-0.19	0.26	0.46	
A-R	0.44	0.49	0.23	0.11	-0.03	0.22	0.13	0.26	0.28	-0.10	0.07	-0.18	0.25	0.48	0.75

A-M = Ability Math, A-R = Ability Read, subscript T indicates questions in the Teacher's Questionnaire.

Table C.1.9: Comparing Mother's and Teacher's Equally Weighted Indexes based on Hierarchical Clustering, Male 34Y

	MC_M	$MC1_T$	$MC2_T$
MC_M	1.00	0.16	0.19
$MC1_T$	0.16	1.00	0.60
$MC2_T$	0.19	0.60	1.00

	HEC_M	HEC_T
HEC_M	1.00	0.08
HEC_T	0.08	1.00

	C_M	$C1_T$	$C2_T$
C_M	1.00	0.38	0.38
$C1_T$	0.38	1.00	0.83
$C2_T$	0.38	0.83	1.00

	H_M	H_T
H_M	1.00	0.16
H_T	0.16	1.00

	A_M	A_T
A_M	1.00	0.24
A_T	0.24	1.00

	ES_M	$ES1_T$	$ES2_T$
ES_M	1.00	0.10	0.11
$ES1_T$	0.10	1.00	0.62
$ES2_T$	0.11	0.62	1.00

	TR_M	TR_T
TR_M	1.00	0.03
TR_T	0.03	1.00

	E_M	E_T
E_M	1.00	0.16
E_T	0.16	1.00

Table C.1.10: Means of Model Covariates

	All		Train		Validation		Test	
	Empl	Unempl	Empl	Unempl	Empl	Unempl	Empl	Unempl
SocialClass10Y1	0.09	0.13	0.08	0.12	0.09	0.18	0.11	0.11
SocialClass10Y2	0.28	0.27	0.27	0.26	0.30	0.29	0.28	0.26
SocialClass10Y3	0.30	0.31	0.31	0.30	0.30	0.27	0.28	0.38
SocialClass10Y4	0.17	0.14	0.18	0.16	0.16	0.09	0.16	0.15
SocialClass10Y5	0.14	0.13	0.14	0.14	0.13	0.15	0.15	0.09
SocialClass10Y6	0.02	0.02	0.03	0.02	0.02	0.02	0.02*	0.00*
AbilityMath10Y	0.26*	−0.02*	0.24*	−0.04*	0.31*	−0.07*	0.28	0.06
AbilityRead10Y	0.13*	−0.06*	0.13*	−0.08*	0.15*	−0.18*	0.13	0.11
eqMC_M	0.01	−0.03	0.00	−0.02	0.02	0.12	0.03	−0.23
eqHEC_M	−0.00	−0.00	−0.00	0.03	−0.03	0.07	0.03	−0.19
eqC_M	0.01	−0.06	−0.00	0.04	−0.00	−0.27	0.04	−0.12
eqH_M	−0.01	0.07	−0.01	0.05	0.03	0.18	−0.05	0.01
eqA_M	0.02*	−0.23*	0.03*	−0.23*	−0.00	−0.33	0.02	−0.14
eqES_M	0.00	−0.11	0.02	−0.14	−0.03	−0.06	0.00	−0.05
eqTR_M	0.01	−0.01	0.00	−0.03	0.04	−0.04	−0.00	0.07
eqE_M	−0.01	−0.04	0.00	−0.04	−0.04	−0.03	−0.03	−0.03
eqSE	0.04	−0.04	0.00	−0.02	0.07	−0.22	0.12	0.06
eqLC	0.02*	−0.21*	0.02*	−0.17*	0.01*	−0.44*	0.05	−0.11
eqC1_T	0.03*	−0.19*	0.02*	−0.19*	0.03*	−0.34*	0.08	−0.04
eqC2_T	0.03*	−0.28*	0.03*	−0.28*	0.02*	−0.53*	0.06	−0.02
eqES1_T	0.02*	−0.18*	0.01	−0.14	0.03*	−0.40*	0.04	−0.07
eqES2_T	0.04*	−0.23*	0.02*	−0.19*	0.03*	−0.37*	0.12	−0.22
eqTR_T	−0.02	0.10	−0.01	0.11	0.03	0.22	−0.08	−0.04
eqA_T	0.02*	−0.21*	0.03*	−0.24*	−0.04	−0.34	0.06	0.01
eqMC1_T	0.03	−0.10	0.00	−0.02	0.10*	−0.23*	0.05	−0.21
eqMC2_T	0.02*	−0.19*	0.01	−0.13	0.02*	−0.42*	0.04	−0.10
eqHEC_T	0.02*	−0.23*	0.01	−0.14	0.01*	−0.43*	0.05*	−0.28*
eqS_T	−0.02*	0.17*	−0.02	0.19	−0.01	0.38	−0.04	−0.11
eqE_T	0.03*	−0.21*	0.02	−0.14	0.07*	−0.37*	0.03	−0.24
eqH_T	−0.03*	0.21*	−0.02*	0.20*	−0.05*	0.46*	−0.05	−0.01
pcaMC_M	0.01	−0.03	0.00	−0.02	0.02	0.13	0.04	−0.23
pcaHEC_M	−0.01	−0.02	−0.00	0.01	−0.03	0.05	0.01	−0.17
pcaC_M	0.00	−0.06	−0.00	0.04	−0.00	−0.27	0.04	−0.12
pcaH_M	−0.01	0.07	−0.01	0.06	0.04	0.19	−0.05	0.01
pcaA_M	0.02*	−0.23*	0.03*	−0.23*	−0.00	−0.32	0.02	−0.15
pcaES_M	0.01	−0.11	0.02	−0.14	−0.03	−0.08	0.01	−0.06
pcaTR_M	0.01	−0.01	0.00	−0.02	0.04	−0.04	−0.00	0.06
pcaE_M	−0.02	−0.02	0.00	−0.02	−0.06	−0.04	−0.03	−0.01
pcaSE	0.04	−0.03	0.00	−0.02	0.07	−0.22	0.12	0.11
pcaLC	0.03*	−0.21*	0.02	−0.15	0.01*	−0.50*	0.08	−0.11
pcaC1_T	0.03*	−0.19*	0.02*	−0.18*	0.03*	−0.34*	0.08	−0.04
pcaC2_T	0.03*	−0.28*	0.03*	−0.28*	0.02*	−0.52*	0.06	−0.02
pcaES1_T	0.02*	−0.18*	0.01	−0.14	0.03*	−0.40*	0.04	−0.06
pcaES2_T	0.04*	−0.22*	0.02*	−0.17*	0.03*	−0.37*	0.13	−0.20
pcaTR_T	−0.02	0.10	−0.01	0.11	0.03	0.22	−0.08	−0.04
pcaA_T	0.02*	−0.22*	0.03*	−0.24*	−0.04	−0.34	0.06	−0.00
pcaMC1_T	0.03	−0.10	0.00	−0.02	0.10*	−0.23*	0.05	−0.21
pcaMC2_T	0.02*	−0.18*	0.01	−0.11	0.02*	−0.44*	0.04	−0.10
pcaHEC_T	0.02*	−0.23*	0.02	−0.14	0.01*	−0.44*	0.05*	−0.30*
pcaS_T	−0.02*	0.17*	−0.02	0.19	−0.01	0.38	−0.04	−0.11
pcaE_T	0.03*	−0.21*	0.02	−0.14	0.07*	−0.37*	0.03	−0.24
pcaH_T	−0.03*	0.21*	−0.02*	0.20*	−0.05*	0.46*	−0.05	−0.02

Stars indicate mean differences which are significant at 5% significance level.

Table C.1.11: Means of Model Covariates - cont.

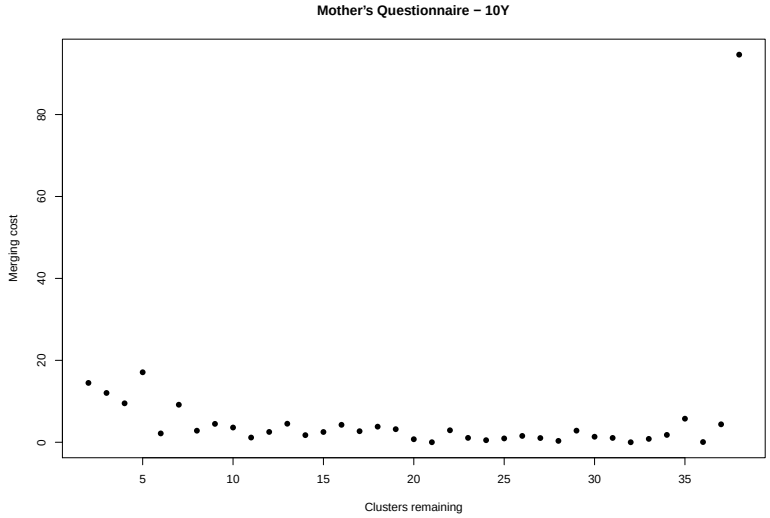
	All		Train		Validation		Test	
	Empl	Unempl	Empl	Unempl	Empl	Unempl	Empl	Unempl
XMC_M								
$XHEC_M$								
XC_M	−0.00	−0.01	−0.01	0.07	0.01	−0.21	0.01	−0.07
XH_M								
XA_M	0.02*	−0.16*	0.02*	−0.19*	0.01	−0.18	0.01	−0.06
XES_M	0.00*	−0.19*	0.03*	−0.25*	−0.06	−0.13	−0.01	−0.06
XTR_M								
XE_M	15.26*	−57.51*	7.86*	−72.64*	32.59	−44.34	20.11	−25.19
XSE	0.01*	0.77*	−0.10*	0.94*	0.23	0.57	0.11	0.45
XLC	0.03*	−0.14*	0.03*	−0.25*	−0.00	−0.01	0.07	0.06
$XC1_T$								
$XC2_T$	0.03*	−0.23*	0.03*	−0.23*	0.02*	−0.42*	0.04	−0.03
$XES1_T$								
$XES2_T$	0.03*	−0.24*	0.03*	−0.23*	0.03	−0.20	0.02	−0.29
XTR_T	−0.02	0.10	−0.01	0.12	0.02	0.17	−0.08	−0.04
XA_T	0.02*	−0.19*	0.02*	−0.22*	−0.01	−0.27	0.05	−0.02
$XMC1_T$	0.03	−0.07	0.00	−0.01	0.09*	−0.17*	0.05	−0.15
$XMC2_T$	−0.01	0.05	−0.03*	0.28*	0.02	−0.57	0.04	−0.00
$XHEC_T$	0.08*	−0.43*	0.08*	−0.71*	0.02	0.31	0.15	−0.34
XS_T	−0.02*	0.17*	−0.02	0.19	−0.01	0.38	−0.04	−0.11
XE_T	0.03*	−0.21*	0.02	−0.14	0.07*	−0.37*	0.03	−0.24
XH_T	−0.03*	0.20*	−0.04*	0.33*	−0.05	0.30	0.00	−0.29

Group lasso indices using weights based on the C.8 specification. Stars indicate mean differences which are significant at 5% significance level.

C.1.2 Figures

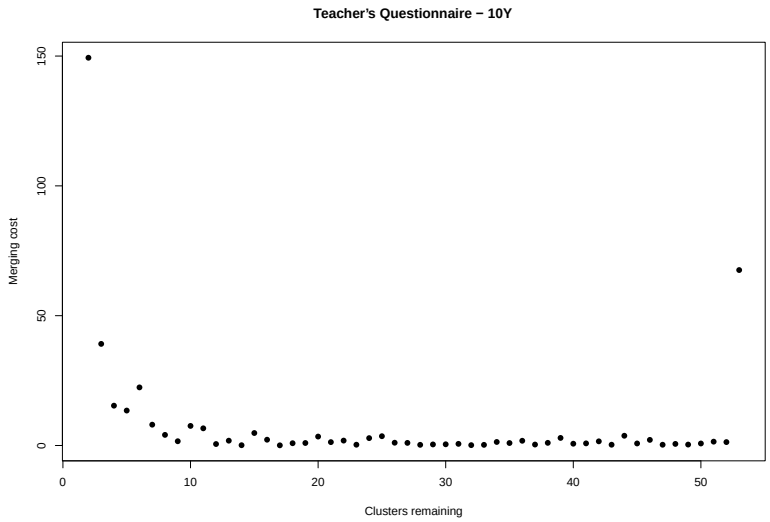
Cluster Analysis

Figure C.1.1: Merging Costs of K clusters - Mother's Questionnaire



Each point represents the merging costs of merging the 'Clusters remaining + 1' to 'Clusters remaining'. The only exception is the point at the very right which represents the within-cluster variance when no merging occured.

Figure C.1.2: Merging Costs of K clusters - Teacher's Questionnaire



Each point represents the merging costs of merging the 'Clusters remaining + 1' to 'Clusters remaining'. The only exception is the point at the very right which represents the within-cluster variance when no merging occured.

C.2 Additional Material

C.2.1 Identification of Group Lasso Weights

By estimating the following logit model under a group lasso penalty

$$\begin{aligned} Y_i &= \Lambda(\alpha + \beta_1 C_{i1} + \beta_2 C_{i2} + \cdots + \beta_K C_{iK_C} + OtherControls_i'\delta) + \varepsilon_i, \\ &= \Lambda(\alpha + IC_i^C \gamma_C + OtherControls_i'\delta) + \varepsilon_i, \\ \text{where } IC_i^C &= \sum_{l=1}^{K_C} \omega_l^C C_{il}, \end{aligned}$$

we can derive situation specific weights (here for the Conscientiousness index) under the summing up to one constraint $\sum_l \hat{\omega}_l = 1$ as follows:

$$\begin{aligned} \hat{\beta}_l &= \widehat{\gamma_C \omega_l}, \\ \sum_{l=1}^{K_C} \hat{\beta}_l &= \sum_{l=1}^{K_C} \widehat{\gamma_C \omega_l}, \\ \sum_{l=1}^{K_C} \hat{\beta}_l &= \hat{\gamma}_C, \qquad \text{summming up to one constraint} \\ \Rightarrow \hat{\omega}_l &= \frac{\hat{\beta}_l}{\sum_l \hat{\beta}_l}. \end{aligned}$$