ARTICLE IN PRESS

Journal of Econometrics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Shrinkage for categorical regressors

Phillip Heiler [a,*], Jana Mareckova [b]

[a] Aarhus University, Department of Economics and Business Economics, CREATES, TrygFonden's Centre for Child Research, Fuglesangs Allé 4, 8210 Aarhus V, Denmark

[b] Schweizerisches Institut für empirische Wirtschaftsforschung, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland

ARTICLE INFO

Article history:
Received 2 November 2018
Received in revised form 7 July 2020
Accepted 23 July 2020
Available online xxxx

JEL classification:
C25
C51
C52

Keywords:
Categorical regressors
Regularization
Smoothing kernels
Model averaging

ABSTRACT

This paper introduces a flexible regularization approach that reduces point estimation risk of group means stemming from e.g. categorical regressors, (quasi-)experimental data or panel data models. The loss function is penalized by adding weighted squared ℓ_2-norm differences between group location parameters and informative first stage estimates. Under quadratic loss, the penalized estimation problem has a simple interpretable closed-form solution that nests methods established in the literature on ridge regression, discretized support smoothing kernels and model averaging methods. We derive risk-optimal penalty parameters and propose a plug-in approach for estimation. The large sample properties are analyzed in an asymptotic local to zero framework by introducing a class of sequences for close and distant systems of locations that is sufficient for describing a large range of data generating processes. We provide the asymptotic distributions of the shrinkage estimators under different penalization schemes. The proposed plug-in estimator uniformly dominates the ordinary least squares estimator in terms of asymptotic risk if the number of groups is larger than three. Monte Carlo simulations reveal robust improvements over standard methods in finite samples. Real data examples of estimating time trends in a panel and a difference-in-differences study illustrate potential applications.

1. Introduction

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

In this paper we propose a flexible penalization approach called pairwise cross-smoothing (PCS) to improve on the issue of point estimation risk. The method penalizes the loss function by adding sums of weighted squared ℓ_2-norm differences between group location (reference) parameters and informative first stage estimates (targets). It nests existing smoothing and model averaging methods for orthogonal regressors, has favorable computational costs due to closed-form solutions for both estimator and penalty or smoothing parameters and can easily be extended to the case of mixed data. Using a simple projection as a first stage estimate, we derive mean squared error optimal smoothing parameters and propose a plug-in approach for estimation. The additional flexibility provides a substantial decrease in the oracle risk bounds compared to more restrictive aggregation methods such as (generalized) ridge regression or kernel smoothing.

* Corresponding author.
E-mail address: pheiler@econ.au.dk (P. Heiler).

https://doi.org/10.1016/j.jeconom.2020.07.051
0304-4076/

Please cite this article as: P. Heiler and J. Mareckova, Shrinkage for categorical regressors. Journal of Econometrics (2020), https://doi.org/10.1016/j.jeconom.2020.07.051.

We further contribute to the literature by analyzing the behavior of both estimated smoothing parameters as well as the PCS estimator in an asymptotic local to zero framework. We introduce a class of sequences for close and distant systems of locations that covers a wide range of data generating processes. We derive the asymptotic distribution of the PCS estimator under fixed, theoretically optimal and estimated smoothing parameters. In addition, we show that the feasible PCS dominates the OLS estimation risk uniformly over the class of sequences if the number of groups is larger than three. We also compare the asymptotic risk to an efficient parametric model averaging estimator. For aggregation problems such as combining two indicator variables or differences-in-differences type of setups the PCS estimator tends to dominate while for sufficiently high-dimensional problems the order reverses.

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

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

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

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

These smooth model averaging or shrinkage methods often have superior asymptotic risk properties over their non-shrunk counterparts. Hansen and Racine (2012) develop a jackknife model averaging estimator 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 averaging estimators in a nested least squares setup in a local to zero framework, i.e. weak partial correlations of additional regressors beyond a correctly specified base model. Liu (2015) derives distributional theory for least squares averaging estimators in the linear framework under different data-dependent weighting schemes and generalized error term structures. He considers a local to zero asymptotic framework for subsets of regressors and shows the nonstandard distributional behavior of the averaging estimators. Hansen (2016a) considers shrinkage of parametric models towards restricted parameter spaces under locally quadratic loss functions and provides sufficient conditions for risk dominance. Cheng et al. (2019) consider averaging between two general method of moments estimators under potential misspecification of the second, overidentified model. They show that their averaging estimator dominates the asymptotic risk of the base estimator uniformly over all degrees of misspecification if the shrinkage dimension is large enough. If applied to cell means or selection of orthogonal dummies many of these methods become special cases of PCS estimators, i.e. PCS estimators with a more restricted shrinkage subspace, and thus behave qualitatively similar in terms of asymptotic distribution and estimation risk. In addition they are closely related to classical shrinkage estimators that shrink parametric estimates towards constant vectors or restricted subspaces (Stein, 1956; Oman, 1982).

There is also a literature on regularization methods, in which coefficient estimates are enhanced by adding ℓ_1-norm penalties of pairwise differences which allow for partial and complete fusion of groups, see e.g. Tibshirani et al. (2005) for linear models and Tutz and Oelker (2017) for group-specific generalized linear models. The main differences to the other methods are nonsmooth aggregation, i.e. groups are set to be identical, and estimation that is done in a single, one-step procedure while e.g. model averaging directly and nonparametric smoothing implicitly use first stage estimates such as submodels or averages. These regularization methods are more suited for sparse high-dimensional applications but suffer from similar criticism as pre-testing or superefficient estimators, i.e. in finite samples actual risk gains can be inferior to standard likelihood or least squares approaches and heavily depend on the magnitude of the coefficients (Hansen, 2016b).

The direct or implicit aggregation that is introduced by all of these methods for regression models leads to the question of what an "optimal" aggregation rule is. Our framework allows for almost any linear aggregation based on a set of smoothing parameters.[1] Thus, comparing PCS to the aforementioned methods, the theoretical potential for a reduction in point estimation risk is generally larger and hence the PCS class also serves as a benchmark for future research.

The organization of the paper is as follows. Section 2 introduces the model, the pairwise cross-smoothing estimator and its connection to established smoothing and regularization methods. Section 3 presents the MSE optimal smoothing parameters, the plug-in estimator, and discusses the connection to frequentist model averaging methods. Section 4 introduces the local asymptotic framework and provides the distributional properties of the PCS estimator under fixed, optimal, and plug-in weights. It also contains the asymptotic risk properties of the feasible PCS estimator. Section 5 provides Monte Carlo evidence on estimation risk in finite samples. Section 6 contains the applications. Section 7 concludes. The proofs and major technical aspects are collected in Appendix A.

2. Pairwise cross-smoothing

In this section we introduce and discuss the model, the penalization strategy and the pairwise cross-smoothing estimator. Column vectors are denoted in boldface letters. Consider independent and identically distributed data $(Y_i, \mathbf{X}_i')$, $i = 1, \dots, n$, where Y_i is a real-valued random variable and $\mathbf{X}_i$ contains ordered and/or unordered discrete random variables.[2] These always uniquely determine J orthogonal groups. For example, two binary discrete random variables determine four orthogonal groups. Let the $J \times 1$ vector $\mathbf{D}_i$ indicate whether an observation i belongs to a group $j \in \{1, \dots, J\}$. In such a case, the jth entry of the vector $\mathbf{D}_i$ contains a one, $D_{ij} = 1$, and the remaining entries are equal to zero, $D_{ij'} = 0$ for all $j' \neq j$. Let $P(D_{ij} = 1) = p_j$. For the remainder we assume that the groups are asymptotically non-empty, i.e. $\inf_j p_j > \underline{p} > 0$.

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

$$Y_i = \mathbf{D}_i'\boldsymbol{\mu} + \varepsilon_i \tag{2.1}$$

with $\boldsymbol{\mu} = (\mu_1, \dots, \mu_J)'$, $E[\varepsilon_i|\mathbf{D}_i] = 0$ and $V[\varepsilon_i|\mathbf{D}_i] = \sigma^2(\mathbf{D}_i)$ allowing for heteroskedasticity. For regularity, we also assume finite $E[\varepsilon_i^4|\mathbf{D}_i]$ with probability one. Let $\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \dots, \hat{\mu}_J)'$ be a consistent first stage estimator for the group means. We propose to estimate the model for the conditional mean of Y_i as a penalized least squares problem:

$$\begin{aligned}(\hat{\mu}_1^{PCS}, \dots, \hat{\mu}_J^{PCS}) &= \underset{\mu_1,\dots,\mu_J}{\arg\min} \sum_{i=1}^{n} (Y_i - \mathbf{D}_i'\boldsymbol{\mu})^2 + Q(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) \\ Q(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) &= \sum_{k=1}^{J}\sum_{j=1}^{J} \lambda_{kj}(\mu_k - \hat{\mu}_j)^2,\end{aligned} \tag{2.2}$$

where $\boldsymbol{\Lambda} = (\lambda_{11}, \lambda_{12}, \dots, \lambda_{1J}, \dots, \lambda_{JJ})'$ are given smoothing or penalty parameters with $\lambda_{jj} = 0$ for all $j \in \{1, \dots, J\}$. PCS stands for pairwise cross-smoothing since geometrically the penalty term $Q(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ can be seen as a set of smooth real quadric iso-hypersurfaces that cross each other in a J-dimensional space.[3] 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 estimates. By allowing for reference and target dependent smoothing parameters λ_{kj}, the penalty provides maximal flexibility for smoothing.

Regarding the choice of the smoothing parameters, the more informative group j is for group k, the larger the smoothing parameter λ_{kj} should be and vice versa. In the special case of $\lambda_{kj} = 0$ for all pairs (k, j), none of the groups uses information from the other groups and the optimization is identical to the ordinary least squares problem. By choosing a large λ_{kj}, $\hat{\mu}_k^{PCS}$ is shrunk towards $\hat{\mu}_j$. Setting all λ_{kj}'s to large values pushes $\hat{\mu}_k^{PCS}$ towards the mean of all $\hat{\mu}_j$ where $j \neq k$. We discuss the issue of selecting risk-optimal smoothing parameters in Section 3.

[1] This is not restrictive relative to nonlinear aggregation as even convex linear aggregation can produce any value between the smallest and largest group mean. However, our method even allows for nonconvex linear aggregation strategies which could be beneficial for decreasing point estimation risk by allowing to put more weight to very similar groups compared to convex aggregation, see Section 3.

[2] For an extension to mixed data consider Online Appendix B.1.

[3] For example, if $J = 3$, then the quadric iso-hypersurfaces are elliptic or hyperbolic cylinders and for each $j \in \{1, 2, 3\}$ the quadric iso-hypersurface is centered around the point $[\hat{\mu}_j, \hat{\mu}_j, \hat{\mu}_j]$.

Let $n_k := \sum_{i=1}^{n} D_{ik}$ denote the number of observations within group k. Existence and uniqueness of the solution to (2.2) are guaranteed if $\sum_{l\neq k} \lambda_{kl} > -n_k$ for all $k \in \{1, \dots, J\}$.[4] Under these conditions, the kth group estimate is given by

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

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

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

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

which is a linear combination of cell means.

It is noteworthy that the smoothing parameters and therefore also the implicit weights $\lambda_{kj}/(n_k + \sum_{l\neq k} \lambda_{kl})$ are not all restricted to be larger or equal than zero. Just the overall smoothing for one reference category cannot be too negative. This fundamentally differentiates our approach from discretized support kernel approaches that are built as weighted averages using probability mass functions (Hall et al., 2004). They lead to weights which are restricted to be larger than zero. Shrinking simultaneously to different targets demands high flexibility from the smoothing parameters. Imposing strict positivity might not necessarily be optimal since smoothing away from distant groups can help to increase the smoothing to closer groups. The actual signs then depend on the absolute distances between group locations. For further discussion of the presence of negative smoothing parameters consider Section 3.

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

$$Q_{GRR}(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{k=1}^{J} \sum_{j\neq k} \lambda_k (\mu_k - \hat{\mu}_j)^2, \tag{2.5}$$

$$\hat{\mu}_k^{GRR}(\lambda_k) = \frac{n_k \bar{Y}_k}{n_k + (J-1)\lambda_k} + \lambda_k \sum_{j\neq k} \frac{\hat{\mu}_j}{n_k + (J-1)\lambda_k}. \tag{2.6}$$

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

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

$$Q_{Kernel}(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \sum_{k=1}^{J} \sum_{j\neq k} \lambda_j (\mu_k - \hat{\mu}_j)^2, \tag{2.7}$$

$$\hat{\mu}_k^{Kernel}(\boldsymbol{\Lambda}) = \frac{n_k \bar{Y}_k}{n_k + \sum_{l\neq k} \lambda_l} + \frac{\sum_{j\neq k} \lambda_j \hat{\mu}_j}{n_k + \sum_{l\neq k} \lambda_l}. \tag{2.8}$$

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

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

$$Q_{RR}(\boldsymbol{\Lambda}, \boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \lambda \sum_{k=1}^{J} \sum_{j\neq k} (\mu_k - \hat{\mu}_j)^2, \tag{2.9}$$

$$\hat{\mu}_k^{RR}(\lambda) = \frac{n_k \bar{Y}_k}{n_k + (J-1)\lambda} + \lambda \frac{\sum_{j\neq k} \hat{\mu}_j}{n_k + (J-1)\lambda}. \tag{2.10}$$

[4] For the derivations consider Appendix A.1.

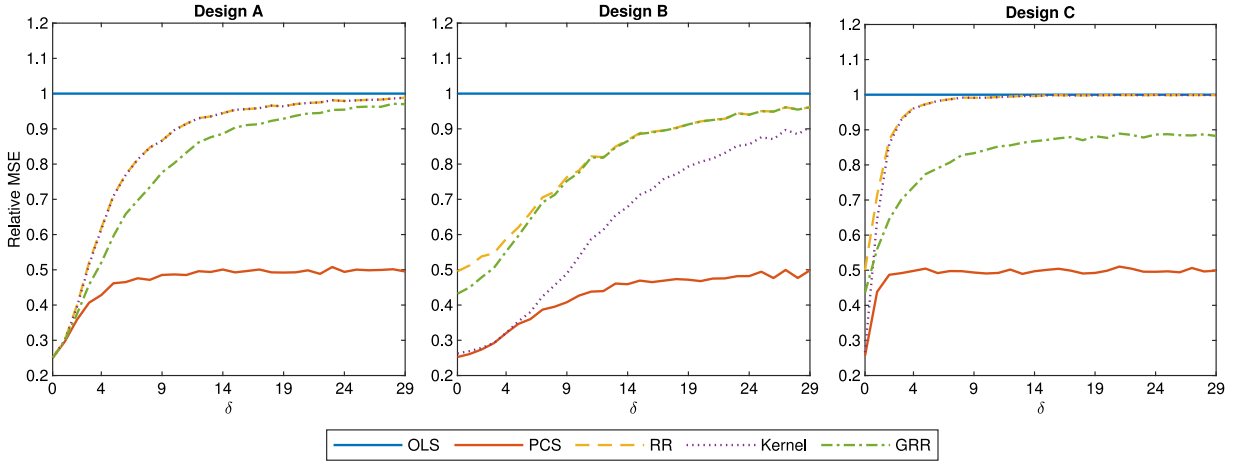

Fig. 2.1. Oracle estimators—Relative mean squared errors. The figure depicts the simulated mean squared error relative to the OLS estimator for the estimators with risk-optimal (oracle) smoothing parameters under normally distributed errors for $n = 400$, equal selection probabilities, and different parameter values δ. The parameter vectors are $\sqrt{n}\boldsymbol{\mu}_A = \sqrt{n}\boldsymbol{\mu}_B = (0, 0, 0, \delta)'$, $\sqrt{n}\boldsymbol{\mu}_C = (0, 3\delta, -2\delta, \delta)'$. The group variances are $\boldsymbol{\sigma}_A^2 = (1, 1, 1, 1)'$ and $\boldsymbol{\sigma}_B^2 = \boldsymbol{\sigma}_C^2 = (1, 1, 1, 10)'$. Simulations are based on 5000 replications.

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

Fig. 2.1 depicts the mean squared error of the different estimators relative to the OLS estimator using mean squared error optimal smoothing parameters[5] in the case of four groups and different levels of heterogeneity regarding both means and variances. One can see that all of the approaches compare favorably to the OLS estimator if the differences in group locations are not too big. The theoretically optimal PCS estimator not only dominates all other approaches but also qualitatively behaves closer to a correctly chosen restricted estimator that has superior risk properties even when locations are not close to identical. In addition, for the very heterogeneous design C, the differences compared to the optimal kernel and optimal (G)RR are most pronounced. Thus, the flexibility of the PCS penalty should in principle be able to generate substantial risk improvements compared to the alternative shrinkage methods.

3. Oracle risk, optimal weighting and plug-in estimation

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

$$\hat{\mu}_k = \frac{\frac{1}{n}\sum_{i=1}^n D_{ik} Y_i}{\frac{1}{n}\sum_{i=1}^n D_{ik} \vee \underline{p}} \tag{3.1}$$

for all k with $a \vee b = \max\{a, b\}$.[6] Note that rewriting (2.3) yields a weight-based representation of the PCS

$$\hat{\mu}_k^{PCS}(\boldsymbol{\Lambda}_k) \equiv \hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k) = \left(1 - \sum_{j \neq k} \omega_{kj}\right)\bar{Y}_k + \sum_{j \neq k} \omega_{kj}\hat{\mu}_j \tag{3.2}$$

with $\boldsymbol{\omega}_k = (\omega_{k1}, \ldots, \omega_{kJ})'$ and $\omega_{kj} = \lambda_{kj}/(n_k + \sum_{l \neq k} \lambda_{kl})$ being a one-to-one correspondence.[7] While the penalized regression representation is insightful for comparison with alternative regularization methods, using the weighted version will simplify further analysis.

The parameter mean squared error is given by the following proposition:

5 The PCS optimal smoothing parameters can be found in Section 3. For the alternative methods consider Appendix A.4.

6 The precise form of the denominator is chosen for technical simplification. Any other constant below p_k in place of $\underline{p}$ yields the same results. Other modified cell averages such as $\hat{\mu}_k = \sum_{i=1}^n D_{ik} Y_i / (\sum_{i=1}^n D_{ik} + \mathbb{1}(\sum_{i=1}^n D_{ik} = 0))$ can also assure existence.

7 See Online Appendix B.2.

Proposition 3.1. *Denote $E[Y_i|D_{ij}=1]=\mu_j$ and $V[\varepsilon_i|D_{ij}=1]=\sigma_j^2$ for all j, and let $\hat{\boldsymbol{\mu}}$ be chosen according to (3.1). Under the assumptions of model (2.1) the MSE of $\hat{\mu}_k(\boldsymbol{\omega}_k)$ for $k=\{1,\dots,J\}$ is given by*

$$\begin{aligned} MSE(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k)) &= \left[\left(\sum_{j\neq k}\omega_{kj}(\mu_k-\mu_j)\right)^2+\left(1-\sum_{j\neq k}\omega_{kj}\right)^2\frac{\sigma_k^2}{np_k}+\sum_{j\neq k}\omega_{kj}^2\frac{\sigma_j^2}{np_j}\right](1+o(1)) \\ &= \left[\boldsymbol{\omega}_k'\Delta_k\boldsymbol{\mu}\boldsymbol{\mu}'\Delta_k'\boldsymbol{\omega}_k+n^{-1}\boldsymbol{\omega}_k'diag(\boldsymbol{\gamma})^{-1}\boldsymbol{\omega}_k\right](1+o(1)) \end{aligned} \tag{3.3}$$

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

Proposition 3.1 collects the first order terms of both squared bias and variance with fixed weights. For the analysis of the estimation risk using estimated weights, we introduce an asymptotic parameterization later in Section 4. Note that the PCS estimator only depends on the smoothing parameters within its own reference category but is independent of the remaining smoothing parameters. Therefore, optimization of the parameter MSEs can be done group by group contrary to kernel smoothing and ridge regression. The following theorem establishes the MSE optimal smoothing parameters:

Theorem 3.1. *For given $k=\{1,\dots,J\}$, the leading criterion in Proposition 3.1 is minimized at $\boldsymbol{\omega}_k^*=(\omega_{k1}^*,\dots,\omega_{kJ}^*)$, where*

$$\begin{aligned} \omega_{kj}^* &= \frac{\frac{p_j}{\sigma_j^2}+n\sum_{m\neq k}(\mu_k-\mu_m)(\mu_j-\mu_m)\frac{p_mp_j}{\sigma_m^2\sigma_j^2}}{\sum_{l=1}^J\frac{p_l}{\sigma_l^2}+n\sum_{l=1}^J\sum_{m\neq k}(\mu_k-\mu_m)(\mu_l-\mu_m)\frac{p_lp_m}{\sigma_l^2\sigma_m^2}} \\ &= \frac{\gamma_j(1+n\boldsymbol{\mu}'\Delta_k'diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\mu})}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J+\frac{n}{2}\boldsymbol{\mu}'\Delta'M_1\Delta\boldsymbol{\mu}} \end{aligned} \tag{3.4}$$

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

The solution is always unique.[8] The minimizers corresponding to the λ_{kj}'s can be found in the Online Appendix B. Note that if the first stage estimates are all identical for a reference group, the corresponding optimal smoothing parameters become strictly positive. This is in line with Hoerl and Kennard (1970) who show for generalized ridge regression that MSE optimal smoothing parameters have to be positive in the case of a common target. For the general case, negative smoothing parameters can be optimal due to different group specific targets.

The properties of the PCS estimator using oracle weights can be found in Section 4. One can construct the oracle weights for the restricted PCS, i.e. (G)RR and kernel regression in a similar fashion, however (weighted) MSE optimal ridge regression weights and kernel weights in general do not have a closed-form solution for larger J, see Appendix A.4.

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

$$\begin{aligned} \hat{\omega}_{kj} &= \frac{\frac{\hat{p}_j}{\hat{\sigma}_j^2}+n\sum_{m\neq k}(\hat{\mu}_k-\hat{\mu}_m)(\hat{\mu}_j-\hat{\mu}_m)\frac{\hat{p}_m\hat{p}_j}{\hat{\sigma}_m^2\hat{\sigma}_j^2}}{\sum_{l=1}^J\frac{\hat{p}_l}{\hat{\sigma}_l^2}+n\sum_{l=1}^J\sum_{m\neq k}(\hat{\mu}_k-\hat{\mu}_m)(\hat{\mu}_l-\hat{\mu}_m)\frac{\hat{p}_l\hat{p}_m}{\hat{\sigma}_l^2\hat{\sigma}_m^2}} \\ &= \frac{\hat{\gamma}_j(1+n\hat{\boldsymbol{\mu}}'\Delta_k'diag(\hat{\boldsymbol{\gamma}})\Delta_j\hat{\boldsymbol{\mu}})}{\hat{\boldsymbol{\gamma}}'\boldsymbol{\iota}_J+\frac{n}{2}\hat{\boldsymbol{\mu}}'\Delta'\hat{M}_1\Delta\hat{\boldsymbol{\mu}}} \end{aligned} \tag{3.5}$$

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

$$\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}})=\sum_{j=1}^J\hat{\omega}_{kj}\hat{\mu}_j. \tag{3.6}$$

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. (2019). In Section 4.3 we show that while oracle performance cannot be obtained for arbitrary data generating processes, the plug-in estimator still uniformly dominates the ordinary least squares estimator in terms of the (weighted) mean squared error.

[8] The optimal MSE smoothing parameters satisfy the existence and uniqueness condition for $\hat{\boldsymbol{\mu}}^{PCS}$. For more details consider Online Appendix B.3 and B.4.

In Section 2 we showed that PCS can be seen as a generalization to kernel and (generalized) ridge regression in the context of orthogonal regressors. The same holds true when comparing PCS estimators to the class of linear model averaging estimators. In particular one can show that, for given model averaging weights, any linear model averaging estimator that combines estimators for the mean vector based on (nested or non-nested) models that impose linear aggregation restrictions on the groups can be written as a constrained PCS estimator.[9] However note that in contrast to the model averaging literature (Hansen and Racine, 2012; Liu, 2015) the PCS weights are not restricted to lie in the unit simplex. Under fixed weights, the model averaging and the PCS estimator are linear in the outcome. For admissibility of linear estimators of the mean of a multivariate normal distribution, Cohen (1966) shows that symmetry and nonnegative eigenvalue bounds have to be met by the linear operator that maps outcomes to predictions, see also Li (1987) in a regression context. Hansen and Racine (2012) show that in the case of nested linear regression models, positivity of the model weights is a necessary condition for admissibility under mean squared error loss. However, if data dependent weights are used, the resulting estimator is no longer linear in the outcome. Furthermore, the less restrictive weighting scheme of the PCS can contradict the nesting requirement by Hansen and Racine (2012) despite the fact that the submodels are effectively linear. As a consequence, the overall shrinkage sum and not each shrinkage parameter is bounded from below and thus inadmissibility of the PCS estimator does not follow. Interestingly, the eigenvalue conditions of Cohen (1966) for admissibility still hold with probability one for both optimal and feasible PCS estimators.[10]

4. Large sample theory

4.1. Local parameterization

In the following, we outline and discuss the large sample properties of the different weighting schemes and of the PCS estimator over a sufficient class of data generating processes relevant in the context of heterogeneous group means. In particular, we would like to distinguish between systems of locations in which the differences between group means are small (*close* systems) and large (*distant* systems) for a given sample size. For simplicity, instead of invoking standard assumptions to assure consistency and asymptotic normality of the first stage through moment conditions or similar, we start from a less rigorous point.[11] Let $\mathcal{F}$ be the set of distribution functions and denote $\boldsymbol{\delta} = (\delta_1, \delta_2, \dots, \delta_J)'$, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, V_0)$, $V_0 = diag(\sigma_j^2/p_j)$, and $\hat{V} = diag(n\hat{\sigma}_j^2/n_j)$.

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

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

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

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

Close systems require that all scaled pairwise differences do not diverge, i.e. their differences depend on the local parameters $(\delta_k - \delta_j)$ for all k, j.[12] This nests the case in which all means are exactly identical and the local parameters are zero. For distant systems we require the scaled differences to go to infinity for at least one pair in the system. The union of these systems is sufficiently rich to describe a wide range of data generating processes.

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

$$T_n = \sqrt{n}\frac{\hat{\mu}_k - \hat{\mu}_j}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} = \sqrt{n}\frac{\mu_k - \mu_j}{\sqrt{\frac{\sigma_k^2}{p_k} + \frac{\sigma_j^2}{p_j}}} + Z_n \tag{4.1}$$

9 See Online Appendix B.6.

10 See Online Appendix B.8.

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

12 Note that since the system effectively depends only on differences in local parameters, constant shifts to $\boldsymbol{\delta}$ do not affect the analysis. In principle, the mean vector $\boldsymbol{\mu}$ could have subscript n as it is allowed to depend on the sample size. We omit the subscript for readability.

Using the local parameterization it follows that

$$T_n(F_n) \xrightarrow{d} \mathcal{N}\left((\delta_k - \delta_j) \Big/ \sqrt{\sigma_k^2/p_k + \sigma_j^2/p_j},\, 1\right) \text{ if } \{F_n\} \in S(\boldsymbol{\delta}, V_0), \tag{4.2}$$

$$P(|T_n(F_n)| > c) \to 1 \text{ for all } c > 0 \qquad \text{if } \{F_n\} \in S(\infty, V_0). \tag{4.3}$$

Therefore, depending on the local parameter difference $\delta_k - \delta_j$, one can obtain a small, moderate or even large mean for the distribution of the test statistic. In the special case of $\delta_k - \delta_j$ being exactly equal to zero, the local parameterization does no longer affect the asymptotic distribution and classical inference can be conducted using the standard normal distribution. It is apparent that in any other case, choosing a model based on such a test might be misleading. If the local parameter is e.g. at a size that centers the limiting distribution around the critical value used for rejection of the null hypothesis, rejection would occur with probability one half. If this pretest is used for model selection, it is likely to suggest an underparameterized model that translates into higher parameter risk. The PCS estimator can be considered as a smooth variant of such a classical pretesting based estimator. Hence, we expect it to perform better exactly in these regions in which type-II errors are relatively large. Standard asymptotic analysis, however, will always favor the more parameterized model except if parameters are exactly equal. Thus, the approximations based on the local asymptotic framework should be closer to the actual finite sample behavior. The intuition can directly be translated to simultaneous tests of equality in locations for more than two groups.[13] The parameterization of the locally close systems is similar to the ones used for modeling locally misspecified parametric models in the literature on frequentist model averaging, see e.g. Hjort and Claeskens (2003), Hansen (2014), Liu (2015), and Cheng et al. (2019). In our context, local misspecification would correspond to a model that aggregates not exactly identical groups while a fully saturated model yields unbiased but potentially high variance estimates. The $n^{-1/2}$-rate allows for the (squared) misspecification biases and the ordinary variances to both affect first-order asymptotic approximations. Thus, they can act as "exchangable currencies" (Hjort and Claeskens, 2003) for the point estimation risk of the shrinkage estimators. See also Raftery and Zheng (2003) for a critical discussion of local parameterization from a Bayesian perspective.

4.2. Distributional theory

For investigation of the large sample properties of the PCS estimator, the behavior of the smoothing parameters along the sequences of DGPs is crucial. We consider PCS with weights ω^f_{kj} that correspond to fixed smoothing parameters λ_{kj} in (2.2), MSE optimal weights ω^*_{kj} and plug-in weights $\hat{\omega}_{kj}$. The following lemma demonstrates the behavior of the different weighting schemes in large samples.

Lemma 4.1. *Let ω^f_{kj}, ω^*_{kj} and $\hat{\omega}_{kj}$ denote the PCS weights in (3.6) corresponding to fixed values,[14] MSE optimal weights according to (3.4) and plug-in weights according to (3.5). Their limiting behavior along the local parameterization is then given by*

$$\begin{aligned}
&\omega^f_{kj} = O_p(n^{-1}) \text{ if } k \neq j \text{ and } \omega^f_{kk} = 1 + O_p(n^{-1}) && \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0),\\
&\omega^*_{kj} \to \bar{w}_{kj} = \frac{\gamma_j(1 + \boldsymbol{\delta}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\delta})}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}\boldsymbol{\delta}'\Delta' M_1 \Delta\boldsymbol{\delta}} && \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0),\\
&\omega^*_{kj} \to \bar{w}_{kj} = 2\gamma_j \frac{\boldsymbol{\mu}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\mu}}{\boldsymbol{\mu}'\Delta' M_1 \Delta\boldsymbol{\mu}} && \text{if } \{F_n\} \in S(\infty, V_0),\\
&\hat{\omega}_{kj} \xrightarrow{d} w^a_{kj} = \frac{\gamma_j(1 + (\mathbf{Z} + \boldsymbol{\delta})'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j(\mathbf{Z} + \boldsymbol{\delta}))}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}(\mathbf{Z} + \boldsymbol{\delta})'\Delta' M_1 \Delta(\mathbf{Z} + \boldsymbol{\delta})} && \text{if } \{F_n\} \in S(\boldsymbol{\delta}, V_0),\\
&\hat{\omega}_{kj} \xrightarrow{p} \bar{w}_{kj} = 2\gamma_j \frac{\boldsymbol{\mu}'\Delta_k' diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\mu}}{\boldsymbol{\mu}'\Delta' M_1 \Delta\boldsymbol{\mu}} && \text{if } \{F_n\} \in S(\infty, V_0).
\end{aligned}$$

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

Note that the MSE optimal smoothing parameters do not in general vanish asymptotically, i.e. there is potential aggregation even in the limit. This, however, does not exclude the possibility to completely smooth out uninformative groups in large samples. It is qualitatively different from the smoothing kernel approach where uninformative, i.e. conditionally independent, regressors are always smoothed to a global average with smoothing parameters converging to their upper bound (Hall et al., 2004, 2007). The estimated smoothing parameters converge in distribution to a function of a normal random vector if groups are locally close. Under a distant system, they converge in probability to the oracle parameters. Thus, adding a single distant parameter to a locally close system is sufficient to obtain convergence in probability. This is

[13] Online Appendix B.7 contains an example using a Wald test for equality of all means.

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

due to the fact that the effective shrinkage targets in the PCS estimator are weighted leave the kth group out averages. If the kth group is the distant one, a weighted combination of locally close locations will be distant enough to pin down the optimal weights in probability. If the kth group is within the set of the locally close groups, the leave the kth group out average will contain the distant group which is sufficient in large samples to distinguish the shrinkage target from the reference mean and thus lead to probabilistic convergence. Only in the case of all groups being locally close, the differences are not sufficient such that the limiting behavior is governed by a continuous function of a random normal vector. However, due to the rate of convergence of the estimated smoothing parameters in the case of distant systems, they will have an effect on the first-order term determining the limiting distribution of the PCS under estimated smoothing parameters compared to the oracle distribution.

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

Theorem 4.1. *Let $\boldsymbol{\omega}_k^f$, $\boldsymbol{\omega}_k^*$ and $\hat{\boldsymbol{\omega}}_k$ denote the vector of fixed values, MSE optimal weights according to (3.4) and plug-in weights according to (3.5) for the PCS. The asymptotic distributions of the PCS estimators are given by*

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

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

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

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

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

Theorem 4.1 contains the asymptotic distributions of the different PCS estimators. Some of the results clearly parallel the literature on (generalized) ridge regression. In particular, a fixed penalty is asymptotically negligible for the distribution and thus the efficiency of the estimator, i.e. PCS with fixed weights converges in distribution to the corresponding OLS limit. The PCS estimator under optimal weights differs in terms of its distribution from the OLS estimator and thus the improvements in MSE in general do not disappear even for large samples. The behavior of the PCS under estimated smoothing parameters is particularly noteworthy. Note that the distribution to the normal is not uniform along all sequences of DGPs. In particular, the PCS with estimated smoothing parameters under locally close systems converges in distribution to a sum of normal random variables and local parameters multiplied by the limiting weights that are themselves functions of the same random normal variables, local parameters, and other features of the DGP. Thus, the limiting distribution in close systems is in general different from the normal. Assessing or estimating that limiting distribution has to be done with caution as the local parameters cannot be estimated consistently due to the $\sqrt{n}$ multiplier. This is similar to other shrinkage and model averaging methods that rely on smooth aggregation methods in the spirit of James–Stein shrinkage and frequentist model averaging (Hjort and Claeskens, 2003; Liu, 2015; Cheng et al., 2019; Hansen, 2016a).[15]

4.3. Asymptotic risk

Theorem 4.1 shows that when evaluating the risk of the feasible PCS estimator, one has to take the additional variation of the weights under locally close systems into account. In the following, we will focus on the risk under close systems as the risk for distant systems can be obtained as a special case by letting $\|\Delta\boldsymbol{\delta}\|_\infty \to \infty$. For the derivation of the oracle risk, recall that due to the flexibility of the PCS, optimization can be done separately for each individual group. When evaluating the risk of the feasible PCS parameters, however, the additional (co)variation introduced by the weighting parameters has to be taken into account since the latter are functions of the same random vector. The choice for the joint loss function will be the (weighted) parameter vector MSE

$$l(\tilde{\boldsymbol{\mu}}, \boldsymbol{\mu}) = (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu})' W (\tilde{\boldsymbol{\mu}} - \boldsymbol{\mu}) \tag{4.4}$$

with the canonical weighting matrix being the inverse of the asymptotic variance of the OLS parameter vector $W = diag(\boldsymbol{\gamma})$. Thus W is proportional to the identity matrix under homoskedasticity and equal group probabilities. The choice

[15] Hjort and Claeskens (2003) propose a bias correction method for constructing confidence intervals that leads to asymptotically correct coverage rates. Liu (2015) uses a similar approach for least squares model averaging estimators. These bias corrections implicitly use unbiased but inconsistent estimates of the local parameter vector. The corresponding intervals tend to either coincide with the ones obtained from the unrestricted model or to undercover the true parameter in finite samples, see e.g. the simulation results in Section 6.3 by Liu (2015). For PCS they are identical to the conventional confidence bounds of the fully saturated model.

of W also renders the evaluation of the risk invariant to rotations of the parameter vector, such that PCS risk properties are preserved even if outcomes are not generated on the same scale across groups.

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

$$\rho(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}) = \lim_{\zeta \to \infty} \liminf_{n \to \infty} E_{F_n}[\min\{nl(\hat{\boldsymbol{\mu}}_n, \boldsymbol{\mu}), \zeta\}]. \tag{4.5}$$

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

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

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

Theorem 4.2. *Let $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. If $\hat{\boldsymbol{\omega}}$ is chosen according to (3.5), then*

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

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

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

Theorem 4.2 allows us to establish sufficient conditions for a *strict uniform dominance* of the PCS estimator compared to the OLS estimator. Uniform in the sense that for all bounded $\boldsymbol{\delta}$ vectors, the risk is strictly smaller than the risk of the OLS estimator. It turns out that an easily interpretable sufficient condition is that the number of groups has to exceed three, i.e. we obtain the following corollary:

Corollary 4.1. *Let $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$ and $\hat{\boldsymbol{\omega}}$ be chosen according to (3.5). If $J \geq 4$, then*

$$\sup_{\boldsymbol{\delta} \in B} \quad \rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) < 0 \tag{4.6}$$

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

Thus, a sufficient (but by no means necessary) condition for uniform dominance is a simple condition on the dimensionality of the mean vector. This is similar to classical James–Stein estimation that as a necessary condition requires at least a three-dimensional multivariate mean vector when shrinking to a fixed target for global risk reduction over the maximum-likelihood estimator (Stein, 1956). Here, the somewhat more flexible shrinkage target requires one additional dimension, i.e. at least four groups to assure a strictly smaller risk for any close system of locations. Consistency follows directly as a corollary. In a similar spirit, the PCS shrinks towards a restricted subspace, i.e. an estimator that equalizes the group locations under a generalized error term structure. The corresponding subspace has exactly dimensionality $l = 1$ thus the minimal condition for superior risk (Oman, 1982) is that $J \geq 3 + l$ which equals the sufficient condition from Corollary 4.1.

4.4. Asymptotic risk comparison to model averaging

The explicit solution in Theorem 4.2 allows for a comparison in terms of point estimation risk to other model averaging methods such as Mallows Model Averaging (Hansen, 2007, 2014), Jackknife Model Averaging (Hansen and Racine, 2012) or other shrinkage methods for parametric models (James and Stein, 1961; Oman, 1982; Hansen, 2016a). These aforementioned methods try to achieve low point estimation risk by shrinking models towards a restricted subspace based on distance measures such as the (weighted) mean squared error criterion. Hansen (2014, 2016a) provide the asymptotic risk for popular linear model averaging and efficient parametric model averaging estimators in a local parameter framework. It is important to note that, in general, the statistical risk of these methods does not only depend on the local parameter values and the number of groups but also on the explicit restrictions and/or ordering of the (sub)models chosen for model averaging. Thus one cannot establish a general relationship in terms of the asymptotic risk between PCS and arbitrary model averaging methods. However in the following we provide some results on the relationship between the

risk of PCS and model averaging estimators that shrink parameter estimates towards a global mean for interesting special cases. In particular, we consider the canonical case of four groups that arises naturally when thinking about interactions of dummy variables or when considering quasi-experimental designs like differences-in-differences. We also consider the case when the number of groups exceeds seven.

We compare the asymptotic risk of the PCS estimator that uses all groups at the same time to the efficient averaging estimator by Hansen (2016a) that relies on shrinkage of the mean parameter towards a common target. Under homoskedasticity, this estimator is asymptotically equivalent to Mallows Model Averaging (Hansen, 2007) with a fully parameterized and a constant model. Our results suggest that for lower dimensional problems PCS dominates model averaging when the overall local parameter differences are not too small while for higher dimensional cases the ordering reverses. For the derivations please consider Appendix A.8.

Let R be a $(J-1)\times J$ dimensional matrix that maps the mean parameter into a restricted subspace that yields an equal mean for restriction $R\boldsymbol{\mu}=0$. The restricted estimator is obtained by the projection onto that restricted subspace, i.e.

$$\begin{aligned}\hat{\boldsymbol{\mu}}^R &= (I_J - \hat{V}R'(R\hat{V}R')^{-1}R)\hat{\boldsymbol{\mu}} \\ &= P_R\hat{\boldsymbol{\mu}}.\end{aligned} \tag{4.7}$$

The model averaging estimator by Hansen (2016a) is then given by

$$\hat{\boldsymbol{\mu}}^{MA} = \hat{w}^{MA}\hat{\boldsymbol{\mu}} + (1-\hat{w}^{MA})P_R\hat{\boldsymbol{\mu}} \tag{4.8}$$

with optimal estimated weight

$$\hat{w}^{MA} = \left(1 - \frac{J-3}{n(\hat{\boldsymbol{\mu}}-P_R\hat{\boldsymbol{\mu}})'\hat{V}^{-1}(\hat{\boldsymbol{\mu}}-P_R\hat{\boldsymbol{\mu}})}\right)_+ \tag{4.9}$$

where $(x)_+ = x\mathbb{1}(x\geq 0)$ denotes the positive part function. The estimator shrinks the parameters monotonically towards the restricted estimator depending on the (weighted) squared differences between restricted and unrestricted model parameters.

Let $\tilde{\boldsymbol{\delta}} = V_0^{-1/2}\boldsymbol{\delta}$ denote the (square root) variance weighted local parameter vector and let $M = I_J - V_0^{-1/2}\iota_J\iota_J' V_0^{-1/2}/tr(V_0^{-1})$. The matrix M is crucial in evaluating the asymptotic risk properties of the model averaging estimators as the latter depends on $\tilde{\boldsymbol{\delta}}'M\tilde{\boldsymbol{\delta}}$. M defines the effective subspace to which the averages are shrunk to. It is an idempotent, positive semidefinite matrix of rank $J-1$ with eigenvalues independent of the entries in V_0^{-1}. It is best understood in the case of homoskedasticity and equal groups selection probabilities, i.e. if $V_0 = J\sigma^2 I_J$. M then simplifies to $M = I_J - \iota_J\iota_J'/J$. The asymptotic risk is then driven by the *average* difference of inverse variance weighted squared local parameter values $\tilde{\boldsymbol{\delta}}'M\tilde{\boldsymbol{\delta}}$.

Consider the canonical four group design for the general case of heteroskedasticity and nonequal selection probabilities. Comparing the asymptotic risk of the PCS to model averaging, we obtain the following (conservative) inequality

$$\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{MA}, \boldsymbol{\mu}) < 0 + O((\tilde{\boldsymbol{\delta}}'M\tilde{\boldsymbol{\delta}})^{-2}) \tag{4.10}$$

pointwise along the local parameter differences. Thus, if the weighted differences in local parameters are sufficiently large, the PCS estimator will dominate the model averaging estimator. Put differently, there is always a large enough local parameter difference such that PCS dominates model averaging for all differences exceeding this threshold. This is also confirmed by our finite sample simulations in Section 5. Similarly, one can establish that the proposed model averaging estimator dominates the PCS estimator if the number of groups exceeds seven. These inequalities are driven by the fact that the risk gains of PCS over OLS estimation scale linearly with the number of groups while for the simple model averaging method the scale is quadratic.

At this point it is important to emphasize again that generally, if there is a prior knowledge, risk can be improved for PCS estimation by only aggregating groups which are likely to be close to each other but also for model averaging by choosing the right (sub)models for averaging ex ante, a point also made by Hansen (2014). Thus, for any data generating process, the two agnostic methods compared in this chapter can both perform better if proper (data-independent) selection has been made beforehand. The exact risk order then depends on the explicit aggregation strategy. Moreover also note that the asymptotic risk approximations are not necessarily exact in finite samples, e.g. if the error distributions are very asymmetric. We consider the case of deviations from normality in finite samples in Section 5.

5. Monte Carlo study

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

1. Ordinary least squares estimator/frequency method (OLS),
2. pairwise cross-smoothing with plug-in smoothing parameters (PCS),

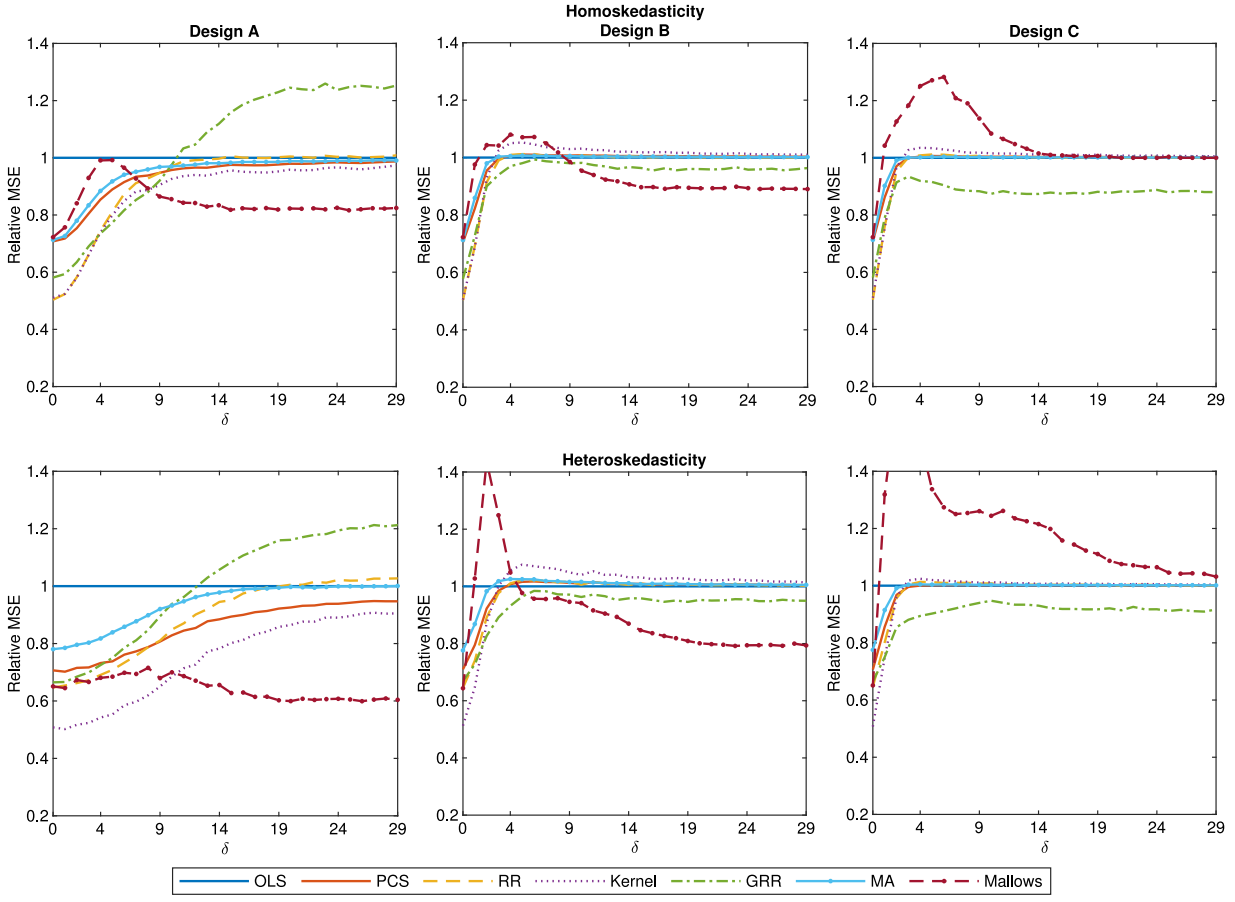

Fig. 5.1. Relative weighted parameter vector mean squared errors. The figure depicts the simulated relative weighted parameter vector mean squared errors for the estimators with plug-in risk-optimal smoothing parameters under log-normally distributed errors for $n = 400$, equal selection probabilities, different parameter values δ, and different structures of the error variance with OLS as a benchmark. The parameter vectors are $\sqrt{n}\boldsymbol{\mu}_A = (0, 0, 0, \delta)'$, $\sqrt{n}\boldsymbol{\mu}_B = (0, 0, -3\delta, \delta)'$ and $\sqrt{n}\boldsymbol{\mu}_C = (0, 2\delta, -3\delta, \delta)'$. The group variances are $\boldsymbol{\sigma}^2_{hom} = (1, 1, 1, 1)'$ and $\boldsymbol{\sigma}^2_{het} = (1, 1, 1, 10)'$. Simulations are based on 5000 replications.

3. ridge regression estimator with a plug-in smoothing parameter (RR),
4. generalized ridge regression estimator with plug-in smoothing parameters (GRR),
5. nonparametric smoothing kernel with plug-in smoothing parameters (Kernel),
6. a selection/pretesting estimator based on Mallows C_p(Mallows),[16]
7. a model averaging estimator combining a full model and a constant model (MA).[17]

We study a setup with a moderate number of groups, i.e. $J = 4$, that are selected with equal likelihood. The three designs A, B and C are set such that mean vectors converge to the origin but vary in the degree of deviations from the origin in finite samples. The mean vectors take following values $\boldsymbol{\mu}_A = (0, 0, 0, \delta/\sqrt{n})'$, $\boldsymbol{\mu}_B = (0, 0, -3\delta/\sqrt{n}, \delta/\sqrt{n})'$ and $\boldsymbol{\mu}_C = (0, 2\delta/\sqrt{n}, -3\delta/\sqrt{n}, \delta/\sqrt{n})'$ with δ varying over a positive grid starting at 0. For $\delta = 0$, we are in the case of identical means, i.e. a global average would be the most efficient estimator. Regarding the error term, we consider homoskedastic and heteroskedastic standardized log-normal distributions. All results are based on 5000 simulations. For PCS we use the plug-in weights from (3.5). The corresponding weights for RR, GRR, and Kernel can be found in Appendix A.4. We use group based sample variances as robust estimators of the residual variances required for the plug-in weights, i.e. we do not assume that homoskedasticity is known ex ante. We also considered selecting smoothing parameters and weights via cross-validation. Results are collected in Online Appendix B.5.

Weighted MSEs relative to OLS for $n = 400$ are reported in Fig. 5.1. Results for other sample sizes follow similar patterns and are therefore omitted. The comparison of the finite sample losses together with the overall performance

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

17 See Section 4.4.

over a wide range of δ values indicates the robustness of PCS over the other methods. Depending on the value of δ one can get up to 30% improvement in the parameter vector MSE by using PCS over OLS. The largest benefits are obtained at lower values of δ as in these settings the shrinkage estimator can benefit from taking information from the other similar groups.

GRR and C_p estimator often perform substantially worse than OLS in the chosen setups. PCS, RR, Kernel and MA estimators seem to mostly dominate OLS. PCS, RR, Kernel, and MA can perform worse than OLS in finite samples for design B and design C in the range of moderate values of δ while RR also exceeds the benchmark risk for larger local parameter values in design A. However, these losses are more pronounced for the Kernel estimator reaching an increase of about 8% in relative MSE compared to PCS. Moreover note that PCS dominates MA for all designs. PCS also outperforms Kernel in designs B and C and RR if δ values that are sufficiently large, i.e. there is a unique crossing point of the relative mean squared errors.

The RR generally performs well especially for small δ yielding MSE improvements up to 50% compared to OLS. In these designs, RR seems to benefit from the smaller amount of smoothing parameters. In design A, however, RR is dominated by PCS for moderate to large δ values and can sometimes even perform slightly worse than OLS.

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

The Kernel estimator performs the best in the design A dominating the OLS and all the other shrinkage methods. If the true DGP has many equal means, the Kernel estimator seems to profit from the close distances towards the shrinkage target that imposes useful soft aggregation. In the more distant designs B and C, however, Kernel performs the worst among all smooth aggregation methods for moderate to large values of δ. In particular, it exceeds the OLS benchmark by up to 8% for sufficiently large local parameter values.

The pretest estimator based on Mallows C_p has its worst performance for moderate values of δ as in this range it is challenging for the model selection criterion to detect the optimal aggregation strategy. It often yields an underfitted model that introduces too much bias into the parameter estimates in line with the discussion in Section 4.1. This is particularly prominent in designs B and C with risk inflations of over 60% compared to OLS. With a higher degree of deviations from the origin (design C), the C_p estimator performs worse than OLS over a larger range of δ parameters. However, for extreme values of δ parameters, it can perform better or close to OLS if there are risk gains from aggregating identical groups (design A and B). In practice, these mean differences are usually unknown and thus using the model selection criterion can be detrimental to the estimation. The presence of data generating processes for which model selection criteria yield inferior risk is a well-known phenomenon in the literature on model selection and post-selection risk, see e.g. Leeb and Pötscher (2008) among many others.

The MA estimator is dominated by PCS for all values of δ and converges to the risk of the OLS for larger values of δ. Given the choice of the restricted model, this is in line with the theory outlined in Section 4.4. For moderate values of δ in designs B and C we see similar but slightly higher finite sample losses over OLS as in the case of RR and PCS.

PCS is virtually never worse than OLS, i.e. it shows uniformly dominant behavior in line with our results in Section 4.3. However, PCS is not always beating all the competitors over all the designs and δ values. For example, for small δ values Kernel, GRR, and RR are up to 20 percentage points superior profiting from an accurate shrinkage target in the design of almost equal means and less smoothing parameters to estimate. For large values of δ, the pretest estimator can perform better than PCS as the risk optimal model can be obtained through strict aggregation. In comparison to the Kernel and MA, PCS shows more robust risk improvements for moderate δ systems.

Therefore, PCS seems to be a robust refinement over OLS for a wide range of DGPs as alternatives are either dominated by PCS (OLS and MA), show worse finite sample properties than OLS in moderate to distant systems (Kernel) or are generally design sensitive (GRR and C_p). Except for Design A under heteroskedasticity and large local parameter values, RR seems also mostly robust and often yields similar risk improvements compared to PCS. Note that in general there is still room for further improvement since the large risk gains that can be obtained by the theoretically optimal PCS (see Section 2, Fig. 2.1) cannot be reached by any method considered in the simulations designs.

6. Applications

6.1. Application I: A fine is a price

Gneezy and Rustichini (2000a) investigate the prediction of the deterrence hypothesis, i.e. that ceteris paribus introducing fines will decrease the likelihood of the associated action or behavior. They run a controlled field experiment at ten day-care centers for young children in Haifa, Israel over a period of twenty weeks. This leads to a small panel with ten observations and twenty time periods. In period five, a fine is introduced for parents that arrive 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

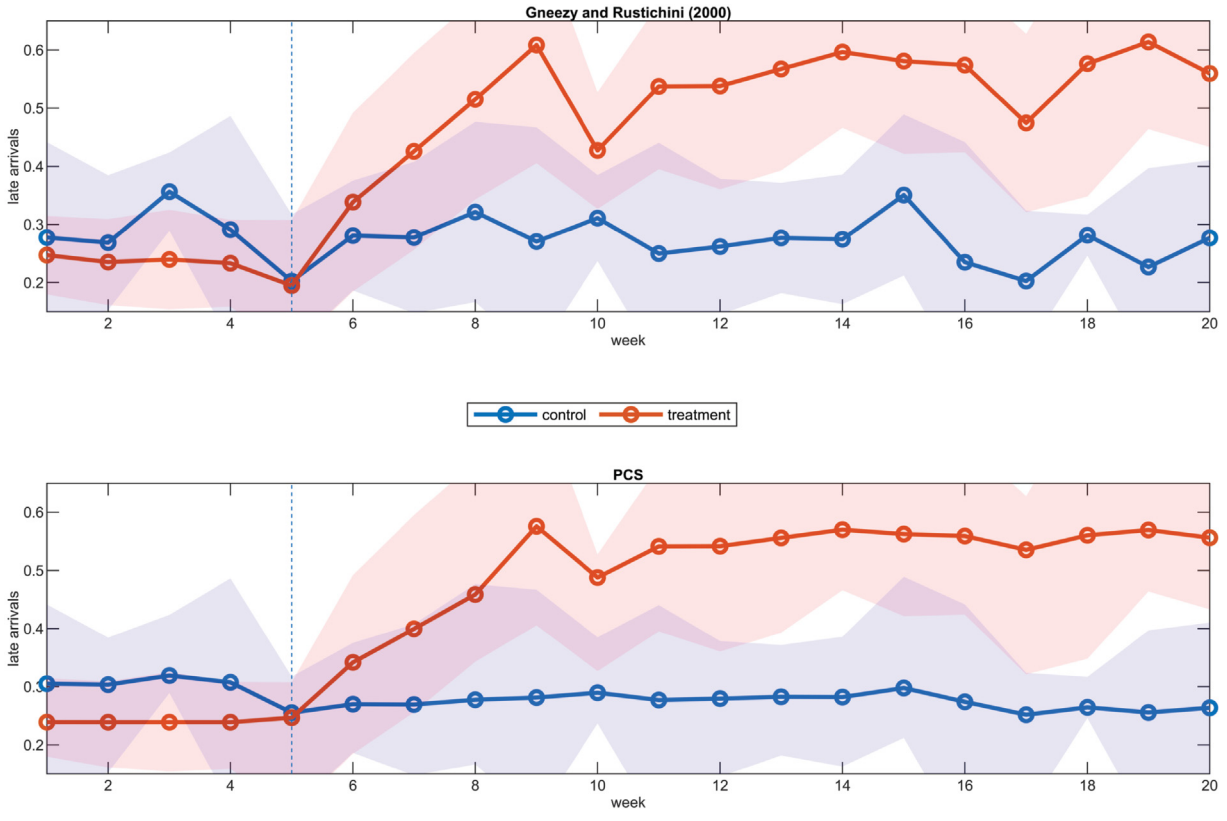

Fig. 6.1. Mean share of late arrivals, OLS and PCS estimates.

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 Fig. 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 simple panel data model.[18] In statistical terms, it contains estimates for the expected share of late arrivals conditional on time period and treatment status. Our method is well-suited for this application since by construction, there are small orthogonal groups that are determined by time and treatment status. We stabilize the estimates of the conditional means by using the plug-in PCS within treatment groups and time periods closely related to Gneezy and Rustichini (2000a), Table 2. Hence we smooth the averages within weeks 1–4, 5–8, 9–16 and 17–20 for both groups using the original means as first stage.[19]

Fig. 6.1 depicts OLS (Gneezy and Rustichini, 2000a) and PCS estimates for the conditional mean over time and treatment status together with 95% confidence intervals. Estimates for the alternative shrinkage methods can be found in Appendix A.9 (see Fig. A.1). 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.

6.2. Application II: Minimum wage study

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

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

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

Table 6.1
Mean and difference-in-differences estimates.

All chains		OLS			PCS	
	NJ (t)		PEN (c)	NJ (t)		PEN (c)
B	20.44	23.33		20.53	22.87	
A	21.03	21.17		21.01	21.12	
DiD		2.75*			2.22*	
Burger king		OLS			PCS	
	NJ (t)		PEN (c)	NJ (t)		PEN (c)
B	22.16	29.42		22.25	29.06	
A	23.63	26.22		23.63	26.06	
DiD		4.67**			4.38**	
KFC		OLS			PCS	
	NJ (t)		PEN (c)	NJ (t)		PEN (c)
B	12.79	10.71		12.76	10.92	
A	13.73	13.00		13.60	12.96	
DiD		−1.35			−1.20	
Roys		OLS			PCS	
	NJ (t)		PEN (c)	NJ (t)		PEN (c)
B	23.14	19.74		22.99	19.80	
A	21.73	15.81		21.68	16.12	
DiD		2.52			2.37	
Wendys		OLS			PCS	
	NJ (t)		PEN (c)	NJ (t)		PEN (c)
B	22.08	24.12		22.43	23.46	
A	23.40	22.10		23.10	22.44	
DiD		3.35			1.69	

The table contains the mean estimates of the full-time employment equivalents. B = before the minimum wage increase, A = after the minimum wage increase, NJ = New Jersey, PEN = Pennsylvania, t = treated, c = control. DiD formula: $(\hat{\mu}_{NJ,A} - \hat{\mu}_{NJ,B}) - (\hat{\mu}_{PEN,A} - \hat{\mu}_{PEN,B})$. All estimates become insignificant if corrected for multiple comparisons using the Holm–Bonferroni method.

*Marginal p-value < 0.1 is computed using conventional confidence bound.

**Marginal p-value < 0.05 is computed using conventional confidence bound.

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

The OLS (Card and Krueger, 1994) and PCS results for pooled data and for each chain separately are reported in Table 6.1.[20] Estimates for the alternative shrinkage methods can be found in Appendix A.10 in Table A.2. We find a positive significant change in employment for the pooled data. As further analysis shows, this result is driven by the significant positive employment change in Burger King. However, when correcting for multiple comparisons at a family-wise error rate of 0.05 using the Holm–Bonferroni method, the null hypothesis of no significant impact of the minimum wage on employment cannot be rejected for any chain. Comparing the results across chains, KFC shows a different pattern from the other stores, as KFC is the only chain with a point estimate that is in line with the theory of increasing labor demand in a less labor costly environment, however the estimate is statistically insignificant.

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

7. Concluding remarks

Pairwise cross-smoothing provides a unifying framework to analyze and compare smoothing methods for categorical data that nests different approaches from the literature on (generalized) ridge regression, nonparametric smoothing

[20] Table A.1 in Appendix A.10 includes means, variances and number of observations for all subgroups.

kernels and model averaging. It penalizes ℓ_2 differences between estimation parameters and first stage estimates. The estimator can be easily implemented with standard software packages using the closed-form solutions in this paper. It has favorable risk properties compared to the ordinary least squares and other commonly used approaches. For future research, relaxing the assumption of a fixed number of groups, i.e. allowing for J to grow with the sample size with closeness restrictions that are related to sparsity in the sense of few different locations and alternative risk functions in the sense of Hansen (2016a) should be considered.

Acknowledgments

The authors would like to thank the Editor, the Associate Editor, two anonymous referees, Lyudmila Grigoryeva, Chu-An Liu, Winfried Pohlmeier, Patrik Guggenberger, Jeffrey S. Racine and the participants of the European Meeting of the Econometric Society 2016 for their time, comments, and discussions that helped to greatly improve the paper. Financial support from the German Research Foundation through FOR 1882 Psychoeconomics, Project 215899445 (Jana Mareckova) and Project 219805061 (Phillip Heiler) and by the Graduate School of Decision Sciences at the University of Konstanz, Germany is gratefully acknowledged. Computations were supported by the Karlsruhe bwHPC-Cluster, Germany. All remaining errors are ours.

Appendix A

A.1. First- and second-order conditions for pairwise cross-smoothing

Let $S_{\boldsymbol{\Lambda}}(\boldsymbol{\mu})$ denote the objective function in (2.2) where $\boldsymbol{\Lambda} = (\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1J}, \lambda_{21}, \ldots, \lambda_{JJ})$ and $\lambda_{kk} = 0$ for all $k = \{1, \ldots, J\}$. Note that:

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

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

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

Setting (A.1) equal to zero to solve for $\hat{\mu}_k^{PCS}$ and rearranging the terms yields

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

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

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

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

An estimator defined as the solution to (2.2) exists and is a unique global minimizer if and only if $\sum_{j=1}^{J}\lambda_{kj} > -n_k$ for all $k \in \{1, \ldots, J\}$.

A.2. Proof of Proposition 3.1

Let $r_k = \frac{1}{n}\sum_{i=1}^{n} D_{ik}Y_i$ and $s_k = \hat{p}_k \vee \underline{p}$ and thus $\hat{\mu}_k = r_k/s_k$. We first establish three auxiliary results regarding moments and a tail probability: First note that for any iid random variable A_i with $E[A_i] = 0$ and finite 4th moment we have that $E[(n^{-1}\sum_{i=1}^{n} A_i)^4] = O(n^{-2})$. Second, the assumption $\inf_k p_k > \underline{p} > 0$ allows us to use a lower-tail Chernoff bound, i.e.

$$\begin{aligned} P(\hat{p}_k \leq \underline{p}) &= P(\hat{p}_k \leq [1 - (p_k - \underline{p})/p_k]p_k) \\ &= P\left(\sum_{i=1}^{n} D_{ik} \leq [1 - (p_k - \underline{p})/p_k]p_k n\right) \\ &\leq \exp\left(-\frac{np_k}{2}\left[\frac{(p_k - \underline{p})}{p_k}\right]^2\right). \end{aligned}$$

Third, the rate for even second and fourth moments of the "censored" variable s_k is of same order as the uncensored $\hat{p}_k$. Consider the centered censored variable

$$s_k - E[s_k] = (\mathbb{1}(\hat{p}_k \leq \underline{p}) - P(\hat{p}_k \leq \underline{p}))\underline{p} + \hat{p}_k - p_k + P(\hat{p}_k \leq \underline{p})E[\hat{p}_k|\hat{p}_k \leq \underline{p}] - \mathbb{1}(\hat{p}_k \leq \underline{p})\hat{p}_k.$$

Thus, for $m = 2, 4, 8, \ldots$

$$\begin{aligned} E[(s_k - E[s_k])^m] \leq C_m \Bigg[& E[(\mathbb{1}(\hat{p}_k \leq \underline{p}) - P(\hat{p}_k \leq \underline{p}))^m]\underline{p}^m + E[(\hat{p}_k - p_k)^m] \\ & + E[(P(\hat{p}_k \leq \underline{p})E[\hat{p}_k|\hat{p}_k \leq \underline{p}] - \mathbb{1}(\hat{p}_k \leq \underline{p})\hat{p}_k)^m] \Bigg]. \end{aligned}$$

It is straightforward to show that $E[(\mathbb{1}(\hat{p}_k \leq \underline{p}) - P(\hat{p}_k \leq \underline{p}))^m] = O(P(\hat{p}_k \leq \underline{p}))$ which follows the exponential decay from the Chernoff bound above. Moreover note that

$$\begin{aligned} E[(P(\hat{p}_k \leq \underline{p})E[\hat{p}_k|\hat{p}_k \leq \underline{p}] - \mathbb{1}(\hat{p}_k \leq \underline{p})\hat{p}_k)^m] &= E[\hat{p}_k|\hat{p}_k \leq \underline{p}]^m P(\hat{p}_k \leq \underline{p})^m(1 - P(\hat{p}_k \leq \underline{p})) \\ &\quad + P(\hat{p}_k \leq \underline{p})E[(\hat{p}_k - P(\hat{p}_k \leq \underline{p})E[\hat{p}_k|\hat{p}_k \leq \underline{p}])^m|\hat{p}_k \leq \underline{p}] \\ &= O(P(\hat{p}_k \leq \underline{p})). \end{aligned}$$

which is again decaying exponentially. However the centered moments $E[(\hat{p}_k - p_k)^m]$ are of slower order $O(1/n)$ for $m = 2$ and $O(1/n^2)$ for $m = 4$, i.e. polynomial. Thus the rate for the centered moments of the censored variable s_k is at most of the same rate as the centered p_k.

A second order mean-value expansion for $\hat{\mu}_k$ now yields

$$\frac{r_k}{s_k} = \frac{E[r_k]}{E[s_k]} + \frac{1}{E[s_k]}(r_k - E[r_k]) - \frac{E[r_k]}{E[s_k]^2}(s_k - E[s_k]) + \frac{\tilde{r_k}}{\tilde{s_k}^3}(s_k - E[s_k])^2 - \frac{1}{\tilde{s_k}^2}(r_k - E[r_k])(s_k - E[s_k])$$

where $\tilde{s_k}$ and $\tilde{r_k}$ are on the line segment between s_k and $E[s_k]$ and r_k and $E[r_k]$ respectively. Taking expectations yields

$$E\left[\frac{r_k}{s_k}\right] = \frac{E[r_k]}{E[s_k]} + E\left[\frac{\tilde{r_k}}{\tilde{s_k}^3}(s_k - E[s_k])^2\right] - E\left[\frac{1}{\tilde{s_k}^2}(r_k - E[r_k])(s_k - E[s_k])\right].$$

Consider the first denominator

$$E[s_k] = E[\hat{p}_k \mathbb{1}(\hat{p}_k > \underline{p})] + \underline{p}P(\hat{p}_k \leq \underline{p}).$$

Now note that $\lim_{n\to\infty} E[s_k] = \lim_{n\to\infty} E[\hat{p}_k \vee \underline{p}] \geq \lim_{n\to\infty} E[\hat{p}_k] = p_k$ and $\lim_{n\to\infty} E[\hat{p}_k \mathbb{1}(\hat{p}_k > \underline{p})] \leq \lim_{n\to\infty} E[\hat{p}_k] = p_k$. Thus it follows that $E[s_k] = p_k + o(1)$. As $E[r_k] = \mu_k p_k$ and $p_k > \underline{p} > 0$ continuity implies that $E[r_k]/E[s_k] = \mu_k(1 + o(1))$.

For the second term note that:

$$E\left[\frac{\tilde{r_k}}{\tilde{s_k}^3}(s_k - E[s_k])^2\right] = E\left[\frac{E[\tilde{r_k}]}{\tilde{s_k}^3}(s_k - E[s_k])^2\right] + E\left[\frac{1}{\tilde{s_k}^3}(\tilde{r_k} - E[\tilde{r_k}])(s_k - E[s_k])^2\right].$$

Using the lower bound $\underline{p} > 0$, we obtain

$$\left|E\left[\frac{E[\tilde{r_k}]}{\tilde{s_k}^3}(s_k - E[s_k])^2\right]\right| \leq \frac{|E[\tilde{r_k}]|}{\underline{p}^3}E[(s_k - E[s_k])^2] = O(1/n)$$

and

$$\begin{aligned} \left|E\left[\frac{1}{\tilde{s_k}^3}(\tilde{r_k} - E[\tilde{r_k}])(s_k - E[s_k])^2\right]\right| &\leq \frac{1}{\underline{p}^3}E[(\tilde{r_k} - E[\tilde{r_k}])^2]^{1/2}E[(s_k - E[s_k])^4]^{1/2} \\ &= O(1/\sqrt{n})O(1/n) \\ &= O(n^{-3/2}) \end{aligned}$$

by convexity of the absolute value with Jensen's inequality and Hölder's inequality. The rates for $\tilde{r}_k$ come from the rate of r_k. The rates for the s_k follow from the auxiliary results above.

The last term is bounded as

$$\left|E\left[\frac{1}{\tilde{s_k}^2}(r_k - E[r_k])(s_k - E[s_k])\right]\right| \leq \frac{1}{\underline{p}^2}E[(r_k - E[r_k])^2]^{1/2}E[(s_k - E[s_k])^2]^{1/2}$$
$$= O(1/n)$$

by convexity of the absolute value with Jensen's inequality, Hölder's inequality, and the $\underline{p} > 0$ bound. Putting everything together yields

$$E[\hat{\mu}_k] = \mu_k(1 + o(1)) + O(1/n)$$
$$= \mu_k + o(1).$$

For the variance first note that

$$\hat{\mu}_k - \mu_k = \frac{\frac{1}{n}\sum_i^n D_{ik}Y_i}{\hat{p}_k \vee \underline{p}} - \frac{\hat{p}_k \vee \underline{p}}{\hat{p}_k \vee \underline{p}}\mu_k$$
$$= \frac{\frac{1}{n}\sum_i^n D_{ik}(Y_i - \mu_k)}{\hat{p}_k \vee \underline{p}} + \frac{(\hat{p}_k - \underline{p})\mathbb{1}(\hat{p}_k \leq \underline{p})}{\hat{p}_k \vee \underline{p}}\mu_k$$
$$= \frac{r_{k,c}}{s_k} + \zeta_k.$$

Thus the variance of the modified first-stage is given by

$$V[\hat{\mu}_k] = V\left[\frac{r_{k,c}}{s_k} + \zeta_k\right].$$

First note that

$$V[\zeta_k] \leq E[\zeta_k^2] \leq \frac{\mu_k^2}{\underline{p}^2}E[\mathbb{1}(\hat{p}_k \leq \underline{p})] = \frac{\mu_k^2}{\underline{p}^2}P(\hat{p}_k \leq \underline{p})$$

which can again be controlled via the lower-tail Chernoff bound. For the first term note that

$$V\left[\frac{r_{k,c}}{s_k}\right] = E\left[\frac{r_{k,c}^2}{s_k^2}\right] - E\left[\frac{r_{k,c}}{s_k}\right]^2.$$

Identically to the proof for the mean of the modified estimator, we have that $E[r_{k,c}/s_k] = E[r_{k,c}]/E[s_k] + O(1/n) = 0 + O(1/n)$ which implies that $E[r_{k,c}/s_k]^2 = O(1/n^2)$. Thus, it is sufficient to show that $E[r_{k,c}^2/s_k^2] = E[r_{k,c}^2]/p_k^2 + o(1/n)$ and $cov(r_{k,c}/s_k, \zeta_k) = o(1/n)$. Expanding function $g(a, b) = a/b^2$ evaluated for $(r_{k,c}^2, s_k)$ around $(E[r_{k,c}^2], E[s_k])$ yields

$$\frac{r_{k,c}^2}{s_k^2} = \frac{E[r_{k,c}^2]}{E[s_k]^2} + \frac{1}{E[s_k]^2}(r_{k,c}^2 - E[r_{k,c}^2]) - 2\frac{E[r_{k,c}^2]}{E[s_k]^3}(s_k - E[s_k]) - \frac{2}{\tilde{s_k}^3}(r_{k,c}^2 - E[r_{k,c}^2])(s_k - E[s_k]) + \frac{6}{2}\frac{\tilde{r_{k,c}^2}}{\tilde{s_k}^4}(s_k - E[s_k])^2$$

where $\tilde{r_{k,c}^2}$ and $\tilde{s_k}$ are on the line segment between $r_{k,c}^2$ and $E[r_{k,c}^2]$ and s_k and $E[s_k]$ respectively. Taking expectations yields:

$$E\left[\frac{r_{k,c}^2}{s_k^2}\right] = \frac{E[r_{k,c}^2]}{E[s_k]^2} - 2E\left[\frac{1}{\tilde{s_k}^3}(r_{k,c}^2 - E[r_{k,c}^2])(s_k - E[s_k])\right] + 3E\left[\frac{\tilde{r_{k,c}^2}}{\tilde{s_k}^4}(s_k - E[s_k])^2\right].$$

Using the results from above for the mean, we have that

$$\frac{E[r_{k,c}^2]}{E[s_k]^2} = \frac{E[r_{k,c}^2]}{p_k^2}(1 + o(1))$$

and

$$\left|E\left[\frac{1}{\tilde{s_k}^3}(r_{k,c}^2 - E[r_{k,c}^2])(s_k - E[s_k])\right]\right| \leq \frac{1}{\underline{p}^3}E[(r_{k,c}^2 - E[r_{k,c}^2])^2]^{1/2}E[(s_k - E[s_k])^2]^{1/2}$$
$$= O(n^{-3/2})$$

by the 4th moment rate for $r_{k,c}$ and variance rate for s_k. Similarly

$$\left|E\left[\frac{\tilde{r_{k,c}^2}}{\tilde{s_k}^4}(s_k - E[s_k])^2\right]\right| \leq \frac{1}{\underline{p}^4}E[(\tilde{r_{k,c}^2})^2]^{1/2}E[(s_k - E[s_k])^4]^{1/2}$$

$$= O(1/n^2)$$

by the 4th moment inequalities for s_k and $r_{k,c}$ and exploiting that $\tilde{r}^2_{k,c}$ is between constant $E[r^2_{k,c}]$ and $r^2_{k,c}$ which implies that

$$\begin{aligned} E[(\tilde{r}^2_{k,c})^2] &\leq 2(E[r^4_{k,c}] + E[r^2_{k,c}]^2) \\ &= O(1/n^2) \end{aligned}$$

by the usual moment rates. Collecting all rates yields

$$\begin{aligned} E\left[\frac{r^2_{k,c}}{s^2_k}\right] &= \frac{E[r^2_{k,c}]}{p^2_k}(1 + o(1)) + O(n^{-3/2}) \\ &= \frac{\sigma^2_k}{np_k} + o(1/n) \end{aligned}$$

as $E[r^2_{k,c}] = n^{-2}\sum_i \sum_j E[D_{ik}D_{jk}(Y_i - \mu_k)(Y_j - \mu_k)] = E[(Y_i - \mu_k)^2 | D_{ik} = 1]p_k/n = \sigma^2_k p_k/n$. Finally note that by the Cauchy–Schwarz inequality

$$cov(r_{k,c}/s_k, \zeta_k) = O(1/\sqrt{n})o(1/\sqrt{n}) = o(1/n)$$

and consequently $V[r_k/s_k] = \sigma^2_k/(p_k n) + o(1/n)$. The covariance part follows from a similar expansion as above, i.e.

$$\begin{aligned} cov(\hat{\mu}_k, \hat{\mu}_l) &= \frac{E[r_{k,c}r_{l,c}]}{E[s_k s_l]} + o(1/n) \\ &= 0 + o(1/n) \end{aligned}$$

as $E[r_{k,c}r_{l,c}] = n^{-2}\sum_i \sum_j E[D_{ik}D_{jl}(Y_i - \mu_k)(Y_j - \mu_l)] = 0$ by independence and the orthogonality of the groups, i.e. $D_{ik}D_{il} = 0$.

A.3. Proof of Theorem 3.1

The problem of the constrained minimization of (3.3) can be rewritten as the following Lagrangian:

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

The first-order conditions (FOCs) are given by

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

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

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

The expression for ω^*_{kj} can be inferred from the jth entry of $\boldsymbol{\omega}^*_k$. For uniqueness conditions consider Online Appendix B.3.

A.4. Weighted mean squared error optimal and plug-in smoothing parameters for kernel and (generalized) ridge regression

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

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

A.4.1. Kernel smoothing

The implicit kernel constraints yield the estimator of the form

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

Now using the identical approximation from the proof of Proposition 3.1 yields the expected risk for group k

$$E[(\hat{\mu}_k^{Kernel}(\mathbf{\Lambda}) - \mu_k)^2] = \left(\frac{(np_k)^2 \frac{1}{\gamma_k n} + \sum_{j\neq k}\sum_{m\neq k,j}\lambda_j\lambda_m(\mu_j-\mu_k)(\mu_m-\mu_k)}{(np_k + \sum_{l\neq k}\lambda_l)^2} + \frac{\sum_{j\neq k}\lambda_j^2[\frac{1}{\gamma_j n} + (\mu_j-\mu_k)^2]}{(np_k + \sum_{l\neq k}\lambda_l)^2}\right)(1+o(1))$$

and overall weighted MSE

$$\sum_{k=1}^{J}\gamma_k E[(\hat{\mu}_k^{Kernel}(\mathbf{\Lambda}) - \mu_k)^2] = \left(\sum_{k=1}^{J}\left[\frac{np_k}{np_k + \sum_{l\neq k}\lambda_l}\right]^2 \frac{1}{n} + \frac{\sum_{k=1}^{J}\sum_{j\neq k}\lambda_j^2 \frac{\gamma_k}{\gamma_j}\frac{1}{n}}{(np_k + \sum_{l\neq k}\lambda_l)^2} + \frac{\sum_{k=1}^{J}\sum_{j\neq k}\sum_{m\neq k}\gamma_k\lambda_j\lambda_m(\mu_j-\mu_k)(\mu_m-\mu_k)}{(np_k + \sum_{l\neq k}\lambda_l)^2}\right)(1+o(1)).$$

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

In the Monte Carlo simulations in Section 5, the numerical optimization required for the Kernel estimate with plug-in smoothing parameters was performed in Matlab using the `fmincon` function under the uniqueness and existence restriction $\sum_{l\neq k}\lambda_{kl} > -n_k$, choosing the `sqp` algorithm and setting the optimality tolerance to 10^{-16}. The analysis of convergence to a global optimum showed that the optimizer converges faster for moderate and large δ values. In case of low δ values the local optima only differ on the third or fourth decimal digit for the different iterations.

A.4.2. Ridge regression

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

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

The weighted MSE of the RR estimator takes form:

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

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

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

The sum notation of the FOC reveals that the smoothing parameter λ is non-trivially intertwined across the reference groups. This implies that a closed-form solution exists only in special cases, e.g. for a balanced design when $p_k = 1/J$ for all k or for a design with 2 groups. In general, the FOC is a polynomial equation of any order between 1 and $3(J-1)+1$. This means that already for some designs with more than two groups, one has to solve a polynomial equation of order larger than 4. According to the Abel–Ruffini theorem, there is no guarantee that a solution in radicals exists for polynomial equations of order five and higher with arbitrary coefficients. In these cases, one has to solve the FOC numerically and find the global minimum. The plug-in estimator can be obtained by replacing the expressions for p_k, σ_k^2 and μ_k for $k = 1\ldots, J$

by the corresponding estimates as for the PCS. Depending on the design, there is either a closed form solution for optimal λ or λ has to be obtained using numerical optimization.

The Monte Carlo simulations in Section 5 and application in Section 6.1 use the closed form solution for a balanced design. For the minimum wage application in Section 6.2, the numerical optimization implemented in Matlab used the `fsolve` solver with optimality tolerance and step tolerance set to 10^{-20}.

A.4.3. Generalized ridge regression

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

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

which can be rewritten in the following weighted form:

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

The GRR estimator depends on the weights within its own reference category k. Therefore, optimization of the parameter vector MSE can be done group by group and is invariant to any MSE weighting. The MSE of the GRR estimator takes form:

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

The optimal solution for ω_k is:

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

The plug-in estimator can be obtained by replacing the expressions for p_k, σ_k^2 and μ_k for $k = 1\ldots,J$ by the corresponding estimates as for the PCS.

A.5. Proof of Lemma 4.1

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

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

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

Proof. $\omega_{kj}^* \to \bar{w}_{kj} = \frac{\gamma_j(1+\boldsymbol{\delta}'\Delta_k'diag(\boldsymbol{\gamma})\Delta_j\boldsymbol{\delta})}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}\boldsymbol{\delta}'\Delta'M_1\Delta\boldsymbol{\delta}}$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. Use the closed-form in (3.4) and continuity together with $\sqrt{n}\Delta_k\boldsymbol{\mu} \to \Delta_k\boldsymbol{\delta}$ and $\sqrt{n}\Delta\boldsymbol{\mu} \to \Delta\boldsymbol{\delta}$.

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

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

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

Proof. $\hat{\omega}_{kj} \xrightarrow{d} w_{kj}^a = \frac{\gamma_j(1+(\mathbf{Z}+\boldsymbol{\delta})'\Delta_k'diag(\boldsymbol{\gamma})\Delta_j(\mathbf{Z}+\boldsymbol{\delta}))}{\boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_1\Delta(\mathbf{Z}+\boldsymbol{\delta})}$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. Take $\hat{\omega}_{kj}$ according to (3.5). Note that $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}$ and thus $\hat{M}_1 \xrightarrow{p} M_1$. Additionally $\sqrt{n}(\hat{\mu}_k - \hat{\mu}_j) = \sqrt{n}(\hat{\mu}_k - \mu_k) - \sqrt{n}(\hat{\mu}_j - \mu_j) + \sqrt{n}(\mu_k - \mu_j) \xrightarrow{d} Z_k - Z_j + \delta_k - \delta_j$ since $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$. Similarly $\sqrt{n}\Delta_k\hat{\boldsymbol{\mu}} \xrightarrow{d} \Delta_k(\mathbf{Z}+\boldsymbol{\delta})$ and $\sqrt{n}\Delta\hat{\boldsymbol{\mu}} \xrightarrow{d} \Delta(\mathbf{Z}+\boldsymbol{\delta})$. The rest follows from continuity of $\hat{\omega}_{kj}$.

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

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

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

A.6. Proof of Theorem 4.1

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

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

Using Lemma 4.1 together with $\sqrt{n}(\hat{\mu}_j - \mu_j) = O_p(1)$ for all $\{F_n\}$ we have that

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

Proof. $\sqrt{n}(\hat{\mu}_k^{PCS}(\boldsymbol{\omega}_k^*) - \mu_k - B_{2k}(\boldsymbol{\omega}_k^*)) \xrightarrow{d} \mathcal{N}\left(0, \sum_{j=1}^{J} \bar{\omega}_{kj}^2 \frac{\sigma_j^2}{p_j}\right)$ if $\{F_n\} \in S(\boldsymbol{\delta}, V_0) \cup S(\infty, V_0)$. Using the definition from the PCS, the CLT for $\sqrt{n}(\hat{\mu}_j - \mu_j)$ together with Lemma 4.1 yields

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

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

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

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

where convergence in distribution follows from joint convergence of the $\hat{\omega}_{kj}$'s and $\sqrt{n}(\hat{\mu}_j - \mu_j)$'s as they are continuous functions of the same random normal vector and using Lemma 4.1 for the weights for $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$ and $\sqrt{n}(\mu_j - \mu_k) \to \delta_j - \delta_k$ by definition of sequences in $S(\boldsymbol{\delta}, V_0)$.

Proof. $\sqrt{n}(\hat{\mu}_k^{PCS}(\hat{\boldsymbol{\omega}}_k) - \mu_k - B_{3k}(\bar{\boldsymbol{\omega}}_k)) \xrightarrow{d} Z_k \sim \mathcal{N}\left(0, \frac{\sigma_k^2}{p_k}\right)$ if $\{F_n\} \in S(\infty, V_0)$. By Lemma 4.1, $\hat{\omega}_{kj} \xrightarrow{p} \bar{\omega}_{kj}$ as $\{F_n\} \in S(\infty, V_0)$. Rewriting the PCS yields

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

or equivalently

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

The right hand side is asymptotically normal as the components are stabilizing transformations of continuous functions of the same random normal vector. In terms of its asymptotic variance, one can either show the equivalence to Z_k using the delta method or simpler by Theorem 4.2. It implies that as $\|\Delta\boldsymbol{\delta}\|_\infty \to \infty$, the PCS risk is converging to the OLS. Since both estimators are asymptotically normal, the asymptotic variances have to coincide.

A.7. Proof of Theorem 4.2 and Corollary 4.1

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

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

which by Lemma 4.1 converge in distribution, i.e.

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

with $d_0 = \boldsymbol{\gamma}'\boldsymbol{\iota}_J + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_1\Delta(\mathbf{Z}+\boldsymbol{\delta})$. By Theorem 4.1, the distributional limit for the PCS under $\{F_n\} \in S(\boldsymbol{\delta}, V_0)$ is given by

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

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

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

with

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

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

$$E[B] = E[\eta(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_3\mathbf{Z}] = E\left[tr\left(\frac{\partial}{\partial\mathbf{x}}\eta(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_3V_0\right)\right]$$

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

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

and hence

$$\begin{aligned}E[B] &= tr(\Delta'M_3V_0)E\left[\frac{1}{d_0}\right] - E\left[\frac{tr(\Delta'M_1\Delta(\mathbf{Z}+\boldsymbol{\delta})(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_3V_0)}{d_0^2}\right]\\ &= \boldsymbol{\gamma}'\boldsymbol{\iota}_J tr(\Delta'M_3V_0)E\left[\frac{1}{d_0^2}\right] + \frac{1}{2}tr(\Delta'M_3V_0)E\left[\frac{(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_1\Delta(\mathbf{Z}+\boldsymbol{\delta})}{d_0^2}\right] - E\left[\frac{(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_3V_0\Delta'M_1\Delta(\mathbf{Z}+\boldsymbol{\delta})}{d_0^2}\right].\end{aligned}$$

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

$$\begin{aligned}\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) =& \rho(\hat{\boldsymbol{\mu}}, \boldsymbol{\mu}) + E\left[\frac{(\mathbf{Z}+\boldsymbol{\delta})'\Delta'C\Delta(\mathbf{Z}+\boldsymbol{\delta})}{(tr(V_0^{-1}) + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_1\Delta(\mathbf{Z}+\boldsymbol{\delta}))^2}\right]\\ &- 2tr(V_0^{-1})tr(\Delta'M_3V_0)E\left[\frac{1}{(tr(V_0^{-1}) + \frac{1}{2}(\mathbf{Z}+\boldsymbol{\delta})'\Delta'M_1\Delta(\mathbf{Z}+\boldsymbol{\delta}))^2}\right]\end{aligned}$$

with $C = M_2 - tr(\Delta'M_3V_0)M_1 + 2M_3V_0\Delta'M_1$.

Simplification:. Note that

$$\begin{aligned}M_1 &= V_0^{-1}\otimes V_0^{-1}, & \Delta &= (I_J\otimes\boldsymbol{\iota}_J) - (\boldsymbol{\iota}_J\otimes I_J)\\ M_2 &= V_0^{-1}\otimes\boldsymbol{\gamma}\boldsymbol{\gamma}', & \Delta' &= (I_J\otimes\boldsymbol{\iota}_J') - (\boldsymbol{\iota}_J'\otimes I_J)\\ M_3 &= V_0^{-1}\otimes\boldsymbol{\gamma}\end{aligned}$$

First simplify all matrices of the asymptotic risk formula:

$$\begin{aligned}\Delta'M_1\Delta &= ((I_J\otimes\boldsymbol{\iota}_J') - (\boldsymbol{\iota}_J'\otimes I_J))[V_0^{-1}\otimes V_0^{-1}]((I_J\otimes\boldsymbol{\iota}_J) - (\boldsymbol{\iota}_J\otimes I_J))\\ &= [(V_0^{-1}\otimes\boldsymbol{\iota}_J'V_0^{-1}) - (\boldsymbol{\iota}_J'V_0^{-1}\otimes V_0^{-1})]((I_J\otimes\boldsymbol{\iota}_J) - (\boldsymbol{\iota}_J\otimes I_J))\\ &= [(V_0^{-1}\otimes\boldsymbol{\iota}_J'V_0^{-1}\boldsymbol{\iota}_J) - (V_0^{-1}\boldsymbol{\iota}_J\otimes\boldsymbol{\iota}_J'V_0^{-1}) - (\boldsymbol{\iota}_J'V_0^{-1}\otimes V_0^{-1}\boldsymbol{\iota}_J) + (\boldsymbol{\iota}_J'V_0^{-1}\boldsymbol{\iota}_J\otimes V_0^{-1})].\end{aligned}$$

Note that $\boldsymbol{\iota}_J'V_0^{-1}\boldsymbol{\iota}_J = tr(V_0^{-1})$ and

$$(V_0^{-1}\boldsymbol{\iota}_J\otimes\boldsymbol{\iota}_J'V_0^{-1}) + (\boldsymbol{\iota}_J'V_0^{-1}\otimes V_0^{-1}\boldsymbol{\iota}_J) = 2V_0^{-1}\boldsymbol{\iota}_J\boldsymbol{\iota}_J'V_0^{-1}$$

which implies that

$$\Delta'M_1\Delta = 2(tr(V_0^{-1})V_0^{-1} - V_0^{-1}\boldsymbol{\iota}_J\boldsymbol{\iota}_J'V_0^{-1}).$$

Now further note that

$$\begin{aligned}\Delta'M_3V_0 &= [(I_J\otimes\boldsymbol{\iota}_J') - (\boldsymbol{\iota}_J'\otimes I_J)](V_0^{-1}\otimes\boldsymbol{\gamma})V_0\\ &= [(V_0^{-1}\otimes\boldsymbol{\iota}_J'\boldsymbol{\gamma}) - (\boldsymbol{\iota}_J'V_0^{-1}\otimes\boldsymbol{\gamma})]V_0\\ &= [tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}']V_0\\ &= tr(V_0^{-1})I_J - V_0^{-1}\boldsymbol{\iota}_J\boldsymbol{\iota}_J'\end{aligned}$$

Thus for the trace we have that

$$\begin{aligned}tr(\Delta'M_3V_0) &= tr(V_0^{-1})tr(I_J) - tr(V_0^{-1}\boldsymbol{\iota}_J\boldsymbol{\iota}_J')\\ &= tr(V_0^{-1})[J-1].\end{aligned}$$

Using these simplifications one obtains

$$\begin{aligned}\Delta'M_3V_0\Delta'M_1\Delta &= 2[tr(V_0^{-1})I_J - \boldsymbol{\gamma}\boldsymbol{\gamma}'V_0][tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}']\\ &= 2[tr(V_0^{-1})^2V_0^{-1} - tr(V_0^{-1})\boldsymbol{\gamma}\boldsymbol{\gamma}' - tr(V_0^{-1})\boldsymbol{\gamma}\boldsymbol{\gamma}' + \boldsymbol{\gamma}\boldsymbol{\gamma}'V_0\boldsymbol{\gamma}\boldsymbol{\gamma}']\\ &= 2tr(V_0^{-1})(tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}')\end{aligned}$$

as $\boldsymbol{\gamma}'V_0\boldsymbol{\gamma} = tr(V_0^{-1})$. Similarly,

$$\begin{aligned}\Delta'M_2\Delta &= ((I_J \otimes \iota_J') - (\iota_J' \otimes I_J))[V_0^{-1} \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}'](I_J \otimes \iota_J) - (\iota_J \otimes I_J))\\ &= [(V_0^{-1} \otimes \iota_J'\boldsymbol{\gamma}\boldsymbol{\gamma}') - (\iota_J'V_0^{-1} \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}')](I_J \otimes \iota_J) - (\iota_J \otimes I_J))\\ &= (V_0^{-1} \otimes \iota_J'\boldsymbol{\gamma}\boldsymbol{\gamma}'\iota_J) - (V_0^{-1}\iota_J \otimes \iota_J\boldsymbol{\gamma}\boldsymbol{\gamma}') - (\iota_J'V_0^{-1} \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}'\iota_J) + (\iota_J'V_0^{-1}\iota_J \otimes \boldsymbol{\gamma}\boldsymbol{\gamma}')\\ &= tr(V_0^{-1})^2V_0^{-1} - tr(V_0^{-1})(V_0^{-1}\iota_J \otimes \boldsymbol{\gamma}') - tr(V_0^{-1})(\iota_J'V_0^{-1} \otimes \boldsymbol{\gamma}) + tr(V_0^{-1})\boldsymbol{\gamma}\boldsymbol{\gamma}'\\ &= tr(V_0^{-1})(tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}')\end{aligned}$$

as $V_0^{-1}\iota_J = \boldsymbol{\gamma}$. Putting these terms together yields

$$\begin{aligned}&\Delta'M_2\Delta - tr(\Delta'M_3V_0)\Delta'M_1\Delta + 2\Delta'M_3V_0\Delta'M_1\Delta\\ &\quad= tr(V_0^{-1})(tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}') - tr(V_0^{-1})[J-1]2(tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}') + 4tr(V_0^{-1})(tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}')\\ &\quad= tr(V_0^{-1})[1 - 2(J-1) + 4](tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}')\\ &\quad= tr(V_0^{-1})[7 - 2J](tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}').\end{aligned}$$

For the last term in the PCS risk note that

$$-2tr(V_0^{-1})tr(\Delta'M_3V_0) = -2tr(V_0^{-1})tr(V_0^{-1})[J-1].$$

For the following, let $x = (\mathbf{Z}+\boldsymbol{\delta})'(tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}')(\mathbf{Z}+\boldsymbol{\delta})$ and thus the overall risk difference of PCS to OLS is given by

$$\begin{aligned}&\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{OLS}, \boldsymbol{\mu})\\ &\qquad= tr(V_0^{-1})E\left[\frac{(7-2J)x - 2tr(V_0^{-1})(J-1)}{(tr(V_0^{-1}) + x)^2}\right]\\ &\qquad= tr(V_0^{-1})E\left[\frac{5x + 2(tr(V_0^{-1}) + x) - 2J(tr(V_0^{-1}) + x)}{(tr(V_0^{-1}) + x)^2}\right]\\ &\qquad= tr(V_0^{-1})\left\{5E\left[\frac{x}{(tr(V_0^{-1}) + x)^2}\right] - 2(J-1)E\left[\frac{1}{tr(V_0^{-1}) + x}\right]\right\}\end{aligned}$$

This can be further simplified by standardizing the $\mathbf{Z}$ vector, i.e.

$$\begin{aligned}x &= (\mathbf{Z}+\boldsymbol{\delta})'V_0^{-1/2}(tr(V_0^{-1})I_J - V_0^{1/2}\boldsymbol{\gamma}\boldsymbol{\gamma}'V_0^{1/2})V_0^{-1/2}(\mathbf{Z}+\boldsymbol{\delta})\\ &= tr(V_0^{-1})(\mathbf{Z}+\boldsymbol{\delta})'V_0^{-1/2}\left[I_J - \frac{V_0^{-1/2}\iota_J\iota_J'V_0^{-1/2}}{tr(V_0^{-1})}\right]V_0^{-1/2}(\mathbf{Z}+\boldsymbol{\delta})\end{aligned}$$

Let $z = (\mathbf{Z}+\boldsymbol{\delta})'V_0^{-1/2}\left[I_J - \frac{V_0^{-1/2}\iota_J\iota_J'V_0^{-1/2}}{tr(V_0^{-1})}\right]V_0^{-1/2}(\mathbf{Z}+\boldsymbol{\delta})$. The overall PCS risk gain is then given by

$$\begin{aligned}\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{OLS}, \boldsymbol{\mu}) &= \left\{5E\left[\frac{z}{(1+z)^2}\right] - 2(J-1)E\left[\frac{1}{1+z}\right]\right\}\\ &= -(2J-7)E\left[\frac{1}{1+z}\right] - 5E\left[\frac{1}{(1+z)^2}\right].\end{aligned}$$

As $I_J - V_0^{-1/2}\iota_J\iota_J'V_0^{-1/2}/tr(V_0^{-1})$ is symmetric and idempotent this is strictly negative for $J \geq 3.5$ implying Corollary 4.1.

A.8. Supplementary material for Section 4.4

Asymptotic risk of the model averaging estimator:. Let R be the restriction matrix that sets all means equal to each other via the linear restrictions $R\boldsymbol{\mu} = 0$. The asymptotic risk relative to OLS of the model averaging (MA) estimator as in Hansen (2016a), equation (57) and (59) with $G = I_J$ then simplifies to

$$\rho(\hat{\boldsymbol{\mu}}^{MA}, \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{OLS}, \boldsymbol{\mu}) = \tau^2E\left[\frac{1}{(\mathbf{Z}+\boldsymbol{\delta})'B(\mathbf{Z}+\boldsymbol{\delta})}\right] - 2\tau tr(A)E\left[\frac{1}{(\mathbf{Z}+\boldsymbol{\delta})'B(\mathbf{Z}+\boldsymbol{\delta})}\right] + 4\tau E\left[\frac{(\mathbf{Z}+\boldsymbol{\delta})'B_1AB_1(\mathbf{Z}+\boldsymbol{\delta})}{((\mathbf{Z}+\boldsymbol{\delta})'B(\mathbf{Z}+\boldsymbol{\delta}))^2}\right]$$

with

$$\begin{aligned} B_1'AB_1 &= R'(RV_0R')^{-1}RV_0WV_0B \\ &= R'(RV_0R')^{-1}RV_0R'(RV_0R')^{-1}R \\ &= R'(RV_0R')^{-1}R \\ &= B \end{aligned}$$

and

$$\begin{aligned} tr(A) &= tr(W^{1/2}V_0R'(RV_0R')^{-1}RV_0W^{1/2}) \\ &= tr((RV_0R')^{-1}RV_0R') \\ &= J-1 \end{aligned}$$

Moreover note that

$$\begin{aligned} B &= R'(RV_0R')^{-1}R \\ &= \frac{1}{tr(V_0^{-1})}\left[tr(V_0^{-1})V_0^{-1} - \boldsymbol{\gamma}\boldsymbol{\gamma}'\right]. \end{aligned}$$

Thus setting $\tau = (J-3)$ as in Hansen (2016a), the overall risk difference for $J \geq 3$ simplifies to

$$\begin{aligned} \rho(\hat{\boldsymbol{\mu}}^{MA}, \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{OLS}, \boldsymbol{\mu}) &= [\tau^2 - 2\tau(J-1) + 4\tau]tr(V_0^{-1})E\left[\frac{1}{x}\right] \\ &= -(J-3)^2E\left[\frac{1}{z}\right] \end{aligned}$$

Asymptotic risk comparison:. First we establish some auxiliary results. Note that $V_0^{-1/2}(\mathbf{Z}+\boldsymbol{\delta}) \sim \mathcal{N}(V_0^{-1/2}\boldsymbol{\delta}, I_J)$ and that

$$tr(I_J - V_0^{-1/2}\boldsymbol{\iota}_J\boldsymbol{\iota}_J'V_0^{-1/2}/tr(V_0^{-1})) = J-1$$

and thus z term follows a noncentral chi squared distribution with $(J-1)$ degrees of freedom and noncentrality parameter determined by the variance weighted sum of local location differences, i.e. $z \sim \chi^2(J-1, \tilde{\boldsymbol{\delta}}'M\tilde{\boldsymbol{\delta}})$ with $\tilde{\boldsymbol{\delta}} = V_0^{-1/2}\boldsymbol{\delta}$ and $M = I_J - V_0^{-1/2}\boldsymbol{\iota}_J\boldsymbol{\iota}_J'V_0^{-1/2}/tr(V_0^{-1})$. Now we can establish a sufficient condition for asymptotic risk dominance of the PCS. Note that since $z/(1+z) < 1$ almost surely, we have that

$$\begin{aligned} \rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{OLS}, \boldsymbol{\mu}) &< 5E\left[\frac{1}{1+z}\right] - 2(J-1)E\left[\frac{1}{1+z}\right] \\ &= -(2J-7)E\left[\frac{1}{1+z}\right] \end{aligned}$$

which implies that the risk difference of PCS to MA is given by

$$\begin{aligned} \rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{MA}, \boldsymbol{\mu}) &< -(2J-7)E\left[\frac{1}{1+z}\right] + (J-3)^2E\left[\frac{1}{z}\right] \\ &= [(J-3)^2 - (2J-7)]E\left[\frac{1}{z}\right] + (2J-7)E\left[\left(\frac{1}{z(1+z)}\right)\right] \end{aligned}$$

which for the canonical case $J = 4$ yields

$$\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{MA}, \boldsymbol{\mu}) < 0 + O((\tilde{\boldsymbol{\delta}}'M\tilde{\boldsymbol{\delta}})^{-2}).$$

This implies that for the canonical design, there is always a large enough local parameter value such that the PCS will dominate the MA estimator. Here, the local parameter should be interpreted relative to the square root of the asymptotic variances of the group means, i.e. increasing the differences relative to them will favor PCS over model averaging.

For the reverse case note that due to the positive semidefiniteness of M

$$\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{OLS}, \boldsymbol{\mu}) > -2(J-1)E\left[\frac{1}{z}\right]$$

Table A.1
Summary statistics of the Card and Krueger (1994) data.

Chain	NJ (treated)						PEN (control)					
	Before			After			Before			After		
	$\hat{\mu}$	σ^2	n	$\hat{\mu}$	σ^2	n	$\hat{\mu}$	σ^2	n	$\hat{\mu}$	σ^2	n
All	20.44	82.92	321	21.03	86.36	319	23.33	140.57	77	21.17	68.5	77
BK	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

Table A.2
Mean and difference-in-differences estimates with plug-in smoothing parameters.

All chains	OLS		PCS		RR		GRR		Kernel		MA		Mallows	
	NJ (t)	PEN (c)	NJ (t)	PEN (c)	NJ (t)	PEN (c)	NJ (t)	PEN (c)	NJ (t)	PEN (c)	NJ (t)	PEN (c)	NJ (t)	PEN (c)
B	20.44	23.33	20.53	22.87	20.61	22.43	20.58	22.77	20.58	22.27	20.58	22.79	20.78	23.33
A	21.03	21.17	21.01	21.12	21.10	21.32	21.20	21.45	21.07	21.19	21.03	21.13	20.78	20.78
DiD		2.75		2.22		1.60		1.93		1.56		2.11		2.55
Burger king														
B	22.16	29.42	22.25	29.06	22.40	28.39	22.26	28.56	22.39	28.38	22.27	29.07	22.90	29.42
A	23.63	26.22	23.63	26.06	23.76	26.01	23.82	25.75	23.77	26.00	23.65	26.07	22.90	26.22
DiD		4.67		4.38		3.74		4.37		3.75		4.37		3.20
KFC														
B	12.79	10.71	12.77	10.92	12.77	11.20	12.64	10.94	12.73	11.11	12.82	10.96	13.24	10.71
A	13.73	13.00	13.60	12.96	13.66	12.88	13.45	12.62	13.61	12.71	13.66	13.01	13.24	13.24
DiD		−1.35		−1.20		−0.79		−0.87		−0.71		−1.21		−2.53
Roys														
B	23.14	19.74	22.99	19.80	22.99	19.81	22.84	20.08	22.89	19.54	23.04	19.85	22.19	22.19
A	21.73	15.81	21.68	16.12	21.65	16.69	21.35	16.22	21.55	16.59	21.73	16.16	22.19	15.81
DiD		2.52		2.37		1.78		2.38		1.61		2.37		6.38
Wendys														
B	22.08	24.12	22.43	23.46	22.56	23.04	22.59	23.24	22.50	22.96	22.83	22.89	22.85	22.85
A	23.40	22.10	23.10	22.44	23.13	22.87	23.01	22.74	22.98	22.71	22.87	22.83	22.85	22.85
DiD		3.35		1.69		0.75		0.92		0.73		0.09		0.00

The table contains the mean estimates of the full-time employment equivalents and the difference-in-differences estimates for the Card and Krueger (1994) data. B = before the minimum wage increase, A = after the minimum wage increase, NJ = New Jersey, PEN = Pennsylvania, t = treated, c = control. DiD formula: $(\hat{\mu}_{NJ,A} - \hat{\mu}_{NJ,B}) - (\hat{\mu}_{PEN,A} - \hat{\mu}_{PEN,B})$.

and thus the risk difference to MA is bounded by

$$\rho(\hat{\boldsymbol{\mu}}^{PCS}(\hat{\boldsymbol{\omega}}), \boldsymbol{\mu}) - \rho(\hat{\boldsymbol{\mu}}^{MA}, \boldsymbol{\mu}) > [(J-3)^2 - 2(J-1)]E\left[\frac{1}{z}\right]$$
$$> (J^2 - 8J + 11)E\left[\frac{1}{z}\right]$$

which is greater than zero if $J \geq 7$.

A.9. Supplementary material for Section 6.1

MA and OLS tend to be less smooth and very close to each other while RR suggests substantially more smoothing. Kernel, GRR, and PCS are in between. Differences between the methods for the control group are less pronounced while reaching up to 11.51% on single periods for the treatment group. For the results using cross-validated smoothing parameters please consider Online Appendix B.5.

A.10. Supplementary material for Section 6.2

Table A.2 contains mean estimates of the full-time employment equivalents before and after the minimum wage increase for each state and the difference-in-differences estimates for: the OLS method, all shrinkage methods with plug-in smoothing parameters and the pretest estimator based on Mallows C_p as in Section 5. Among the shrinkage methods, PCS favors less biased estimates. Meanwhile RR, GRR, and Kernel shrink the effect of the minimum wage increase stronger away from the OLS estimate. The MA estimates are close to the PCS estimates except of the minimum wage effect for

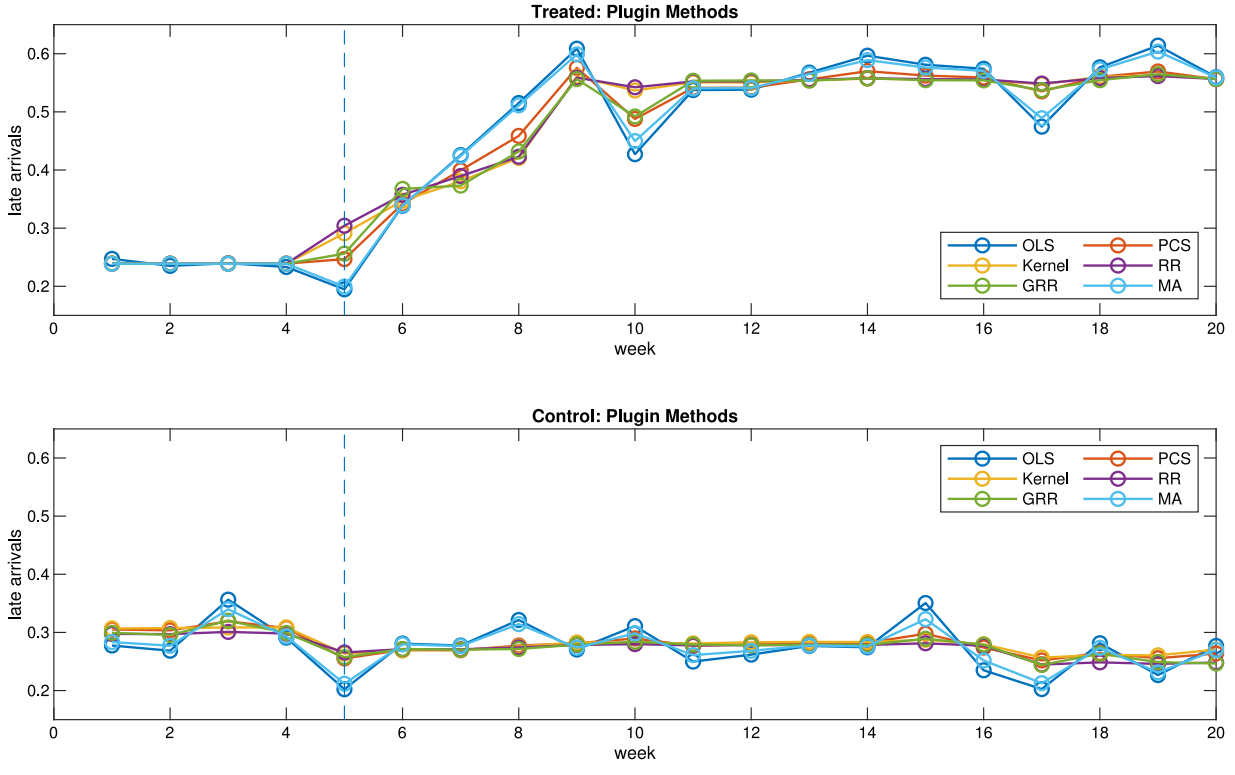

Fig. A.1. Mean share of late arrivals: All plugin methods.

Wendys which is shrunk to zero. The pretest estimator yields models with low number of parameters exhibiting the behavior discussed in Sections 4.1 and 5. For the results using cross-validated smoothing parameters please consider Online Appendix B.5.

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.07.051.

References

Aitchison, J., Aitken, C.G., 1976. Multivariate binary discrimination by the kernel method. Biometrika 63 (3), 413–420.
Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: 2nd International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973. pp. 267–281.
Buckland, S.T., Burnham, K.P., Augustin, N.H., 1997. Model selection: An integral part of inference. Biometrics 53 (2), 603–618.
Burnham, K.P., Anderson, D., 2003. Model selection and multimodel inference. In: A Practical Information-Theoretic Approach. Springer. Springer New York.
Card, D., Krueger, A.B., 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. Amer. Econ. Rev. 84 (4), 772–793.
Cheng, X., Liao, Z., Shi, R., 2019. On uniform asymptotic risk of averaging GMM estimators. Quant. Econ. 10 (3), 931–979.
Claeskens, G., Hjort, N.L., 2008. Model Selection and Model Averaging. In: Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
Cohen, A., 1966. All admissible linear estimates of the mean vector. Ann. Math. Stat. 37 (2), 458–463.
Gneezy, U., Meier, S., Rey-Biel, P., 2011. When and why incentives (don't) work to modify behavior. J. Econ. Perspect. 25 (4), 191–209.
Gneezy, U., Rustichini, A., 2000a. A fine is a price. J. Legal Stud. 29 (1), 1–17.
Gneezy, U., Rustichini, A., 2000b. Pay enough or don't pay at all. Quart. J. Econ. 115 (3), 791–810.
Hall, P., 1981. On nonparametric multivariate binary discrimination. Biometrika 68 (1), 287–294.
Hall, P., Li, Q., Racine, J.S., 2004. Cross-validation and the estimation of conditional probability densities. J. Amer. Statist. Assoc. 99 (468), 1015–1026.
Hall, P., Li, Q., Racine, J.S., 2007. Nonparametric estimation of regression functions in the presence of irrelevant regressors. Rev. Econ. Stat. 89 (4), 784–789.
Hansen, B.E., 2007. Least squares model averaging. Econometrica 75 (4), 1175–1189.
Hansen, B.E., 2014. Model averaging, asymptotic risk, and regressor groups. Quant. Econ. 5 (3), 495–530.
Hansen, B.E., 2016a. Efficient shrinkage in parametric models. J. Econometrics 190 (1), 115–132.
Hansen, B.E., 2016b. The risk of James–Stein and lasso shrinkage. Econometric Rev. 35 (8–10), 1456–1470.
Hansen, B.E., Racine, J.S., 2012. Jackknife model averaging. J. Econometrics 167 (1), 38–46.

Hjort, N.L., Claeskens, G., 2003. Frequentist model average estimators. J. Amer. Statist. Assoc. 98 (464), 879–899.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.
James, W., Stein, C., 1961. Estimation with quadratic loss. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. University of California Press, Berkeley, Calif., pp. 361–379.
Leeb, H., Pötscher, B.M., 2008. Sparse estimators and the oracle property, or the return of Hodges' estimator. J. Econometrics 142 (1), 201–211.
Li, K.-C., 1987. Asymptotic optimality for c_p, c_L, cross-validation and generalized cross-validation: Discrete index set. Ann. Statist. 15 (3), 958–975.
Li, Q., Racine, J., 2003. Nonparametric estimation of distributions with categorical and continuous data. J. Multivariate Anal. 86 (2), 266–292.
Liang, H., Zou, G., Wan, A.T., Zhang, X., 2011. Optimal weight choice for frequentist model average estimators. J. Amer. Statist. Assoc. 106 (495), 1053–1066.
Liu, C.-A., 2015. Distribution theory of the least squares averaging estimator. J. Econometrics 186 (1), 142–159.
Mallows, C.L., 1973. Some comments on C_p. Technometrics 15 (4), 661–675.
Oman, S.D., 1982. Shrinking towards subspaces in multiple linear regression. Technometrics 24 (4), 307–311.
Ouyang, D., Li, Q., Racine, J.S., 2009. Nonparametric estimation of regression functions with discrete regressors. Econometric Theory 25 (1), 1–42.
Raftery, A.E., Zheng, Y., 2003. Discussion: Performance of Bayesian model averaging. J. Amer. Statist. Assoc. 98 (464), 931–938.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6 (2), 461–464.
Simonoff, J.S., 1996. Smoothing Methods in Statistics. In: Springer Series in Statistics, Springer New York.
Stein, C., 1956. Efficient nonparametric testing and estimation. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. University of California Press.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. J. R. Statist. Soc.: Ser. B (Statist. Methodol.) 67 (1), 91–108.
Tutz, G., Oelker, M.-R., 2017. Modelling clustered heterogeneity: Fixed effects, random effects and mixtures. Internat. Statist. Rev. 85 (2), 204–227.
Zhang, X., Liang, H., 2011. Focused information criterion and model averaging for generalized additive partial linear models. Ann. Statist. 39 (1), 174–200.