Graduate School of Decision Sciences

Universität Konstanz

GSDS
Working Paper
No. 2017-02

We Belong Together - A Cross-Smoothing Approach for Non-overlapping Group Mean Estimates

Phillip Heiler
Jana Mareckova

January 2017

Graduate School of Decision Sciences

All processes within our society are based on decisions – whether they are individual or collective decisions. Understanding how these decisions are made will provide the tools with which we can address the root causes of social science issues.

The GSDS offers an open and communicative academic environment for doctoral researchers who deal with issues of decision making and their application to important social science problems. It combines the perspectives of the various social science disciplines for a comprehensive understanding of human decision behavior and its economic and political consequences.

The GSDS primarily focuses on economics, political science and psychology, but also encompasses the complementary disciplines computer science, sociology and statistics. The GSDS is structured around four interdisciplinary research areas: (A) Behavioural Decision Making, (B) Intertemporal Choice and Markets, (C) Political Decisions and Institutions and (D) Information Processing and Statistical Analysis.

GSDS – Graduate School of Decision Sciences
University of Konstanz
Box 146
78457 Konstanz

Phone: +49 (0)7531 88 3633
Fax: +49 (0)7531 88 5193
E-mail: gsds.office@uni-konstanz.de

-gsds.uni-konstanz.de

ISSN: 2365-4120
January 2017

We Belong Together - A Cross-Smoothing Approach for Non-overlapping Group Mean Estimates[*]

Phillip Heiler[†] Jana Mareckova

University of Konstanz University of Konstanz

January 17, 2017

We propose a general framework for estimating means of orthogonal groups stemming from e.g. categorical regressors or (quasi-)experimental data. It penalizes the loss function by adding squared L_2-norm differences between group location parameters and a first stage estimate for potentially all other groups. Under quadratic loss, the penalized estimation problem has a simple interpretable closed form solution that is related to methods established in the literature on discretized support smoothing kernels and model averaging methods. We provide optimal smoothing parameters that serves as a benchmark method, propose a plug-in approach and study their comparative statics. The behavior of both methods is analyzed in an asymptotic local to zero framework that allows for the presence of moderate test statistics for arbitrary sample sizes. We introduce a class of sequences for close and distant systems that is sufficient for describing a large range of data generating processes. We show consistency, derive the asymptotic distribution under fixed, theoretically optimal and estimated smoothing parameters and provide upper limits for the statistical complexity of the estimators. The method is applied to the estimation of time trends in a short panel based on the Haifa field experiment by Gneezy and Rustichini (2000a) and to the difference-in-differences minimum wage study by Card and Krueger (1994).

Keywords: Categorical data analysis, shrinkage, smoothing

JEL classification: ....

[*] Financial support by the Graduate School of Decision Sciences (University of Konstanz) is gratefully acknowledged. All remaining errors are ours.

[†] Graduate School of Decision Sciences, Universitätsstraße 1, D-78462 Konstanz, Germany. Phone: +49-7531-88-5287, email: phillip.heiler@uni-konstanz.de.

Contents

List of Figures

List of Tables

1. Introduction

Modern data sets such as (quasi-)experimental data, survey data or medical records typically contain many ordered categorical explanatory variables that can be used to estimate conditional mean functions. Without restricting the functional form of the conditional mean, one problem is having small or even empty groups sizes for distinct combinations of categorical predictors that form orthogonal groups or *cells*. In general, the question is how to appropriately setup a statistical model for the conditional mean of an outcome variable of interest in such cases. For simplicity, imagine a very simple example with two potential groups and a small sample. A fully saturated model would allow one parameter per group while a global model would yield the simple average over all observations. If means are truly identical, aggregating will yield a lower variance for parameter estimate and predictions since they are based on a larger sample size. However, the true means are unknown. A classical test would be a two sample t-test for equality of the means. Say, the researcher decides on a 5% confidence level. The test will lead to a zero-one decision based on the value of the test statistics. However, what if test sizes are "moderate", i.e. we have a t-statistics of 1.96 leading to a rejection of the equality of two means? Aggregating both groups might still be beneficial in terms of the predictive accuracy since the differences in the means are relatively small for a given sample size and bias introduced by the aggregation might be offset by reduction in variance. In addition, choosing a significance level creates a somewhat arbitrary discontinuity around the critical value. However, e.g. two t-statistics of 1.95 and 1.96 might stand for rather similar processes in finite samples. Therefore, considering smooth variants of aggregation can potentially be beneficial in finite samples.

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 (Hall, 1983; Simonoff, 1996) or for mixed data (Li and Racine, 2003). In the nonparametric regression framework Hall et al. (2004, 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.

For probability distribution functions there is also a literature on empirical Bayes with data driven shrinkage parameters under appropriate priors for multinomial data, see e.g. Fienberg and Holland (1973), Titterington and Bowman (1985) or Simonoff (1996) for a comprehensive review with particular focus on sparse asymptotics.

In the frequentist model averaging literature, aggregation across parameters or predictions of different models has a similar effect in the cell context. Methods proposed by e.g. Hjort and Claeskens (2003), Hansen (2007) and Liu (2015) implicitly aggregate different estimates for cell means and effectively smooth across estimates.

In the shrinkage literature, the smoothness between the coefficient estimates is enhanced by adding an L_1-norm penalty of all the pairwise differences to a loss function. The L_1-norm allows for partial and complete fusion of the groups depending on the choice of the smoothing parameter. Such an approach is e.g. implemented in Tibshirani et al. (2005) for linear models with ordered categorical data and in Tutz and Oelker (2016) for group-specific generalized linear models. The main difference to the other methods is that aggregation and estimation 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.

From a theoretical point of view, the direct or implicit aggregation that is introduced by all of these methods for regression models leads to the question of what an overall “good” aggregation rule is. A proper aggregation rule should be theoretically optimal under a reasonable metric and if possible allow for conducting valid asymptotic inference.

We propose a general framework for orthogonal groups that penalizes the loss function by adding squared L_2-norm differences between group location parameters and first stage estimates for potentially all other group location parameters in the model. First stage estimates can be cell means or other arbitrary consistent estimators. Adding a squared L_2-norm penalty has several advantages. Namely, under a quadratic loss, the penalized estimation problem has a simple interpretable closed form solution, which we call the pairwise cross-smoothing estimator (PCS) that is closely related to the methods established in the literature on discretized support smoothing kernels for categorical data and model averaging methods. Additionally, the maximum degree of smoothing flexibility can be achieved by introducing individual smoothing parameters for all squared L_2-norm differences.

One of the main questions is how to choose the smoothing parameters optimally. In the case of having exclusively categorical regressors, a closed form solution of the mean squared error (MSE) optimal smoothing parameters can be derived. This allows for an extensive study of their comparative statics. Additionally, they can be used as a benchmark method in future research. Since the MSE optimal smoothing parameters depend on unknown parameters, we propose a plug-in approach for their estimation. We further contribute to the literature by analyzing the behavior of both estimated smoothing parameters as well as estimator based on the estimated smoothing parameters in an asymptotic local to zero framework that allows for the presence of moderate test statistics for arbitrary sample sizes. We introduce a class of sequences for close and distant systems that is sufficient for describing a wide range of data generating processes. We show consistency and derive the asymptotic distribution of the PCS estimator under fixed, theoretically optimal and estimated smoothing parameters. In addition, we provide upper limits on the effective degrees of freedom of the linear operator associated with the PCS.

Monte Carlo evidence suggests that the theoretically optimal smoothing parameters outperform the other considered competitors by a large factor. In contrast to existing methods, the feasible PCS seems to be a more conservative but robust refinement over ordinary least squares which is more significant for smaller group sizes and closer systems of locations. In fact, we observe a uniformly dominant behavior of the PCS over the ordinary least squares.

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 well-known difference-in-differences study about the effect of minimum wages on employment by Card and Krueger (1994) illustrating potential applications.

Section 2 introduces the pairwise cross-smoothing model and its relation to conventional smoothing kernels and model averaging methods. Section 3 presents the MSE optimal smoothing parameters, discusses their small and large sample behavior and gives some preliminary results on the large sample behavior of the PCS under fixed and optimal smoothing. Section 4 introduces the local asymptotic framework, demonstrates the large sample properties of the feasible PCS estimator and discusses the degree of model complexity reduction inherent in the pairwise cross smoothing. Section 5 provides some Monte Carlo evidence. Section 6 and 7 contain the applications. Section 8 concludes.

2. Pairwise Cross-Smoothing

2.1. The Model

All the notation is based on scalars, sums and vectors. For the more matrix affine reader, consulting Appendix A might turn out to be useful. Consider independent and identically distributed data (Y_i, X_i), $i = 1, \dots, n$, where the vector X_i contains ordered and/or unordered discrete random variables. The discrete variables in X_i uniquely determine J orthogonal groups. For example, two binary discrete random variables determine four orthogonal groups. Let a vector D_i represent whether an observation i belongs to a group $j \in \{1, \dots, J\}$. In such a case, the j-th entry of the vector D_i contains one, $D_{ij} = 1$, and the rest of the entries is equal to zeros, $D_{ij'} = 0$ for all $j' \neq j$, i.e. $D_i \in \{e_j, 1 \leq j \leq J\}$, which together form the standard basis of $\mathbb{R}^J$.

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

$$Y_i = D_i'\mu + \varepsilon_i \tag{2.1}$$

with $\mu = (\mu_1, \dots, \mu_J)'$ and $E[\varepsilon_i | D_i] = 0$. Let $\hat{\mu} = (\hat{\mu}_1, \dots, \hat{\mu}_J)'$ be a consistent first-stage estimator for the conditional group means. We propose to estimate the model for the conditional mean of Y_i as a penalized least squares problem, i.e.

$$(\hat{\mu}_1^{PCS}, \dots, \hat{\mu}_J^{PCS}) = \underset{\mu_1,\dots,\mu_J}{\arg\min} \sum_{i=1}^{n} (Y_i - D_i'\mu)^2 + \sum_{j=1}^{J} \sum_{s \neq j} \lambda_{js} (\mu_j - \hat{\mu}_s)^2, \tag{2.2}$$

where λ_{js} are given smoothing parameters with $\lambda_{jj} = 0$ for all $j \in \{1, \dots, J\}$ and PCS stands for pairwise cross-smoothing since the penalties form hyperrectangles that geometrically overlap in $\mathbb{R}^{J-1}$. 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 $\hat{\mu}$.

Regarding the optimal choice of the smoothing parameters, the more informative group s is for group j, the larger the smoothing parameter λ_{js} should be and vice versa. In the special case of $\lambda_{js} = 0$ for all pairs (j, s), i.e. none of the groups uses information from the other groups, the optimization problem (2.2) becomes a standard OLS and $\hat{\mu}^{PCS}$ is equal to OLS group mean estimates. By setting one of the λ_{js} arbitrarily large, i.e. group s uses (full) information of the group j, $\hat{\mu}_j^{PCS}$ is shrunken to $\hat{\mu}_s$. For a fixed j, setting all λ_{js}'s to large values, i.e. all other groups are very informative for the group j, would make $\hat{\mu}_j^{PCS}$ to be shrunken to the mean of all $\hat{\mu}_s$ where $s \neq j$.

The uniqueness of the solution of (2.2) is guaranteed if $\sum_{s\neq k} \lambda_{ks} > -n_k$ for all $k \in \{1, \ldots, J\}$. For a complete derivation see Appendix B.1. Under this condition, the k-th group estimate is given by

$$\hat{\mu}_k^{PCS}(\Lambda_k) = \frac{n_k \bar{Y}_k}{n_k + \sum_{l\neq k} \lambda_{kl}} + \sum_{j\neq k} \frac{\lambda_{kj}\hat{\mu}_j}{n_k + \sum_{l\neq k} \lambda_{kl}}, \tag{2.3}$$

with $\Lambda_k = (\lambda_{k1}, \ldots, \lambda_{kJ})$ and $\bar{Y}_k$ denoting the sample mean of group k, $\hat{\mu}_j$ being the first stage estimate for group j and $n_k = \sum_{i=1}^n D_{ik}$. By looking at the k-th group estimate, we can illustrate how the penalty works. One can see that the k-th group location estimator is a linear combination of the different first-stage group estimates and its own cell mean.

Regarding $\hat{\mu}$, a possible choice is the linear (cell based) projection of Y_i on D_i, i.e. $\hat{\mu} = (\sum_{i=1}^n D_i D_i')^{-1} \sum_{i=1}^n D_i Y_i$, the vector of cell means. This is also referred to as frequency approach in the literature since it uses cell probabilities as weights. The k-th group estimator can be easily decomposed into:

$$\hat{\mu}_k^{PCS}(\Lambda_k) = \frac{n_k \bar{Y}_k}{n_k + \sum_{l\neq k} \lambda_{kl}} + \sum_{j\neq k} \frac{\lambda_{kj}\bar{Y}_j}{n_k + \sum_{l\neq k} \lambda_{kl}}, \tag{2.4}$$

which is a linear combination of the cell means. It is noteworthy that the smoothing parameters λ_{kj}'s and therefore also the implicit weights are not all restricted to be larger than or equal to zero. Just the overall smoothing for one baseline category can't be too negative based on the condition for the uniqueness of (2.3). This fundamentally differentiates our approach from e.g. discretized support kernel approaches that are built as weighted averages using probability mass functions. They lead to weights which are restricted to be larger than zero. We will come back to this point in the Section 2.2.

2.2. Relation to Kernel Estimators

The pairwise cross smoothing approach is directly linked to traditional smoothing kernels for binary data in the sense of Aitchison and Aitken (1976). These methods are originally intended to deal with the small to empty cell problem in the context of multivariate discrete (Hall, 1983; Simonoff, 1996) or mixed data (Li and Racine, 2003). They try to get reasonable estimates of the joint or conditional probability distributions of the discrete random variables. The kernel methods effectively smooth the probability estimates of the cells towards *close* cells, i.e. they smoothly aggregate information across the cells.

The direction of the results on the methods discussed above carries over to the nonparametric estimation of regression functions. Prominent sources concerned with discrete or mixed regressors are Li and Racine (2007), Hall et al. (2004, 2007) and Ouyang et al. (2009) who put particular emphasis on cross-validated smoothing parameters and their behavior under the presence of irrelevant regressors.

For the case of exclusively discrete data, Ouyang et al. (2009) show qualitatively different behavior under cross-validated smoothing parameters that cannot be replicated as a special case of the mixed data framework. They consider the case of arbitrary ordered or unordered discrete regressors $X_i = (X_{i1}, \ldots, X_{ir})'$ and a model

$$Y_i = g(X_i) + \varepsilon_i$$

with $E[\varepsilon_i|X_i] = 0$. The discretized support kernels for unordered regressors are defined by

$$l(X_{is}, x_s, \lambda_s) = \lambda_s^{\mathbb{1}(X_{is} \neq x_s)}$$

which for the general case of $s \in \{1, \ldots, r\}$ leads to a product kernel in the shape of

$$L(X_i, x, \lambda) = \prod_{s=1}^{r} l(X_{is}, x_s, \lambda_s) = \prod_{s=1}^{r} \lambda_s^{\mathbb{1}(X_{is} \neq x_s)} \tag{2.5}$$

with $\lambda_s \in [0, 1]$ for all s being the bandwidth parameters and $\lambda = (\lambda_1, \ldots, \lambda_r)$. The point $x = (x_1, \ldots, x_r)$ is a fixed point of interest. For ordered discrete data, they propose to implement the same principle but with exponentially decaying weights in the absolute differences between the discrete variables, i.e. putting higher weights on close cells or

Table 2.1: Kernel Weights: Example with $\lambda_1 = 0.3$ and $\lambda_2 = 0.7$

$(x_{i1}, x_{i2}) = (0, 0)$

$x_{j1} \vert x_{j2}$	0	1	2
0	1	0.7	0.49
1	0.3	0.21	0.15
2	0.09	0.06	0.04

$(x_{i1}, x_{i2}) = (1, 1)$

$x_{j1} \vert x_{j2}$	0	1	2
0	0.21	0.3	0.21
1	0.7	1	0.7
2	0.21	0.3	0.21

formally

$$\begin{aligned} l(X_{is}, x_s, \lambda_s) &= \lambda_s^{|X_{is}-x_s|} \\ L(X_i, x, \lambda) &= \prod_{s=1}^{r} \lambda_s^{|X_{is}-x_s|}. \end{aligned} \tag{2.6}$$

In the case of $\lambda_s \to 0$ for all s, the kernel will degenerate towards a cell based indicator function equivalent to the frequency based approach while for $\lambda_s \to 1$ the results will be shrunken completely towards the global mean. The resulting estimator for $g(x)$ is then given by

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i L(X_i, x, \lambda)}{\sum_{i=1}^{n} L(X_i, x, \lambda)}.$$

For illustration consider the case of a univariate regression model (i.e. $r = 1$) with $X_i \in \{0, 1, 2\}$. A smooth nonparametric estimate for $g(0)$ is then given by

$$\hat{g}(0) = \frac{\sum_{X_i=0} Y_i + \sum_{X_i=1} \lambda_1 Y_i + \sum_{X_i=2} \lambda_1^2 Y_i}{\sum_{X_i=0} + \sum_{X_i=1} \lambda_1 + \sum_{X_i=2} \lambda_1^2} \tag{2.7}$$

$$= \sum_{X_i=0} w_0 Y_i + \sum_{X_i=1} w_1 Y_i + \sum_{X_i=2} w_2 Y_i \tag{2.8}$$

with $w_j = \lambda_1^j / (\sum_{X_i=0} + \sum_{X_i=1} \lambda_1 + \sum_{X_i=2} \lambda_1^2)$ which is still a weighted average but over a larger range of cells than in the frequency approach.

Since the bandwidths are bounded between 0 and 1, more distant cells will automatically receive a lower weight. Table 2.1 illustrates this relationship for the case of two discrete variables taking on values from 0 to 2 for two different points of interest. The degree of smoothing depends on the smoothing parameter related to a specific covariate and is independent of the group that is considered as a target. Cell distances are not smoothed differently *within* one direction as illustrated by the second table in Table 2.1. This approach puts a particular, i.e. exponential structure on the kernel decay over the support of the ordered discrete variables. A somewhat more flexible approach would be to recode the two ordered regressors into nine orthogonal groups and allow for a separate smoothing parameter for each category. This is still in line with the product kernel idea. An even more flexible approach would introduce different smoothing parameters depending on the target category. This is what PCS allows for.

Therefore, kernel estimators introduced in the literature can be seen as restricted versions of PCS. This is also illustrated by equations (2.4) and (2.8). Equation (2.8) is very reminiscent of the closed form solution of the PCS when cell based projections serve as first stage estimates in equation (2.4). In particular, for the case of only binary regressors or for transformed data that consists of non-overlapping groups only, there exists a one to one relation between PCS and a traditional smoothing kernel, i.e. the latter one is a restricted version of PCS. By restricting $\lambda_{kj} = \lambda_j n_j$ for all $k \neq j$, one obtains the traditional smoothing kernel. Hence, for the case of orthogonal groups there exists a penalized regression representation in the fashion of equation (2.3) that has the identical solution as the smoothing kernels if cell based means are used as first stage estimators.

2.3. Relation to Model Averaging

The question of how to aggregate across distinctive groups can also be rephrased from a model or variable selection perspective, i.e. which group deserves its own location parameter and which groups can be put into one? Hence 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. There is a large and growing literature on model selection and model averaging. At first, note that the similarities between model averaging and the pairwise cross smoothing estimator mainly come from the fact that for models containing location parameters only, predictions and parameters are the same notion. Here, we focus on the papers in the frequentist model averaging literature that are more closely related to this work, i.e. we omit relevant contributions in the literature especially in the context of Bayesian model averging, forecast combination, instrumental variables and others. For a comprehensive list of relevant papers in these fields, we refer the reader to Section 1 of Hansen (2014).

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,

1970) or Schwarz-Bayes Criterion (BIC, Schwarz, 1978) or traditional multivariate testing procedures. More recently, frequentist model averaging methods have become more popular. 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 different least squares estimates based on a 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. Hansen and Racine (2012) develop a jacknife 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 under penalization of the averaging weights in a local to zero $n^{-1/2}$ framework. 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, i.e. weak partial correlations of additional regressors beyond a correctly specified base model. He shows the nonstandard behavior of the averaging estimators and proposes alternative procedures for inference. In the context of point estimation risk, Cheng et al. (2016) 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 dominates the asymptotic risk of the base estimator uniformly over all degrees of misspecification. Their results indicate a more robust increase in performance than classical pretesting for overidentification conditions.

Papers that contain mean squared error optimal plug-in weights such as Hjort and

Claeskens (2003), Liu (2015) and Cheng et al. (2016) provide closed-form solutions for the optimal weights in the case of two models only. Beyond that, there are no closed-form solutions in general, i.e. solutions can only be obtained by numerical optimization. It turns out that there is a one-to-one correspondence between their solution and the PCS estimator in the case of two groups. We contribute to the literature by exploiting the orthogonal structure of our data to achieve an interpretable closed-form solution for an arbitrary number of groups. This allows for a comprehensive discussion of the comparative statics with respect to the relevant population and sample information and has negligible computational costs. We adopt some of the ideas from the local asymptotic framework to be flexible with respect to the data generating process. In contrast to weak partial correlations, we consider *close* and *distant* systems of location parameters in Section 4.3.

Additionally to the summability constraint, it is common in the model averaging literature (Hansen and Racine, 2012; Liu, 2015) to restrict the weights to lie in the unit simplex. Li (1987) and Hansen and Racine (2012) argue that for admissibility of the averaging estimator in the linear case all eigenvalues of the corresponding projection matrix have to lie in the unit interval. Since the submodel projection matrices have eigenvalues in $[0, 1]$, positivity of the weights is sufficient for admissibility. Hansen and Racine (2012) further show than in the case of nested linear regression models, positivity is a necessary condition for admissibility under mean squared error loss.

One might ask, if there is a one-to-one correspondence between the PCS approach and conventional model averaging methods why do the results on admissibility not apply here? In particular, positive smoothing parameters should be obtained by admissibility under MSE loss since the submodels are effectively linear. Consider the simple case of three groups. If one chooses the nested sequence of models that yields predictions $\bar{Y}_1$, $\bar{Y}_{12}$ and $\bar{Y}$ where the subscript denotes the groups used for the average[1]. By rewriting the averaging estimator for $\hat{\mu}_1$ with model weights $\{(1-\omega_1-\omega_2), \omega_1, \omega_2\}$, one obtains that the following

[1] Note that any other arbitrary combination could be chosen here as long as they are distinguishable by one group per nesting step.

must hold between the model averaging weights and the PCS smoothing parameters:

$$\begin{aligned} \omega_1\left(\frac{n_2}{n_1+n_2}\right) \ + \ \omega_2\left(\frac{n_2}{n}\right) &= \omega_{12}, \\ \omega_2\left(\frac{n_3}{n}\right) &= \omega_{13} \end{aligned}$$

with $\omega_{kj} = \lambda_{kj}/(n_k + \sum_{l\neq k}\lambda_{kl})$. However, since there are only two weights overall in the classical model averaging framework, it must also hold $\omega_{12} = \omega_{32}$ and $\omega_{13} = \omega_{23}$ which clearly adds additional restrictions to the optimization problem of the MSE. The difference is that in the model averaging context, the general outcome for all observations is considered while we focus on group specific predictions. The difference can also be seen through the number of smoothing parameters that are $J-1$ in the model averaging and $(J-1)^2$ in the PCS framework. Thus, only within the more restrictive averaging framework positivity is a sufficient and necessary condition for admissibility and hence there is no contradiction between the results from the literature and the optimal solutions for the PCS.

Naturally, under some data generating processes, the restrictions imposed by the more constrained averaging method are actually helpful to stabilize the estimated weighting parameters such that the resulting estimator might perform better in finite samples. We discuss some of these issues in Section 3.2 and Section 5.

3. Optimal Aggregation

3.1. Mean Squared Error Optimal Smoothing

To derive MSE optimal smoothing parameters, we first rewrite (2.3) as:

$$\hat{\mu}_k^{PCS}(\Lambda_k) = (1 - \sum_{j\neq k}\omega_{kj})\bar{Y}_k + \sum_{j\neq k}\omega_{kj}\hat{\mu}_j \tag{3.1}$$

with $\omega_{kj} = \lambda_{kj}/(n_k + \sum_{l\neq k}\lambda_{kl})$. The k-th group representation (3.1) will allow us to get MSE optimal finite sample smoothing parameters as a function of the true group means and the variance of the error terms. We choose the predictive mean squared error as optimality criterion for the quality of the estimator. Note that due to the diagonal structure of the gram matrix and the base category dependence of the smoothing parameters, minimization of the predictive MSE is equivalent to minimization of the parameter MSE, see also Appendix C.1.

Let $E[Y_i|D_{ij}=1]=\mu_j$, $V[\varepsilon_i|D_{ij}=1]=\sigma_j^2$ be finite moments and $\Delta\mu_{kj}\equiv\mu_k-\mu_j$. From now on, assume that the first stage estimator is a linear projection which implies zero covariances across the elements of $\hat{\mu}_k^{PCS}$ by construction. Under these assumptions, one can derive a leading term of a first order approximation of the MSE using a first stage estimator which is based on random sampling over the different cells[2]. The approximation can be seen as the exact finite sample MSE in the case of fixed regressors, i.e. deterministic selection into cells.

Proposition 3.1 *Under the assumptions given above, the leading term of the MSE of $\hat{\mu}_k(\Lambda_k)$ is then given by*

$$\begin{aligned} MSE(\hat{\mu}_k^{PCS}(\Lambda_k)) &= bias(\hat{\mu}_k^{PCS})^2 + V[\hat{\mu}_k^{PCS}] \\ &= \Big(\sum_{j\neq k}\omega_{kj}(\Delta\mu_{kj})\Big)^2 + (1-\sum_{j\neq k}\omega_{kj})^2\frac{\sigma_k^2}{np_k} + \sum_{j\neq k}\omega_{kj}^2\frac{\sigma_j^2}{np_j}. \end{aligned} \tag{3.2}$$

with $p_j=P(D_{ij}=1)>0$, $j=1,\dots,J$.

Proof: The proof can be found in the Appendix C.2. ■

Theorem 3.1 *The MSE in (3.2) is minimized at*

$$\omega_{kj}^* = \frac{\sigma_k^2 p_j/\sigma_j^2 p_k}{a_{kj}} \tag{3.3}$$

or equivalently

$$\lambda_{kj}^* = \frac{\sigma_k^2 np_j/\sigma_j^2}{a_{kj}-\sum_{l\neq k}\frac{\sigma_k^2 p_l a_{kj}}{\sigma_l^2 p_k a_{kl}}} \quad \textit{for all } k\neq j, \quad \lambda_{kk}^*=0 \tag{3.4}$$

where $a_{kj}=\left(1+\frac{\sigma_k^2/np_k}{1+b_{kj}}\sum_{l\neq k}\frac{1+b_{kl}}{\sigma_l^2/np_l}+\frac{\Delta\mu_{kj}}{1+b_{kj}}\sum_{l\neq k}\frac{\Delta\mu_{kl}}{\sigma_l^2/np_l}\right)$ *and* $b_{kj}=\sum_{m\neq k}\frac{\Delta\mu_{km}\Delta\mu_{jm}}{\sigma_m^2/np_m}$.

Proof: The proof can be found in the Appendix C.3. ■

Note that the results in Theorem 3.1 are for any $n\in\mathbb{N}$. The uniqueness of (3.3) for finite n is established in A.2. For $n\to\infty$, a unique solution exists only for a case of two groups with unequal means. Otherwise, the (3.2) has weak minimizers. See A.2 for the complete proof. Intuitively, one can see that once the variance stops contributing to the

[2] $\hat{\mu}_k=\mu_k+\frac{1}{np_k}\sum_{i=1}^n D_{ik}\varepsilon_i+O_p(n^{-1})$.

MSE (i.e. $n \to \infty$), there are more ways how to achieve the same values of a squared bias for more than two groups and two groups with the same mean. The uniqueness of the solution stems from a unique trade-off between the bias and the variance which appears in finite samples or in the special case of two unequal means.

3.2. Comparative Statics in Small Samples

For a qualitative discussion, the small sample behavior of the smoothing parameters will be analyzed under different DGPs. The following framework of 4 groups $\{1, 2, 3, 4\}$ is considered. Group 1 is chosen as a base category, i.e. the group whose mean is shrunken to the other groups. Therefore, the following MSE optimal smoothing parameters (weights)[3]: ω_{11}, ω_{12}, ω_{13} and ω_{14} will be analyzed 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 the group means: $\Delta\mu_{12}$, $\Delta\mu_{23}$ and $\Delta\mu_{13}$, since the smoothing parameters depend on the differences in means, not on the levels themselves, see (3.3).

Table 3.1 contains the parameter values for all potential design combinations. Only design (A) allows to disentangle the effects of individual inputs from the closed form solution of ω_{kj} directly. The solutions under the designs (B) and (C) are rather convoluted and therefore the effects are illustrated graphically. For the sake of brevity, only the most important results are mentioned. All the other graphs and discussions can be found in Appendix C.4.

Table 3.1: Design Values

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

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

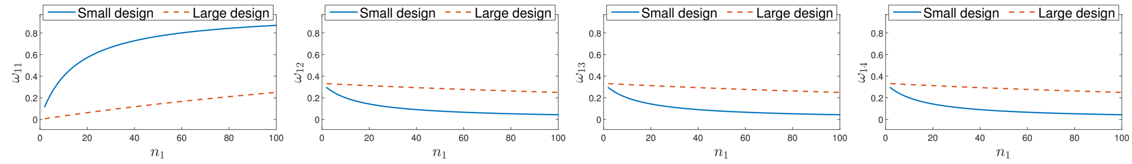
Equal Means Design (A) In the design (A), we analyze the effect of number of observations under both types of error variances and the effect of variance under heteroscedasticity in a small and large sample design.

The closed form solution for the MSE optimal smoothing parameters under equal means boils down to

$$\omega_{kj} = \frac{n_j \prod_{l \neq j} \sigma_l^2}{\sum_{m=1}^{4} n_m \prod_{l \neq m} \sigma_l^2}. \tag{3.5}$$

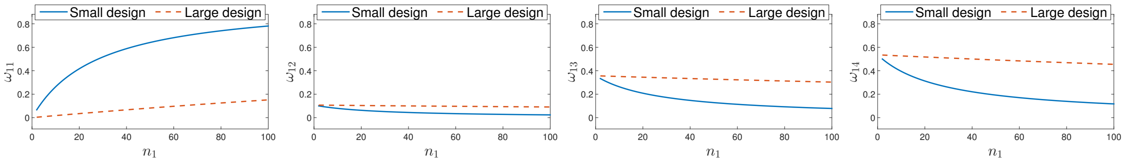
Keeping the other parameters constant, the smoothing parameter ω_{kj} increases in n_j and decreases in n_m where $m \neq j$. In other words, the mean estimate of the group with more observations is relatively more informative and therefore can stabilize the estimates of the other groups with the same mean by getting a higher smoothing weight, see Figure 3.1. The other smoothing weights then decrease correspondingly. The degree of the changes in ω_{kj} depends on the overall number of observations in the sample. In a small sample design, additional observations play an important role for the shrinkage intensity and help to smooth the mean of base category towards the largest group strongly, while the other intensities become relatively less informative. In a large sample design, the additional observations do not have much importance, since the whole system is already stabilized towards large groups, however more observations still lead to an increase in shrinkage intensity. The effects for n_2 and n_3 are qualitatively the same and can be found in Figure C.1 in Appendix C.4.

Figure 3.1: Effect of n_1 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Equal Means, Homoscedasticity



When changing the number of observations under heteroscedasticity, the qualitative results do not change. Quantitatively, target groups with larger variances have lower smoothing levels, see Figure 3.2. Note that under equal means in a small sample design, the number of observations is more important to get a higher smoothing weight in comparison to a large design where the variance is the more important factor. Compare group 1 and group 4 in Figure 3.2: Group 1 gets a high weight in a small sample design and group 4 gets a very high smoothing weight in a large sample design. This implies that a small variance can be exploited well only when the number of observations in the group is reasonably large.

Figure 3.2: Effect of n_1 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Equal Means, Heteroscedasticity



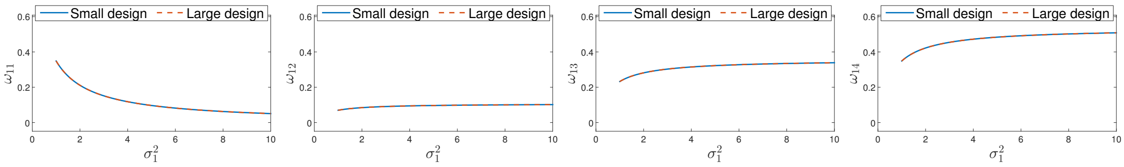
Regarding the variance, the following results can be derived:

$$\frac{\partial \omega_{kj}}{\partial \sigma_k^2} = \frac{n_k n_j \prod_{l \neq \{k,j\}} \sigma_l^2 \prod_{l \neq k} \sigma_l^2}{[\sum_{m=1}^4 n_m \prod_{l \neq m} \sigma_l^2]^2} \qquad \text{for } j \neq k, \tag{3.6}$$

$$\frac{\partial \omega_{kj}}{\partial \sigma_j^2} = -\frac{n_j \prod_{l \neq j} \sigma_l^2}{[\sum_{m=1}^4 n_m \prod_{l \neq m} \sigma_l^2]^2} \left(\sum_{m \neq j} n_m \prod_{l \neq \{k,m\}} \sigma_l^2 \right). \tag{3.7}$$

Given the non-negativity of n and σ^2, the ω_{kk} decreases in its own variance σ_k^2. Meanwhile, the other smoothing parameters ω_{kj} increase in σ_k^2 and decrease in σ_j^2. Intuitively, groups with lower variance can provide more precise mean estimates and therefore the smoothing towards them is higher and the other weights decrease correspondingly, see Figure 3.3 and the plots in Figure C.3 in Appendix C.4. At high variance levels, smoothing weights stay relatively stable as can be seen from (3.6) and (3.7) by taking limits with $\sigma_k^2 \to \infty$ and $\sigma_j^2 \to \infty$. This stability comes from the fact that the group with a large variance is basically ignored at large levels of variance by getting an almost zero weight and the information for the mean estimate is taken from more informative groups. The lower the variance of the target group, the higher smoothing weight the group gets. Since group 4 has always the lowest variance, it has the largest smoothing weight. On Figure 3.3 one can see that the results are independent of the sample size. That is because the number in all groups are scaled by the same constant and by (3.5), the values coincide in this case.

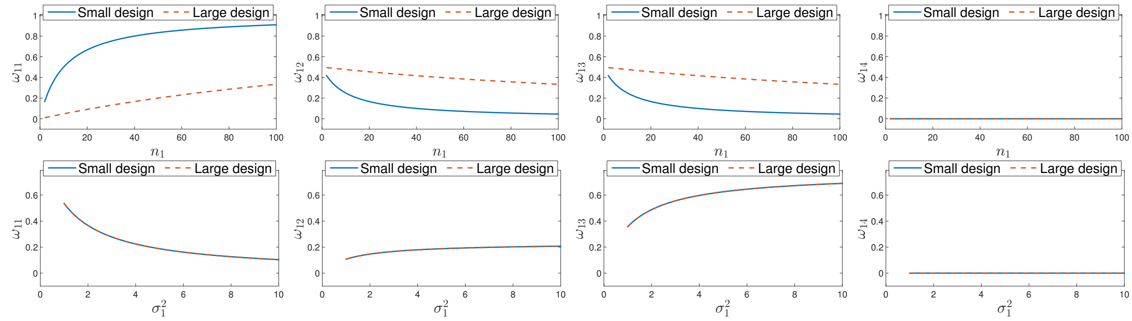
Figure 3.3: Effect of σ_1^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Equal Means, Heteroscedasticity



Unequal Means Design (B) The mean design (B) represents a situation of three equal means and one distant mean, in which the effect of number of observations and effect of the mean differences were analyzed under both types of variances and the effect of variance was added for the heteroscedasticity in small and large sample design.

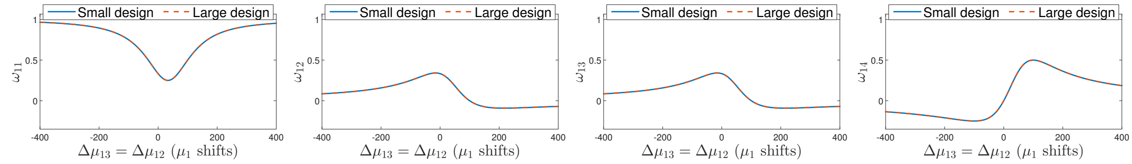
Unless one of the means is shifted, group 1 is basically not smoothed towards group 4, see Figure 3.4. It is not surprising that the large distance between μ_4 and the other group means decreases the smoothing towards the distant mean considerably. As group 4 is ignored in the smoothing scheme, the changes 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 groups in a similar manner as in the design (A). The only difference is the higher level of the weights for these three groups, since group 4 is not contributing. For a more detailed discussion of the effects of error variances and number of observations consider Appendix C.4.

Figure 3.4: Effect of n_1 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means, Homoscedasticity and Effect of σ_1^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means, Heteroscedasticity



The analysis of mean differences reveals that under homoscedasticity the impact on the smoothing parameters depends on: (1) which group mean changes, (2) the size of the mean difference and (3) the presence of another close group mean, see Figure 3.5. Introducing heteroscedasticity does not alter the results qualitatively, see Appendix C.4.

Figure 3.5: Effect of $\Delta\mu_{12}$ and $\Delta\mu_{13}$ on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means, Homoscedasticity

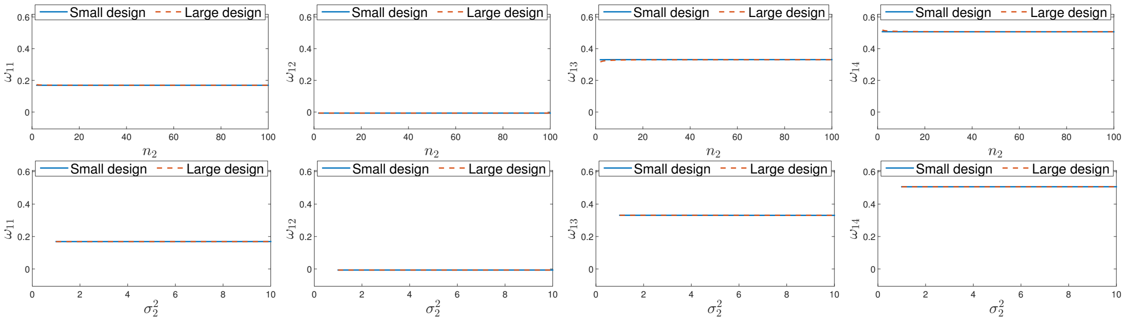


The presence of another close mean affects the shapes of the curves. The main effects are illustrated shifting μ_1. The effects of shifting the μ_2 and μ_3 follow a similar logic and are discussed in Appendix C.4. Shifting μ_1 far away from all the other group means causes the most weight to be put into its own smoothing weight ω_{11} as there is no other close group mean to which it would be sensible to smooth. However, when $\mu_1 \in [0, 100]$, the ω_{11} is the lowest as there are other close means towards which it pays off to smooth, see the peaks of ω_{12} and ω_{13} at 0 (i.e. $\mu_1 = 0$) and ω_{14} at 100 (i.e. $\mu_1 = 100$). Shifting μ_1 from 0 towards negative values causes a milder decrease in smoothing weights towards group 2 and 3 in comparison to group 4, since there is no other close group mean for μ_2 and μ_3 to compete with as μ_4 is even further. Shifting μ_1 towards positive values causes a steeper decrease in smoothing weights towards groups 2 and 3 until μ_1 reaches 100, since μ_4 is a relatively close group mean for μ_1 and therefore ω_{14} increases as it competes with groups 2 and 3 for a smoothing weight. Beyond $\mu_1 = 100$, ω_{14} decreases as μ_1 is getting further from μ_4. Notice that for large negative mean differences in $\Delta\mu_{12}$ (when μ_1 shifts) the smoothing weights towards groups 2, 3 and 4 are close to zero or even in negative values, because their group means are very far from μ_1. For large positive differences, groups 2 and 3 are ignored by having a very low (even negative) smoothing weight, while group 4 still has a positive smoothing weight as μ_4 is the closest mean to μ_1.

Unequal Means Design (C) The mean design (C) represents a situation of three close means (2 of them are equal to each other) and one distant mean and heteroscedasticity in small and large samples.

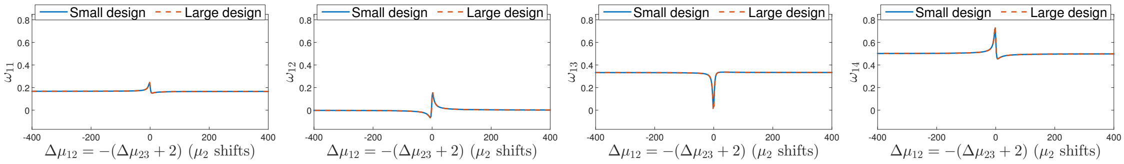
In this design, group 2 gets almost zero or even slightly negative smoothing weight because μ_2 is too far from μ_1 in comparison to the other means. An increase in n_2 or σ_2^2 has therefore almost no effect on the smoothing weights, see Figure 3.6. The reason is that the smoothing weights decide to ignore group 2 because the $\Delta\mu_{21}$ is so large that it is simply not sensible to smooth no matter how big the group is or how small the variance is. As a consequence, all the smoothing weights stay stable because any change in n_2 or σ_2^2 is simply not taken into account. Since μ_1, μ_3 and μ_4 are very close to each other, group 1 is smoothed to them. The shrinkage intensity then depends on the number of observations and the error variance. 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. it is almost as being in the equal mean design situation, see Figures C.9 and C.10 in Appendix C.4 for n_1, n_3, σ_1^2 and σ_3^2.

Figure 3.6: Effect of n_2 and σ_2^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means, Heteroscedasticity



Regarding the effect of mean differences, the results depend again on: (1) which group mean changes, (2) the size of the mean difference, (3) the presence of another close group mean and (4) error variances, see Figure 3.7 for shifting μ_2 and graphs in Figure C.11 in the Appendix C.4 for shifting μ_1 and μ_3. The effect of shifting μ_1 are 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". The same happens in the case of shifting μ_3. Then we are directly back in design (B) in which group 2 plays a role of a "distant group mean".

Figure 3.7: Effect of $\Delta\mu_{12}$ and $\Delta\mu_{23}$ when μ_2 shifts on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Unequal Means, Heteroscedasticity



Since groups 1, 3 and 4 have very close means, shifting μ_2 to more extreme values than -2 or 4 makes μ_2 to be a distant group mean, i.e. the smoothing towards group 2 is very close to zero or even slightly negative out of the [-4,2] interval for $\Delta\mu_{12}$. Within this interval, we can see relatively big changes in the smoothing parameters. The logic behind the behavior of the smoothing parameters is similar to design (B) only on a smaller scale, since there is no large stabilizing mean. For a detailed description consider Appendix C.4.

As mentioned above, in the case when all the means are very close to each other the MSE optimal smoothing parameters are changing their values sharply in a narrow interval covering the close distance between the means. Once, there is one distant mean, the smoothing parameters stabilize around certain levels depending on the distances to the other means and their error variances. This behavior could potentially cause problems for any estimate of the smoothing parameter that is subject to small sample variation. Finite samples deviations from the true parameters might yield smoothing parameters far away from the optimal ones leading to unfavorable aggregation.

3.3. Large Sample Properties with Fixed and MSE Optimal Smoothing

To learn more about the large sample behavior of the MSE optimal smoothing parameters and the corresponding PCS estimator, one can establish the following propositions:

Proposition 3.2 *If the optimal smoothing parameters according to (3.3) or (3.4) are chosen, then*

$$MSE(\hat{\mu}_k^{PCS}(\Lambda_k^*)) = O(n^{-1}) \tag{3.8}$$

with $\Lambda_k^* = (\lambda_{k1}^*, \lambda_{k2}^*, \dots, \lambda_{kJ}^*) \in \mathbb{R}^J$ *being the MSE optimal smoothing parameters.*

Proof: The proof can be found in the Appendix C.5. ■

Together with closed form of the theoretical MSE, one can establish a rate for the theoretical optimal smoothing parameters. In fact, one obtains that $\lambda_{kj}^* = O(n)$ and $\omega_{kj}^* = O(1)$, see the conclusion in Appendix C.5. This means that asymptotically the MSE optimal smoothing parameters do not vanish in general. Hence there is potential aggregation even in the limit. This is qualitatively different from the smoothing kernel approach where informative, i.e. conditionally independent, regressors are smoothed to a global average with a smoothing parameter converging to its upper bound.

The result (3.8) implies that the MSE optimal PCS estimator is consistent. In fact, one can establish two generic asymptotic normality results under fixed and MSE optimal smoothing.

Theorem 3.2 *If the smoothing parameters are fixed, i.e.* Λ_k *does not vary with sample size,* $n_k/n \xrightarrow{p} p_k > 0$ *for all* k *and* $\hat{\mu}$ *is a linear (cell based) projection, then*

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\Lambda_k) - \mu_k + B_k(\Lambda_k)) \xrightarrow{d} N\left(0, \frac{\sigma_k^2}{p_k}\right) \tag{3.9}$$

with $B_k(\Lambda_k) = \sum_j^J \omega_{kj} \Delta\mu_{kj}/\sqrt{n} = O(n^{-1/2})$.

Proof: The proof can be found in the Appendix A.3. ■

Hence, we see that $\hat{\mu}_k^{PCS}(\Lambda_k)$ is asymptotically normally distributed and we know an analytic expression for the small sample bias under fixed smoothing parameters. Note that similar to ridge regression with fixed tuning parameter, the asymptotic variance is equivalent to the OLS solution under heteroskedasticity of unknown form[4].

Theorem 3.3 *If the optimal smoothing parameters according to (3.4) or (3.3) are chosen and* $n_k/n \xrightarrow{p} p_k > 0$ *for all* k *and* $\hat{\mu}$ *is a linear (cell based) projection, then*

$$\sqrt{n}(\hat{\mu}_k^{PCS}(\Lambda_k^*) - \mu_k + B_k(\Lambda_k^*)) \xrightarrow{d} N\Big(0, \sum_{j=1}^{J} \bar{\omega}_{kj}^2 \frac{\sigma_j^2}{p_j}\Big) \tag{3.10}$$

with $B_k(\Lambda_k^*) = \sum_j^J \omega_{kj}^* \Delta\mu_{kj}/\sqrt{n} = O(n^{-1/2})$ *and* $\bar{\omega}_{kj} = \lim\limits_{n\to\infty} \omega_{kj}^*$.

Proof: The proof can be found in the Appendix A.3. ■

By construction, this estimator is infeasible since it depends on true population quantities.

4. Plug-In Estimation and Large Sample Properties

4.1. Plug-In Estimation

To get a feasible estimator, we propose plug-in estimation of the MSE optimal smoothing parameters, i.e.

$$\hat{\omega}_{kj} = \frac{\hat{\sigma}_k^2 n_j/\hat{\sigma}_j^2 n_k}{\hat{a}_{kj}} \quad \text{for all } k \neq j, \tag{4.1}$$

or equivalently

$$\hat{\lambda}_{kj} = \frac{\hat{\sigma}_k^2 n_j/\hat{\sigma}_j^2}{\hat{a}_{kj} - \sum_{l\neq k} \frac{\hat{\sigma}_k^2 n_l \hat{a}_{kj}}{\hat{\sigma}_l^2 n_k \hat{a}_{kl}}} \quad \text{for all } k \neq j, \quad \hat{\lambda}_{kk} = 0 \tag{4.2}$$

where $\hat{a}_{kj} = \Big(1 + \frac{\hat{\sigma}_k^2/n_k}{1+\hat{b}_{kj}} \sum_{l\neq k} \frac{1+\hat{b}_{kl}}{\hat{\sigma}_l^2/n_l} + \frac{\Delta\hat{\mu}_{kj}}{1+\hat{b}_{kj}} \sum_{l\neq k} \frac{\Delta\hat{\mu}_{kl}}{\hat{\sigma}_l^2/n_l}\Big)$, $\hat{b}_{kj} = \sum_{m\neq k} \frac{\Delta\hat{\mu}_{km}\Delta\hat{\mu}_{jm}}{\hat{\sigma}_m^2/n_m}$, $\Delta\hat{\mu}_{kj} = \hat{\mu}_k - \hat{\mu}_j$ and $\hat{\sigma}_k^2 = \frac{1}{n_k-1}\sum_{i=1}^n D_{ik}(Y_i - \hat{\mu}_k)^2$.

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. (2016).

[4] In the case of orthogonal group regressors unknown heteroskedasticity is equivalent to grouped heteroskedasticity.

4.2. Local Parameterization

In the following, we derive and discuss the fundamental properties of the PCS estimator under estimated weights, i.e. we address their connection to the infeasible theoretically MSE optimal smoothing estimator and the uniform behavior over a sufficient class of data generating processes. In particular, we would like to distinguish between systems of locations in which the differences are *small* for a given sample size (close systems) and where differences are *large* (distant systems). Formally,

$$\begin{aligned} &\text{(close systems)} \quad \sqrt{n}(\mu_k - \mu_j) \to \delta_{kj} \in \mathbb{R} \quad \forall k, j \text{ or } \sqrt{n}\Delta \to \delta \in \mathbb{R}^{J^2} \\ &\text{(distant systems)} \quad ||\sqrt{n}(\iota_J \mu_k - \mu)|| \to \infty \quad \forall k \text{ or } ||\sqrt{n}\Delta|| \to \infty \end{aligned}$$

with $\Delta = (\mu_1 - \mu_1, \mu_1 - \mu_2, \ldots, \mu_J - \mu_J)$ and $\delta = (\delta_{11}, \delta_{12}, \ldots, \delta_{JJ})$. Note that close systems require that all scaled pairwise differences do not diverge, i.e. their differences depend on the local parameters δ_{kj} while for the second specification it is sufficient if just one group in the system is different from the rest in the limit.

To further motivate these classes of sequences and in particular the rate of the local parameter differences, consider a system of J different locations that are estimated via least squares. Assume that the asymptotic variances are known. Let Z be a random variable which obeys a classical central limit theorem. A simple test for equality of two means μ_k and μ_j can be rewritten as follows:

$$T = \sqrt{n}\frac{\hat{\mu}_k - \hat{\mu}_j}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} = \sqrt{n}\frac{\mu_k - \mu_j}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} + Z + O_p(n^{-1/2})$$

$$\begin{cases} \to \infty, & \text{if systems are distant,} \\ \xrightarrow{d} \mathcal{N}\Big(\delta_{kj} \Big/ \sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}, 1\Big), & \text{if systems are close.} \end{cases}$$

Therefore, depending on the local parameter δ_{kj}, one can obtain a small, moderate or even large mean of the test statistics' distribution. In the special case of δ_{kj} being exactly equal to zero, the local parameterization does not longer affect the asymptotic distribution and standard 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 distribution around the critical value used for rejection of the null hypothesis. The PCS estimator can be considered as a smooth variant of a classical pre-testing based estimator. Hence, we expect it to perform better exactly in these *medium* type of regions in which the test has both a rather high type-I and type-II error. So far, this intuition only applies to a single pairwise difference.

For a more general statement, consider a Wald statistics for equality between all possible pairs in the system. Again, we assume knowledge of the variances. Let Z_{kj} be a random variables which obey a classical central limit theorem for all k, j. Under abuse of notation, the test statistics can be rewritten as follows:

$$W = n\sum_k \sum_{j>k} \frac{(\hat{\mu}_k - \hat{\mu}_j)^2}{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}} = \sum_k \sum_{j>k} \left(\frac{n(\mu_k - \mu_j)^2}{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}} + 2\frac{\sqrt{n}(\mu_k - \mu_j)}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} Z_{kj} + (Z_{kj})^2 \right) + O_p(n^{-1/2})$$

$$\begin{cases} \to \infty, \text{ if systems are distant,} \\ \xrightarrow{d} \sum_k \sum_{j>k} \left(\frac{\delta_{kj}^2}{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}} + 2\frac{\delta_{kj}}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} \mathcal{N}(0,1) + \mathcal{X}_1^2 \right), \text{ if systems are close.} \end{cases}$$

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 test statistics can be very different from the classical critical value. As in the case of two groups, one can expect that for moderate sizes of the local parameters, a test will reject the equality of means even if they are identical in the limit. Therefore, the local asymptotic framework allows for a better representation of the finite sample behavior, in particular when test statistics are *moderate*.

The specification for close and distant systems is somewhat reminiscent of Cheng et al. (2016) sequences for locally misspecified models up to order $n^{-1/2}$ and severely misspecified models for the sequences of the second type. Misspecification in their context refers to the validity or, in the intermediate case, to the speed of convergence of valid moment conditions. In this paper, it simply describes setups with at least one location that is either globally different from the rest or the intermediate case, in which it is approaching the others at a slower than $n^{-1/2}$ speed. Again, there is a one-to-one correspondence in the case of two groups, i.e. the overidentified model is a moment condition yielding the global average while the base model only consists of the group specific means. Hence, misspecification in this context refers to differences in mean locations along drifting sequences of data generating processes.

4.3. Estimated Optimal Smoothing Parameters

Before considering the asymptotic behavior of the PCS estimator under drifting sequences of parameters, we first establish some basic convergence results for the plug-in smoothing parameters. Note again that the mean squared error minimization for $\hat{\mu}_k^{PCS}$ depends only on $\omega_{kj}, j = 1, \dots, J$. For any $n \in \mathbb{N}$, minimization of the sample equivalent to (3.2) is a quadratic optimization problem with a positive definite hessian and hence convex. It is useful to consider the following form of $\hat{\omega}_{kj}$ (see Appendix C.6 for the derivation):

$$\hat{\omega}_{kj} = \left[\frac{Z'\hat{m}_1 Z + \sqrt{n}\Delta'\hat{m}_2 Z + n\Delta'\hat{m}_3\Delta + \hat{c}_0}{Z'\hat{M}_1 Z + \sqrt{n}\Delta'\hat{M}_2 Z + n\Delta'\hat{M}_3\Delta + 1} + 1 \right]^{-1} \frac{n_j \hat{\sigma}_k^2}{n_k \hat{\sigma}_j^2} \tag{4.3}$$

with Z being a vector of random variables which converge to a standard normal distribution and $\hat{m}_1, \hat{m}_2, \hat{m}_3, \hat{M}_1, \hat{M}_2, \hat{M}_3, (\hat{c}_0)$ being random matrices (scalar) that converge(s) in probability, i.e. they depend only on ratio of cell observations and estimated as well as true cell variances. Equation (4.3) allows us to study the behavior of the plug-in estimator under different sequences of data generating processes.

1. Case: Locally close groups

Consider sequences $\sqrt{n}\Delta \to \delta$ where δ is a constant vector in $\mathbb{R}^{J^2}$. Along these sequences, the differences between groups are sufficiently small for a given sample size. This includes the degenerate case in which all locations are identical to the global mean. Since Z is a vector of asymptotically standard normally distributed random variables, we obtain that for an arbitrary matrix M:

$$\begin{aligned} n\Delta' M \Delta &\to c_{M,\delta} \in \mathbb{R} \\ \sqrt{n}\Delta' M Z &\overset{d}{\to} \mathcal{N}_{\delta,M} \sim \mathcal{N}(0, \Sigma_{\delta,M}) \\ Z' M Z &\overset{d}{\to} \mathcal{X}_M \end{aligned}$$

where $\Sigma_{\delta,M}$ is a variance covariance matrix depending on M and δ and $\mathcal{X}_M$ is a weighted sum of chi square random variables with positive and negative weights. Note that all elements in the estimated smoothing parameter depend on the same random vector Z, hence we have joint convergence in distribution. Using this together with an application of the portmanteau lemma and the continuous mapping theorem it follows that:

$$\hat{\omega}_{kj} \xrightarrow{d} \left[\frac{\mathcal{X}_a + \mathcal{N}_{\delta,b} + c_{e,\delta}}{\mathcal{X}_A + \mathcal{N}_{\delta,B} + c_{E,\delta}} + 1\right]^{-1} \frac{p_j \sigma_k^2}{p_k \sigma_j^2} \equiv \tilde{\omega}_{kj} \tag{4.4}$$

with $c_{e,\delta}$, $c_{E,\delta} \in \mathbb{R}$. Therefore, the estimated smoothing parameters have a nonstandard limiting distribution if all group means are locally close. Note that the limiting distribution is a function of the local parameter δ through $\mathcal{N}_{\delta,b}, \mathcal{N}_{\delta,B}$ as well as $c_{e,\delta}$ and $c_{E,\delta}$. δ cannot be estimated consistently due to the $\sqrt{n}$ multiplier. From (4.4) it follows that the estimated smoothing parameters $\hat{\omega}_{kj}$ are $O_p(1)$.

2. Case: Locally different groups

Consider distant sequences $||\sqrt{n}\Delta|| \to \infty$, i.e. at least two groups differ along the sequences of data generating processes. It follows that:

$$n\Delta' M \Delta = O(n)$$
$$\Delta' M Z \xrightarrow{d} \mathcal{N}_{\Delta,M}$$
$$Z' M Z \xrightarrow{d} \mathcal{X}_M.$$

For the random weights, one obtains that:

$$\hat{\omega}_{kj} = \left[\frac{O_p(n^{-1}) + O_p(n^{-1/2}) + c_{e,\Delta}}{O_p(n^{-1}) + O_p(n^{-1/2}) + c_{E,\Delta}} + 1\right]^{-1} \frac{p_j \sigma_k^2}{p_k \sigma_j^2} = \omega_{kj}^* + O_p(n^{-1/2}).$$

with $c_{e,\Delta} c_{E,\Delta} \in \mathbb{R}$ being constants depending on the vector of pairwise differences Δ. The plug-in smoothing parameters are no longer random in the limit as in the locally close case. The plug-in approach estimates the optimal smoothing parameters consistently as long as there are at least two arbitrary groups for which the root-n multiplied pairwise difference goes to infinity. Interestingly, if one enriches a system under locally close groups with just one additional locally different group, the behavior of all smoothing parameters is affected accordingly, see also Section 4.3.2.

4.3.1. Consistency

Consider the case of locally close groups, Recall that the PCS is given by $\hat{\mu}_k^{PCS}(\hat{\Lambda}_k) = \sum_{j=1}^J \hat{\omega}_{kj} \hat{\mu}_j$. Plugging in the first stage estimator yields

$$\hat{\mu}_k^{PCS}(\hat{\Lambda}_k) - \mu_k = \sum_{j=1}^J \frac{\hat{\omega}_{kj}}{\sqrt{n}} \sqrt{n} \Delta\mu_{jk} + \sum_{j=1}^J \hat{\omega}_{kj} \frac{\sigma_j}{\sqrt{np_j}} z_j + O_p(n^{-1}).$$

The first term converges to zero in probability since $\hat{\omega}_{kj}/\sqrt{n}$ are $O_p(n^{-1/2})$ by the previous results and the $\sqrt{n}\Delta\mu_{jk}$'s just converge to finite constants δ_{jk}. Similarly, the second term consists of a sum of product of an $O_p(1)$ and an $O_p(n^{-1/2})$ term and hence converges to zero in probability.

For the case of distinct groups, note that the estimation noise of the weights is asymptotically negligible, i.e. one can write the PCS as the optimal PCS plus an $o_p(1)$ term. The first part has the asymptotic risk of the PCS estimator under true optimal weights, or the limiting expression (3.2). Together with Markov's inequality this implies that under at least one different group $P(|\hat{\mu}_k(\hat{\Lambda}_k) - \mu_k| > \varepsilon)$ can by bounded in the limit by (3.2) (divided by ε^2) with weights being replaced by the theoretically optimal ones. Proposition 3.2 shows that the MSE under theoretically optimal weights is $O(n^{-1})$. This implies that the PCS estimator is consistent. Thus, consistency is achieved under all sequences of DGPs.

4.3.2. Asymptotic Distribution of the Plug-In Estimator and Valid Confidence Interval

One can use the results from above directly to get the distributional properties of the PCS estimator using the estimated (plug-in) MSE optimal weights under drifting sequences of data generating processes. Using the distributional results for the estimated smoothing parameter, it follows in locally close systems that

$$\sqrt{n}\Big(\hat{\mu}_k^{PCS}(\hat{\Lambda}_k) - \mu_k\Big) \xrightarrow{d} \sum_{j=1}^{J} \tilde{\omega}_{kj}\delta_{jk} + \sum_{j=1}^{J} \tilde{\omega}_{kj}\frac{\sigma_j}{\sqrt{p}_j}\mathcal{N}_j$$

with $\mathcal{N}_j$ being an independent standard normally distributed random variable for all j. Note that since the smoothing parameters are $O_p(1)$ it follows that the first component is a random function, i.e. it is $O_p(1)$. The limiting distribution of the stabilizing transformation is nonstandard since it is a random function plus a weighted combination of standard normally distributed random variables with random weights. Hence in the case of all groups being locally close, the limiting behavior of the PCS is nonstandard. Additionally, the asymptotic distribution depends on the unknown parameter δ through both δ directly as well as the smoothing parameters as described in equation (4.4).

In the case of at least one different group, recall that the estimated smoothing parameters converge in probability to a constant at the root-n rate. Thus it follows that

$$\sqrt{n}(\hat{\mu}^{PCS}(\hat{\Lambda}_k) - \mu_k - \sum_{j=1}^{J} \omega^*_{kj} \Delta\mu_{jk}) = \sum_{j=1}^{J} \omega^*_{kj} \sqrt{n}(\hat{\mu}_j - \mu_j)$$
$$+ \sum_{j=1}^{J} \sqrt{n}(\hat{\omega}_{kj} - \omega^*_{kj}) \Delta\mu_{jk} + O_p(n^{-1/2})$$

where by the following argument, the "bias" term is actually of order $O(n^{-1/2})$. If one looks at the proof of Proposition 3.2, one can see that since the optimal smoothing parameters are $O(1)$, it follows that the variance part is $O(n^{-1})$. The squared bias part has to have the same rate of convergence. Otherwise, there exists an n_0 for which some smoothing parameters would yield a smaller MSE than the optimal MSE smoothing parameters for $n > n_0$ which violates the fact that the smoothing parameters are the minimizers. Since the estimated smoothing parameters are continuous functions of the same random vector as the OLS estimator for the location, variance and cell probabilities, asymptotic normality and joint convergence in distribution is guaranteed. Hence, one can apply the Delta method to obtain asymptotically valid confidence bounds. In general, let Σ be the joint asymptotic variance covariance matrix of the first stage estimated means, variances and cell probabilities. And Let $G_k = \nabla(\hat{\mu}^{PCS}(\hat{\Lambda}_k) - \mu_k - \sum \omega^*_{kj} \Delta\mu_{j,k})$ be the gradient with respect to first stage estimated means, variances and cell probabilities at the true parameters. We have that

$$\sqrt{n}(\hat{\mu}^{PCS}(\hat{\Lambda}_k) - \mu_k - \sum_{j=1}^{J} \omega^*_{kj} \Delta\mu_{jk}) \xrightarrow{d} N(0, G'\Sigma G). \tag{4.5}$$

More details about a structure of the asymptotic variance under homoscedasticity can be found in the Appendix D.

The result in (4.5) suggests the following bias-corrected confidence bound for μ_k

$$CI_{\alpha,n}(\mu_k) = \left[\hat{\mu}^{PCS}_k(\hat{\Lambda}_k) - \sum_{j=1}^{J} \hat{\omega}_{kj} \Delta\hat{\mu}_{jk} \pm \frac{z_{1-\alpha/2}}{\sqrt{n}} \sqrt{\hat{G}'_k \hat{\Sigma} \hat{G}_k}\right],$$

where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution and $\hat{G}_k$ and $\hat{\Sigma}$ are consistent estimates of the corresponding population quantities. It turns out that under homoskedasticity the CI above is equivalent to the classical OLS confidence bounds. By the distributional theory, it has asymptotically correct coverage rates. We do not recommend to rely on this approximation in very small samples. For that purpose, there is large literature on robust confidence intervals in small samples with and without distributional assumptions. Since the main focus of this paper is point estimation risk, we do not pursue these ideas on inference any further.

Note that comparing the asymptotic variance of the limiting distributions of PCS under true and estimated smoothing parameters, i.e. equations (4.5) and (3.10), it seems that under distant systems the refinement that is due to optimal PCS is actually of order $n^{-1/2}$ an hence present in the first order asymptotic distribution whilst for the plug-in PCS under homoskedasticity it is of order n^{-1}. Our simulations in Section 5 seem not to contradict these results. For close systems, the same comparison based on the asymptotic normality is not possible due to the nonstandard behavior of plug-in PCS. However, for the optimal PCS the same intuition applies.

Figures 4.1 and 4.2 illustrate the asymptotic behavior using a simple Monte Carlo experiment[5]. They contain a histogram of the estimated smoothing parameters as well as the PCS estimator and a normal density for comparison. None of the theoretical results can be rejected in these four group setups. Figure 4.1a depicts the distribution of an estimated smoothing parameter under close groups and 4.1b contains a corresponding cross-smoothing estimator for $n = 25, 50, 100$, and 400. One can see that the distribution of the estimated smoothing parameter is nonstandard, even for large n. Similarly, the distribution of the PCS estimator is also not equal to the normal even for large n, i.e. there is more mass concentrated around the true parameter and less in the tails.

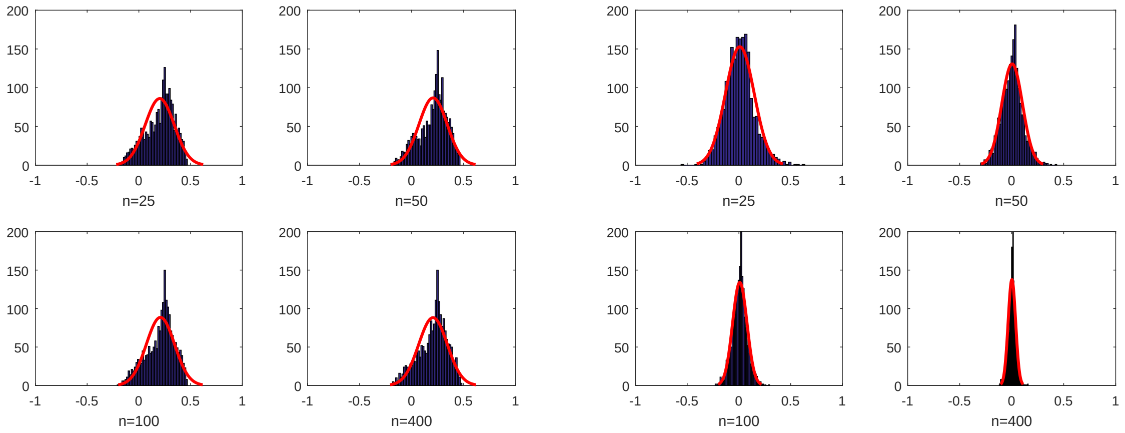
Figure 4.2a shows the distribution of the estimated smoothing parameter under distant groups. While the distribution is closer to the normal in this particular case, the general conclusion about convergence in probability can be seen by the more and more concentrated distribution around one third which is the theoretically optimal solution. Figure 4.2b illustrates the distribution of the corresponding averaging estimator. In contrast to the close group design, its distribution is very close to the normal, even for smaller sample sizes. This confirms the theoretical finding that a single distant group in the system is sufficient for asymptotic normality of the averaging estimator.

4.4. Effective Degrees of Freedom

Many estimators for μ have a corresponding linear map that maps Y to its predictions $\hat{Y}$. In particular, these estimators determine a matrix Π such that $\Pi Y = \hat{Y}$. In the case of the standard projection in the first stage one obtains that $\Pi = D(D'D)^{-1}D'$. The

[5] The designs are homoskedastic with unit variance, standardized log-normal errors, $\mu = (0, 0, n^{-1/2}, 0)$ for the close design and $\mu = (0, 0, 1, 0)$ for the distant design. All simulations use 2000 replications.

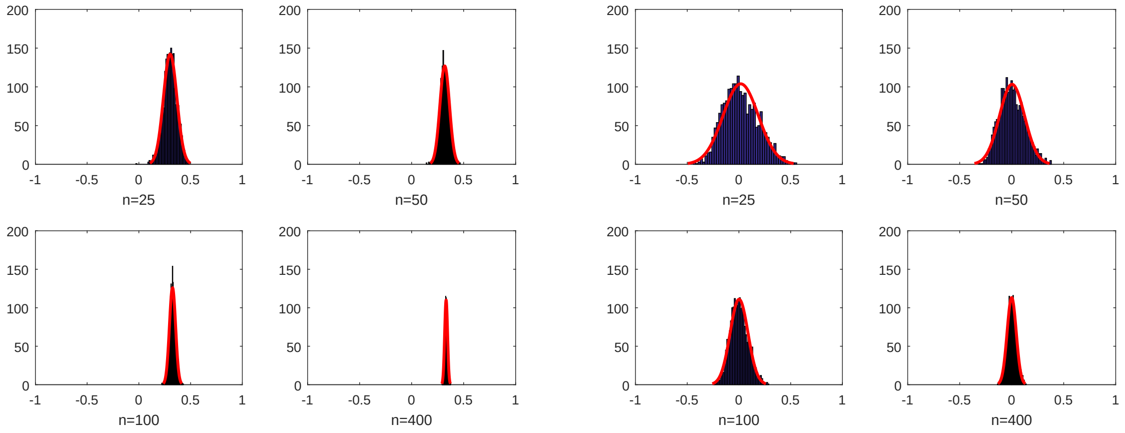
Figure 4.1: Distributional Plots under Close Groups



(a) Distributional Plot $\hat{\omega}_{12}$

(b) Distributional Plot $\hat{\mu}_1^{PCS}(\hat{\Lambda}_1)$

Figure 4.2: Distributional Plots under Distant Groups



(a) Distributional Plot $\hat{\omega}_{12}$

(b) Distributional Plot $\hat{\mu}_1^{PCS}(\hat{\Lambda}_1)$

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 of the matrix 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 tuning parameter κ yields a corresponding projection 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 on 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 4.1 *Let $\hat{\mu}$ be a (cell-based) projection and let $\Pi(\Lambda^*)$ and $\Pi(\hat{\Lambda})$ denote the linear operator based on the MSE optimal smoothing parameters and the plug-in estimator respectively. If $\#\{(k,j) : \mu_k \neq \mu_j, k = 1, \ldots, J-1, j > k\} > 0$, then*

$$tr(\Pi(\Lambda^*)) = 2 + O(n^{-1}) \quad \textit{and} \quad tr(\Pi(\hat{\Lambda})) = 2 + O_p(n^{-1/2}) \tag{4.6}$$

else

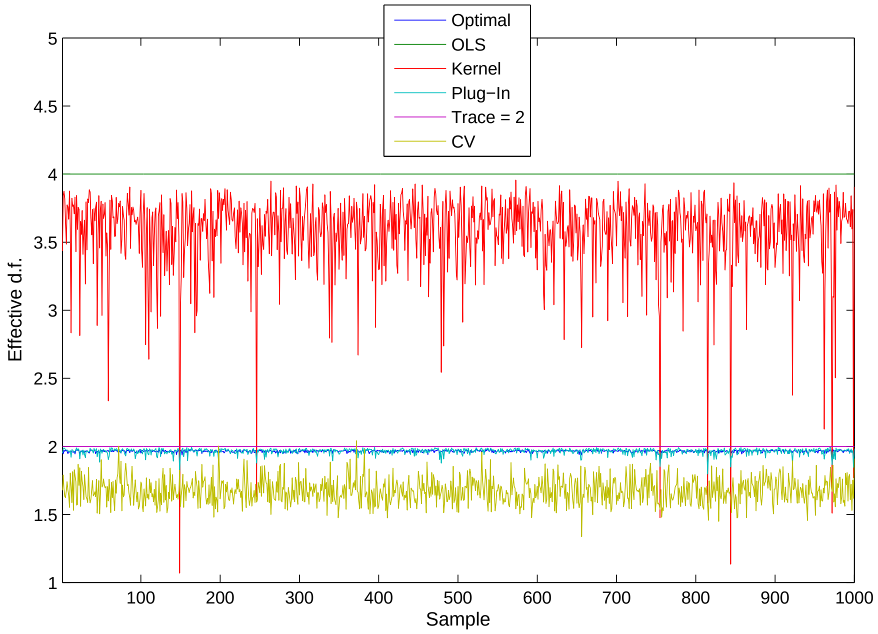
$$tr(\Pi(\Lambda^*)) = 1 + O(n^{-1}) \quad \textit{and} \quad tr(\Pi(\hat{\Lambda})) = 1 + O_p(n^{-1/2}). \tag{4.7}$$

Proof: The proof can be found in Appendix 4. ■

In other words, 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. Trivially, if there is only one location, i.e. the global mean, it converges to one. Hence independently of the total number of different groups under the true DGP, the effective degrees of freedom obtained by the PCS estimator will always converge to the same fixed number two or one. This seems to imply that, for large samples, any system described by countable many different locations should optimally (in a MSE sense) be modelled by two effective parameters only. We are not aware of any comparable result in the literature.

For illustrative purpose, consider Figure 4.3. It depicts the effective the dregrees of freedom over a grid of 1000 Monte Carlo repetitions for a variety of estimators in the case of four groups. For more details on the DGP consider Section 5. By construction, the highest complexity is obtained by the frequency approach (OLS). Since it estimates

Figure 4.3: Effective Degrees of Freedom, Distant Systems



exactly one parameter per cell, its trace is at a constant four over all replications. The PCS methods are slightly below two effective degrees of freedom. Note that the theoretical optimal PCS as well as the plug-in approach have similar overall complexities that are very stable over all replications. However, this does not necessarily translate into equal performance that is subject of investigation in the subsequent section. The cross validated PCS has a somewhat stronger variation in the effective degrees of freedom and seems to be smoothing more than the other PCS methods as well. Kernel smoothing (Ouyang et al., 2009) has by far the most volatile complexity over all replications and, except for the frequency method, the largest average complexity being above three in the majority of cases.

5. Monte Carlo Study

The following simulations are meant to give further insights into the small sample behavior of the pairwise cross smoothing estimator and potential alternatives over a large range of data generating processes. In particular, the performance gains for both distant and close systems under a large range of local parameter values will be analyzed. The following estimators are considered:

1. Ordinary least squares/frequency method (OLS),
2. pairwise cross smoothing with theoretically optimal smoothing (PCS$_{opt}$),
3. pairwise cross smoothing with plugin smoothing parameters (PCS$_{est}$),
4. pairwise cross smoothing with leave-one-out cross-validated smoothing parameters (PCS$_{CV}$),
5. a pretesting estimator based on Mallow's C_p[6],
6. nonparametric smoothing kernels with cross-validated bandwidths[7].

We consider a setup with a moderate number of groups, i.e. $J = 4$ that are selected with equal likelihood. We consider the following two baseline scenarios $\mu_{close} = (0, 0, 0, \delta/\sqrt{n})'$ and $\mu_{distant} = (-1, 0, 0, \delta)'$ where δ varies over a positive grid starting at 0. Note that in the close system, the limiting parameter for the fourth group is identical to zero independently of the local parameter value. For $\delta = 0$, we are in the case of identical means, i.e. a global model would be the most efficient one. For the distant system, we let δ vary over the same grid, however it is already a distant at $\delta = 0$. There is no convergence to the same mean but there are potential gains of aggregation due to mean two and three being identical over the whole grid. For the error, we assume a homoskedastic, standardized log-normal distribution[8].

Relative predictive mean squared errors with OLS as reference for $n = 16$, 32 and 48 are reported in Figure 5.1. Larger sample sizes are not relevant in this moderate group setup since all estimators except for the infeasible PCS with theoretically optimal weights are virtually identical to OLS or slightly worse (Mallow's) depending on the data generating process.

First, note that the pairwise cross smoothing with theoretically optimal weights dominates all approaches over all DGPs by a substantial which is in line with the theoretical results. It even shows improvement in distant systems for larger sample when all empirical approaches are as good as OLS or worse. Note that for all cases, all estimators get closer

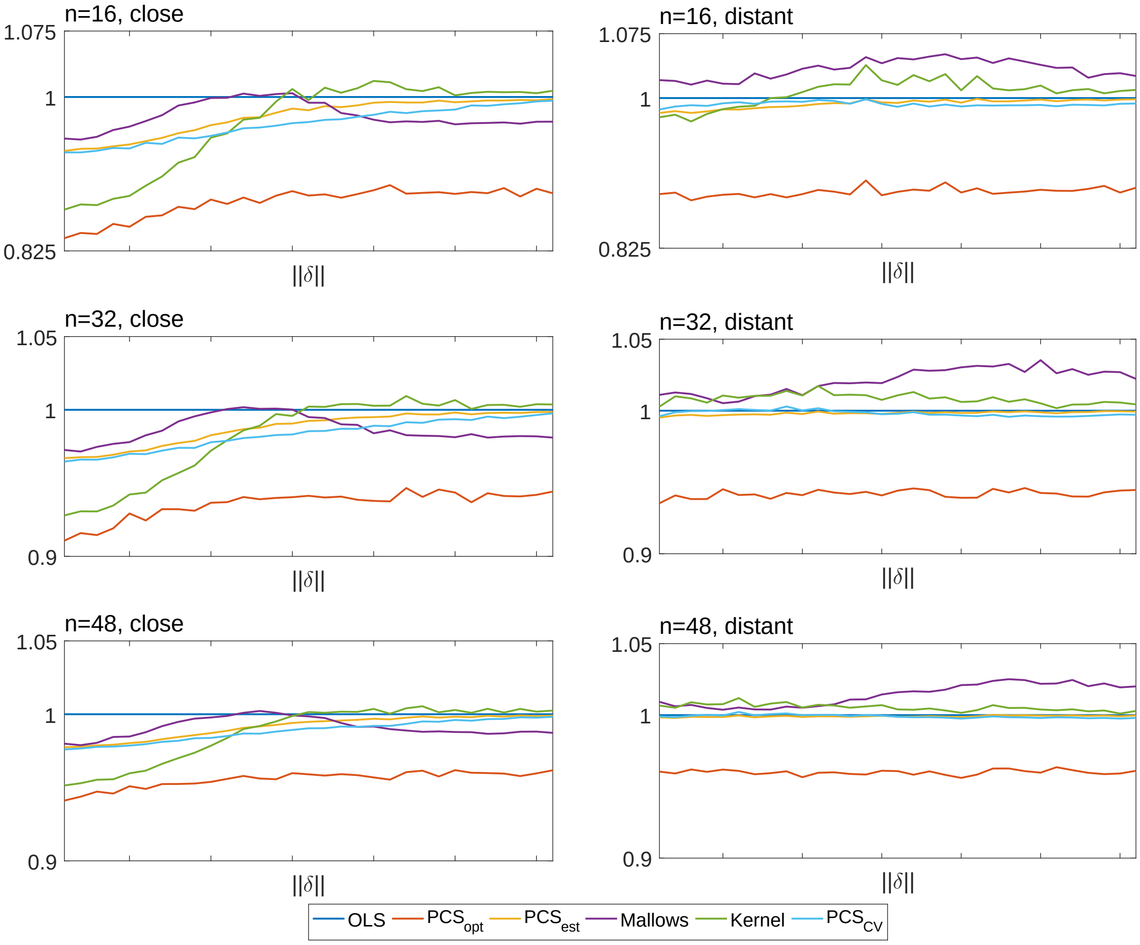
[6] We consider all possible submodels and choose the one with the lowest criterion value. Note that in our setup, the criterion by Mallows (1973) does selection identical to the Akaike (1970) information criterion.

[7] Ouyang et al. (2009).

[8] All results are robust with respect to the error distribution, i.e. changing symmetry and heteroskedasticity. Results for the normal distribution and heteroskedasticity do not differ qualitatively and are available on request.

to the OLS as the sample size increases.

Figure 5.1: Relative Predictive Mean Squared Errors



For the feasible PCS, improvements are largest in small samples and range from 0% to 13% with most values being between 2% to 6%. The more distant a system gets or the larger the local parameters become, the closer the PCS is to the OLS. It is noteworthy that the PCS is virtually never worse than OLS, i.e. it shows uniformly dominant behavior in terms of the predictive risk. However, it is not always beating all the competitors, i.e. for some DGPs, the kernel method and the pretesting estimator are superior. Note that e.g. for close systems at $\delta = 0$, the restrictions implied by the nonparametric kernel method (see Section 2.2) are actually not conflicting with the theoretically optimal aggregation hence it has some advantage due to the smaller dimensionality of the smoothing parameter vector. This translates to a better performance for that particular DGPs up to moderate deviations from it. For close systems with large local parameters, the estimator based on Mallow's criterion sometimes beats the other methods by a small margin. However, both of these alternative methods and in particular the Mallow's show poor behavior for distant systems, i.e. kernel shows worse performance than the OLS for large sample sizes and Mallow's shows worse performance over all DGPs considered while the PCS has small improvements for small sample sizes and is virtually identical for larger samples. Therefore PCS seems to be a robust refinement over OLS which is more significant for smaller group sizes and closer systems.

However, there is still room for improvement since the large gains that are shown by the theoretically optimal PCS cannot be reached in most DGPs.

6. 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 ran 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 was introduced for parents that came 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 6.1 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 panel data model[9]. In statistical terms, it presents 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

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

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 orignal means as first stage.

Figure 6.1: Mean Share of Late Arrivals, OLS and PCS estimates

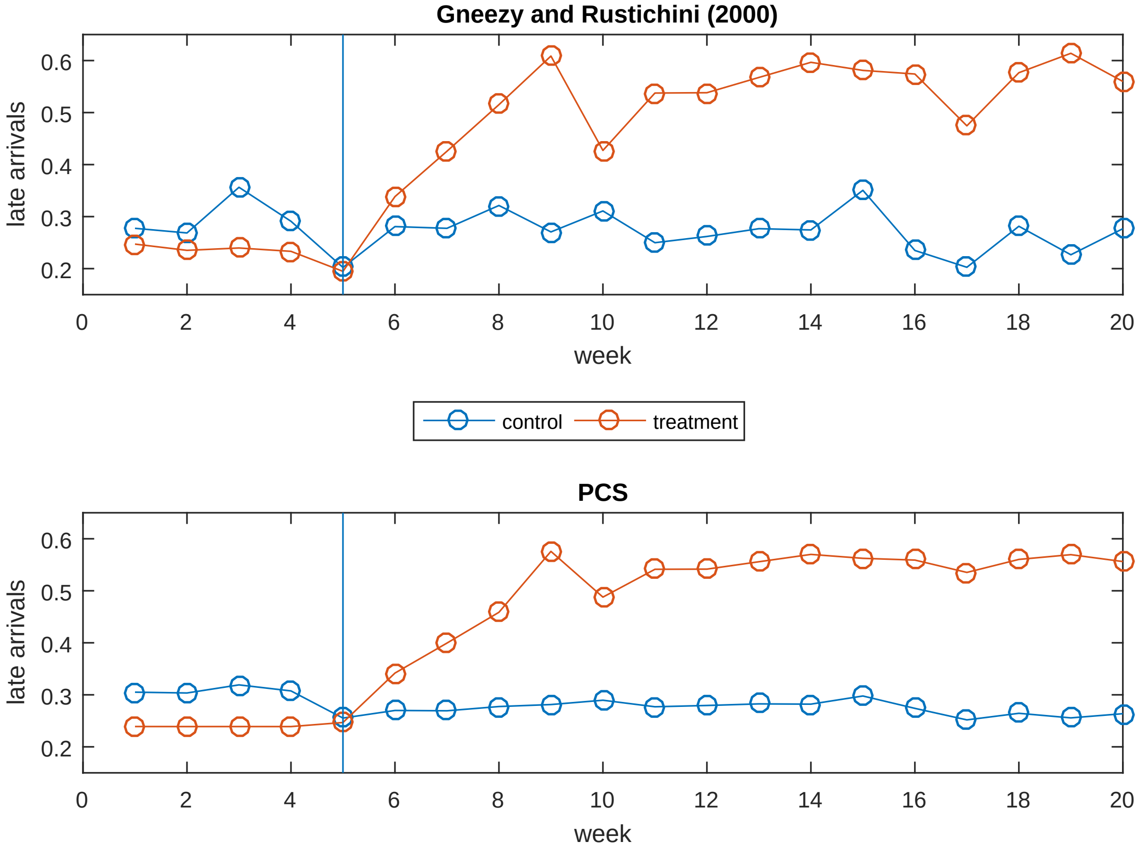


Figure 6.1 depicts OLS (Gneezy and Rustichini, 2000a) and PCS estimates for the conditional mean over time and treatment status. The major findings of the original visualization are confirmed. In fact, our estimates reveal the pattern much clearer since the PCS suggests a more stable share of the control group and a less fluctuating mean of the treatment group before and after the time of treatment.

7. 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. In reaction to a critique from Neumark and Wascher (2000) on the quality of their data, they re-estimated the models using an administrative employment data in Card and Krueger (2000). The results confirmed the conclusion from their previous paper that minimum wage increase in New Jersey has no significant effect on the total employment in New Jersey's fast-food industry or possibly even has a positive effect contradicting the findings of Neumark and Wascher (2000). No adverse employment effects are also confirmed by a meta-study of Doucouliagos and Stanley (2009).

The setup is well-suited for our method since there are four orthogonal groups by construction that are determined by state and time. We applied the PCS estimator in the difference-in-differences (DiD) model on the original Card and Krueger (1994) data and for each fast food chain separately to account for potential 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 full-time equivalent employment is measured as the number of full-time workers plus 0.5 times the part-time workers.

The OLS (Card and Krueger, 1994) and PCS results for pooled data and for each chain separately are in Tables 7.1 - 7.5. Table E.1 in the Appendix E includes means, variances and number of observations for all subgroups. The results of the KFC chain in Table 7.3 showed a different pattern from the other stores, as KFC was the only chain which confirmed the theory of increasing the labor demand in a less labor costly environment. Other chains in the data set did not follow this pattern. All the estimated effects of the minimum wage on the employment are closer to 0 for the PCS in comparison to the OLS. In the case of pooled data, Burger King, KFC and Roys, the difference between OLS and PCS are not so big. 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.

Table 7.1: Mean 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

The table contains the mean estimates of the full-time employment. B = before the minimum wage increase, A = after the minimum wage increase, NJ = New Jersey, PEN = Pennsylvania, t = treated, c = control. DiD formula: $(\hat{\mu}_{NJ,A} - \hat{\mu}_{NJ,B}) - (\hat{\mu}_{PEN,A} - \hat{\mu}_{PEN,B})$.

Table 7.2: Mean Estimates - 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

The table contains the mean estimates of the full-time employment. B = before the minimum wage increase, A = after the minimum wage increase, NJ = New Jersey, PEN = Pennsylvania, t = treated, c = control. DiD formula: $(\hat{\mu}_{NJ,A} - \hat{\mu}_{NJ,B}) - (\hat{\mu}_{PEN,A} - \hat{\mu}_{PEN,B})$.

Table 7.3: Mean Estimates - 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

The table contains the mean estimates of the full-time employment. B = before the minimum wage increase, A = after the minimum wage increase, NJ = New Jersey, PEN = Pennsylvania, t = treated, c = control. DiD formula: $(\hat{\mu}_{NJ,A} - \hat{\mu}_{NJ,B}) - (\hat{\mu}_{PEN,A} - \hat{\mu}_{PEN,B})$.

Table 7.4: Mean Estimates - 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

The table contains the mean estimates of the full-time employment. B = before the minimum wage increase, A = after the minimum wage increase, NJ = New Jersey, PEN = Pennsylvania, t = treated, c = control. DiD formula: $(\hat{\mu}_{NJ,A} - \hat{\mu}_{NJ,B}) - (\hat{\mu}_{PEN,A} - \hat{\mu}_{PEN,B})$.

Table 7.5: Mean Estimates - 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. B = before the minimum wage increase, A = after the minimum wage increase, NJ = New Jersey, PEN = Pennsylvania, t = treated, c = control. DiD formula: $(\hat{\mu}_{NJ,A} - \hat{\mu}_{NJ,B}) - (\hat{\mu}_{PEN,A} - \hat{\mu}_{PEN,B})$.

8. Concluding Remarks

Pairwise cross smoothing provides a unifying framework to analyze and compare smoothing methods for exclusively categorical data that nests different approaches from the nonparametric smoothing kernel and model averaging literature. It penalizes L_2 differences between estimation targets and a first stage estimator or fixed value that serves as a reference target. The estimator can be easily implemented with standard software packages using the closed form solutions derived in this paper. For future research, refined inference of the estimated PCS under close systems should be tackled. In addition, relaxing the assumption of a fixed number of groups, i.e. allowing for J to grow with the sample size with some closeness restrictions that are related to sparsity in the sense of few different locations should be considered. The Monte Carlo simulations are also highly suggestive of a uniform dominance property over the ordinary least squares in the sense of Cheng et al. (2016).

References

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

Akaike, H. (1970). Statistical predictor identification. *Annals of the Institute of Statistical Mathematics*, 22(1):203–217.

Buckland, S. T., Burnham, K. P., and Augustin, N. H. (1997). Model selection: an integral part of inference. *Biometrics*, pages 603–618.

Burnham, K. P. and Anderson, D. (2003). Model selection and multi-model inference. *A Pratical informatio-theoric approch. Sringer.*

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.

Card, D. and Krueger, A. B. (2000). Minimum wages and employment: a case study of the fast-food industry in New Jersey and Pennsylvania: reply. *The American Economic Review*, 90(5):1397–1420.

Cheng, X., Liao, Z., and Shi, R. (2016). Uniform Asymptotic Risk of Averaging GMM Estimator Robust to Misspecification. *University of Pennsylvania and UCLA.*

Claeskens, G., Hjort, N. L., et al. (2008). *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge.

Doucouliagos, H. and Stanley, T. D. (2009). Publication Selection Bias in Minimum-Wage Research? A Meta-Regression Analysis. *British Journal of Industrial Relations*, 47(2):406–428.

Fienberg, S. E. and Holland, P. W. (1973). Simultaneous estimation of multinomial cell probabilities. *Journal of the American Statistical Association*, 68(343):683–691.

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). Fine is a price, a. *J. Legal Stud.*, 29:1.

Gneezy, U. and Rustichini, A. (2000b). Pay enough or don't pay at all. *Quarterly journal of economics*, pages 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.

Hall, P. (1983). Large Sample Optimality of Least Squares Cross-Validation in Density Estimation. *The Annals of Statistics*, pages 1156–1174.

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. and Martin, M. (1988). On the Bootstrap and Two-Sample Problems. *Australian Journal of Statistics*, 30(1):179–192.

Hall, P., Racine, J., and Li, Q. (2004). Cross-Validationa 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. and Racine, J. S. (2012). Jackknife model averaging. *Journal of Econometrics*, 167(1):38–46.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association*, 98(464):879–899.

Li, K.-C. (1987). Asymptotic optimality for Cp, CL, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, pages 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.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

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.*

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.

Neumark, D. and Wascher, W. (2000). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania: Comment. *The American Economic Review*, 90(5):1362–1396.

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

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

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer Science & Business Media.

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. (2016). Modelling Clustered Heterogeneity: Fixed Effects, Random Effects and Mixtures. *International Statistical Review*, pages n/a–n/a.

Welch, B. L. (1947). The Generalization of Student's' Problem when several Different Population Variances are Involved. *Biometrika*, 34(1/2):28–35.

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.

9. Mixed Data

The original framework is rather restrictive beyond applications that form orthogonal groups by construction. Imagine k-dimensional continuous regressors X and an additive linear model:

$$Y = X\beta + D\mu + \varepsilon$$

with $E[\varepsilon_i|X_i, D_i] = 0$. We propose to use the PCS in a two step procedure to estimate both the location parameters as well as the parameter on the continuous regressor. The idea is similar to **?**, i.e. works by partialling out the expectations conditional on the set of orthogonal dummies. Note that

$$Y - E[Y|D] = (X - E[X|D])'\beta + \varepsilon.$$

Replacing the conditional expectations by PCS estimators that only rely on the orthogonal data yields the following estimator for β:

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

with

$$\begin{aligned}\hat{E}[X|D] &= D(D'D + U'W_yU)^{-1}(I + U'W_yV(D'D)^{-1})D'Y\\ \hat{E}[Y|D] &= D(D'D + U'W_xU)^{-1}(I + U'W_xV(D'D)^{-1})D'X\end{aligned}$$

and W_x, W_y being the diagonal matrix of PCS smoothing parameters for the regression model of X on D and Y on D respectively. To obtain an estimate for μ one can substract the continuous component and use a projection, i.e.

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

One can show that both estimators are $\sqrt{n}$-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 is now achieved with respect to the risk of the first stage. A better approach would be to use the residuals $Y^* \equiv (Y - X\hat{\beta}^{PCS})$ in another PCS step, i.e. estimate the model

$$Y^* = D\mu + u$$

which yields

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

where $u = X(\beta - \hat{\beta}) + \varepsilon$. Simulations reveal that this positively affects the parameter risk for both the continuous and the discrete part. In fact, the improvements on the continuous part are non negligible. There are huge gains for the parameter risk of the discrete part which seems to dominate the simple OLS, in particular in the presence of small and/or very volatile groups. The gains are more pronouneced if D and X are correlated. Detailed results are available on request.

A. Matrix Notation Appendix

A.1. General Notation

In a matrix notation, the model looks as follows:

$$Y = D\mu + \varepsilon$$

with $Y = (Y_1, \ldots, Y_n)'$, D is a $n \times J$ matrix collecting D_i' vectors in i-th rows, $\mu = (\mu_1, \ldots, \mu_J)'$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ and $E[\varepsilon|D] = 0$. The optimization problem for the PCS can be written as

$$\hat{\mu}^{PCS}(W) = \arg\min_{\mu}(Y - D\mu)'(Y - D\mu) + (U\mu - V\hat{\mu})'W(U\mu - V\hat{\mu}) \tag{A.1}$$

with $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_J)$, $U = (I_J \otimes \iota_J)$, $V = (\iota_J \otimes I_J)$ and $W = diag(\Lambda)$ with $\Lambda = (\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1J}, \lambda_{21}, \ldots, \lambda_{JJ})$ and $\lambda_{jj} = 0$ for all $j \in \{1, \ldots, J\}$.

Under $D'D + UWU$ being a positive definite matrix, the global minimizer of (A.1), i.e. the PCS estimator as a function of the smoothing parameters, is

$$\hat{\mu}^{PCS}(W) = (D'D + U'WU)^{-1}(D'Y + U'WV\hat{\mu}). \tag{A.2}$$

Regarding $\hat{\mu}$, a possible choice is the linear (cell based) projection of Y on D, i.e. $\hat{\mu} = (D'D)^{-1}D'Y = \bar{Y}$, where $\bar{Y}$ is a vector of cell means. This is also referred to as "frequency approach" in the literature. Under this choice, the pairwise cross smoothing estimator simplifies to

$$\hat{\mu}^{PCS}(W) = (D'D + U'WU)^{-1}(I + U'WV(D'D)^{-1})D'Y. \tag{A.3}$$

The estimator is linear in Y. However, note that the "projection" matrix $D(D'D + Q'WQ)^{-1}(I + Q'WP(D'D)^{-1})D'$ that maps from outcome to prediction is neither symmetric nor idempotent.

A.2. Uniqueness of the MSE Optimal Regularization Parameters

The problem of minimizing (3.2) can be rewritten in a matrix form as follows:

$$\begin{aligned}
\min_{\Omega_k} MSE &= \min_{\Omega_k} \Omega_k' H_k \Omega_k + 2\gamma_k(1 - \iota'\Omega_k), \qquad \text{(A.4)}\\
\text{where } \Omega_k &= (\omega_{k1}, \omega_{k2}, \ldots, \omega_{kJ})',\\
H_k &= \Delta_k \Delta_k' + V,\\
\Delta_k &= (\mu_k - \mu_1, \mu_k - \mu_2, \ldots, \mu_k - \mu_J)',\\
V &= \text{diag}(\sigma_j^2/p_j n), \quad j \in \{1, \ldots, J\},\\
\gamma_k &\ldots \text{ Lagrange multiplier},\\
\iota &= (1, \ldots, 1)'.
\end{aligned}$$

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

$$\gamma_k^* = [\iota'(H_k' H_k)^{-1} H_k' \iota]^{-1} \qquad \text{(A.5)}$$

$$\Omega_k^* = [\iota'(H_k' H_k)^{-1} H_k' \iota]^{-1} (H_k' H_k)^{-1} H_k' \iota \qquad \text{(A.6)}$$

To investigate if the Ω_k^* is a unique global minimizer of (3.2), we first rewrite the optimization problem (3.2) into a form used in null-space methods to solve equality quadratic problems (see 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 matrix Z such that $\iota' Z = 0$ and $\text{rank}(\iota \vdots Z) = J$, solving the following null-space method problem yields the same solution as solving (3.2):

$$\min_{\Omega_{Z,k} \in \mathbb{R}^{J-1}} \Omega_{Z,k}' Z' H_k Z \Omega_{Z,k} + \Omega_{Z,k}' Z' H_k \iota \Omega_{\iota,k} \qquad \text{(A.7)}$$

$$\text{where } \iota' \iota \Omega_{\iota,k} = 1 \qquad \text{(A.8)}$$

and then $\Omega_k^* = Z\Omega_{Z,k}^* + \iota \Omega_{\iota,k}$ and $\gamma_k^* = (\iota'\iota)^{-1} \iota' H_k \Omega_k^*$. Note that (A.8) just determines the value of $\Omega_{\iota,k}$ and (A.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 (Gould, 1985). For completeness, theorem quoted below from (Gould, 1985, Theorem 1.1(i)) enables to test for the uniqueness of the solution.

Theorem A.1 *Suppose (A.7) is as given with ι' of full row rank and Z is constructed so that $\iota' Z = 0$ and rank$(\iota \vdots Z) = J$. Then (A.7) has a strong minimizer if and only if $Z'H_kZ$ is positive definite.*

It is easy to see that ι' has a full row rank $= 1$. 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 $\iota' Z = 0$ and rank$(\iota \vdots Z) = J$. Note that Z has a full column rank $= J - 1$. This implies that if H_k is positive definite, then $Z'H_kZ$ is also positive definite.

We know that $H_k = \Delta_k \Delta_k' + V$. Since V is a diagonal matrix with positive elements for a finite n, V is a positive definite matrix. Since $\Delta_k \Delta_k'$ gives a matrix which is rank deficient, it can happen that for a non-zero vector x we get that $x'\Delta_k\Delta_k'x = 0$ but it cannot be negative as illustrated below:

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

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

$$x'H_kx = x'(\Delta_k\Delta_k' + V)x = \underbrace{x'\Delta_k\Delta_k'x}_{\geq 0} + \underbrace{x'Vx}_{>0} > 0 \quad \text{for all } x \neq 0.$$

This means that H_k is a positive definite matrix and that Ω_k^* is a unique minimizer of (3.2).

Case 2: $n \to \infty$: For $n \to \infty$, V converges to a zero matrix and $Z'H_kZ$ converges then to $Z'\Delta_k\Delta_k'Z$.

Subcase - $J = 2$, non-equal means ($\mu_1 \neq \mu_2$): One can derive that $Z'\Delta_k\Delta_k'Z = (\Delta\mu_{k1} - \Delta\mu_{k2})^2 = (\mu_2 - \mu_1)^2 > 0$, i.e. $Z'\Delta_k\Delta_k'Z$ is a positive definite matrix. And according to Theorem A.1, (3.2) then has a unique minimizer Ω_k^*.

Subcase - $J = 2$, equal means ($\mu_1 = \mu_2$): One can derive that $Z'\Delta_k\Delta_k'Z = (\Delta\mu_{k1} - \Delta\mu_{k2})^2 = (\mu_2 - \mu_1)^2 = 0$, i.e. $Z'\Delta_k\Delta_k'Z$ is a positive semi-definite and singular matrix. The following theorem, quoted from (Gould, 1985, Theorem 1.1(ii)) states conditions for the existence of weak minimizers for the problem (3.2).

Theorem A.2 *Suppose (A.7) is as given with ι' of full row rank and Z is constructed so that $\iota'Z = 0$ and rank($\iota \vdots Z$) $= J$. Then (A.7) has weak minimizers if $Z'H_kZ$ is positive semi-definite with $Z'H_kZ$ singular and $Z'H_kZ\Omega_{Z,k} = -Z'H_k\iota\Omega_{\iota,k}$ compatible.*

It is easy to see that ι' has a full row rank $= 1$ and Z is chosen as in the case of finite n. We showed that $Z'H_kZ$ is a positive semi-definite and singular matrix in the limit. Now we check that the last equality in the Theorem A.2 holds.

$$
\begin{aligned}
0 \cdot \Omega_{Z,k} &= -\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} \Delta\mu_{k1}^2 & \Delta\mu_{k1}\Delta\mu_{k2} \\ \Delta\mu_{k1}\Delta\mu_{k2} & \Delta\mu_{k2}^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Omega_{\iota,k} \\
0 &= -\begin{pmatrix} \Delta\mu_{k1}^2 - \Delta\mu_{k1}\Delta\mu_{k2} & \Delta\mu_{k1}\Delta\mu_{k2} - \Delta\mu_{k2}^2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \Omega_{\iota,k} \\
0 &= -(\Delta\mu_{k1}^2 - \Delta\mu_{k2}^2)\Omega_{\iota,k} \\
0 &= -(\Delta\mu_{k1} + \Delta\mu_{k2}) \underbrace{(\Delta\mu_{k1} - \Delta\mu_{k2})}_{=0} \Omega_{\iota,k} \\
0 &= 0
\end{aligned}
$$

In this case, there exist weak minimizers Ω_k^* of (3.2).

Subcase - $J > 2$: As for finite n, $\Delta_k\Delta_k'$ is a positive semi-definite matrix. Since Z has a full column rank, this implies that $Z'H_kZ$ is also positive semi-definite.

We also know that $\operatorname{rank}(\Delta_k \Delta_k') = 1$ because each row of $\Delta_k \Delta_k'$ is just a scaled Δ_k'. Further, we use rank inequalities to determine $\operatorname{rank}(Z'\Delta_k \Delta_k' Z)$.

Note that

$$\begin{aligned}
&\operatorname{rank}(Z'\Delta_k\Delta_k') \leq \min(\operatorname{rank}(Z'), \operatorname{rank}(\Delta_k\Delta_k')) \\
&\operatorname{rank}(Z'\Delta_k\Delta_k') \leq \min(J-1, 1) \\
&\operatorname{rank}(Z'\Delta_k\Delta_k') \leq 1 \quad \Rightarrow \quad \operatorname{rank}(Z'\Delta_k\Delta_k') = 1
\end{aligned}$$

Then,

$$\begin{aligned}
&\operatorname{rank}(Z'\Delta_k\Delta_k'Z) \leq \min(\operatorname{rank}(Z'\Delta_k\Delta_k'), \operatorname{rank}(Z)) \\
&\operatorname{rank}(Z'\Delta_k\Delta_k'Z) \leq \min(1, J-1) \\
&\operatorname{rank}(Z'\Delta_k\Delta_k'Z) \leq 1 \quad \Rightarrow \quad \operatorname{rank}(Z'\Delta_k\Delta_k'Z) = 1 < J-1
\end{aligned}$$

This means that matrix $Z'\Delta_k\Delta_k'Z$ is rank deficient and singular.

To use the Theorem A.2, we still need to check if $Z'\Delta_k\Delta_k'Z$ is compatible with

$$Z'H_kZ\Omega_{Z,k} = -Z'H_k\iota\Omega_{\iota,k},$$

i.e. if any solutions $\Omega^*_{Z,k}$ exist where $H_k = \Delta_k\Delta_k'$. First we solve for $\Omega_{\iota,k}$ based on (A.8), i.e. $\Omega_{\iota,k} = 1/J$. Now, the question is if the following system of linear equations is solvable:

$$Z'H_kZ\Omega_{Z,k} = -Z'H_k\iota\frac{1}{J}$$

$$\begin{pmatrix}
(\mu_2-\mu_1)^2 & (\mu_2-\mu_1)(\mu_3-\mu_2) & \dots & (\mu_2-\mu_1)(\mu_J-\mu_{J-1}) \\
(\mu_3-\mu_2)(\mu_2-\mu_1) & (\mu_3-\mu_2)^2 & \dots & (\mu_3-\mu_2)(\mu_J-\mu_{J-1}) \\
\vdots & & \ddots & \vdots \\
(\mu_J-\mu_{J-1})(\mu_2-\mu_1) & (\mu_J-\mu_{J-1})(\mu_3-\mu_2) & \dots & (\mu_J-\mu_{J-1})^2
\end{pmatrix}\Omega_{Z,k} =$$

$$\begin{pmatrix}
-(1/J)(\mu_2-\mu_1)\sum_{j=1}^J \Delta\mu_{kj} \\
-(1/J)(\mu_3-\mu_2)\sum_{j=1}^J \Delta\mu_{kj} \\
\vdots \\
-(1/J)(\mu_J-\mu_{J-1})\sum_{j=1}^J \Delta\mu_{kj}
\end{pmatrix}$$

As we know $\text{rank}(Z'H_kZ) = 1 < J-1$. In that case, the system above will have infinitely many solutions if and only if the rank of the coefficient matrix $Z'H_kZ$ is equal to the rank of the augmented matrix $[Z'H_kZ \mid -Z'H_k\iota(1/J)]$. Now we need to check if $\text{rank}([Z'H_kZ \mid -Z'H_k\iota\frac{1}{J}]) = 1$.

$$[Z'H_kZ \mid -Z'H_k\iota(1/J)] =$$

$$= \begin{pmatrix} (\mu_2-\mu_1)^2 & \dots & (\mu_2-\mu_1)(\mu_J-\mu_{J-1}) & -(1/J)(\mu_2-\mu_1)\sum_{j=1}^J \Delta\mu_{kj} \\ (\mu_3-\mu_2)(\mu_2-\mu_1) & \dots & (\mu_3-\mu_2)(\mu_J-\mu_{J-1}) & -(1/J)(\mu_3-\mu_2)\sum_{j=1}^J \Delta\mu_{kj} \\ \vdots & & \vdots & \vdots \\ (\mu_J-\mu_{J-1})(\mu_2-\mu_1) & \dots & (\mu_J-\mu_{J-1})^2 & -(1/J)(\mu_J-\mu_{J-1})\sum_{j=1}^J \Delta\mu_{kj} \end{pmatrix}$$

Multiplying each column q by $-\frac{(1/J)\Delta\mu_{kq}}{\mu_{q+1}-\mu_q}$ for $q \in \{1,\dots,k-1\}$ and each column r by $-\frac{(1/J)\Delta\mu_{k,r+1}}{\mu_{r+1}-\mu_r}$ for $r \in \{k,\dots,J-1\}$ in $[Z'H_kZ \mid -Z'H_k\iota(1/J)]$, we get the following matrix:

$$\begin{pmatrix} -(1/J)\Delta\mu_{k1}(\mu_2-\mu_1) & \dots & -(1/J)\Delta\mu_{kJ}(\mu_2-\mu_1) & -(1/J)(\mu_2-\mu_1)\sum_{j=1}^J \Delta\mu_{kj} \\ -(1/J)\Delta\mu_{k1}(\mu_3-\mu_2) & \dots & -(1/J)\Delta\mu_{kJ}(\mu_3-\mu_2) & -(1/J)(\mu_3-\mu_2)\sum_{j=1}^J \Delta\mu_{kj} \\ \vdots & & \vdots & \vdots \\ -(1/J)\Delta\mu_{k1}(\mu_J-\mu_{J-1}) & \dots & -(1/J)\Delta\mu_{kJ}(\mu_J-\mu_{J-1}) & -(1/J)(\mu_J-\mu_{J-1})\sum_{j=1}^J \Delta\mu_{kj} \end{pmatrix}$$

Since $\Delta\mu_{kk} = 0$, summing the first $J-1$ columns will yield the last column, i.e. $\text{rank}([Z'H_kZ \mid -Z'H_k\iota(1/J)]) = 1$ because $-Z'H_k\iota(1/J)$ is a linear combination of columns in $Z'H_kZ$. This implies then that $Z'H_kZ$ is compatible with (A.8) and weak minimizers Ω_k^* of (3.2) exist.

A.3. Large Sample Properties

To learn more about the large sample behavior of the optimal smoothing parameters and the corresponding PCS estimator, one can establish the following proposition:

Proposition A.1 *If the optimal smoothing parameters according to (3.3) or (3.4) are chosen, then*

$$MSE(\hat{\mu}^{PCS}(W^*)) = O(n^{-1}) \tag{A.9}$$

with $W^* = diag(\Lambda^*)$ *and* $\Lambda^* = (\lambda^*_{11}, \lambda^*_{12}, \ldots, \lambda^*_{JJ}) \in \mathbb{R}^{J^2}$ *being the MSE optimal smoothing parameters.*

Together with closed form of the theoretical MSE, one can establish a rate for the theoretical optimal smoothing parameters. In fact, one obtains that $\Lambda^* = O(n)$ and $\Omega^* = (\Omega_1^{*\prime}, \ldots, \Omega_J^{*\prime})' = O(1)$. This is qualitatively different from the smoothing kernel approach where informative, i.e. conditionally independent, regressors are smoothed to a global average with a smoothing parameter converging to its upper bound. This means that asymptotically the MSE optimal smoothing parameters do not vanish in general. Hence there is potential aggregation even in the limit.

This implies that the optimal PCS estimator is consistent. In fact, one can establish two generic asymptotic normality results under fixed and MSE optimal smoothing.

Theorem A.3 *If the smoothing parameters are fixed, i.e.* W *does not vary with sample size,* $n^{-1}D'D \overset{p}{\rightarrow}= E[D_iD_i']$, $rank(E[D_iD_i']) = J$ *and* $\hat{\mu}$ *is a linear (cell based) projection, we get*

$$\sqrt{n}(\hat{\mu}^{PCS}(W) - \mu - B(W)) \xrightarrow{d} N(0, E[D_iD_i']^{-1}E[D_iD_i'\varepsilon_i^2]E[D_iD_i']^{-1}) \tag{A.10}$$

with $B(W) = (D'D + U'WU)^{-1}U'W(V - U)\mu$.

Proof: From exogeneity it follows that $E[D_i\varepsilon_i] = 0$. Moreover, $n^{-1}W = 0$. If the first stage estimator is a linear projection it follows that:

$$\hat{\mu}^{PCS}(W) = (D'D + U'WU)^{-1}(D'D + U'WV)\mu + (D'D + U'WU)^{-1}(I + U'WV(D'D)^{-1})D'\varepsilon$$

$$\hat{\mu}^{PCS}(W) - \mu = (D'D + U'WU)^{-1}U'W(V - U)\mu + (D'D + U'WU)^{-1}(I + U'WV(D'D)^{-1})D'\varepsilon$$

Retransforming and using LL-CLT yields:

$$\sqrt{n}(\hat{\mu}^{PCS}(W) - \mu - B(W)) \xrightarrow{d} N(0, E[D_iD_i']^{-1}E[D_iD_i'\varepsilon_i^2]E[D_iD_i']^{-1})$$

with $B(W) = (D'D + U'WU)^{-1}U'W(V - U)\mu$. ■

Theorem A.4 *If the optimal smoothing parameters according to (3.3) or (3.4) are chosen and* $n^{-1}D'D \xrightarrow{p} E[D_iD_i']$, $rank(E[D_iD_i']) = J$ *and* $\hat{\mu}$ *is a linear (cell based) projection, we get*

$$\sqrt{n}(\hat{\mu}^{PCS}(W^*) - \mu - B(W^*)) \xrightarrow{d} N\Big(0, (E[D_iD_i'] + U'\bar{W}^*U)^{-1}(I_J + U'\bar{W}^*VE[D_iD_i']^{-1})E[D_iD_i'\varepsilon_i^2] \\ (I_J + E[D_iD_i']^{-1}V'\bar{W}^*U)(E[D_iD_i'] + U'\bar{W}^*U)^{-1}\Big) \tag{A.11}$$

with $B(W^*) = (D'D + U'W^*U)^{-1}U'W^*(V - U)\mu$ *and* $\bar{W} = plim\ n^{-1}W^*$.

Proof: From the convergence results on the MSE optimal smoothing parameters, i.e. $\Lambda^* = O(n)$ we know that W^* which is the matrix of MSE optimal smoothing parameters is also $O(n)$. Let $\bar{W}$ denote the probability limit of $n^{-1}W^*$. For the estimated parameter using these lambdas we obtain

$$\begin{aligned}\hat{\mu}^{PCS}(W^*) - \mu = (D'D + U'W^*U)^{-1}U'W^*(V - U)\mu \\ + (D'D + U'W^*U)^{-1}(I_J + U'W^*V(D'D)^{-1})D'\varepsilon\end{aligned}$$

From this it follows that:

$$\begin{aligned}\sqrt{n}(\hat{\mu}^{PCS}(W^*) - \mu - B(W^*)) \xrightarrow{d} N\Big(0,(E[D_iD_i'] + U'\bar{W}U)^{-1}(I_J + U'\bar{W}VE[D_iD_i']^{-1})E[D_iD_i'\varepsilon_i^2] \\ (I_J + E[D_iD_i']^{-1}V'\bar{W}U)(E[D_iD_i'] + U'\bar{W}U)^{-1}\Big)\end{aligned}$$

with $B(W^*) = (D'D + U'W^*U)^{-1}U'W^*(V - U)\mu$ and $\bar{W} = \lim\ n^{-1}W^*$. ■

We propose to estimate the smoothing parameters using a plug-in approach, i.e.

$$\begin{aligned}\hat{\mu}_k^{PCS} &= \hat{\mu}'\hat{\Omega}_k^* \\ \hat{\Omega}_k^* &= [\iota'(\hat{H}_k'\hat{H}_k)^{-1}\hat{H}_k'\iota]^{-1}(\hat{H}_k'\hat{H}_k)^{-1}\hat{H}_k'\iota \\ \hat{H}_k &= \hat{\Delta}_k\hat{\Delta}_k' + \hat{V} \\ \hat{\Delta}_k &= (\hat{\mu}_k - \hat{\mu}_1, \ldots, \hat{\mu}_k - \hat{\mu}_J)', \\ \hat{V} &= \text{diag}(\hat{\sigma}_j^2/p_jn),\end{aligned}$$

where $\hat{\mu}$ is a linear cell based projection and $\hat{\sigma}_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n} D_{ij}(Y_i - \hat{\mu}_j)^2$.

B. Supplementary Material for Section 2

B.1. SOC Conditions for (2.2)

Let $S_\lambda(\mu)$ denote the objective function in (2.2). Note that:

$$\frac{\partial S_\lambda(\mu)}{\partial \mu_k} = -2\sum_{i=1}^{n}(Y_i - D_i'\mu)D_{ik} + 2\sum_{s\neq k}\lambda_{ks}(\mu_k - \bar{\mu}_s) \tag{B.1}$$

$$\frac{\partial^2 S_\lambda(\mu)}{\partial \mu_k^2} = 2n_k + 2\sum_{s\neq k}\lambda_{ks} \tag{B.2}$$

$$\frac{\partial^2 S_\lambda(\mu)}{\partial \mu_k \partial \mu_l} = 0 \qquad l \neq k \tag{B.3}$$

Hence the matrix of second derivatives of $S_\lambda(\mu)$ is a diagonal matrix that leads to a strictly convex penalty if and only if

$$\sum_{s\neq k}\lambda_{ks} > -n_k \text{ for all } k \in \{1, \ldots, T\}.$$

An estimator defined as the solution to (2.2) is then a unique global minimizer. ■

C. Supplementary Material for Section 3

C.1. Predictive MSE = Regular MSE

$$\begin{aligned} E[(Y-\hat{Y})'(Y-\hat{Y})] &= E[(D(\mu-\hat{\mu})+\varepsilon)'(D(\mu-\hat{\mu})+\varepsilon)] \\ &= E[(\mu-\hat{\mu})'D'D(\mu-\hat{\mu})] + E[\varepsilon'\varepsilon] \end{aligned}$$

which in the case of the PCS is proportional to

$$\sum_{k=1}^{J} E[(\mu_k - \hat{\mu}_k^{PCS})^2 n_k] = \sum_{k=1}^{J} E[(\mu_k - \sum_{j\neq k}\omega_{kj}\hat{\mu}_j - \omega_{kk}\bar{Y}_k)^2 n_k],$$

where $\omega_{kk} = 1 - \sum_{j\neq k}\omega_{kj}$. Since the n_k just scale up the kth squared difference, minimization is not affected. Removing the n_k leaves us with parameter MSE. ■

C.2. Proof of Proposition 3.1

By using a linear projection, we have $\hat{\mu}_j = \bar{Y}_j$ for all j. Then,

$$\hat{\mu}_k^{PCS} = (1 - \sum_{j \neq k} \omega_{kj})\bar{Y}_k + \sum_{j \neq k} \omega_{kj} \bar{Y}_j.$$

For cell averages holds: $E[\bar{Y}_k] = \mu_k$, $V[\bar{Y}_k] = \sigma_k^2 / n_k$. Then,

$$\begin{aligned} MSE(\hat{\mu}_k) &= bias(\hat{\mu}_k)^2 + V[\hat{\mu}_k] \\ &= \Big((1 - \sum_{j \neq k} \omega_{kj})\mu_k + \sum_{j \neq k} \omega_{kj}\mu_j - \mu_k \Big)^2 + (1 - \sum_{j \neq k} \omega_{kj})^2 \frac{\sigma_k^2}{n_k} + \sum_{j \neq k} \omega_{kj}^2 \frac{\sigma_j^2}{n_j} \\ &= \Big(- \sum_{j \neq k} \omega_{kj}(\Delta\mu_{kj}) \Big)^2 + (1 - \sum_{j \neq k} \omega_{kj})^2 \frac{\sigma_k^2}{n_k} + \sum_{j \neq k} \omega_{kj}^2 \frac{\sigma_j^2}{n_j} \\ &= \Big(\sum_{j \neq k} \omega_{kj}(\Delta\mu_{kj}) \Big)^2 + (1 - \sum_{j \neq k} \omega_{kj})^2 \frac{\sigma_k^2}{np_k} + \sum_{j \neq k} \omega_{kj}^2 \frac{\sigma_j^2}{np_j} + O_p(n^{-1}). \end{aligned}$$

■

C.3. Proof of Theorem 3.1

Only the most important steps are mentioned here. A more detailed can be provided upon a request.

Let $V_k = \sigma^2 / np_k$. Note that only the smoothing parameters with baseline group k are relevant in (3.2). Then, optimizing with respect to ω_{kl} yields[10]:

$$\begin{aligned} \frac{\partial MSE(\hat{\mu}_k^{PCS})}{\partial \omega_{kl}} &= -2(1 - \sum_{j \neq k} \omega_{kj})V_k + 2\omega_{kl}V_l + 2\sum_{j \neq k} \omega_{kj}\Delta\mu_{kj}\Delta\mu_{kl} \overset{!}{=} 0 \\ \Leftrightarrow \omega_{kl}V_l &= (1 - \sum_{j \neq k} \omega_{kj})V_k - \Delta\mu_{kl} \sum_{j \neq k} \omega_{kj}\Delta\mu_{kj} \end{aligned}$$

Using this result one can derive:

$$\frac{\omega_{kl}V_l}{1 + \sum_{m \neq k} \frac{\Delta\mu_{km}\Delta\mu_{lm}}{V_m}} = \frac{(1 - \sum_{j \neq k} \omega_{kj})V_k}{1 + \sum_{m \neq k} \frac{\Delta\mu_{km}^2}{V_m}}.$$

Note that the right hand side does not depend on l. Using this equation for two indeces

[10] All ω_{kl}'s in the following text should have a star superscript which is left out unless necessary for readability.

l and j yields:

$$\frac{\omega_{kj}V_j}{1+\sum_{m\neq k}\frac{\Delta\mu_{km}\Delta\mu_{jm}}{V_m}} = \frac{\omega_{kl}V_l}{1+\sum_{m\neq k}\frac{\Delta\mu_{km}\Delta\mu_{lm}}{V_m}}$$

$$\omega_{kj} = \frac{\omega_{kl}V_l}{1+\sum_{m\neq k}\frac{\Delta\mu_{km}\Delta\mu_{lm}}{V_m}} \frac{1+\sum_{m\neq k}\frac{\Delta\mu_{km}\Delta\mu_{jm}}{V_m}}{V_j}.$$

Plugging it back into FOC equation leads to:

$$\omega^*_{kj} = \frac{V_k}{V_j a_{kj}},$$

where $a_{kj} = \left(1 + \frac{\sigma_k^2/np_k}{1+b_{kj}}\sum_{l\neq k}\frac{1+b_{kl}}{\sigma_l^2/np_l} + \frac{\Delta\mu_{kj}}{1+b_{kj}}\sum_{l\neq k}\frac{\Delta\mu_{kl}}{\sigma_l^2/np_l}\right)$ with $b_{kj} = \sum_{m\neq k}\frac{\Delta\mu_{km}\Delta\mu_{jm}}{\sigma_m^2/np_m}$.

The rest comes from a simple algebra. ■

C.4. Comparative Statics of Smoothing Parameters in Small Samples

This subsection contains all additional material regarding the small sample analysis of the MSE optimal smoothing parameters.

Discussion - Design (B) The effects for the change in n under homoscedasticity are plotted in Figure C.4 and one can see the exactly the same effect for n_2 and n_3 which was described in the main text. Introducing heteroscedasticity has no effect on almost zero smoothing of group 1 towards group 4. The effect on smoothing towards groups 1, 2 and 3 is the same as in the case of equal means, i.e. the higher the variance, the lower the smoothing weight, see Figure C.6. The effects for the change in σ^2 under heteroscedasticity are in Figure C.7and were described in the main text.

Shifting μ_2 away from zero leads to a situation in which only groups 1 and 3 are groups with zero means. Therefore, these two groups get high smoothing weights, see Figure C.5. Shifting μ_2 towards positive values has the following effect. Parameter ω_{12} decreases sharply as the means of group 1 and 3 are a way more sensible target groups. After reaching $\mu_2 = 100$, ω_{12} becomes even negative. Smoothing towards groups 1 and 3 increases until μ_2 reaches 100, then it starts slightly decreasing. The intuition is that $\mu_2 \geq 100$ is so extreme that it even pays off to smooth towards the group 4 for a small cost of a lower smoothing weight towards group 1 and 3. Parameter ω_{14} has a U-shape on the interval of $\mu_2 \in [0, 100]$, i.e. until $\mu_2 = 50$, group 2 is a better target. Afterwards they become similar and after μ_2 exceeds 100, the group 4 even gains higher smoothing weight due to being closer to group 1.

Shifting μ_2 towards negative values causes parameter ω_{12} to decrease. But not as sharply as in the opposite direction, since there is no other competing group with a negative mean to which it would be more sensible to smooth. Smoothing towards groups 1 and 3 does not change much. Parameter ω_{14} gains slowly on importance as μ_2 becomes more and more negative and gets a higher weight than group 2 after μ_2 exceeds -100.

By replacing the group 2 and μ_2 by group 3 and μ_3 respectively, all the results in the last two paragraphs are valid for shifting μ_3. Notice also that, when the target groups 1, 2 and 3 have a zero mean at the same time, the group 1 is shrunken to each of them equally and has almost a zero smoothing weight to the group 4.

Under heteroscedasticity, relatively larger error variances lead to flatter curves and relatively smaller variances lead to more amplified curves but the qualitative results described above do not change. This effect is illustrated in Figure C.8.

Discussion - Design (C) Since groups 1, 3 and 4 have very close means, shifting μ_2 to more extreme values than -2 or 4 makes μ_2 to be a distant group mean, i.e. the smoothing towards group 2 is very close to zero or even slightly negative on these two intervals. Once the smoothing weight towards group 2 is close to zero, pushing μ_2 to more extreme values does not have any effect on any of the smoothing weights and all the smoothing weights stabilize as before by ignoring the changes in the distant mean. Therefore, all the interesting effects are on the interval where $\Delta\mu_{12} \in [-4, 2]$. Beyond these values the smoothing weights just converge to a stable level of smoothing, see Figure C.11.

Within the framework of shifting the mean of group 2, if $\Delta\mu_{12} = 0$, i.e. $\mu_2 = 0$, then the parameter ω_{13} loses the smoothing weight because it is the only unequal mean. Parameter ω_{12} gets a relatively high weight and the other parameters get high smoothing weights accordingly to the error variances of the target groups. When $\Delta\mu_{12} = -2$, i.e. $\mu_2 = 2$, groups 1 and 4 get their maximum weights since they are more informative then the other two groups. Even though the μ_2 and μ_3 are not that far from μ_1, the group 1 is not smoothed to them that much because they compete with each other by having the same mean. Group 2 is mainly punished for its large variance and therefore it gets a lower weight than group 3. When $\Delta\mu_{12} = -4$, i.e. $\mu_2 = 4$, then the group 3 starts being more important for group 1, as it is a more sensible target then group 2 and it has a lower variance. When $\Delta\mu_{12} = 2$, i.e. $\mu_2 = -2$, group 3 gains smoothing weight since the mean of group 2 is not 0 anymore. The increase in ω_{13} is compensated by a decrease in weights for groups 1 and 4, since these profited before from 3 equal means and reduced the weight for group 3, the only unequal mean. The mean of group 2 gains a little bit in this case in comparison to $\Delta\mu_{12} = -2$, since it is not now a direct competitor to group 3 as they have now different means. So, now it contributes more to the smoothing. Of course, as a group with the closest mean and lowest variance, group 4 gets the largest smoothing weight.

Discussion - Summary All the results are shortly repeated here. This text can be skipped without any influence on understanding of the rest of the paper.

For target groups whose group mean is close to the mean of the base category, more observations and lower variances in the target group contribute to a large own smoothing weight and decrease the other smoothing weights. If the target group mean is far from the mean of the base category, then it is ignored by getting a close to zero or even negative own smoothing weight and other smoothing weights are not affected by changes in the number of observations or variances of the target group.

Regarding the changes in the mean differences, the smoothing parameters are mainly influenced by the size of the difference and by a presence of another close group mean. If the base category is far away from all the other group means, the most smoothing weight goes into its own smoothing parameter and the rest of the groups is ignored. Surprisingly, a presence of one distant mean (design (B)) helps to stabilize the weights in a sense that the base category is strongly smoothed to other close or equal means and weights change rather smoothly with a shift in any mean. However, when all the means are very close to each other (design (C) - shifting μ_2), the smoothing parameters are sensitive in the narrow interval covering a close distance between the means and then with a larger distance they stabilize around certain levels depending on the distances to the other means and their error variances. This behavior could potentially cause problems for any estimate of the smoothing parameter that is subject to small sample variation. Finite samples deviations from the true parameters might yield smoothing parameters far away from the optimal ones leading to unfavorable aggregation.

Figure C.1: Effect of n_1, n_2 and n_3 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (A), Homoscedasticity

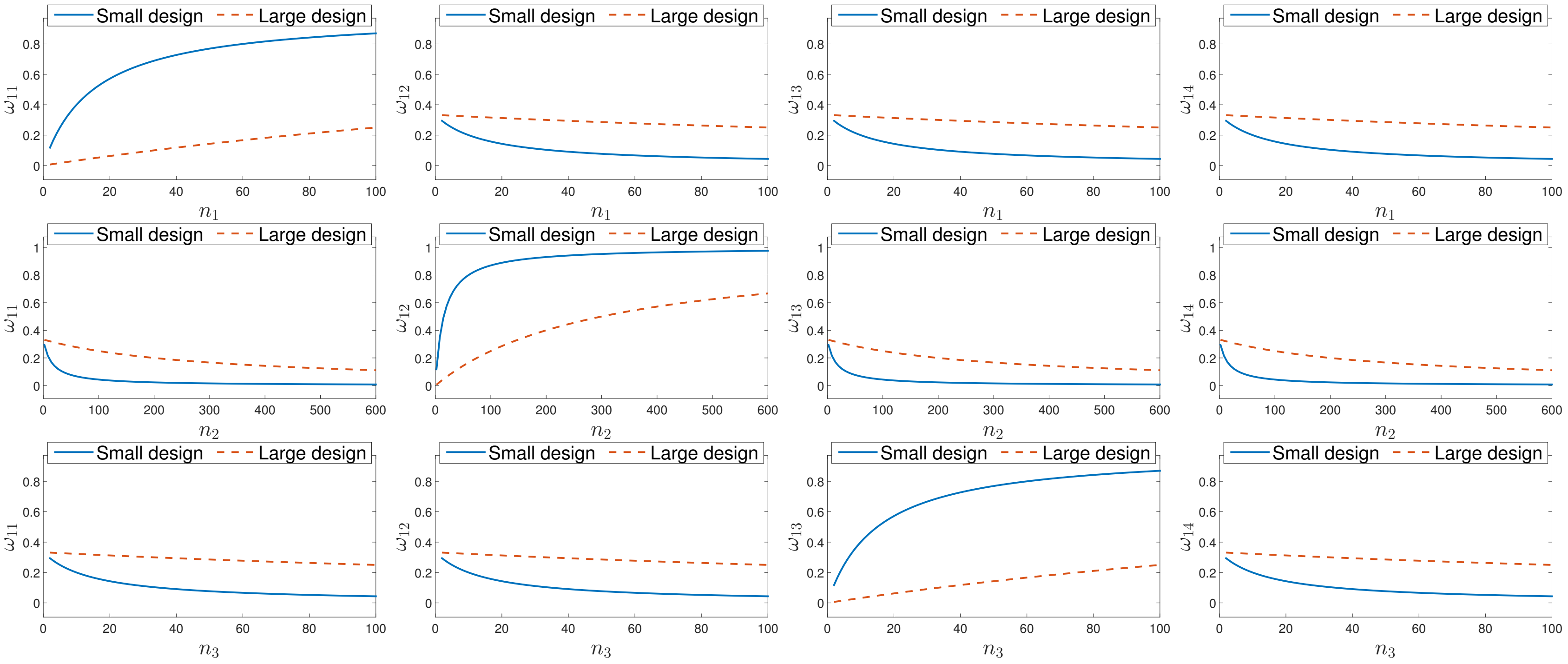


Figure C.2: Effect of n_1, n_2 and n_3 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (A), Heteroscedasticity

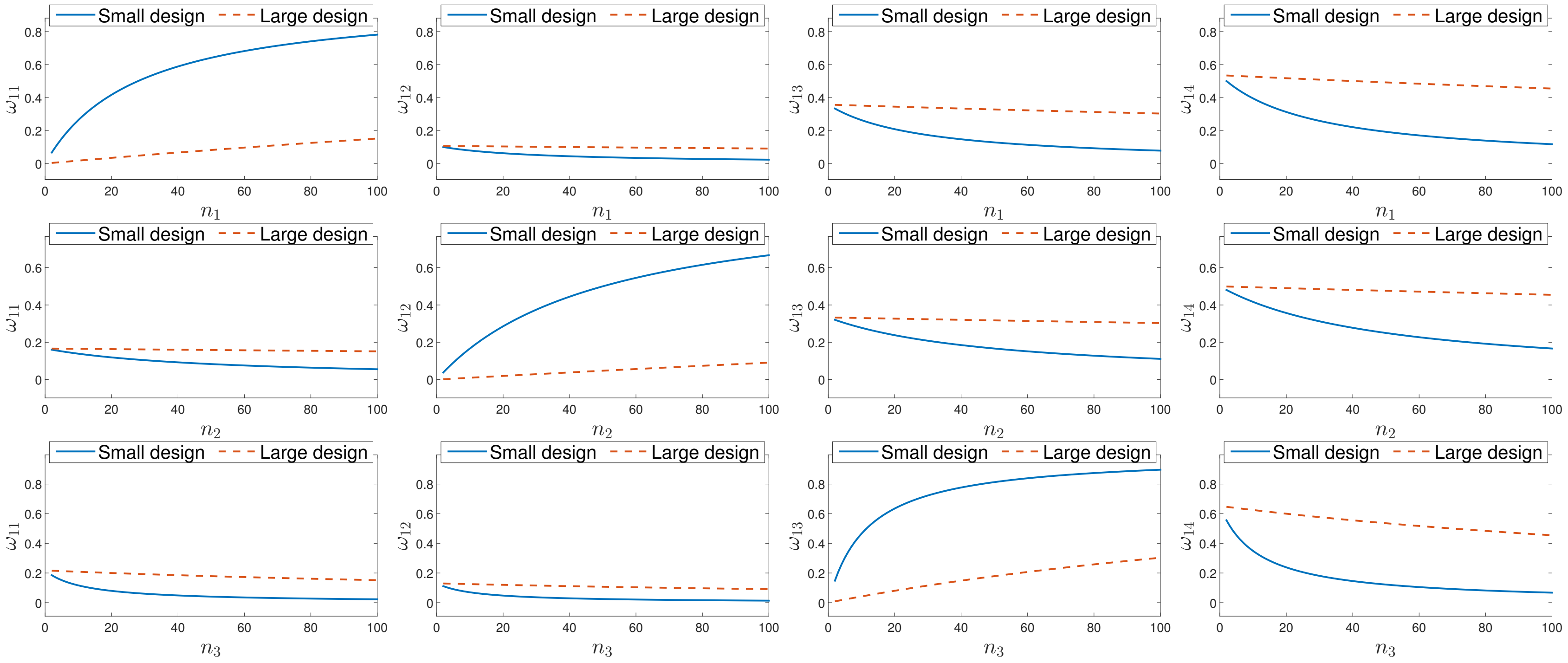


Figure C.3: Effect of σ_1^2, σ_2^2 and σ_3^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (A), Heteroscedasticity

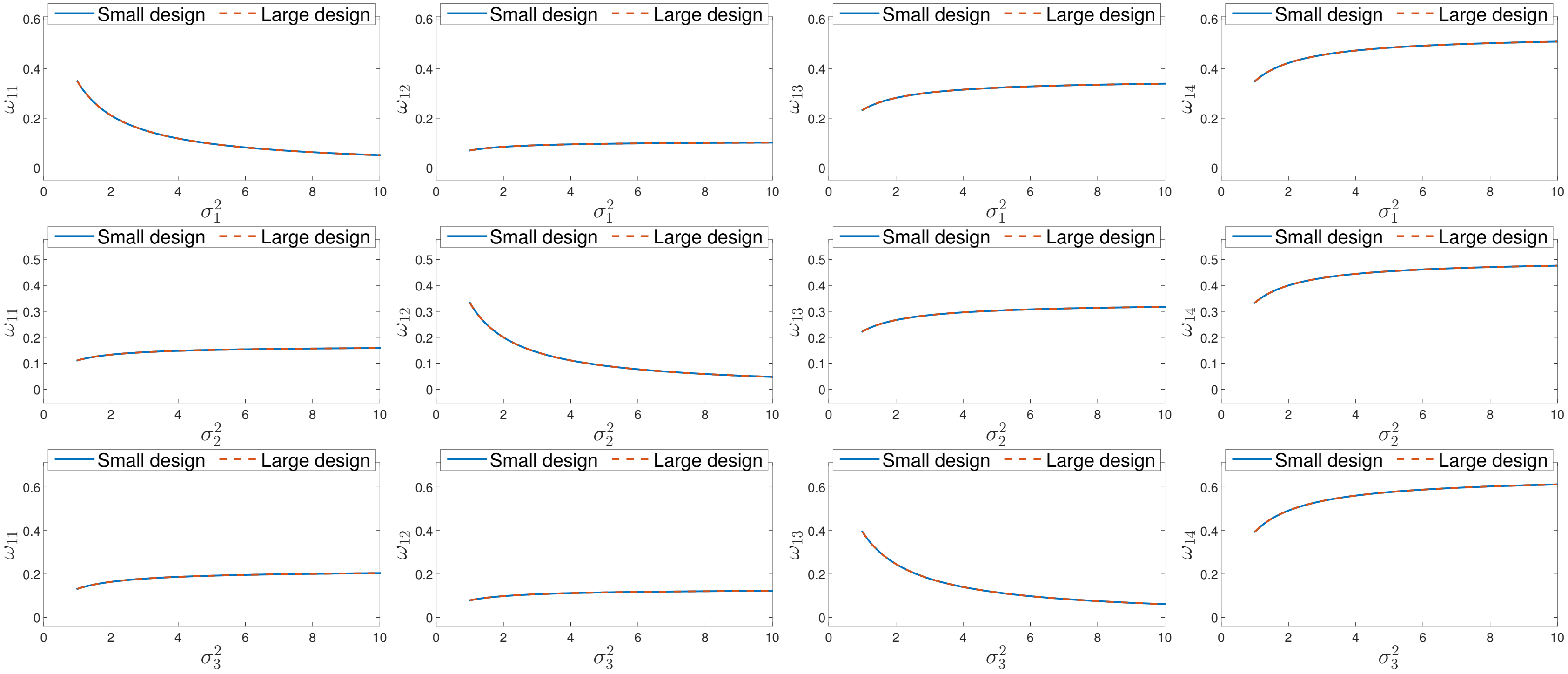


Figure C.4: Effect of n_1, n_2 and n_3 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (B), Homoscedasticity

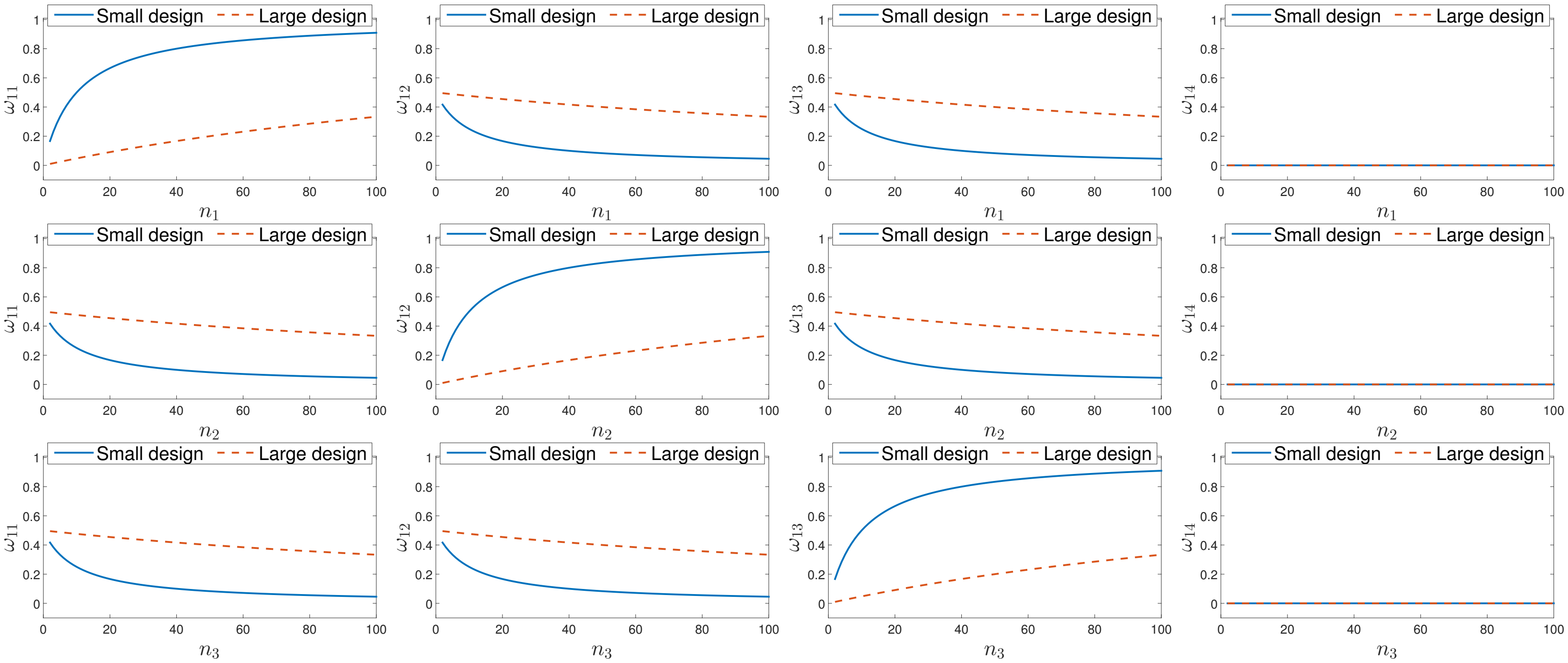


Figure C.5: Effect of $\Delta\mu_{12}$, $\Delta\mu_{23}$ and $\Delta\mu_{13}$ on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (B), Homoscedasticity

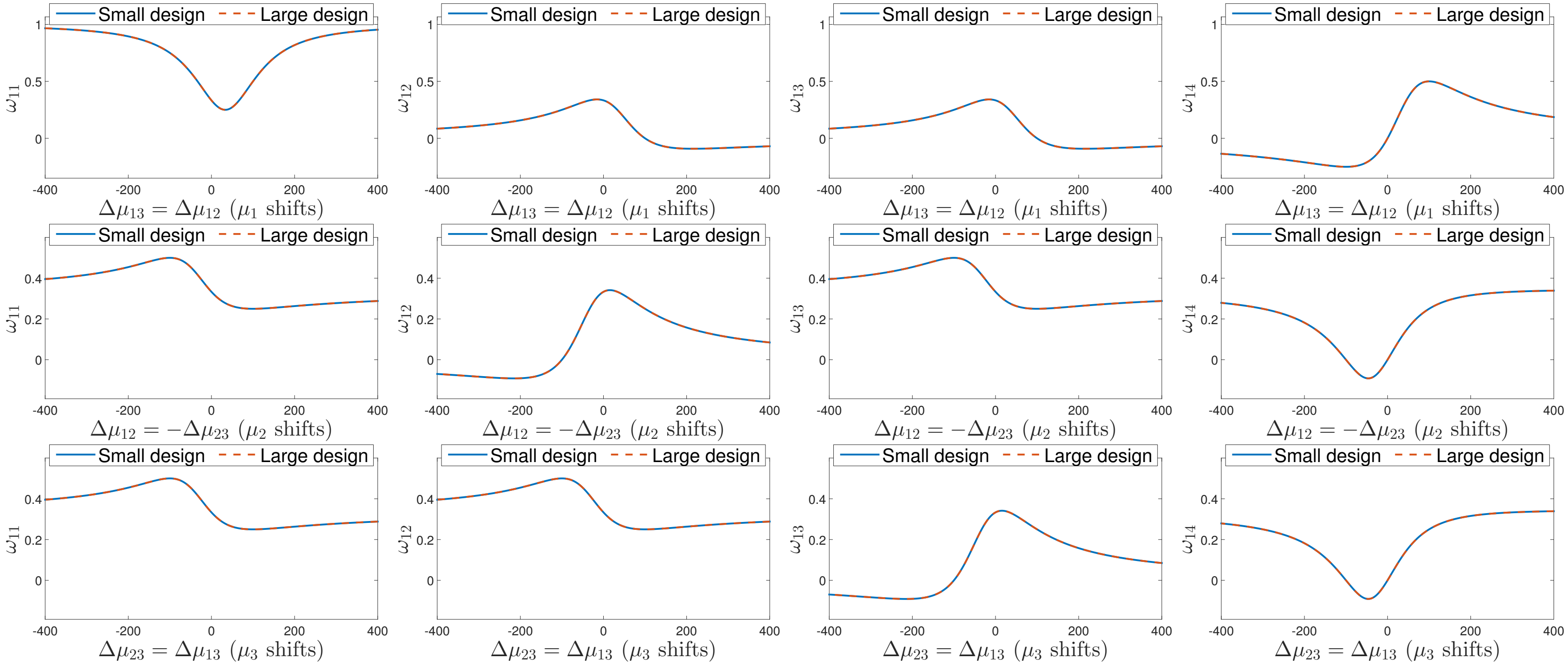


Figure C.6: Effect of n_1, n_2 and n_3 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (B), Heteroscedasticity

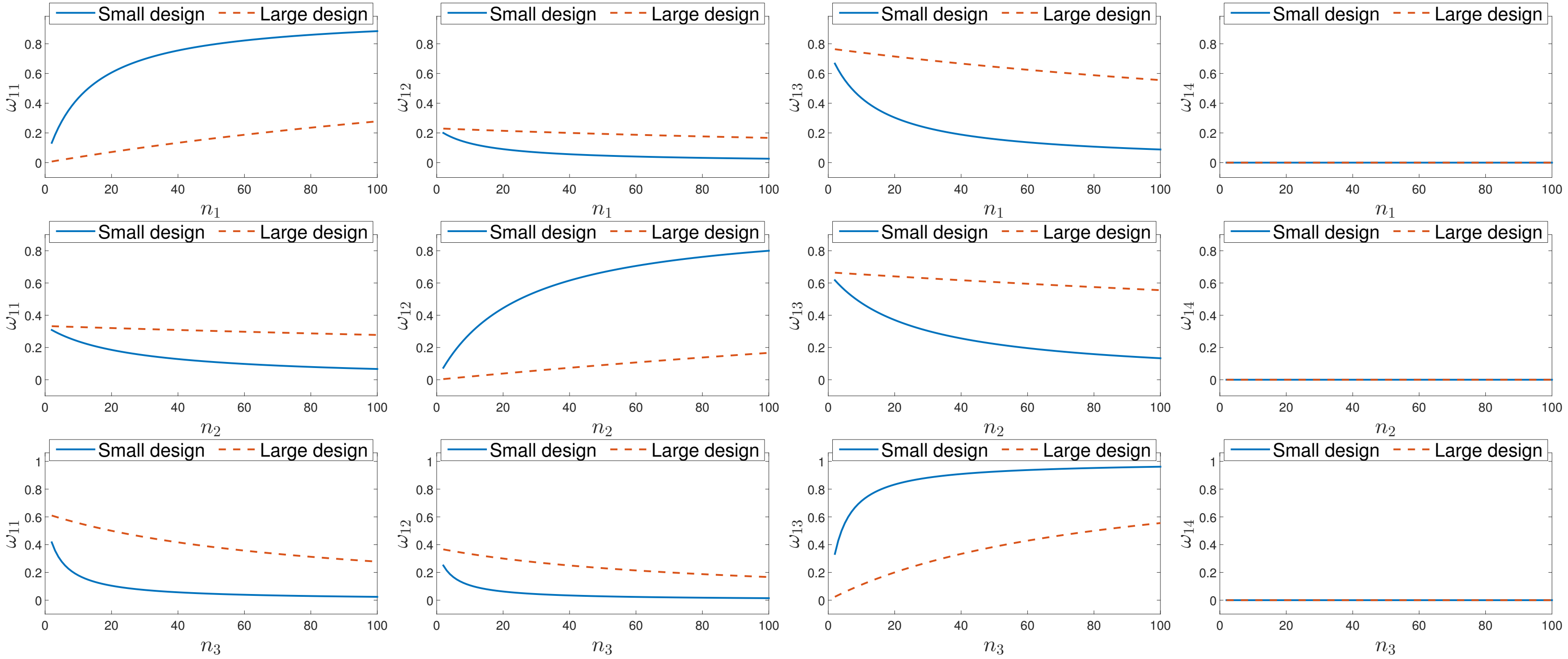


Figure C.7: Effect of σ_1^2, σ_2^2 and σ_3^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (B), Heteroscedasticity

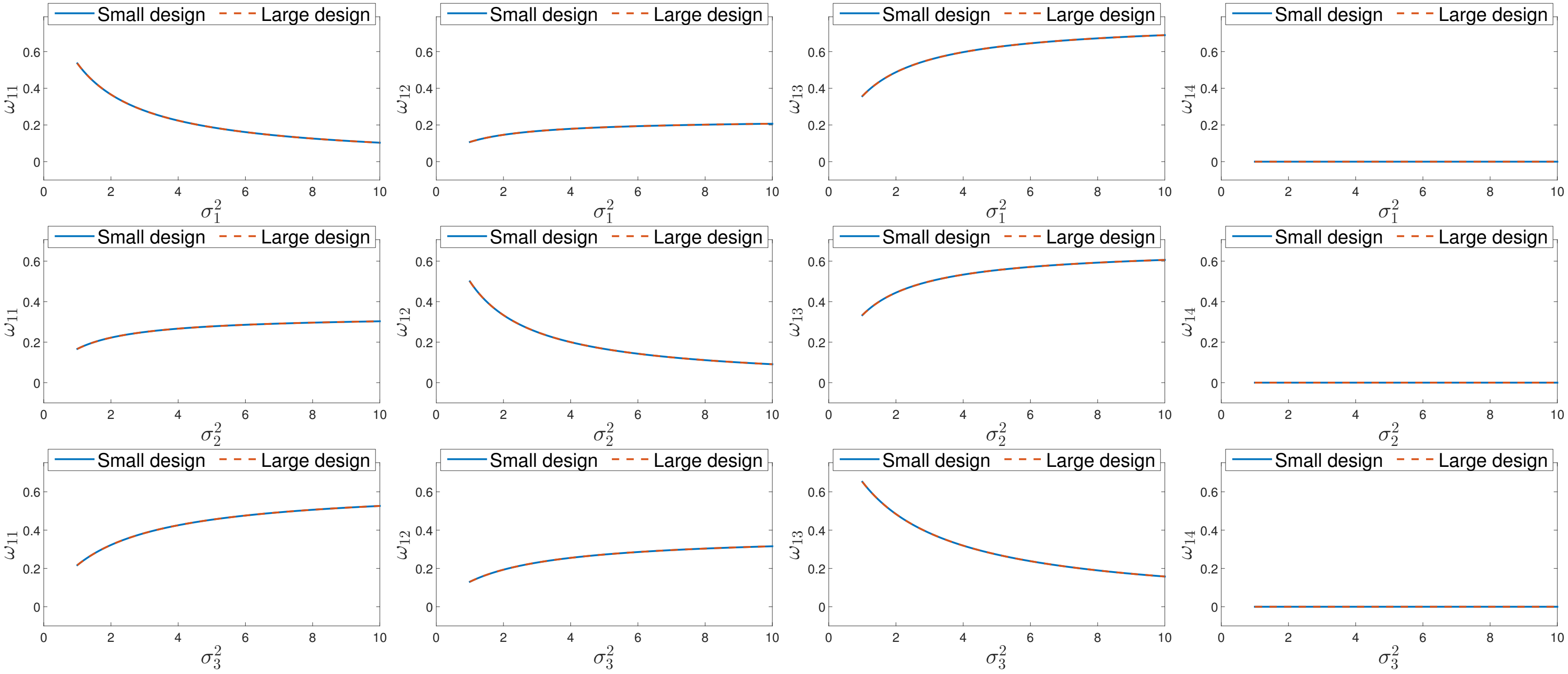


Figure C.8: Effect of $\Delta\mu_{12}$, $\Delta\mu_{23}$ and $\Delta\mu_{13}$ on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (B), Heteroscedasticity

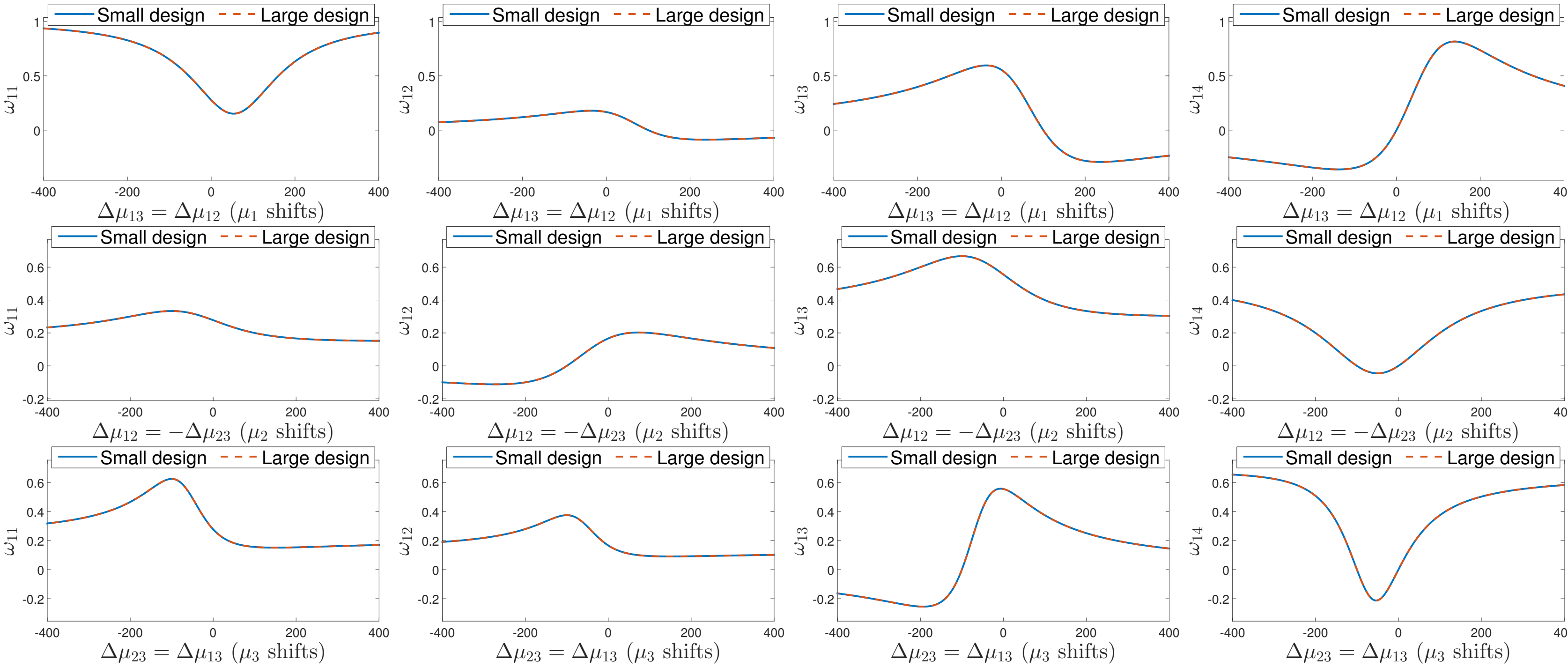


Figure C.9: Effect of n_1, n_2 and n_3 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (C), Heteroscedasticity

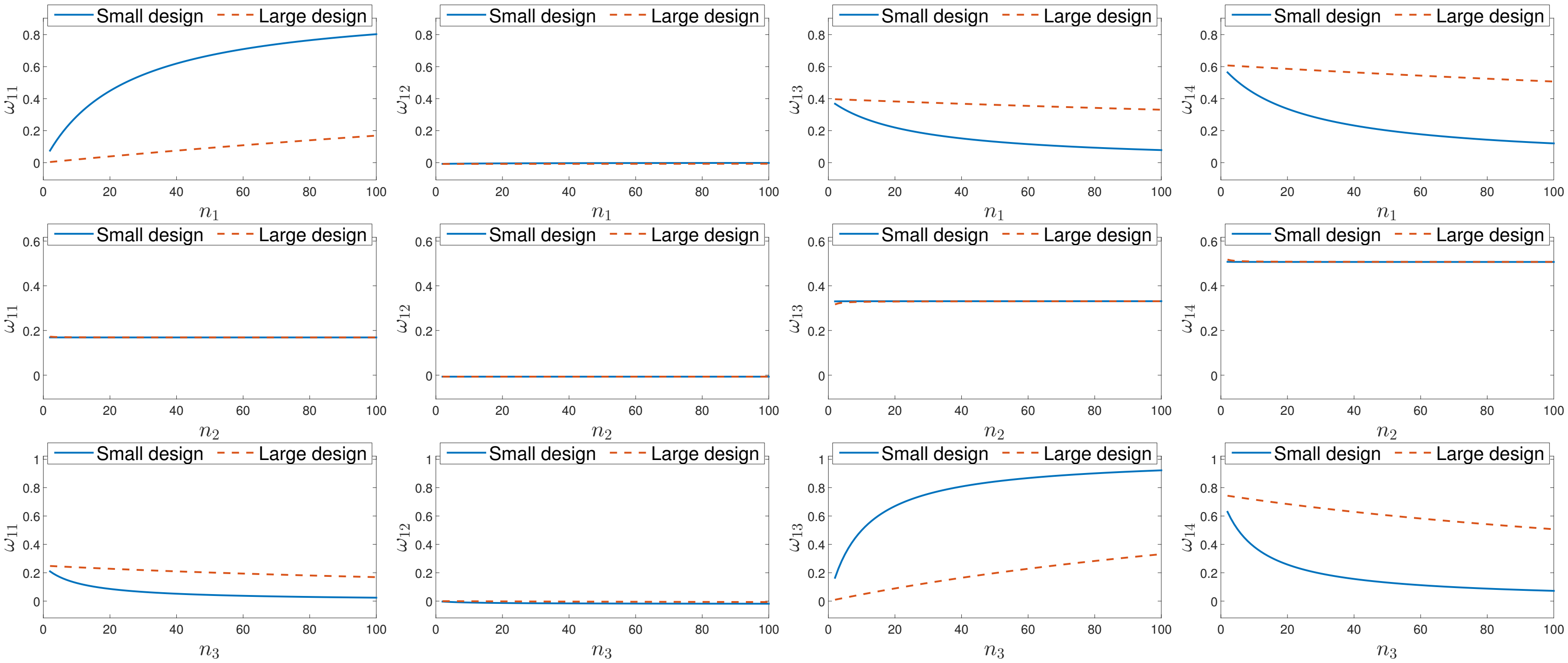


Figure C.10: Effect of σ_1^2, σ_2^2 and σ_3^2 on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (C), Heteroscedasticity

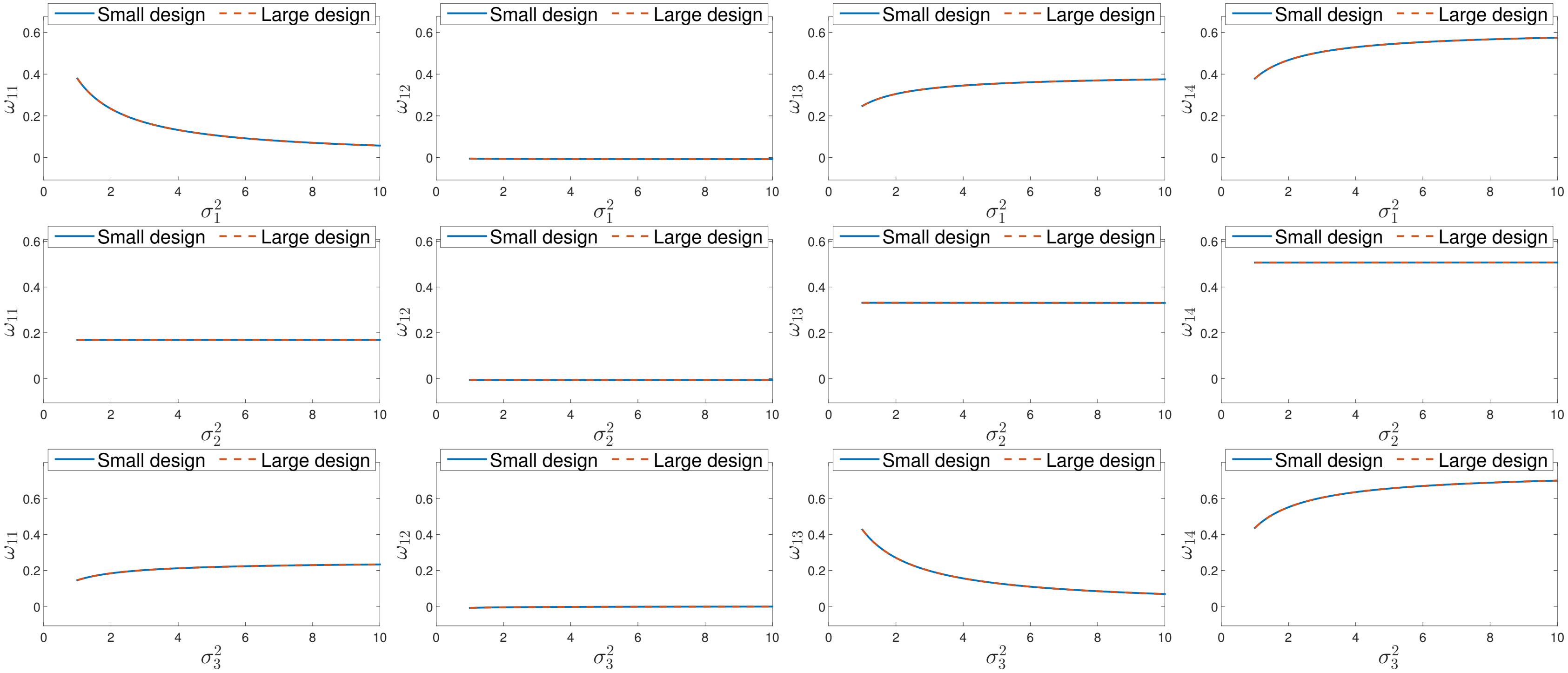
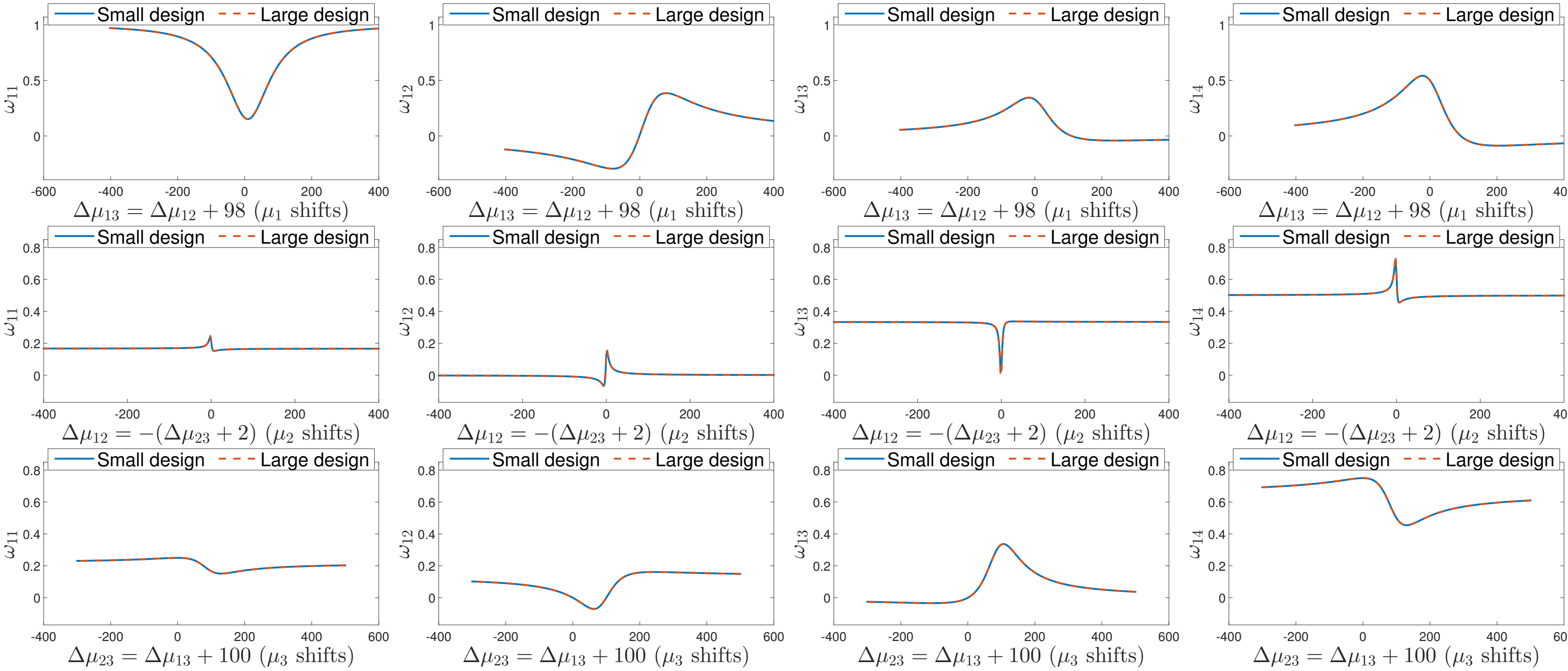


Figure C.11: Effect of $\Delta\mu_{12}$, $\Delta\mu_{23}$ and $\Delta\mu_{13}$ on ω_{11}, ω_{12}, ω_{13}, ω_{14}, Design (C), Heteroscedasticity



C.5. Proof of Proposition 3.2

Let e_k be the k-th unit vector. Note that by definition of the minimization problem and well-known properties of the OLS, we have that

$$0 \leq MSE(\hat{\mu}_k^{PCS}(\Lambda_k^*)) \leq MSE(\hat{\mu}_k^{PCS}(e_k)) = MSE(\hat{\mu}_k) = O(n^{-1}).$$

Additional conclusion: Hence there exists an N, M such that $MSE(\hat{\mu}_k) \leq M/n$ for all $n \geq N$, since the left hand side is an upper bound for $MSE(\hat{\mu}_k^{PCS}(\Lambda_k^*))$ we have to have that the latter is $O(n^{-1})$. From the closed form of the MSE, we know that

$$\begin{aligned} MSE(\hat{\mu}_k^{PCS}(\Lambda_k^*)) &= O\Big((\sum_{j=1}^{J} \omega_{kj}^* \Delta\mu_{kj})^2 + \frac{1}{n} \sum_{j=1}^{J} {\omega_{kj}^*}^2 \frac{\sigma_j^2}{p_j} \Big) \\ &= O\Big(\max_{j,l} \omega_{kj}^* \omega_{kl}^* \Delta\mu_{kj} \Delta\mu_{kl} \Big) + O\Big(\frac{1}{n} \max_{j} {\omega_{kj}^*}^2 \Big) \end{aligned}$$

and hence a necessary condition in line with the PCS MSE rate is that $O(\max_j {\omega_{kj}^*}^2) = O(1)$ which implies that $\omega_{kj}^* = O(1)$ for all k, j. ■

C.6. Representation of $\hat{\omega}_{kj}$

For this derivation, we work with the representation of the weighted average directly. Recall that $\hat{\omega}_{kj} = \frac{\hat{\sigma}_k^2 n_j / \hat{\sigma}_j^2 n_k}{\hat{a}_{kj}}$. The estimators using the consistent plug-in yield the following terms:

$$\begin{aligned} \Delta\hat{\mu}_{km}\Delta\hat{\mu}_{jm} = &(\hat{\mu}_k - \mu_k)(\hat{\mu}_j - \mu_j) - (\hat{\mu}_k - \mu_k)(\hat{\mu}_m - \mu_m) - (\hat{\mu}_m - \mu_m)(\hat{\mu}_j - \mu_j) \\ &+ (\hat{\mu}_m - \mu_m)^2 + (\hat{\mu}_k - \mu_k)(\mu_j - \mu_m) - (\hat{\mu}_m - \mu_m)(\mu_j - \mu_m) \\ &+ (\mu_k - \mu_m)(\hat{\mu}_j - \mu_j) - (\mu_k - \mu_m)(\hat{\mu}_m - \mu_m) + (\mu_k - \mu_m)(\mu_j - \mu_m) \end{aligned}$$

Let $z_m = \sqrt{n_m}(\hat{\mu}_m - \mu_m)/\sigma_m$. It follows that:

$$\begin{aligned}\frac{\Delta\hat{\mu}_{km}\Delta\hat{\mu}_{jm}}{\hat{\sigma}_m^2} n_m = {} & z_k z_j \frac{\sigma_k \sigma_j n_m}{\hat{\sigma}_m^2 \sqrt{n_k n_j}} - z_k z_m \frac{\sigma_k \sigma_m \sqrt{n_m}}{\hat{\sigma}_m^2 \sqrt{n_k}} - z_m z_j \frac{\sigma_m \sigma_j \sqrt{n_m}}{\hat{\sigma}_m^2 \sqrt{n_j}} \\ & + z_m^2 \frac{\sigma_m^2}{\hat{\sigma}_m^2} + z_k \sqrt{n}(\mu_j - \mu_m) \frac{\sigma_k n_m}{\hat{\sigma}_m^2 \sqrt{n_k n}} - z_m \sqrt{n}(\mu_j - \mu_m) \frac{\sigma_m \sqrt{n_m}}{\hat{\sigma}_m^2 \sqrt{n}} \\ & + z_j \sqrt{n}(\mu_k - \mu_m) \frac{\sigma_j n_m}{\hat{\sigma}_m^2 \sqrt{n_j n}} - z_m \sqrt{n}(\mu_k - \mu_m) \frac{\sigma_m \sqrt{n_m}}{\hat{\sigma}_m^2 \sqrt{n}} + n(\mu_k - \mu_m)(\mu_j - \mu_m) \frac{n_m}{\hat{\sigma}_m^2 n}.\end{aligned}$$

Applying the same manipulations to $\Delta\hat{\mu}_{kj}\Delta\hat{\mu}_{kl}$ and plugging all results into the estimate for a_{kl}, one obtains the estimated weights:

$$\hat{\omega}_{kj} = \left[\frac{Z'\hat{m}_1 Z + \sqrt{n}\Delta'\hat{m}_2 Z + n\Delta'\hat{m}_3\Delta + \hat{c}_0}{Z'\hat{M}_1 Z + \sqrt{n}\Delta'\hat{M}_2 Z + n\Delta'\hat{M}_3\Delta + 1} + 1\right]^{-1} \frac{n_j \hat{\sigma}_k^2}{n_k \hat{\sigma}_j^2}$$

with $Z = (z_1, \ldots, z_J)$, $\Delta = (\mu_1 - \mu_1, \mu_1 - \mu_2, \ldots, \mu_J - \mu_J)$ and $\hat{m}_1, \hat{m}_2, \hat{m}_3, \hat{M}_1, \hat{M}_2, \hat{M}_3, (\hat{c}_0)$ being random matrices (scalar) that converge(s) in probability.

D. Supplementary Material for Section 4

D.1. Distribution under Distant Systems

$$\begin{aligned}\hat{\mu}^{PCS}(\hat{\Lambda}_k) &= \sum \hat{\omega}_{kj}\hat{\mu}_j \\ &= \sum \omega_{kj}^* \hat{\mu}_j + \sum (\hat{\omega}_{kj} - \omega_{kj}^*)\mu_j + \sum (\hat{\omega}_{kj} - \omega_{kj}^*)(\hat{\mu}_j - \mu_j)\end{aligned}$$

$$\begin{aligned}\Leftrightarrow \sqrt{n}(\hat{\mu}^{PCS}(\hat{\Lambda}_k) - \mu_k - \sum \omega_{kj}^* \Delta\mu_{jk}) &= \sum \omega_{kj}^* \sqrt{n}(\hat{\mu}_j - \mu_j) \\ &\quad + \sum \sqrt{n}(\hat{\omega}_{kj} - \omega_{kj}^*)\Delta\mu_{jk} + O_p(n^{-1/2})\end{aligned}$$

Let $[] \equiv (\hat{\mu}^{PCS}(\hat{\Lambda}_k) - \mu_k - \sum \omega_{kj}^* \Delta\mu_{j,k})$ be the debiased estimator. Since the estimated smoothing parameters are continuous functions of the same random vector as the OLS estimator for the location variance and cell probabilities, asymptotic normality and joint convergence in distribution is guaranteed. Hence, one can apply the Delta method to obtain asymptotically valid confidence bounds. In general, let Σ be the joint asymptotic variance covariance matrix of the first stage estimated means, variances and cell probabilites. We require that

$$\sqrt{n}\begin{pmatrix}\hat{\mu}_1-\mu_1\\ \vdots\\ \hat{\mu}_J-\mu_J\\ \hat{\sigma}_1^2-\sigma_1^2\\ \vdots\\ \hat{\sigma}_J^2-\sigma_J^2\\ \hat{p}_1-p_1\\ \vdots\\ \hat{p}_J-p_J\end{pmatrix}\xrightarrow{d} N(0,\Sigma).$$

It then follows that

$$\sqrt{n}(\hat{\mu}^{PCS}(\hat{\Lambda}_k)-\mu_k-\sum\omega^*_{kj}\Delta\mu_{jk})\xrightarrow{d} N(0,G'\Sigma G)$$

where $G=\nabla[]$ is the gradient of the debiased estimator with respect to all the elements $(\hat{\mu}_1-\mu_1,\ldots,\hat{p}_j-p_j)'$.

Under homoskedasticity and first stage estimator being OLS, it simplifies to

$$\sqrt{n}\begin{pmatrix}\hat{\mu}_1-\mu_1\\ \vdots\\ \hat{\mu}_J-\mu_J\\ \hat{\sigma}^2-\sigma^2\\ \hat{p}_1-p_1\\ \vdots\\ \hat{p}_J-p_J\end{pmatrix}\xrightarrow{d} N\left(0,\begin{bmatrix}\sigma^2 diag(p_j)^{-1} & [0] & [0]\\ 0 & \sigma^4(\kappa-1) & 0\\ [0] & [0] & diag(p_j(1-p_j))\end{bmatrix}\right).$$

where *kappa* is a kurtosis of Y_i. Applying the delta method, yields an asymptotic variance covariance matrix that is identical to the OLS under homoskedasticity, i.e. $G'\Sigma G=\sigma^2 diag(p_j)^{-1}$.

D.2. Proof of Theorem 4.1

Note that $tr(\Pi(\Lambda^*)) = \sum_{k=1}^{J} \omega_{kk}^*$. First assume equality of all means, it follows that

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

If $\#\{(k,j) : \mu_k \neq \mu_j, k = 1, \ldots, J-1, j > k\} > 0$, then

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

Note in general it holds that

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

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

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

For the estimated smoothing parameters however the equality only holds for probability limits and hence we have

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

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

E. Supplementary Material for Section 7

Table E.1: Means, Variances and Number of Observations in Card and Krueger (1994) Data

	NJ (treated)						PEN (control)					
Chain	Before			After			Before			After		
	$\bar{\mu}$	σ^2	n	$\bar{\mu}$	σ^2	n	$\bar{\mu}$	σ^2	n	$\bar{\mu}$	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