Efficient estimation of regression models with user-specified parametric model for heteroskedasticty*

Saraswata Chaudhuri† and Eric Renault‡

This version: September 13, 2023.

Abstract

Several modern textbooks report that, thanks to the availability of heteroskedasticity robust standard errors, one observes the near-death of Weighted Least Squares (WLS) in cross-sectional applied work. We argue in this paper that it is actually possible to estimate regression parameters at least as precisely as Ordinary Least Squares (OLS) and WLS, even when using a misspecified parametric model for conditional heteroskedasticity. Our analysis is valid for a general regression framework (including Instrumental Variables and Nonlinear Regression) as long as the regression is defined by a conditional expectation condition. The key is to acknowledge, as first pointed out by Cragg (1992) that, when the user-specific heteroskedasticity model is misspecified, WLS has to be modified depending on a choice of some univariate target for estimation. Moreover, targeted WLS can be improved by properly combining moment equations for OLS and WLS respectively. Efficient GMM must be regularized to take into account the possible multicollinearity of estimating equations when errors terms are actually nearly homoscedastic.

JEL Classification: C12; C13; C21.

Keywords: asymptotic optimality; misspecification; nuisance parameters; weighted least squares

*We thank F. Bugni, J. Galbraith, S. Goncalves, J-M. Dufour, J. MacKinnon, P.C.B. Phillips, P. Rilstone, C. Rothe, P. Sant'Anna, A. Santos, R. Startz, Y. Shin, K. Xu, V. Zinde-Walsh, and various seminar and conference participants for useful suggestions. Older versions of the paper are available from the first author's webpage.

†Department of Economics, McGill University & Cireq, Montreal. Email: saraswata.chaudhuri@mcgill.ca.

‡Department of Economics, University of Warwick. Email: Eric.Renault@warwick.ac.uk.

1 Introduction

It has been widely accepted since the seminal work of White (1980) that, as bluntly announced by Stock and Watson (2011)'s popular textbook of econometrics, "despite the theoretical appeal of Weighted Least Squares (WLS), heteroskedasticity robust standard errors provide a better way to handle potential heteroskedasticity in most applications". Angrist and Pischke (2010) go as far as reporting the "near-death of generalized least squares in cross sectional applied work". The maintained point of view of our paper is that "White-washing" as coined by Leamer (2010) is misleadingly overlooking the need of "doing the hard work of modelling the heteroskedasticity (...) to determine if sensible reweighting of observations materially changes the locations of the estimator of interest" (and not only "the widths of the confidence intervals") (Leamer (2010)).

The "hard work of modelling the heteroskedasticity" by the practitioner is arguably the main reason why Eicker-White robust inference is so popular. The main contribution of this paper is to document that, irrespective of the user-specified parametric model for heteroskedasticity, we always have a well-defined optimal weighting strategy that dominates OLS. Up to regularity conditions, the only constraint on the specification of the parametric heteroskedasticity model is that it must contain conditional homoskedasticity as a particular case. In other words, even though we agree with Stock and Watson (2011) that the functional form of conditional variance is rarely ever known, it is not a sufficient reason to throw out the baby with the bath water and to overlook the improvements in accuracy brought by proper reweighting of the observations. Even though we could put forward a broader scope for validity (see Section 3), our analysis is valid with a general regression framework (including instrumental variables or nonlinear regressions) as soon as the regression is defined by a conditional expectation condition.

By referring to regression defined by a conditional expectation with a conditional heteroskedasticity model that nests the conditional homoskedasticity case, we remain true to the new research agenda of "resurrecting weighted least squares" as put forward by Romano and Wolf (2017).

Romano and Wolf (2017) propose an adaptive estimator such that one relies on WLS only if a test of conditional homoskedasticity based on the same conditional variance model is able to reject the null of homoskedasticity. By doing so, one will use WLS only when it is worthwhile doing so, while, for sake of finite-sample performance, avoiding (by using OLS) a noisy estimation of the variance function when homoskedasticity is not rejected. The resulting new estimator, dubbed adaptive least squares (ALS), can provide finite sample improvements over both OLS and WLS. However, the asymptotic variance of the ALS estimator can still exceed that of the OLS estimator when the regression error is conditionally heteroskedastic and the parametric model used to capture conditional variance is misspecified. This is because the ALS estimator is not designed for any kind of asymptotic optimality in such cases. DiCiccio et al. (2019) address this issue of lack of optimality by setting the focus on asymptotic variance of estimator of a given component of the parameter vector and choosing accordingly between OLS and WLS or some

optimal linear combination of these two.

In our paper, we wish to further advance the dissenting view of Romano and Wolf (2017) and others on the neglect of WLS by proposing a Targeted parametric WLS (TWLS) estimator that takes a direct route to optimality. The key idea is twofold:

On the one hand, we follow DiCiccio et al. (2019) to acknowledge the need of targeting a univariate function of the unknown parameters to define our optimality criterion.

On the other hand, by contrast with DiCiccio et al. (2019), we do not limit the search for optimality to a choice between OLS and WLS. Since there is in general no strong argument to assume that the parametric heteroskedasticity model is well-specified, observations reweighting should be given by an estimated skedastic function that is simply chosen for the sake of minimization of the asymptotic variance of the estimator of the target. The goal is accurate estimation of the target and not necessarily optimal approximation of the true skedastic function. It is worth noting that we are not the first to follow this route. In an important but unfortunately overlooked in the empirical literature paper, Cragg (1992) proposed to minimize over the heteroskedasticity parameters the trace or determinant of the asymptotic variance of the corresponding version of the weighted least squares estimator (different from standard WLS) of the regression parameters.

While Cragg (1992) does not discuss it, his minimization strategies of trace or determinant correspond respectively to the well-known notions of A and D optimality in the Design of Experiments literature. More generally, several concepts of optimality including E-optimality, L-optimality, etc. are compared in Kiefer (1974) who states a "General Equivalence Theory for Optimum Designs". Resorting to such targets in our context is a possibly appealing compromise due to the fact that, unless the parametric heteroskedasticity model is well-specified, there is no guarantee of existence of a minimized asymptotic variance matrix for the full vector of regression parameters. However, without that existence, for some of the regression coefficients, the standard errors from using Cragg (1992) may exceed that from using standard WLS, and it is evident in hindsight that A-optimality, D-optimality and other similar notions may not be attractive in empirical work. This is the reason why, while we remain true to Cragg's idea to optimize over heteroskedasticity parameters the estimation accuracy of a one-dimensional target, we argue that this target must be suggested by structural ideas (like estimation of a slope coefficient or others) rather than from the optimum designs literature.

It is worth noting that our concept of targeted estimation is germane with the strategy put forward by van der Laan and Rubin (2006) for "Targeted Maximum Likelihood Learning". While they stress that "the density estimator was targeted to be a good estimator of the density and might therefore result in a poor estimator of a particular smooth function of the density", we echo their remark in the case of WLS. While WLS was targeted to a be a good estimator of the vector of regression parameters in case of a well-specified parametric model of the skedastic function, it might result in a poor estimator of a particular smooth function of the regression parameters when the heteroskedasticity model is misspecified.

We address this issue of optimizing the asymptotic variance of the estimator of a scalar target in three ways that are, roughly speaking, in increasing order of efficiency but also of computational complexity. A common feature of all the estimators we consider is to refer to some given user-specified parametric model for heteroskedasticity, where the skedastic function is a known function of observations and of a vector $\gamma \in \Gamma \subset \mathbb{R}^{d_\gamma}$ of unknown parameters. To fix ideas without introducing too many notation at this point, for any generic parameter let us denote by $WLS(\gamma)$ generically the weighted least squares estimator computed with weights defined by the value γ of heteroskedasticity parameters. By this notation, the standard/naive WLS will be $WLS(\gamma_{WLS})$ where γ_{WLS} is the value of γ that makes the user-specified skedastic function the closest in terms of mean squared error to the true unknown skedastic function.

The first strategy is simply plugging in a Targeted WLS (TWLS) estimator, the term "Targeted" being used to stress the fact that the choice of weights is target-driven by contrast with a naive use of WLS that is based on matching with the observed heteroskedasticity.

The second strategy improves upon plug in estimators by considering all convex (or more generally affine) combinations of plug in OLS and plug in alternative estimators $WLS(\gamma)$. The idea of convex combination (CC) of estimators has already been pushed forward by DiCiccio et al. (2019) who characterize the most efficient CC of OLS and the naive WLS estimators. By contrast, as in the case of TWLS, here also we do not limit ourselves to naive WLS. Our preferred TCC (Targeted CC) is obtained as a result of a double optimization of asymptotic variance detailed in Section 4. The key idea is first, for given $WLS(\gamma)$, to define an optimal convex combination with OLS, and second to choose optimally the heteroskedasticity parameters γ entering this CC. By definition, our estimator must be not only at least as accurate in terms of asymptotic variance as OLS (since OLS is one of the two inputs of the convex combination) but also at least as accurate as naive WLS and also DiCiccio et al. (2019)'s CC estimator (since those estimators correspond to one particular value γ_{WLS}).

It is worth acknowledging that, even though we improve upon DiCiccio et al. (2019) by considering as one of the two inputs of the convex combination any estimator $WLS(\gamma)$ (and not only naive $WLS(\gamma_{WLS})$), our estimator may still be suboptimal by remaining true to OLS as the second input of the combinaion. The conditional moment restrictions that define the regression function allow one to define nonparametrically optimal instruments that may be more closely proxied by convex combinations of estimators $WLS(\gamma^1)$ and $WLS(\gamma^2)$ which do not include OLS. In this respect, this paper sets the focus on an approach that may be suboptimal but always user-friendly since it is a direct correction to OLS.

Finally, following Chen et al. (2016), we also consider matrix extensions of convex/affine combinations of estimators by considering matrix weights that sum to identity matrix.

Revisiting one of the main results of Chen et al. (2016) (see their Proposition 1, page 49), we show that efficient matrix combinations is tantamount to efficient use by GMM of the complete set of valid moment conditions at hands. As for previous estimators TWLS and TCC, the parameters γ for our $WLS(\gamma)$ to combine with OLS through GMM are chosen by minimization of the asymptotic variance of the GMM

estimator of the target. Defined in this way, the efficient use of the complete set of moment conditions delivers what we call TGMM (Targeted GMM) estimator. Our TGMM is shown to be asymptotically the most efficient estimator but is flawed by some finite sample issues that may warrant a maintained interest in TCC. In particular, our TGMM can be seen as a corrected version (and with modifications delivering additional improvements) of an estimator recently proposed by Lu and Wooldridge (2020). We put forward a necessary regularization to take into account the possibility of near-multicollinearity between moment conditions when the data generating process is only slightly heteroskedastic.

Our three categories of estimators do not encompass the methodologies devised in papers like Gourieroux et al. (1996), Spady and Stouli (2019), Papadopoulosa and Tsionas (2021) and others who impose some additional structure, like additional moment conditions, to propose alternative efficient estimators in presence of heteroskedasticity of unknown form.

Our paper is organized as follows.

We define in Section 2 the concept of targeted estimation in the context of a regression model. We make explicit the need of setting the focus on a given target parameter that we take as a smooth scalar function of the regression coefficients. We characterize our TWLS, its feasible version and its asymptotic variance. While the general simulation evidence and empirical illustrations for all estimators proposed in this paper and their main contenders are presented in Appendix A, we present in Section 2 a small scale Monte Carlo work to compare the finite sample performance of our TWLS with Cragg's D-optimal estimator. We run this comparison in the context of DGP 2 of Romano and Wolf (2017) while the user-specified heteroskedasticity model is given by a standard exponential affine skedastic function.

Section 3 follows Chen et al. (2016)'s general idea of estimating the parameters from just-identifying subsets of the available moment conditions and then combining the resulting estimators in a linear fashion. However, we address more thoroughly the comparison between affine combinations (with scalar coefficients) and combinations with matrix coefficients. The analysis is provided in a general context of moment estimation and illustrated with the case of weighted least square estimators.

We characterize in Section 4 our TCC estimator, its feasible version and its asymptotic variance. In the same Monte Carlo setting as in Section 2, we compare the finite sample performance of our TCC with two alternative CC estimators: ALS (Adaptive Least Squares) of Romano and Wolf (2017) and the MIN estimator proposed by DiCiccio et al. (2019). We also compare with two machine learning estimators, Support Vector Regression (SVR) and LASSO, that have been recently studied with the same DGP 2 of Romano and Wolf (2017) by Gonzales Coya and Perron (2022).

Besides the general comparison between TCC and TGMM provided in Section 3, we show in Section 5 that the framework of efficient GMM is convenient to provide a practical way to get our TGMM estimator. It also allows us to address several issues that appear to be important in practice (see also Lu and Wooldridge (2020)). The efficient weighting matrix for GMM must be regularized to avoid asymptotic explosion in case of near-homoskedasticity, that is due to near-singularity of the variance matrix of the

moment conditions. In addition, the lack of convexity of the objective function is addressed by bounding the set of possible values of heteroskedastic parameters. We keep these parameters in a convenient ball around the naive value $\bar{\gamma}$ estimated by standard WLS. This provides a useful hedge against non-convexity and in particular the perverse behaviour of weights that, as inversely related to conditional variance, may be pushed to infinity for a spurious minimization of sum of weighted squared residuals.

Section 6 concludes. Appendix A provides an extensive numerical study of all the estimators using the simulation designs and empirical examples from Romano and Wolf (2017) and Lu and Wooldridge (2020). Technical discussions and proofs of results are collected in Appendix B.

2 Targeted estimation

2.1 The basic regression model

The linear model is of the form:

$$y_i = x_i'\beta + u_i(\beta), i = 1, ..., n \tag{1}$$

where $\beta \in \mathbb{R}^p$ is the vector of unknown parameters, the observed sample $(y_i, x_i')_{i=1}^n$ is independent and identically distributed, and the p explanatory variables components of x_i are not redundant:

$$E[x_i x_i'] \quad \text{is nonsingular.}$$

As a result, the true unknown value β^0 of regression parameters is well-defined by the conditional expectation condition:

$$E[u_i(\beta^0)|x_i] = 0. \tag{2}$$

Remark 1: The asymptotic theory devised in this paper could be easily extended to the case of serial dependence in the sequence $u_i(\beta^0) = u_i, i = 1, \ldots, n$. However, this possible extension is not very useful since, in the case of serial dependence, it would be worthwhile to study not only the optimal re-weighting of observations (as in this paper) but also to discuss improvements of estimators that take serial correlation into account which is beyond the scope of this paper.

Remark 2: The various strategies of improvement of estimators accuracy based on a user-specified parametric model for heteroskesdasticity apply not only to the linear regression model:

$$E[y_i - x_i'\beta|x_i] = 0$$

but also to the nonlinear regression model:

$$E[y_i - m(x_i, \beta)|x_i] = 0$$

where $m(.,.)$ is a known scalar function and to the IV regression model:

$$E[y_i - x_i'\beta \,|z_i] = 0$$

where z_i is a vector of instrumental variables able to identify the IV regression parameters β. Of course, in this latter case, conditional heteroskedasticity of the error term $u_i\left(\beta^0\right) = y_i - x_i'\beta^0$ must be understood given z_i. We will consider in Section 3 an even more general framework of averaging a number of moment conditions estimators as in Chen et al. (2016).

While setting the focus on the regression model (1), we define the true skedastic function as the non-random function:

$$E\left[u_i^2(\beta^0)|x_i\right] = \omega_0^2(x_i) > 0.$$

On the other hand, the user's model of the true skedastic function $\omega_0^2(x_i)$ is given by a parametric family:

$$\omega^2(x_i, \gamma) > 0, \gamma \in \Gamma \subset \mathbb{R}^{d_\gamma}. \tag{3}$$

As is standard, we will always assume that this family nests the case of conditional homoskedasticity; e.g., the parameter γ satisfies this condition if all entries except for the first one are equal to zero, i.e.,

$$\text{there exist } \gamma^{\text{hom}} = \left(\gamma_1^{\text{hom}}, 0, ..., 0\right)' \in \Gamma \text{ such that } \omega^2\left(x_i, \gamma^{\text{hom}}\right) \equiv \omega_{\text{hom}}^2.$$

Examples: Following the extant literature, Romano and Wolf (2017) point out the three parametric families below (Ex 1, 2 and 3) that can be seen as the user's model of conditional heteroskedasticity. It is important to keep in mind that none of these families is assumed to contain the true Data Generating Process (DGP). Let $x_i = (x_{i,1}, \ldots, x_{i,p})'$ with $x_{i,1} = 1$.

$$\textbf{Ex1:} \quad \omega^2(x_i, \gamma) = \exp\left(\gamma_1 + \sum_{j=2}^{p} \gamma_j \log\left(|x_{i,j}|\right)\right),$$

$$\textbf{Ex2:} \quad \omega^2(x_i, \gamma) = \gamma_1 + \sum_{j=2}^{p} \gamma_j |x_{i,j}|,$$

$$\textbf{Ex3:} \quad \omega^2(x_i, \gamma) = \exp\left(\gamma_1 + \sum_{j=2}^{p} \gamma_j x_{i,j}\right).$$

2.2 Weighted Least Squares

For any $\gamma \in \Gamma$ we define a weighted-by-$\omega^2(x_i, \gamma)$ estimator of β as:

$$\hat{\beta}(\gamma) = \left(\sum_{i=1}^{n} \frac{x_i x_i'}{\omega^2(x_i, \gamma)}\right)^{-1} \sum_{i=1}^{n} \frac{x_i y_i}{\omega^2(x_i, \gamma)}. \tag{4}$$

Under standard regularity conditions, $\sqrt{n}\left(\hat{\beta}(\gamma)-\beta^0\right)$ is asymptotically normal with asymptotic variance matrix $\Sigma_\beta(\gamma)$ given by the sandwich formula:

$$\Sigma\left(\beta^0,\gamma\right)=\left(E\left[\frac{x_ix_i'}{\omega^2(x_i,\gamma)}\right]\right)^{-1}E\left[\frac{x_ix_i'\omega_0^2(x_i)}{\omega^4(x_i,\gamma)}\right]\left(E\left[\frac{x_ix_i'}{\omega^2(x_i,\gamma)}\right]\right)^{-1}. \tag{5}$$

Remark: The parametric model (3) for conditional heteroskedasticity is defined as well-specified if and only if for some parameter value $\gamma^0\in\Gamma$:

$$\omega^2\left(x_i,\gamma^0\right)=\omega_0^2(x_i). \tag{6}$$

In this case:

$$\Sigma\left(\beta^0,\gamma^0\right)=\left(E\left[\frac{x_ix_i'}{\omega^2(x_i,\gamma^0)}\right]\right)^{-1}\ll\Sigma\left(\beta^0,\gamma\right),\forall\gamma\in\Gamma$$

with the notation $A\ll B$ for meaning "$B-A$ positive semi-definite". Note that it is the case in particular if $u_i\left(\beta^0\right)$ is conditionally homoskedastic. In this case, $\gamma^0=\gamma^{\text{hom}}$ and $\omega^2\left(x_i,\gamma^0\right)=\omega_{\text{hom}}^2=\omega_0^2(x_i)$.

As already explained, our focus of interest is precisely the case when there is some conditional heteroskedaticity and the parametric model for heteroskedasticity is misspecified, meaning that we cannot define a true unknown value γ^0 as the solution of (6). We can only define a pseudo-true value, denoted by γ_{WLS}, as solution of:

$$\gamma_{WLS}=\arg\min_{\gamma\in\Gamma}E\left[\left(\omega_0^2(x_i)-\omega^2(x_i,\gamma)\right)^2\right]=\arg\min_{\gamma\in\Gamma}E\left[\left(u_i^2\left(\beta^0\right)-\omega^2(x_i,\gamma)\right)^2\right]. \tag{7}$$

The notation is motivated by the fact that the standard definition of (infeasible) WLS (see, e.g., Romano and Wolf (2017)) corresponds to $\hat{\beta}(\gamma_{WLS})$ while a feasible version is given by $\hat{\beta}(\hat{\gamma}_{WLS})$ for any consistent estimator $\hat{\gamma}_{WLS}$ of γ_{WLS}, for instance obtained by a sample counterpart of the minimization program (7), with $u_i\left(\beta^0\right)$ replaced by the OLS residual $\hat{u}_i$.

It is however worth keeping in mind than when the first minimization in (7) does not provide a value function equal to zero (that is when the heteroskedasticity model is misspecified), there is no strong argument in favor of the choice of the value γ_{WLS} of γ. Ideally, one would like to define an optimal value γ^* such that:

$$\Sigma\left(\beta^0,\gamma^*\right)\leq\Sigma\left(\beta^0,\gamma\right),\forall\gamma\in\Gamma. \tag{8}$$

However, except in the well-specified case (then $\gamma^*=\gamma^0$), such an optimal value γ^* does not exist in general, and there is no compelling argument to consider that γ_{WLS} is "closer to optimality" than contenders. The only way to escape this deadlock will be to replace the matrix-minimization problem (8) by a program of minimization of a scalar function. This will be the purpose of the next subsection.

Remark: We keep the terminology WLS for the estimator $\hat{\beta}(\gamma_{WLS})$. An estimator $\hat{\beta}(\gamma)$ for some user-

specified value of γ, leading to weights $\omega^2(x_i, \gamma)$ in (4), will be dubbed "User-specified WLS" (UWLS) and, when computed with our preferred value of γ (see below) "Targeted WLS" (TWLS).

2.3 Targeted Regression

As explained above, we need to set the focus on estimation of a parameter of interest that is a known and univariate smooth function of β:

$$h = h(\beta), \quad \beta \in \mathbb{R}^p.$$

We consider substitution estimators based on UWLS estimators $\widehat{\beta}(\gamma)$ of β:

$$\widehat{h}(\gamma) = h(\widehat{\beta}(\gamma)).$$

All these estimators are consistent for $h(\beta^0)$ and their asymptotic variance is given by:

$$\sigma_h^2(\beta^0, \gamma) = \delta'\left(\beta^0\right)\Sigma(\beta^0, \gamma)\delta\left(\beta^0\right), \quad \delta(\beta) = \frac{\partial h(\beta)}{\partial \beta}. \tag{9}$$

As pointed out by Cragg (1992), the estimator WLS (or a feasible version based on $\widehat{\gamma}_{WLS}$) "can lead to larger diagonal elements of $\Sigma(\beta^0, \gamma)$ than those of OLS". In other words, defining as "target of interest" a specific component of the vector β (i.e., with a vector δ having all its components equal to 0 except one equal to 1), it is not optimal to use the supposedly optimal estimator $h(\widehat{\beta}(\widehat{\gamma}_{WLS}))$. This direct plugging in of the WLS estimator $\widehat{\gamma}_{WLS}$ is actually theoretically inferior to other more targeted estimators of the parameter of interest. While this concept of targeted estimator had been put forward by van der Laan and Rubin (2006) in the context of maximum likelihood, because "substitution estimators will often fail to be asymptotically efficient due to the bias caused by the curse of dimensionality", misspecification of the user-specified model of the skedastic function also points out the need of considering alternative substitution estimators. For example, Cao et al. (2009) take a similar strategy in the context of doubly robust estimation under the assumption that one set of nuisance parameters is correctly specified.

Surprisingly, while as quoted above, Cragg (1992) was concerned by estimation of specific components of β, he did not choose to minimize the asymptotic variance of a specific component but rather an aggregate of the asymptotic variances of all the components of β, aggregate that could be either the determinant or the trace of the variance matrix $\Sigma(\beta^0, \gamma)$.

More generally, our TWLS for a target $h(\beta)$ will be defined as a feasible version of $h(\widehat{\beta}(\gamma_h^*))$ with:

$$\gamma_h^* = \arg\min_{\gamma \in \Gamma} \sigma_h^2(\beta^0, \gamma).$$

Remark: Formulas (5) and (9) clearly show that TWLS is seeking for an optimal tradeoff between two objective functions. On the one hand, the first objective function (similar in spirit to WLS, albeit with a

different loss function) tries to choose γ in order that $\omega^2(x_i;\gamma)$ tracks as accurately as possible the true skedastic function $\omega_0^2(x_i)$. This first objective function does not depend on the target, as summarized by δ. On the other hand, a second objective function heavily depends on the target δ through a norm of the vector $E\left[\frac{x_i x_i'}{\omega^2(x_i;\gamma)}\right]^{-1}\delta$. While this second objective function does not play any role in the case of a well-specified heteroskedasticity model ($\omega^2(x_i;\gamma^*)=\omega_0^2(x_i)$), it is more and more at stake when the heteroskedasticity model becomes more misspecified.

2.4 Feasible TWLS

The first task is to estimate for any $(\beta,\gamma)\in\mathbb{R}^p\times\Gamma$ the asymptotic variance matrix:

$$\Sigma(\beta^0,\gamma)=B_2^{-1}(\gamma)V_{22}(\beta^0,\gamma)B_2^{-1}(\gamma)$$

where:

$$B_2(\gamma)=E\left[\frac{x_i x_i'}{\omega^2(x_i;\gamma)}\right]\quad\text{and}\quad V_{22}(\beta^0,\gamma)=E\left[\frac{x_i x_i'\omega_0^2(x_i)}{\omega^4(x_i;\gamma)}\right].$$

The notations $B_2(\gamma)$ and $V_{22}(\beta^0,\gamma)$ have been chosen to be consistent with the definition by blocks of the following two bigger matrices defined in Section 4:

$$B(\gamma)=[B_1(\gamma),B_1(\gamma)]\quad\text{and}\quad V(\beta^0,\gamma)=\left[\begin{array}{cc}V_{11}(\beta^0,\gamma) & V_{12}(\beta^0,\gamma)\\ V_{21}(\beta^0,\gamma) & V_{22}(\beta^0,\gamma)\end{array}\right].$$

Under standard regularity conditions, we have consistent estimators (uniformly in $\gamma\in\Gamma$):

$$\begin{aligned}\widehat{B}_{2,n}(\gamma) &= \frac{1}{n}\sum_{i=1}^n\frac{1}{\omega^2(x_i;\gamma)}x_i x_i'\xrightarrow{P}B_2(\gamma)\\ \widehat{V}_{22,n}(\gamma) &= \frac{1}{n}\sum_{i=1}^n\frac{(y_i-x_i'\widehat{\beta}_{OLS})^2}{\omega^4(x_i;\gamma)}x_i x_i'\xrightarrow{P}V_{22}(\beta^0,\gamma).\end{aligned}$$

Note that, while the consistency of the estimator $\widehat{B}_{2,n}(\gamma)$ is simply implied by the uniform law of large numbers, the consistency of $\widehat{V}_{22,n}(\gamma)$ can be seen as a consequence of (uniform) consistency of an Eicker-White estimator when the vector x_i of explanatory variables is replaced by the pseudo-sphericized one $\widetilde{x}_i(\gamma)$:

$$\widetilde{x}_i(\gamma)=\frac{1}{\omega^2(x_i,\gamma)}x_i\Longrightarrow\widehat{V}_{22,n}(\gamma)=\frac{1}{n}\sum_{i=1}^n(y_i-x_i'\widehat{\beta}_{OLS})^2\widetilde{x}_i(\gamma)\widetilde{x}_i'(\gamma).$$

We can then define:

$$\widehat{\sigma}_h^2(\beta,\gamma)=\frac{\partial h(\beta)}{\partial\beta'}[\widehat{B}_{2,n}(\gamma)]^{-1}\widehat{V}_{22,n}(\gamma)[\widehat{B}_{2,n}(\gamma)]^{-1}\frac{\partial h(\beta)}{\partial\beta}$$

and a consistent estimator of the targeted γ:

$$\hat{\gamma}_{h,TWLS} = \arg\min_{\gamma \in \Gamma} \widehat{\sigma}_h^2(\widehat{\beta}_{OLS}, \gamma).$$

Thus, we obtain our feasible TWLS estimator:

$$\widehat{h}_{TWLS} \equiv h(\widehat{\beta}(\widehat{\gamma}_{h,TWLS}))$$

and its estimated standard error:

$$se_{h,TWLS,n} = \left[\frac{1}{n}\widehat{\sigma}_h^2(\widehat{\beta}_{OLS}, \hat{\gamma}_{h,TWLS})\right]^{1/2}.$$

Remarks: Two important remarks are in order.

First, the calculation of the feasible TWLS implies that we plug in the estimator of γ_h^*, denoted by $\widehat{\gamma}_{h,TWLS}$. One might be afraid that this additional source of randomness modifies the asymptotic variance of our estimator, leading to moving target for minimization. However, as explained in more details in Section 3, the fact that our regression equation (1)/(2) is defined by a conditional expectation (and not just an L^2 projection) ensures that the asymptotic variance of interest, $\sigma_h^2(\beta^0, \gamma_h^*)$, is not modified when plugging in a consistent estimator of γ_h^* for performing our targeted WLS.

Second, one might consider plugging in $\widehat{\beta}(\widehat{\gamma}_{h,TWLS})$ instead of $\widehat{\beta}_{OLS}$ to compute the estimated standard error $se_{h,TWLS,n}$. By doing so, one may even imagine a kind of iterated TWLS. However, it is worth keeping in mind that there is no reason to believe that $\widehat{\beta}(\widehat{\gamma}_{h,TWLS})$ is a better estimator than $\widehat{\beta}_{OLS}$ to estimate functions of β that are not the target $h(\beta)$.

2.5 The importance of targeting

In accordance to the latter remark above, it is worth keeping in mind that, by definition, the optimal weights for TWLS depend upon the target of interest. If one is interested in the complete set of regression parameters β, one must indeed estimate each of them with different optimal weights. The result of the regression would then come as follows:

$$y_i = \underset{(se_{\beta_1,TWLS,n})}{\widehat{\beta}_1} + \underset{(se_{\beta_2,TWLS,n})}{\widehat{\beta}_2} x_{i,2} + \ldots + \underset{(se_{\beta_p,TWLS,n})}{\widehat{\beta}_p} x_{i,p} + \widehat{u}_{i,n}$$

Of course, this way of presenting regression results overlooks the cross-correlation between estimators of different components of β. But there is nothing new or restrictive in this respect by comparison with standard practice of separate check of the Student t-values of each regression parameter. Of course, one

might also be interested in some out of sample prediction:

$$h(\beta) = \beta_1 + \beta_2 x_{n+1,2} + \ldots + \beta_p x_{n+1,p}.$$

But, then, it defines a new target for estimation that must itself be treated as a target according to the TWLS approach. Obviously, in this case, the alleged possible non-zero correlations between estimators of different components of β will play a role.

We now provide a simple numerical example (inspired by Romano and Wolf (2017)) to display the quantitative impact of targeting.

Example: Regression model: $y_i = \beta_1 + \beta_2 x_i + u_i$ where (x_i, u_i) is i.i.d. with $x_i \sim$ Uniform$(1, 4)$ and $u_i|x_i \sim N(0, \omega_0^2(x_i))$ for $i = 1, 2, \ldots, n$. We consider three possible DGPs in increasing order of magnitude of conditional heteroskedasticity: $\omega_0^2(x_i) = (\log(x_i))^{2j}$ for $j = 1, 2, 3$, while the user-specified heteroskedasticity model is misspecified and given by the popular exponential affine form: $\omega^2(x_i, \gamma) = \exp(\gamma_1 + \gamma_2 x_i)$. We compare by Monte Carlo (across 10000 trials) the average standard errors (SE) of three estimators of the intercept parameter β_1 and the slope parameter β_2:

- the WLS estimator (as a benchmark),
- the D-optimal estimator of Cragg (1992) (ratio of its SE to the one of WLS),
- our TWLS estimator for β_1 and β_2 (ratio of its SE to the one of WLS).

The results for sample sizes $n = 50, 100, 200, 400$ respectively are given in Table 1 below.

n	β_1						β_2					
	$\omega_0^2(x) = (\log(x))^2$		$\omega_0^2(x) = (\log(x))^4$		$\omega_0^2(x) = (\log(x))^6$		$\omega_0^2(x) = (\log(x))^2$		$\omega_0^2(x) = (\log(x))^4$		$\omega_0^2(x) = (\log(x))^6$	
	Cragg	TWLS	Cragg	TWLS	Cragg	TWLS	Cragg	TWLS	Cragg	TWLS	Cragg	TWLS
50	.924	.906	.935	.843	.773	.519	1.032	.975	1.076	.921	.826	.594
100	.909	.899	.856	.798	.910	.490	1.017	.980	.961	.914	.933	.566
200	.950	.934	1.008	.783	1.270	.476	1.000	.991	1.006	.909	1.201	.542
400	1.009	.930	1.272	.784	1.767	.493	1.004	.992	1.151	.917	1.533	.563

Table 1: Ratios of the standard errors of Cragg's estimator and TWLS with respect to that of WLS. A ratio smaller than one implies the concerned estimator is more precise than WLS. Similar ratios of the means squared errors have similar values.

Then, three main conclusions are in order.

First, although Cragg's estimator is targeted, the fact that its target is the determinant of the variance matrix of the estimator of $\beta = (\beta_1, \beta_2)$ leads to estimators of the intercept and the slope that are actually worse than the un-targeted WLS in almost half (5/12) of the cases for both β_1 and β_2 and in three quarter (9/12) of the cases for β_2 alone. Moreover, it is worrying to notice that it is for the larger sample sizes ($n = 200$ or 400) that Cragg's relative performance with respect to WLS deteriorates a lot when conditional heteroskedasticity is made more severe.

Second, the results for TWLS are much more encouraging. It always performs better than both Cragg and WLS. Moreover, it is compelling to notice that the comparative advantage of TWLS with respect to WLS increases significantly when conditional heteroskedasticity is made more severe.

Third, for moderate heteroskedasticity, WLS does a job almost as good as TWLS. There is not much room for improvement on WLS in this case. An interpretation of that result is to remember that the value γ_{WLS} of the parameter γ used for WLS is actually such that $\omega^2(x, \gamma_{WLS})$ is the best possible approximation of the true skedastic function. It is not surprising to find that for smooth enough variations of the skedastic function, that function can be satisfactorily tracked by the misspecified exponential affine heteroskedasticity model.

3 Combining Estimators

As discussed in Section 2, for any user-specified value γ of heteroskedasticity parameters, we can define a UWLS estimator $\hat{\beta}(\gamma)$. This estimator can be interpreted as a GMM estimator provided by the following just-identified set of moment conditions:

$$E\left[\frac{x_i}{\omega^2(x_i, \gamma)}(y_i - x_i'\beta)\right] = 0.$$

$\hat{\beta}(\gamma)$ is a consistent estimator of β^0 with asymptotic variance $\Sigma\left(\beta^0, \gamma\right)$. Beyond the TWLS estimator $h\left(\hat{\beta}\left(\gamma_h^*\right)\right)$ defined in Section 2, it may make sense, for the sake of asymptotic variance minimization, to build new estimators by convex combinations (CC) of plug-in UWLS estimators $h\left(\hat{\beta}(\gamma)\right)$ for different values of $\gamma \in \Gamma$. As announced in the Introduction, we will consider only CC of two such estimators. We first discuss such CC of estimators in the general setting of just-identified sets of moment conditions.

3.1 A general framework

We consider two sets of just-identified moment conditions that both identify the true unknown value β^0 of a p dimensional parameter vector β.

- A first set of p moments conditions identifies β^0:

$$E[g_1(y_i, x_i; \beta)] = 0 \iff \beta = \beta^0.$$

  This first set of moments may for instance be orthogonality conditions for OLS, UWLS, two stage least squares (2SLS) or nonlinear least squares (NLLS). In the UWLS case, for some given value γ^1 of the heteroskedasticity parameters:

$$g_1(y_i, x_i; \beta) = \frac{x_i}{\omega^2(x_i, \gamma^1)}(y_i - x_i'\beta). \tag{10}$$

- A second set of p moments conditions also identifies β^0:

$$E[g_2(y_i, x_i; \beta)] = 0 \iff \beta = \beta^0.$$

This second set may for instance re-weight differently orthogonality conditions through another value γ^2 of the heteroskedasticity parameters:

$$g_2(y_i, x_i; \beta) = \frac{x_i}{\omega^2(x_i, \gamma^2)}(y_i - x_i'\beta) \tag{11}$$

Since the two sets of moment conditions are just-identified, each of them defines without ambiguity a GMM estimator as follows:

$$\widehat{\beta}^{(j)} = \arg\min_{\beta} \|\bar{g}_{j,n}(\beta)\|, \quad j = 1, 2$$

where:

$$\bar{g}_{j,n}(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_j(y_i, x_i; \beta).$$

We maintain throughout the standard assumptions for the asymptotic theory of GMM. In particular, since the two sets of moment conditions are just-identified, we are led to assume that the two Jacobian matrices:

$$G_j = G_j(\beta^0) \quad \text{where} \quad G_j(\beta) = E\left[\frac{\partial}{\partial \beta'} g_j(\beta)\right], \quad j = 1, 2$$

are non-singular matrices. As a consequence, we have asymptotically a one-to-one mapping between the GMM estimators and the corresponding sample moments:

$$\sqrt{n}(\widehat{\beta}^{(j)} - \beta^0) = -\left[G_j\right]^{-1}\sqrt{n}\bar{g}_{j,n}(\beta^0) + o_p(1). \tag{12}$$

In all this section, we will run affine regressions based on the joint asymptotic normal distribution of $\left[\sqrt{n}(\widehat{\beta}^{(j)} - \beta^0)\right]_{1 \le j \le 2}$ implied by the central-limit theorem:

$$\begin{bmatrix} \sqrt{n}\bar{g}_{1,n}(\beta^0) \\ \sqrt{n}\bar{g}_{2,n}(\beta^0) \end{bmatrix} \xrightarrow{d} \aleph\left(0, \Omega = \Omega(\beta^0)\right) \quad \text{where} \quad \Omega(\beta) = \begin{bmatrix} \Omega_{11}(\beta) & \Omega_{12}(\beta) \\ \Omega_{21}(\beta) & \Omega_{22}(\beta) \end{bmatrix}.$$

3.2 Convex combinations of estimators

As already announced, we dub CC estimators all estimators which, extending an initial proposal of DiCiccio et al. (2019), are based on a convex combination (CC) of the two GMM estimators and thus can be written as:

$$\widehat{h}_\lambda = (1 - \lambda)h(\widehat{\beta}^{(1)}) + \lambda h(\widehat{\beta}^{(2)})$$

for some $\lambda \in \mathbb{R}$. Note that we do not introduce any sign constraint on the scalar weight λ, so that the terminology "convex combination" is an abuse of language and we should rather say "affine combination". Asymptotically:

$$\begin{aligned} \sqrt{n}\left(\widehat{h}_\lambda - h(\beta^0)\right) &= \sqrt{n}\left(h(\widehat{\beta}^{(1)}) - h(\beta^0)\right) - \lambda\sqrt{n}\left(h(\widehat{\beta}^{(1)}) - h(\widehat{\beta}^{(2)})\right) \\ &= \sqrt{n}\delta'(\widehat{\beta}^{(1)} - \beta^0) - \lambda\sqrt{n}\delta'(\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}) + o_p(1) \end{aligned} \tag{13}$$

where:

$$\delta = \delta\left(\beta^0\right) \quad \text{and} \quad \delta(\beta) = \frac{\partial}{\partial\beta} h\left(\beta\right).$$

Hence, we minimize the asymptotic variance of $\widehat{h}_\lambda$ by choosing $\lambda = \lambda^*(h)$ that is the asymptotic regression coefficient in the regression of $\sqrt{n}\delta'(\widehat{\beta}^{(1)} - \beta^0)$ on $\sqrt{n}\delta'(\widehat{\beta}_n^{(1)} - \widehat{\beta}_n^{(2)})$:

$$\lambda^*(h) = \lim_{n\to\infty} \frac{\delta' Cov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\delta}{\delta' Var\left(\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\delta}.$$

The optimal CC estimator of the target $h(\beta)$ is thus:

$$\widehat{h}_{\lambda^*(\delta)} = \left[1 - \lambda^*(h)\right] h(\widehat{\beta}^{(1)}) + \lambda^*(h))h(\widehat{\beta}^{(2)}).$$

It leads us to our first result.

Proposition 1 *The optimal TCC (targeted CC) estimator of $h(\beta)$ based on the couple of estimators $\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}\right)$ is given by:*

$$\widehat{h}_{\lambda^*(h)} = \left[1 - \lambda^*(h)\right] h(\widehat{\beta}^{(1)}) + \lambda^*(h)h(\widehat{\beta}^{(2)})$$

with:

$$\lambda^*(h) = \lim_{n\to\infty} \frac{\delta' Cov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\delta}{\delta' Var\left(\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\delta}.$$

Remark 1: The asymptotic expansion in (13) shows that we can also interpret our CC estimators as follows:

$$\begin{aligned} \sqrt{n}\left(\widehat{h}_\lambda - h(\beta^0)\right) &= \sqrt{n}\delta'\left[(1-\lambda)\widehat{\beta}^{(1)} + \lambda\widehat{\beta}^{(2)} - \beta^0\right] + o_p(1) \\ &= \sqrt{n}\left\{h\left((1-\lambda)\widehat{\beta}^{(1)} + \lambda\widehat{\beta}^{(2)}\right) - h\left(\beta^0\right)\right\} + o_p(1). \end{aligned}$$

In other words, the CC estimator can also be interpreted as a plug-in estimator where it is a convex combination $\left[(1-\lambda)\widehat{\beta}^{(1)} + \lambda\widehat{\beta}^{(2)}\right]$ of the two estimators of β that is plugged in.

Remark 2: The proof of Proposition 1 (see Appendix B) shows that the optimal TCC estimator based

on $\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}\right)$ is $h\left(\widehat{\beta}^{(1)}\right)$, for all possible target $h(\beta)$, if and only if:

$$Cov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right) = 0$$

that is, by virtue of (12) and with obvious simplified notations:

$$Cov\left(g_1, G_1^{-1} g_1 - G_2^{-1} g_2\right) = 0.$$

Elementary calculation (see proof of Proposition 1 in Appendix B) shows that this property is tantamount to the identity:

$$G_2 = \Omega_{21} \Omega_{11}^{-1} G_1. \tag{14}$$

Breusch et al. (1999) have shown that the condition (14) characterizes the fact that the set of moment conditions g_2 is "redundant" with respect to g_1, meaning that the complete set (g_1, g_2) of moment conditions does not deliver a GMM estimator (asymptotically) more accurate than $\widehat{\beta}^{(1)}$. It is not surprising to find that this condition characterizes the case where CC based on the couple of estimators $\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}\right)$ does not deliver better estimators (of any target) than estimators based on $\widehat{\beta}^{(1)}$ only.

Example: In the UWLS (10)/(11) example:

$$\Omega_{11} = E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^4(x_i, \gamma^1)}\right], \quad \Omega_{21} = E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^2(x_i, \gamma^1)\, \omega^2(x_i, \gamma^2)}\right].$$

Let us consider the particular case where the user-specified heteroskedasticity model matches perfectly the true skedastic function for the value γ^1 of the heteroskedasticity parameter:

$$\omega_0^2(x_i) \equiv \omega^2\left(x_i, \gamma^1\right).$$

In this case:

$$\Omega_{11} = E\left[\frac{x_i x_i'}{\omega^2(x_i, \gamma^1)}\right] = -G_1 \quad \text{and} \quad \Omega_{21} = E\left[\frac{x_i x_i'}{\omega^2(x_i, \gamma^2)}\right] = -G_2$$

so that the condition (14) is automatically fulfilled. This is relevant for our study in two cases:

1st case: The user-specified heteroskedasticity model is well–specified so that $\gamma_{WLS} = \gamma^1$ and $\widehat{\beta}^{(1)} = \hat{\beta}(\gamma_{WLS})$ is the optimal WLS estimator. In this case, there is no relevant additional information for estimation of β brought by any other UWLS estimator $\hat{\beta}(\gamma_{UWLS})$.

2nd case: $\widehat{\beta}^{(1)} = \hat{\beta}_{OLS}$ is the OLS estimator and this estimator is optimal because the DGP is homoskedastic: $\omega_0^2(x_i) \equiv \omega_{\text{hom}}^2$. In this case, irrespective of the heteroskedasticity model, there is no relevant additional information for estimation of β brought by any other UWLS estimator $\hat{\beta}(\gamma_{UWLS})$.

Remark 3: The concept of CC is more obvious when the two estimators $\widehat{\beta}^{(1)}$and $\widehat{\beta}^{(2)}$ are asymptotically independent. Then:

$$\lambda^*(h) = \lim_{n\to\infty} \frac{Var\left(\delta'\widehat{\beta}^{(1)}\right)}{Var\left(\delta'\widehat{\beta}^{(1)}\right) + Var\left(\delta'\widehat{\beta}^{(2)}\right)}.$$

In this case, $\lambda^*(h)$ is a weight in $[0,1]$ that gives more weight to $\widehat{\beta}^{(1)}$ (resp. to $\widehat{\beta}^{(2)}$) if and only if the plug-in variance $Var\left(\delta'\widehat{\beta}^{(1)}\right)$ is smaller (resp. larger) than $Var\left(\delta'\widehat{\beta}^{(2)}\right)$.

Note that, by virtue of (12), the asymptotic independence of $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ is tantamount to the asymptotic independence of the moment functions $\sqrt{n}\bar{g}_{1,n}(\beta^0)$ and $\sqrt{n}\bar{g}_{2,n}(\beta^0)$. One may even consider that this condition can be maintained without loss of generality, since one can replace the second set $\sqrt{n}\bar{g}_{2,n}(\beta^0)$ of moment conditions by a set $\sqrt{n}\bar{g}_{2/1,n}(\beta^0)$ that has been previously orthogonalized with respect to $\sqrt{n}\bar{g}_{1,n}(\beta^0)$:

$$\sqrt{n}\bar{g}_{2/1,n}(\beta) = \sqrt{n}\bar{g}_{2,n}(\beta) - \Omega_{21}\Omega_{11}^{-1}\sqrt{n}\bar{g}_{1,n}(\beta).$$

However, for these moment conditions, the Jacobian matrix is:

$$G_{2/1} = G_2 - \Omega_{21}\Omega_{11}^{-1}G_1.$$

Of course, this Jacobian matrix is nil in the case (14) of redundant moment conditions. By contrast, in many circumstances (see e.g. the example below), we can assume that the matrix $G_{2/1}$ is non-singular, such that our general theory of CC applies to orthogonal moment functions $\sqrt{n}\bar{g}_{1,n}(\beta^0)$ and $\sqrt{n}\bar{g}_{2/1,n}(\beta^0)$.

Example: Let us consider the UWLS (10)/(11) example in the case of a well-specified heteroskedasticity model, with $\gamma^1 = \gamma^{\text{hom}}$ and $\gamma^2 = \gamma_{WLS}$. Then:

$$G_2 = -E\left[\frac{x_i x_i'}{\omega^2\left(x_i, \gamma_{WLS}\right)}\right] = -E\left[\frac{x_i x_i'}{\omega_0^2\left(x_i\right)}\right].$$

$\left[-G_2\right]^{-1}$ is the variance matrix of the WLS estimator. On the other hand:

$$\begin{aligned}\Omega_{21}\Omega_{11}^{-1}G_1 &= -E\left[\frac{x_i x_i'\omega_0^2\left(x_i\right)}{\omega^2\left(x_i,\gamma_{WLS}\right)\omega_{\text{hom}}^2}\right]\left\{E\left[\frac{x_i x_i'\omega_0^2\left(x_i\right)}{\omega_{\text{hom}}^4}\right]\right\}^{-1}E\left[\frac{x_i x_i'\omega_0^2\left(x_i\right)}{\omega_{\text{hom}}^2}\right] \\ &= -E\left[x_i x_i'\right]\left\{E\left[x_i x_i'\omega_0^2\left(x_i\right)\right]\right\}^{-1}E\left[x_i x_i'\right].\end{aligned}$$

$\left[-\Omega_{21}\Omega_{11}^{-1}G_1\right]^{-1}$ is the variance matrix of the OLS estimator. Therefore, the WLS estimator $a'\hat{\beta}\left(\gamma_{WLS}\right)$ is strictly more accurate than the OLS estimator $a'\hat{\beta}\left(\gamma^{\text{hom}}\right)$ for any linear combination $a'\beta$ if and only if the matrix $G_{2/1} = \left[G_2 - \Omega_{21}\Omega_{11}^{-1}G_1\right]$ is negative definite. Hence, it is reasonable to maintain the assumption that the matrix $G_{2/1}$ is nonsingular.

Remark 4: As exemplified in Section 3.1, we may be led to consider moment conditions that are indexed by some nuisance parameters $\gamma \in \Gamma$. While the notations did not make explicit the dependence on γ, we

can for instance revisit the second set of moment functions as:

$$g_2(y_i, x_i, \beta) = g_2(y_i, x_i, \beta, \gamma).$$

In particular, for our regression application:

$$g_2(y_i, x_i, \beta, \gamma) = \frac{x_i}{\omega^2(x_i, \gamma)}(y_i - x_i'\beta). \tag{15}$$

We then want to resort to a condition of local robustness: replacing in $g_2()$ a specific value $\bar{\gamma}$ of nuisance parameters by a $\sqrt{n}$-consistent estimator $\hat{\gamma}$ ($\sqrt{n}(\hat{\gamma} - \bar{\gamma}) = O_P(1)$) has no impact on the asymptotic distribution of any GMM estimator of β based on moment conditions including:

$$E[g_2(y_i, x_i, \beta)] = 0 \Longleftrightarrow \beta = \beta^0.$$

This robustness property will be necessary for defining feasible versions of optimal CC estimators by using a first step consistent estimator of γ that will have no effect.

The standard assumption to ensure this robustness is:

$$E\left[\frac{\partial}{\partial \gamma'} g_2(y_i, x_i, \beta^0, \gamma)\right] = 0, \forall \gamma \in \Gamma. \tag{16}$$

It is the simplest case of Chernozhukov et al. (2022). It is obvious that the robustness condition (16) is valid for our regression example (15), insofar as we maintain the assumption of zero conditional expectation:

$$E[y_i - x_i'\beta^0 | x_i] = 0.$$

It is also the case if one wants to extend our study to 2SLS or NLLS.

3.3 Matricial Combinations of Estimators

Following Chen et al. (2016), we introduce MCC (Matricial CC) estimators. While Remark 1 above has shown that our CC estimators can be interpreted as plugging in the target $h(\beta)$ an estimator of β that is a convex combination of $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$, we now consider the possibility to plug in a matrix combination of estimators by considering:

$$\hat{\beta}_A = (I_p - A)\widehat{\beta}^{(1)} + A\widehat{\beta}^{(2)}$$

for any square matrix A of size p.

Hence, we minimize the asymptotic variance matrix of $\sqrt{n}\hat{\beta}_A$ by choosing $A = A^*$ that is the matrix of regression coefficients in the asymptotic regression of $\sqrt{n}(\widehat{\beta}^{(1)} - \beta^0)$ on $\sqrt{n}(\widehat{\beta}_n^{(1)} - \widehat{\beta}_n^{(2)})$:

$$A^* = \lim_{n\to\infty} Cov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\left[Var\left(\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\right]^{-1}.$$

It leads us to our second result.

Proposition 2 *The optimal TMCC (targeted matricial CC) estimator of $h(\beta)$ based on the couple of estimators $\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}\right)$ is given by:*

$$\widehat{h}_{A^*} = h\left((I_p - A^*)\widehat{\beta}^{(1)} + A^*\widehat{\beta}^{(2)}\right).$$

Remark 5: An asymptotic expansion shows that we can also interpret our TMCC estimator as follows:

$$\begin{aligned}\sqrt{n}\left[\widehat{h}_{A^*} - h\left(\beta^0\right)\right] &= \sqrt{n}\delta'\left(\beta^0\right)\left\{(I_p - A^*)\widehat{\beta}^{(1)} + A^*\widehat{\beta}^{(2)} - \beta^0\right\} + o_P(1) \\ &= \sqrt{n}\delta'\left(\beta^0\right)(I_p - A^*)\left(\widehat{\beta}^{(1)} - \beta^0\right) + \sqrt{n}\delta'\left(\beta^0\right)A^*\left(\widehat{\beta}^{(2)} - \beta^0\right) + o_P(1).\end{aligned}$$

This expansion does not allow to interpret the TMCC estimator $\widehat{h}_{A^*}$ as a CC estimator. It is only if the vector $\delta\left(\beta^0\right)$ is an eigenvector of the matrix $A^{*\prime}$ with an eigenvalue λ^*, that we can write:

$$\begin{aligned}\sqrt{n}\left[\widehat{h}_{A^*} - h\left(\beta^0\right)\right] &= \sqrt{n}\delta'\left(\beta^0\right)\left[(1-\lambda^*)\widehat{\beta}^{(1)} + \lambda^*\widehat{\beta}^{(2)} - \beta^0\right] + o_P(1) \\ &= \sqrt{n}\left[h\left((1-\lambda^*)\widehat{\beta}^{(1)} + \lambda^*\widehat{\beta}^{(2)}\right) - h\left(\beta^0\right)\right] + o_p(1).\end{aligned}$$

This result suggests that the set of CC estimators is a strict subset of the set of MCC estimators. Therefore, we expect in general that no CC estimator of the target $h(\beta)$ can be asymptotically as accurate as the TMCC estimator, except when the target is such that its gradient vector (computed at the true value of β) is an eigenvector of the matrix $A^{*'}$.

Remark 6: The case (in some sense without loss of generality as explained in Remark 3) of asymptotically independent estimators $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ is helpful to figure out the efficiency gain obtained by moving from TCC to TMCC. In this case:

$$A^* = \lim_{n\to\infty} Var\left(\widehat{\beta}^{(1)}\right)\left[Var\left(\widehat{\beta}^{(1)}\right) + Var\left(\widehat{\beta}^{(2)}\right)\right]^{-1}.$$

For the sake of notational simplicity, we will write hereafter in this remark asymptotic (co)variances without the lim symbol. The asymptotic expansion in Remark 5 shows that asymptotically the TMCC estimator $\widehat{h}_{A^*}$ depends on the matrix A^* only through:

$$A^{*\prime}\delta = \left[Var\left(\widehat{\beta}^{(1)}\right) + Var\left(\widehat{\beta}^{(2)}\right)\right]^{-1} Var\left(\widehat{\beta}^{(1)}\right)\delta. \tag{17}$$

Therefore, the eigenvector condition discussed in Remark 5 to make TCC and TMCC estimators asymptotically equivalent is tantamount to imposing that (17) can be rewritten as:

$$\left[Var\left(\widehat{\beta}^{(1)}\right) + Var\left(\widehat{\beta}^{(2)}\right)\right]\lambda^*\delta = Var\left(\widehat{\beta}^{(1)}\right)\delta. \tag{18}$$

When left-multiplying this condition by δ', we see that λ^* must be the weight elicited by TCC. However, this necessary condition is obviously not sufficient in general to ensure the eigenvalue conditions (18). This simply confirms that, in general, no CC estimator can be asymptotically as accurate as the TMCC.

Remark 7: An alternative estimation strategy would be to estimate β by over-identified GMM based on the two sets of moment conditions $\bar{g}_n(\beta) = [\bar{g}'_{1,n}(\beta), \bar{g}'_{2,n}(\beta)]'$ stacked together. For any given weighting matrix, we would define a GMM estimator:

$$\hat{\beta}(W) = \arg\min_{\beta} \bar{g}_n(\beta)' W \bar{g}_n(\beta).$$

By standard asymptotic GMM theory:

$$\sqrt{n}\left[\hat{\beta}(W) - \beta^0\right] = -\left[G'WG\right]^{-1} G'W\sqrt{n}\bar{g}_n(\beta^0) + o_P(1) \tag{19}$$

with:

$$G = G\left(\beta^0\right) \quad \text{and} \quad G(\beta^0) = \begin{bmatrix} G_1(\beta) \\ G_2(\beta) \end{bmatrix}.$$

Then, with obvious notations:

$$\begin{aligned} G'W\sqrt{n}\bar{g}_n(\beta^0) &= [G'_1W_{11} + G'_2W_{21}]\sqrt{n}\bar{g}_{1,n}(\beta^0) + [G'_1W_{12} + G'_2W_{22}]\sqrt{n}\bar{g}_{2,n}(\beta^0) \\ &= [G'_1W_{11} + G'_2W_{21}]G_1\sqrt{n}(\widehat{\beta}^{(1)} - \beta^0) + [G'_1W_{12} + G'_2W_{22}]G_2\sqrt{n}(\widehat{\beta}^{(2)} - \beta^0) + o_p(1). \end{aligned}$$

We define a square matrix $A(W)$ of dimension p by:

$$I_p - A(W) = \left[G'WG\right]^{-1}[G'_1W_{11} + G'_2W_{21}]G_1.$$

With easy calculations, we can check that:

$$A(W) = \left[G'WG\right]^{-1}[G'_1W_{12} + G'_2W_{22}]G_2$$

so that the asymptotic expansion (19) of the GMM estimator can be rewritten:

$$\sqrt{n}\left[\hat{\beta}(W) - \beta^0\right] = [I_p - A(W)]\sqrt{n}(\widehat{\beta}^{(1)} - \beta^0) + A(W)\sqrt{n}(\widehat{\beta}^{(2)} - \beta^0) + o_p(1).$$

Therefore, for any weighting matrix W, the GMM estimator $\hat{\beta}(W)$ associated to this matrix is an MCC estimator with a matricial weight A defined by $A(W)$ given above. Hence, the class of MCC estimators asymptotically encompasses not only the CC estimators but also all GMM estimators based on the complete set of moment conditions. Not surprisingly though, MCC does not allow us to beat efficient GMM since we can prove the following result.

Proposition 3 *The optimal TMCC estimator $\widehat{h}_{A^*}$ is asymptotically equivalent to the optimal plug in*

GMM estimator $h\left(\hat{\beta}\left(W^{*}\right)\right)$*, that is computed with the optimal weighting matrix* $W^{*}=\left[\Omega\left(\beta^{0}\right)\right]^{-1}$.

Albeit implicitly contained in the general result of Chen et al. (2016), the result of Proposition 3 is proved in Appendix B for the sake of self-containedness. While Chen et al. (2016) seem to suggest the contrary, the proof shows that the validity of this result heavily rests upon the fact that we are combining only just identified sets of moment conditions.

4 CC Estimators for the Regression Model

4.1 Adaptive convex combinations

Following the general Section 3.2, we revisit in this section all CC $\hat{h}_{\lambda}$ of two GMM estimators $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$:

$$\hat{h}_{\lambda}=(1-\lambda)\,h\left(\hat{\beta}^{(1)}\right)+\lambda h\left(\hat{\beta}^{(2)}\right). \tag{20}$$

The focus of our interest is regression estimators:

(i) $\hat{\beta}^{(1)}$ will always be the OLS estimator $\hat{\beta}_{OLS}$ defined by the moment functions:

$$g_{1}\left(y_{i},x_{i},\beta\right)=x_{i}\left(y_{i}-x_{i}^{\prime}\beta\right);$$

(ii) $\hat{\beta}^{(2)}=\hat{\beta}\left(\gamma\right)$ is the UWLS estimator defined by the moment functions:

$$g_{2,\gamma}\left(y_{i},x_{i},\beta\right)=\frac{x_{i}\left(y_{i}-x_{i}^{\prime}\beta\right)}{\omega^{2}\left(x_{i},\gamma\right)}.$$

Two adaptive CC estimators have been proposed in the extant literature. They both use $\gamma=\gamma_{WLS}$, meaning that they consider CC estimators that are adaptive mixtures of OLS $\hat{h}_{OLS}=h\left(\hat{\beta}_{OLS}\right)$ and WLS $\hat{h}_{WLS}=h\left(\hat{\beta}_{WLS}\right)$. Another more classical CC estimator has been proposed by DiCiccio et al. (2019), again based on convex combination of OLS and WLS. We describe below the three existing estimators in chronological order. To avoid confusion, it is worth stressing a fundamental difference between the adaptive estimators and the CC and TCC estimators considered hereafter:

On the one hand, strictly speaking, the adaptive estimators do not belong to the category of convex combination estimators defined by (20). The weight λ for convex combination will not be a given number but a random variable for sake of adaptiveness.

On the other hand, the CC estimator of DiCiccio et al. (2019) is conformable to (20) with $\hat{\beta}^{(1)}=\hat{\beta}_{OLS}$ and $\hat{\beta}^{(2)}=\hat{\beta}_{WLS}$, but with a choice of weight λ that minimizes the asymptotic variance of $\hat{h}_{\lambda}$ where $h\left(\beta\right)=\beta_{k}$ for one specific component β_{k} of β. This choice is obviously far from optimality in general. Aiming to a minimum asymptotic variance for some given target $h\left(\beta\right)$, our TCC estimator will be based on the choice of a targetingly-optimal weight $\lambda^{*}\left(\gamma\right)$ for any given γ and eventually an optimal choice γ^{*}

for γ, that does not coincide in general with γ_{WLS}.

4.1.1 The ALS estimator of Romano and Wolf (2017)

This estimator is dubbed *adaptive least squares* (ALS) because the final form of the estimator, OLS ($\lambda = 0$) or WLS ($\lambda = 1$) adapts itself to the data at hand. The trick is to choose for λ the value χ_n^{ALS} of the indicator function of a test of the null hypothesis H_0 of conditional homoskedasticity:

$$\chi_n^{ALS} = 1 \iff H_0 \text{ of conditional homoskedasticity is rejected.}$$

We will choose in this paper to apply the Breusch and Pagan (1979) test of homoskedascity with an user-specified heteroskedasticity model: $\omega^2(x_i, \gamma) = \exp(x_i'\gamma)$, $x_{i,1} = 1$. Koenker (1981) shows that a correctly studentized version of the test statistic leads to asymptotically correct significance levels (and a consistent test) under very general regularity conditions. Consistency of the test even does not require studentization of the test and is actually sufficient (see Section 4.3 in Romano and Wolf (2017)) to show that the WLS estimator $\hat{h}_{WLS}$ is asymptotically efficient to the ALS estimator:

$$\hat{h}_{ALS} = \left(1 - \chi_n^{ALS}\right)\hat{h}_{OLS} + \chi_n^{ALS}\hat{h}_{WLS}.$$

However, since the ALS estimator coincides to simple OLS when the test of homoskedasticity does not show a significant need to use WLS, one may expect better finite sample behaviour for $\hat{h}_{ALS}$ than for $\hat{h}_{WLS}$. On the other hand, ALS may be less efficient than OLS under heteroskedasticity since ALS is asymptotically equivalent to WLS which, when based on an incorrect model for conditional variance, can be less efficient than OLS under heteroskedasticity.

4.1.2 The MIN estimator of DiCiccio et al. (2019)

Like for the ALS estimator, we can define a Bernoulli variable χ_n^{MIN} leading to the MIN estimator:

$$\hat{h}_{MIN} = \left(1 - \chi_n^{MIN}\right)\hat{h}_{OLS} + \chi_n^{MIN}\hat{h}_{WLS}.$$

The binary variable χ_n^M is defined by:

$$\chi_n^{MIN} = 1 \iff \sigma_h^2\left(\beta^0, \gamma_{OLS}\right) > \sigma_h^2\left(\beta^0, \gamma_{WLS}\right)$$

where, for the sake of feasibility, we must replace asymptotic variances $\sigma_h^2\left(\beta^0, \gamma_{OLS}\right)$ and $\sigma_h^2\left(\beta^0, \gamma_{WLS}\right)$ by their consistent estimators. Note that this estimator generalizes the case of DiCiccio et al. (2019) since they only consider the case of a target $h(\beta) = \beta_k$ for some given $k = 1, \ldots, p$.

4.1.3 The CC estimator of DiCiccio et al. (2019)

This estimator is defined by the CC:

$$\hat{h}_{\lambda^{CC}} = \left(1-\lambda^{CC}\right) h\left(\hat{\beta}_{OLS}\right) + \lambda^{CC} h\left(\hat{\beta}_{WLS}\right)$$

where the weight λ^{CC} is optimally chosen λ to minimize the asymptotic variance of $\hat{h}_{\lambda}$. To be more explicit, we can now give the complete formulas of matrices whose blocks have already been used in Section 2.4 for feasible TWLS. More precisely, we now consider the joint asymptotic variance of $\hat{\beta}_{OLS}$ and any given UWLS estimator $\hat{\beta}(\gamma)$. Since:

$$\sqrt{n}\begin{bmatrix} \hat{\beta}_{OLS} - \beta^0 \\ \hat{\beta}(\gamma) - \beta^0 \end{bmatrix} = \begin{bmatrix} B_{1,n}^{-1} & 0 \\ 0 & B_{2,n}^{-1}(\gamma) \end{bmatrix} \frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{i=1}^{n} x_i u_i \\ \sum_{i=1}^{n} \frac{x_i u_i}{\omega^2(x_i,\gamma)} \end{bmatrix}$$

with:

$$B_{1,n} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \text{ and } B_{2,n}(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i x_i'}{\omega^2(x_i,\gamma)},$$

the joint asymptotic variance is:

$$\begin{bmatrix} \Sigma\left(\beta^0,\gamma^{\text{hom}}\right) & C_{12}\left(\beta^0,\gamma\right) \\ C_{21}\left(\beta^0,\gamma\right) & \Sigma\left(\beta^0,\gamma\right) \end{bmatrix} = \begin{bmatrix} B_1^{-1} & 0 \\ 0 & B_2^{-1}(\gamma) \end{bmatrix} V\left(\beta^0,\gamma\right) \begin{bmatrix} B_1^{-1} & 0 \\ 0 & B_2^{-1}(\gamma) \end{bmatrix}'$$

with:

$$B_1 = E\left[x_i x_i'\right], \quad B_2(\gamma) = E\left[\frac{x_i x_i'}{\omega^2(x_i,\gamma)}\right]$$

and:

$$V\left(\beta^0,\gamma\right) = \begin{bmatrix} V_{11}\left(\beta^0\right) = E\left[x_i x_i' \omega_0^2(x_i)\right] & V_{12}\left(\beta^0,\gamma\right) = E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^2(x_i,\gamma)}\right] \\ V_{21}\left(\beta^0,\gamma\right) = V_{12}\left(\beta^0,\gamma\right) & V_{22}\left(\beta^0,\gamma\right) = E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^4(x_i,\gamma)}\right] \end{bmatrix}.$$

Therefore, following Proposition 1 of Section 3, we get:

$$\lambda^{CC} = \frac{\delta' Var\left(\hat{\beta}_{OLS}\right)\delta - \delta' Cov\left(\hat{\beta}_{OLS},\hat{\beta}_{WLS}\right)\delta}{\delta' Var\left(\hat{\beta}_{OLS}\right)\delta + \delta' Var\left(\hat{\beta}_{WLS}\right)\delta - 2\delta' Cov\left(\hat{\beta}_{OLS},\hat{\beta}_{WLS}\right)\delta}$$

with:

$$\begin{aligned} Var\left(\hat{\beta}_{OLS}\right) &= \Sigma\left(\beta^0,\gamma^{\text{hom}}\right) &&= B_1^{-1} V_{11}\left(\beta^0\right) B_1^{-1\prime} \\ Var\left(\hat{\beta}_{WLS}\right) &= \Sigma\left(\beta^0,\gamma_{WLS}\right) &&= B_2^{-1} V_{22}\left(\beta^0,\gamma_{WLS}\right) B_2^{-1\prime} \\ Cov\left(\hat{\beta}_{OLS},\hat{\beta}_{WLS}\right) &= C_{12}\left(\beta^0,\gamma_{WLS}\right) &&= B_1^{-1} V_{12}\left(\beta^0,\gamma_{WLS}\right) B_2^{-1\prime}. \end{aligned}$$

Following DiCiccio et al. (2019)' Theorem 4.4 this optimal estimator is feasible in the sense that λ^{CC}

can be replaced by any consistent estimator without modifying the asymptotic distribution of $\hat{h}_{\lambda^{CC}}$ as characterized above. To see that, we just note that for any consistent estimator $\hat{\lambda}$ of any given $\lambda \in [0,1]$:

$$\sqrt{n}\left[\hat{h}_{\hat{\lambda}} - \hat{h}_{\lambda}\right] = o_P(1) \quad \text{since} \quad \frac{\partial \hat{h}_{\lambda}}{\partial \lambda} = h\left(\hat{\beta}_{WLS}\right) - h\left(\hat{\beta}_{OLS}\right) = o_P(1).$$

4.2 Nonparametric approaches

As already announced, we include in this section nonparametric extensions of UWLS, where "User-specified WLS" is no longer based on a choice γ^* of the parameter value γ that implies the choice $\omega^2(x, \gamma^*)$ for the skedastic function but on a nonparametric estimation of the true skedastic function $\omega_0^2(x)$.

Three preliminary remarks are in order.

First, one important reason why the extant literature has proposed adaptive convex combinations of OLS and WLS is that the WLS estimator may be less efficient than OLS when the user-specified model of heteroskedasticity is misspecified. In this respect, the nonparametric approach may be a valid alternative to CC since it will not depart from OLS when the skedastic function is flat.

Second Gonzales Coya and Perron (2022) argue that machine learning strategies generally outperform some classical nonparametric methods, applied in this context by Robinson (1987) (with Nearest Neighbor) or Fan and Yao (1998) (with Local Linear smoothing). A clear advantage of machine learning is that it does not require a tight pre-specification of the nature and number of covariates. Gonzales Coya and Perron (2022) compare the performance of Support Vector Regression (SVR) as applied in this context by Miller and Startz (2019) with Lasso. Gonzales Coya and Perron (2022) conclude that even though the SVR approach is a close contender to Lasso, the latter is preferred in particular because it provides "a sparse model with few non zero coefficients, which can inform us about the nature of the true skedastic function", while by contrast to the output of SVR lacks economic interpretation.

Third, while this issue is not addressed by Gonzales Coya and Perron (2022), it is worth realizing that even though the Lasso strategy is popular because it allows as input variables an arbitrary number of regressors, it may be problematic to include in the machine learning of the skedastic function some variables z that are not deterministic functions of the explanatory variables x because it would assume that the IV condition:

$$E\left[y_i - x_i'\beta | z_i\right] = 0,$$

which is a strong condition, is also fulfilled.

Following Gonzales Coya and Perron (2022), Lasso is the nonparametric method on which we set the focus. Gonzales Coya and Perron (2022) propose an adaptive Lasso estimator of the skedastic function with candidate inputs $z_{i,j},\ j = 1, \ldots, d$, defined by the minimization over the vector γ of coefficients in:

$$\frac{1}{2}\sum_{i=1}^{n}\left[\log\left(\hat{u}_i^2\right) - \gamma_0 - \sum_{j=1}^{d}\gamma_j z_{i,j}\right]^2 + \lambda^A \sum_{j=1}^{d}\hat{\omega}_j\left|\gamma_j\right| \tag{21}$$

where $\hat{u}_i,\ i=1,\ldots,n$ are OLS residuals, $\hat{\omega}_j = |\hat{\gamma}_j|^{-\psi},\ j=1,\ldots,d$ are adaptive weights, λ^A and ψ are two tuning parameters selected via cross-validation, $\hat{\gamma}_j,\ j=1,\ldots,d$ are first step $\sqrt{n}$-consistent estimators of $(\gamma_j)_{1\leq j\leq d}$. To get these first step estimators, Gonzales Coya and Perron (2022) resort to a first step Ridge regression. Then, the adaptive Lasso estimator of regression parameters β proposed by Gonzales Coya and Perron (2022) is the UWLS estimator obtained from the estimation of the skedastic function deduced from the minimization of (21).

The performance of this adaptive Lasso estimator will be compared with CC estimators in Section 4.4. We already note that none of these estimators is targeted towards the estimation of a specific function $h(\beta)$. For this reason, when the user-specified model for the skedastic function is misspecified, it should be the case that a targeted CC estimator has a better performance. We note that even for machine learning methods "the skedastic function is, in general, not consistently estimated" (as acknowledged by Gonzales Coya and Perron (2022)), so that our TCC may outperform Lasso (see Section 4.4 for a discussion). By the way, even within the framework of Gonzales Coya and Perron (2022), a more accurate estimator of the target $h(\beta)$ may be obtained by choosing tuning parameters λ and ψ according to the asymptotic variance of the resulting estimator of $h(\beta)$, instead of the skedastic function fitting (21).

4.3 Our targeted CC (TCC) estimator

By contrast with the CC estimator of DiCiccio et al. (2019), we consider any possible CC of the OLS estimator and some UWLS estimator:

$$\hat{h}_\lambda = (1-\lambda)\, h\left(\hat{\beta}_{OLS}\right) + \lambda h\left(\hat{\beta}\left(\gamma\right)\right).$$

From Proposition 1 of Section 3, we know that the optimal TCC estimator of $h\left(\beta\right)$ based on the couple of estimators $\left(\hat{\beta}_{OLS}, \hat{\beta}\left(\gamma\right)\right)$ is $\hat{h}_{\lambda^*(\gamma)}$ (note that with this notation, $\lambda^{CC} \equiv \lambda^*(\gamma_{WLS})$) with:

$$\lambda^*\left(\gamma\right) = \frac{\delta' Var\left(\hat{\beta}_{OLS}\right)\delta - \delta' Cov\left(\hat{\beta}_{OLS}, \hat{\beta}(\gamma)\right)\delta}{\delta' Var\left(\hat{\beta}_{OLS}\right)\delta + \delta' Var\left(\hat{\beta}(\gamma)\right)\delta - 2\delta' Cov\left(\hat{\beta}_{OLS}, \hat{\beta}\left(\gamma\right)\right)\delta}$$

and:

$$\begin{aligned}
Var\left(\hat{\beta}_{OLS}\right) &= \Sigma\left(\beta^0, \gamma^{\text{hom}}\right) = B_1^{-1} V_{11}\left(\beta^0\right) B_1^{-1\prime} \\
Var\left(\hat{\beta}\left(\gamma\right)\right) &= \Sigma\left(\beta^0, \gamma\right) = B_2^{-1} V_{22}\left(\beta^0, \gamma\right) B_2^{-1\prime} \\
Cov\left(\hat{\beta}_{OLS}, \hat{\beta}\left(\gamma\right)\right) &= C_{12}\left(\beta^0, \gamma\right) = B_1^{-1} V_{12}\left(\beta^0, \gamma\right) B_2^{-1\prime}.
\end{aligned}$$

By construction, the variance of the estimator $\hat{h}_{\lambda^*(\gamma)}$ is the residual variance:

$$\begin{aligned} & \delta' \left\{ Var\left(\hat{\beta}_{OLS}\right) - \left(\lambda^*\left(\gamma\right)\right)^2 Var\left(\hat{\beta}_{OLS} - \hat{\beta}\left(\gamma\right)\right) \right\} \delta \\ = \; & \delta' \Sigma\left(\beta^0, \gamma^{\text{hom}}\right) \delta - \left(\lambda^*\left(\gamma\right)\right)^2 \delta' \left\{ \Sigma\left(\beta^0, \gamma^{\text{hom}}\right) + \Sigma\left(\beta^0, \gamma\right) - 2C_{12}\left(\beta^0, \gamma\right) \right\} \delta. \end{aligned}$$

Therefore, the TCC approach leads to choose $\gamma = \gamma_{TCC}$ as the solution of:

$$\gamma_{TCC} = \arg\max_{\gamma \in \Gamma} \left(\lambda^*\left(\gamma\right)\right)^2 \delta' \left\{ \Sigma\left(\beta^0, \gamma^{\text{hom}}\right) + \Sigma\left(\beta^0, \gamma\right) - 2C_{12}\left(\beta^0, \gamma\right) \right\} \delta. \tag{22}$$

We note that to make this estimator feasible, we must get consistent estimators of the matrices B_1, $B_2\left(\gamma\right)$, $V_{11}\left(\beta^0\right)$,$V_{22}\left(\beta^0, \gamma\right)$, and $V_{12}\left(\beta^0, \gamma\right)$. As already explained in Section 2.4 (for feasible TWLS), we estimate:

(i) the matrices B_1 and $B_2\left(\gamma\right)$ by sample counterparts $B_{1,n}$ and $B_{2,n}\left(\gamma\right)$,

(ii) the matrices $V_{11}\left(\beta^0\right)$,$V_{22}\left(\beta^0, \gamma\right)$, and $V_{12}\left(\beta^0, \gamma\right)$ by using in addition the principle of Eicker-White estimators to prove consistency of estimators $V_{11,n}$,$V_{22,n}\left(\gamma\right)$, and $V_{12,n}\left(\gamma\right)$.

When resorting to such consistent estimators, we eventually get consistent counterparts of the two key quantities of TCC: (i) the maximization (22) leads to an estimated value $\hat{\gamma}_{TCC}$ of γ_{TCC}; and (ii) the function $\lambda^*\left(\gamma\right)$ is consistently estimated by a function $\hat{\lambda}^*\left(\gamma\right)$.

Then our final TCC estimator is:

$$\hat{h}_{TCC} = \left[1 - \hat{\lambda}^*\left(\hat{\gamma}_{TCC}\right)\right] h\left(\hat{\beta}_{OLS}\right) + \hat{\lambda}^*\left(\hat{\gamma}_{TCC}\right) h\left(\hat{\beta}\left(\hat{\gamma}_{TCC}\right)\right).$$

This feasible version $\hat{h}_{TCC}$ is asymptotically equivalent to the oracle estimator:

$$h_{TCC} = \left[1 - \lambda^*\left(\gamma_{TCC}\right)\right] + \lambda^*\left(\gamma_{TCC}\right) h\left(\hat{\beta}\left(\gamma_{TCC}\right)\right).$$

To see that, we resort to: (i) the robustness to consistent estimation $\hat{\lambda}^*\left(\gamma\right)$ of the optimal weight $\lambda^*\left(\gamma\right)$ as already discussed in Section 4.1.3 for the CC estimator of DiCiccio et al. (2019); and (ii) the robustness to consistent estimation of γ_{TCC}. We can then apply the general discussion in Remark 4 of Section 3.2. Thus, the asymptotic distribution of $\sqrt{n}\left[\hat{h}_{TCC} - h\left(\beta^0\right)\right]$ is normal, with mean zero and variance:

$$\delta' \Sigma\left(\beta^0, \gamma^{\text{hom}}\right) \delta - \left(\lambda^*\left(\gamma_{TCC}\right)\right)^2 \delta' \left\{ \Sigma\left(\beta^0, \gamma^{\text{hom}}\right) + \Sigma\left(\beta^0, \gamma_{TCC}\right) - 2C_{12}\left(\beta^0, \gamma_{TCC}\right) \right\} \delta.$$

4.4 A first small scale Monte Carlo experiment

We consider the same regression model as in Section 2.5. We study the finite sample MSE of different estimators for two possible DGPs $\omega_0^2\left(x_i\right) = \left[\log\left(x_i\right)\right]^{2j}$, $j = 1, 2, 3$, while the user-specified regression model is $\omega^2(x_i, \gamma) = \exp(\gamma_1 + \gamma_2 x_i)$. We compare by Monte Carlo (across 10000 trials) the ratio of MSE

of estimators of the slope β_2 with respect to MSE of OLS for four estimators of the slope coefficient β_2:

- the WLS estimator,
- our TWLS estimator with target β_2,
- DiCiccio et al. (2019)'s CC estimator with target β_2,
- our TCC estimator with target β_2.

We recall that the difference between the CC and the TCC estimator is as follows: while the CC estimator is an optimal (for target β_2) convex combination between the OLS and the WLS estimators, the TCC estimator is optimal (for the same target) convex combination between the OLS and the UWLS (User-Specified WLS) for an optimal value of the nuisance parameters γ.

The results for sample sizes $n = 50, 100, 200, 400$ respectively are given in Table 2 below.

n	$\omega_0^2(x) = (\log(x))^2$				$\omega_0^2(x) = (\log(x))^4$				$\omega_0^2(x) = (\log(x))^6$			
	WLS	TWLS	CC	TCC	WLS	TWLS	CC	TCC	WLS	TWLS	CC	TCC
50	.618	.615	.620	.547	.409	.417	.416	.335	.274	.101	.278	.060
100	.594	.579	.594	.483	.357	.336	.360	.249	.197	.060	.198	.036
200	.589	.588	.588	.491	.333	.300	.334	.203	.170	.048	.170	.026
400	.569	.563	.567	.461	.315	.282	.315	.201	.153	.048	.153	.028

Table 2: Ratios of the MSEs of WLS, TWLS, CC and TCC estimators with respect to that of OLS.

Then, two main conclusions are in order:

First, although the CC estimator is computed with the right target β_2, it does not outperform our TWLS estimator. In fact there is almost no difference in the performance of CC and naive WLS and that, in turn, highlights the importance of not stopping at the CC strategy but to further exploit the availability of the γ parameters in attaining further precision for the estimates for targets of interest.

Second, and even more importantly, our TCC estimator largely outperforms both the CC and our TWLS estimators. This dominance is even more striking in case of large samples and severe heteroskedasticity. It is worth noting that this better performance is all the more significant that our target is based on asymptotic variance, while the above Monte Carlo results are about MSE.

It is worth knowing that the DGP $\omega_0^2(x_i) = [\log(x_i)]^2$ has also been used for Monte Carlo experiments by Gonzales Coya and Perron (2022) to assess the performance of machine learning techniques. These authors revisit the performance of the SVR method due to Miller and Startz (2019) and propose an adaptive Lasso estimator. Their Monte Carlo MSE results (ratio with respect to MSE of OLS) are as follows for $n = 100, 200, 400$ respectively: .54, .46 and .46 for Lasso, .56, .48 and .46 for SVR. By comparison, the performance of TCC for the sample sizes $n = 100, 200, 400$ (ratio of MSE with respect to OLS: .48, .49, .46) is actually better which is arguably compelling since the results of Gonzales Coya and Perron (2022) are somewhat biased in favor of machine learning techniques as they assume that the true skedastic function belongs to the space of functions considered for learning.

5 Targeted GMM Estimator

5.1 GMM estimation of the regression model:

As in Section 4, we set the focus on the following two sets of moment conditions.

On the one hand, the just identified moment conditions that define the OLS estimator $\hat{\beta}_{OLS}$:

$$g_1\left(y_i, x_i, \beta\right) = x_i\left(y_i - x_i'\beta\right).$$

On the other hand, for any given value of the nuisance parameters γ, the weighted moment conditions that define the UWLS estimator $\hat{\beta}\left(\gamma\right)$:

$$g_{2,\gamma}\left(y_i, x_i, \beta\right) = \frac{x_i\left(y_i - x_i'\beta\right)}{\omega^2\left(x_i, \gamma\right)}.$$

By application of Proposition 3, one may consider that the optimal MCC estimator is asymptotically equivalent to efficient GMM:

$$\hat{\beta}\left(W^*\right) = \arg\min_{\beta} \bar{g}_n(\beta)' W^* \bar{g}_n(\beta) \tag{23}$$

computed with the optimal weighting matrix:

$$W^* = \left[\Omega\left(\beta^0\right)\right]^{-1} \quad \text{where} \quad \Omega\left(\beta^0\right) = Var\left[\sqrt{n}\bar{g}_n(\beta^0)\right].$$

This is the approach followed by Lu and Wooldridge (2020) who make it feasible by using a consistent estimator W_n of W^* based on OLS residuals and a consistent estimator γ_n of some pseudo-true value γ^* of nuisance parameters:

$$W_n^{-1} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 z_i z_i' \quad \text{where} \quad \hat{u}_i = y_i - x_i'\hat{\beta}_{OLS} \quad \text{and} \quad z_i = \left[\begin{array}{cc} x_i' & \frac{x_i'}{\omega^2(x_i,\gamma_n)} \end{array}\right]'.$$

However, this approach may be flawed in our case since, as already mentioned, we follow Romano and Wolf (2017) to consider only user-specified models of conditional heteroskedasticity that contain the homoskedastic setting as a particular case:

$$\gamma = \gamma^{\text{hom}} \implies \omega^2\left(x_i, \gamma\right) = \omega_{\text{hom}}^2 \implies g_{2,\gamma}\left(y_i, x_i, \beta\right) = \frac{1}{\omega_{\text{hom}}^2} g_1\left(y_i, x_i, \beta\right).$$

Therefore, when $\gamma = \gamma^{\text{hom}}$, the two sets of moment functions are proportional and the variance matrix is singular. In other words, there is no such thing as the inverse of the variance matrix $\Omega\left(\beta^0\right)$ to define an optimal variance matrix as above. Fortunately, we can easily prove the following.

Proposition 4 *For a given pseudo-true $\gamma^* \in \Gamma$, under the maintained assumption that for $\gamma^* \neq \gamma^{\text{hom}}$,*

the variance matrix $\Omega_{\gamma^*}\left(\beta^0\right)$ *of:*

$$g_{\gamma^*}\left(y_i, x_i, \beta^0\right) = \begin{bmatrix} g_1\left(y_i, x_i, \beta^0\right) \\ g_{2,\gamma^*}\left(y_i, x_i, \beta^0\right) \end{bmatrix}$$

is non-singular, then, an optimal weighting matrix W^* *for the GMM problem* (23) *is given by:*

$$[W^*]^{-1} = \begin{bmatrix} \Omega_{11}\left(\beta^0\right) & 1_{[\gamma^* \neq \gamma^{\text{hom}}]}\Omega_{12,\gamma^*}\left(\beta^0\right) \\ 1_{[\gamma^* \neq \gamma^{\text{hom}}]}\Omega_{21,\gamma^*}\left(\beta^0\right) & \Omega_{22,\gamma^*}\left(\beta^0\right) \end{bmatrix}. \tag{24}$$

Remark 1: The proof of Proposition 4 is straightforward. When $\gamma^* \neq \gamma^{\text{hom}}$, then $W^* = \left[\Omega_{\gamma^*}\left(\beta^0\right)\right]^{-1}$ is the weighting matrix for efficient GMM based on moment functions $g_{\gamma^*}\left(y_i, x_i, \beta\right)$. On the other hand, when $\gamma^* = \gamma^{\text{hom}}$:

$$\Omega_{22,\gamma^*}\left(\beta^0\right) = E\left[\frac{x_i x_i' u_i^2\left(\beta^0\right)}{\omega_{\text{hom}}^4}\right] = \frac{1}{\omega_{\text{hom}}^4}\Omega_{11}\left(\beta^0\right)$$

so that:

$$\begin{aligned} \bar{g}_n(\beta)' W^* \bar{g}_n(\beta) &= \left[\bar{g}_{1n}(\beta)', \frac{\bar{g}_{1n}(\beta)'}{\omega_{\text{hom}}^2}\right] \begin{bmatrix} \Omega_{11}^{-1}\left(\beta^0\right) & 0 \\ 0 & \omega_{\text{hom}}^4 \Omega_{11}^{-1}\left(\beta^0\right) \end{bmatrix} \begin{bmatrix} \bar{g}_{1n}(\beta) \\ \frac{\bar{g}_{1n}(\beta)}{\omega_{\text{hom}}^2} \end{bmatrix} \\ &= 2\bar{g}_{1n}(\beta)' \Omega_{11}^{-1}\left(\beta^0\right) \bar{g}_{1n}(\beta) \end{aligned}$$

which leads to efficient GMM based on the moment function $g_1\left(y_i, x_i, \beta\right)$ or equivalently on the moment functions $g_{2,\gamma^*}\left(y_i, x_i, \beta\right)$.

Remark 2: So far, for convex combination of estimators, we have maintained the assumption that the two matrices $\Omega_{11}\left(\beta^0\right) = Var\left[g_1\left(y_i, x_i, \beta^0\right)\right]$ and $\Omega_{22,\gamma^*}\left(\beta^0\right) = Var\left[g_{2,\gamma^*}\left(y_i, x_i, \beta^0\right)\right]$ are non-singular. :This means that the p components of the random vector $(x_i u_i)$ are linearly independent, and also the p components of the random vector $\left(x_i u_i / \omega^2\left(x_i, \gamma^*\right)\right)$ are linearly independent.

We now assume that when $\omega^2\left(x_i, \gamma^*\right)$ is not degenerate ($\gamma^* \neq \gamma^{\text{hom}}$), the matrix $\Omega_{\gamma^*}\left(\beta^0\right)$ is non-singular, meaning that the $2p$ components of the vector $\left[x_i' u_i, \frac{x_i' u_i}{\omega^2(x_i, \gamma^*)}\right]'$ are linearly independent.

Of course, Proposition 4 will deliver a feasible estimator only if we can build a consistent estimator of the optimal weighting matrix W^*. The various blocks of the variance matrix $\Omega_{\gamma^*}\left(\beta^0\right)$ can be estimated by sample counterparts computed by plugging in the OLS estimator $\hat{\beta}_{OLS}$:

$$\hat{\Omega}_{hk,\gamma^*} = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i^2 z_{ih} z_{ik}' \text{ for } h,k = 1,2 \text{ where } \hat{u}_i = y_i - x_i'\hat{\beta}_{OLS}, \quad z_{i1} = x_i, \quad z_{i2} = \frac{x_i}{\omega^2\left(x_i, \gamma_n\right)}.$$

The estimation issue of the indicator term $1_{[\gamma^* \neq \gamma^{\text{hom}}]}$ is tightly related to a test of conditional homoskedasticity, since in the context of Proposition 4, $\omega^2\left(x_i, \gamma^*\right)$ is degenerate if and only if $\gamma^* = \gamma^{\text{hom}}$. Similarly to the ALS estimator of Romano and Wolf (2017), we consider the indicator function χ_n of a

test of the null hypothesis H_0 of conditional homoskedasticity:

$$\chi_n = 1 \Longleftrightarrow H_0 \text{ of conditional homoskedasticity is rejected.}$$

Let us first assume that the test is consistent with asymptotic size α:

$$\begin{aligned} \gamma^* \neq \gamma^{\text{hom}} &\implies \lim_{n \longrightarrow \infty} \Pr\left[\chi_n = 1\right] = 1 \\ \gamma^* = \gamma^{\text{hom}} &\implies \lim_{n \longrightarrow \infty} \Pr\left[\chi_n = 0\right] = 1 - \alpha. \end{aligned} \tag{25}$$

Following Proposition 4, we would be led to consider the following estimator $\hat{W}$ for the optimal weighting matrix W^*:

$$\widehat{W}^{-1} = \begin{bmatrix} \hat{\Omega}_{11,\gamma^*} & \chi_n \hat{\Omega}_{12,\gamma^*} \\ \chi_n \hat{\Omega}_{21,\gamma^*} & \hat{\Omega}_{22,\gamma^*} \end{bmatrix}. \tag{26}$$

However, under the null hypothesis H_0 of conditional homoskedasticity, $\widehat{W}$ would be block-diagonal only with asymptotic probability $(1-\alpha)$ instead of the requested unit probability. A solution is to consider a drifting test size $\alpha = \alpha(n)$ that goes to zero with the sample size n.

Proposition 5 *Under the conditions of Proposition 4, the optimal weighting matrix W^* (24) is consistently estimated by $\hat{W}$ defined in (26), when we use a test for conditional homoskedasticity, whose indicator function χ_n is conformable with (25) with a drifting size $\alpha = \alpha_n \to 0$ when $n \to \infty$. Therefore, we get an efficient GMM estimator $\hat{\beta}_{MCC}(\gamma^*)$ as:*

$$\hat{\beta}_{MCC}(\gamma^*) = \arg\min_{\beta} \bar{g}'_{\gamma^*,n}(\beta)' \widehat{W} \bar{g}_{\gamma^*,n}(\beta)$$

that is asymptotically efficient for the set $\bar{g}_{\gamma^*,n}(\beta)$ of moment conditions.

Remark 1: Proposition 5 defines an MCC estimator $\hat{\beta}_{MCC}(\gamma^*)$ that shares some similarities with the ALS estimator $\hat{\beta}_{ALS}$ of Romano and Wolf (2017). When the null hypothesis of conditional homoskedasticity is not rejected, both estimators coincide with OLS. When the null hypothesis of conditional homoskedasticity is rejected, the ALS estimator $\hat{\beta}_{ALS}$ of Romano and Wolf (2017) coincides with WLS. It is worth comparing it with $\hat{\beta}_{MCC}(\gamma_{WLS})$ that combines efficiently (by virtue of semiparametric efficiency of efficient GMM) the informational content of OLS moment conditions and WLS moment conditions. We stress that, except in the case where the user's model of heteroskedasticity is well-specified ($\omega^2(x_i, \gamma_{WLS}) = \omega_0^2(x_i)$), it is not optimal in general to overlook the informational content of the OLS moment conditions. The rare case of redundancy of the OLS conditions is described in the example of UWLS (10)/(11) in Section 3.2.

Remark 2: For the purpose of consistent estimation of the optimal weighting matrix, we need to introduce the test indicator function χ_n for an empirical version of the population indicator of $\left[\gamma^* \neq \gamma^{\text{hom}}\right]$. It is the simplest way to circumvent the issue of possible (asymptotic) singularity of the variance matrix $\Omega_{\gamma^*}(\beta^0)$.

We note that the use of a generalized inverse of the empirical variance matrix would not do the job, because, when the rank is not constant along the sequence, the limit of the generalized inverses sequence may not coincide with the generalized inverse of the limit.

Remark 3: Our Monte Carlo experiments suggest that, even in the case of tiny conditional heteroskedasticity, the regularization of the sequence of weighting matrices that we propose in Proposition 5 could be useful in finite samples due to near-singularity. A comprehensive theory of efficient GMM in case of near-singularity is beyond the scope of this paper. It is germane with the issue of efficient estimation in case of nearly weak identification; see Dovonon et al. (2022).

Remark 4: The result of Proposition 5 can also be interpreted through the lenses of the general Section 3.3 about matrical combination of estimators. We know that any GMM estimator (23) can be seen as a matricial combination of OLS and UWLS estimators:

$$\hat{\beta}_{MCC}(\gamma^*) = (I_p - A)\,\hat{\beta}_{OLS} + A\hat{\beta}\,(\gamma^*)$$

with:

$$A = [G'WG]^{-1}\,[G_1'W_{12} + G_2'W_{22}]\,G_2.$$

When the weighting matrix W is block-diagonal, then:

$$A \;=\; [G_1'W_{11}G_1 + G_2'W_{22}G_2]^{-1}\,G_2'W_{22}G_2 \quad \text{and} \quad I_p - A \;=\; [G_1'W_{11}G_1 + G_2'W_{22}G_2]^{-1}\,G_1'W_{11}G_1,$$

that is:

$$A \;=\; \left[\Sigma_1^{-1} + \Sigma_2^{-1}\right]^{-1}\Sigma_2^{-1} \quad \text{and} \quad I_p - A \;=\; \left[\Sigma_1^{-1} + \Sigma_2^{-1}\right]^{-1}\Sigma_1^{-1}$$

where Σ_1 and Σ_2 stand for the asymtotic variances of $\hat{\beta}_{OLS}$ and $\hat{\beta}\,(\gamma^*)$ respectively, so that the matricial weights are completely defined by the relative variances of the two estimators. In this case, the covariances between estimators are not taken into account. Proposition 5 shows that this case may be relevant for efficient GMM even when the two estimators (or the corresponding moment conditions) are correlated.

Remark 5: The generic MCC estimator of our target of interest will be $\hat{h}_{MCC}(\gamma) = h\left(\hat{\beta}_{MCC}(\gamma)\right)$. We will refer to the specific MCC estimator $\hat{h}_{MCC}(\widehat{\gamma}_{WLS})$ evaluated at $\gamma = \widehat{\gamma}_{WLS}$ (and its $\sqrt{n}$-asymptotically equivalent counterpart $\hat{h}_{MCC}(\gamma_{WLS})$ by the argument of local robustness in Section 3.2.) as the GMM estimator, $\hat{h}_{GMM}$, adopting the terminology from Lu and Wooldridge (2020). We should note that apart from the regularization of the weighting matrix, as described above, there is another difference with Lu and Wooldridge (2020) — instead of $\widehat{\gamma}_{WLS}$ these authors used a Gamma quasi maximum likelihood estimator $\widehat{\gamma}_{QML}$ of γ as in Section 18.4, Wooldridge (2010). As we have argued in our paper, there is no compelling reason to choose either $\widehat{\gamma}_{WLS}$ or $\widehat{\gamma}_{QML}$ for plug-in except in the unlikely scenario where the user's model for conditional heteroskedasticity is known to be correct. In our simulations $\hat{h}_{MCC}(\widehat{\gamma}_{WLS})$

performs somewhat better than $\hat{h}_{MCC}(\widehat{\gamma}_{QML})$, and hence we focus on the former MCC estimator dubbing it as GMM. The MCC estimator that chooses the plug-in γ by the optimal targeting principle of our paper, as described in the next subsection, will be referred to as the targeted GMM (TGMM) estimator.

5.2 Our targeted MCC estimator: TGMM

For any possible pseudo-true γ, we consider the efficient GMM estimator $\hat{\beta}_{MCC}(\gamma)$ defined by Proposition 5 and the resulting estimated target:

$$\hat{h}_{MCC}(\gamma) = h\left(\hat{\beta}_{MCC}(\gamma)\right).$$

For $\gamma \neq \gamma^{\text{hom}}$, the asymptotic variance of $\hat{h}_{MCC}(\gamma)$ is, with $\delta = \delta\left(\beta^0\right)$ and obvious notations,

$$\sigma^2_{MCC}(\gamma) = \delta'\left[G'_\gamma(\beta^0)\Omega_\gamma^{-1}\left(\beta^0\right)G_\gamma(\beta^0)\right]^{-1}\delta.$$

This may be compared with $\sigma^2_{MCC}\left(\gamma^{\text{hom}}\right)$ defined by:

$$\sigma^2_{MCC}\left(\gamma^{\text{hom}}\right) = \delta'\left[G'_1(\beta^0)\Omega_{11}^{-1}\left(\beta^0\right)G_1(\beta^0)\right]^{-1}\delta.$$

Our choice of the pseudo-true value γ^* is defined by:

$$\gamma^* = \begin{cases} \gamma^{\text{hom}} & \text{if } \sigma^2_{MCC}\left(\gamma^{\text{hom}}\right) \leq \min\left\{\sigma^2_{MCC}(\gamma), \gamma \in \Gamma, \gamma \neq \gamma^{\text{hom}}\right\} \\ \arg\min_{\gamma\in\Gamma,\gamma\neq\gamma^{\text{hom}}} \sigma^2_{MCC}(\gamma) & \text{otherwise.} \end{cases}$$

We get a consistent estimator γ_n of γ^* by using the sample counterparts of $G_\gamma(\beta^0)$ and $\Omega_\gamma^{-1}\left(\beta^0\right)$. Then, by the argument of local robustness (see Section 3.2.), $\hat{h}_{MCC}(\gamma_n)$ has the same asymptotic variance as the infeasible optimal $\hat{h}_{MCC}(\gamma^*)$. Therefore, $\hat{h}_{MCC}(\gamma_n)$ is our optimal targeted MCC estimator.

However, the minimization is not convex and may lead to spurious results in finite samples. This is the reason why we will rather define our estimator of γ^* as follows:

$$\tilde{\gamma}_n(\Gamma_n) = \arg\min_{\gamma\in\Gamma_n\cup\gamma^{\text{hom}}} \sigma^2_{MCC}(\gamma)$$

where Γ_n is a compact neighborhood of $\hat{\gamma}_{WLS}$ in Γ. Hence, for all practical purpose, $\hat{h}_{MCC}(\tilde{\gamma}_n(\Gamma_n))$ will be used as our preferred targeted MCC estimator. We will call $\hat{h}_{MCC}(\tilde{\gamma}_n(\Gamma_n))$ as our TGMM estimator.

With an increasing sequence Γ_n of compact sets such that

$$\lim_{n\longrightarrow\infty} \uparrow \Gamma_n = \Gamma,$$

our TGMM estimator $\hat{h}_{MCC}(\tilde{\gamma}_n(\Gamma_n))$ is also (like $\hat{h}_{MCC}(\gamma_n)$; see above Proposition 4) a consistent

version of our optimal MCC estimator. We will denote our TGMM estimator $\hat{h}_{MCC}\left(\tilde{\gamma}_n\left(\Gamma_n\right)\right)$ by the notation $\hat{h}_{TGMM}$. We do find simulation evidence (available from authors) that $\hat{h}_{MCC}\left(\tilde{\gamma}_n\left(\Gamma_n\right)\right)$ provides valuable hedge against possible instability of the naive estimator $\hat{h}_{MCC}\left(\gamma_n\right)$ (see above Proposition 4).

Expanding on the little simulation evidence in Sections 2.5 and 4.4, Appendix A will provide extensive simulation evidence of the benefit of targeting under all ten cases considered in Romano and Wolf (2017), their empirically relevant DGP, the four cases considered in Lu and Wooldridge (2020), and their empirical illustration. More results are available from us. For brevity we report the results for the eight estimators: OLS, WLS, ALS, MIN, TWLS, CC, TCC, GMM and TGMM. The results are all reported in Tables 3-10. The results corresponding to those presented in Sections 2.5 and 4.4 can be found in Table 3.

The main message of the numerical results in Appendix A is that if the user's model $\omega^2(x;\gamma)$ for $V(u|x)$ allows for improvement in precision over the existing estimators then our proposed targeted estimators TWLS, TCC and TGMM achieve it without any major cost in terms of empirical bias, empirical size, etc. This improvement can be dramatically big, and thus suggests that it will be imprudent to overlook the benefit of targeting in practice. The improvements over the existing estimators including the recently proposed ones are even more compelling because we document them under the same setup considered in the papers that actually proposed these recent estimators.

Comparison among TWLS, TCC and TGMM does not however give a clear winner in terms of empirical MSE and empirical size across all the numerical results. Based on these observations and the simplicity of the estimators, we recommend all three proposed estimators in practice.

6 Conclusion

We have shown in this paper that user-specified parametric models for heteroskedasticity can always allow to improve upon both OLS and WLS irrespective of the amount of misspecification of the unknown skedastic function. The core idea is to question the two most common practices of either robust OLS-based inference (Eicker-White approach) or "Resurrecting Weighted Least Squares" to improve estimators (Romano and Wolf (2017)). We argue that these two practices imply some significant and unfortunate efficiency loss, because they resort to a suboptimal criterion function:

(i) The heteroskedasticity-robust confidence sets provided by the Eicker-White approach are wider than necessary since a more efficient estimator would deliver less conservative inference.

(ii) The WLS criterion is based on minimization of the quadratic distance between the true unknown skedastic function and the set of skedastic functions provided by the user-specified parametric model for heteroskedasticity. In case of a severely misspecified heteroskedasticity model, this minimization may have little to do with the relevant criterion of minimization of variance of estimators.

By contrast, when targeting our optimality criterion on the asymptotic variance of estimator of some

scalar smooth function of the regression parameters, we realize that there are efficiency gains with respect to both OLS and WLS to be drawn from a proper choice of a weighting function in the user-specified heteroskedasticity model (UWLS). We illustrate this claim by studying three new estimators:

(i) TWLS: A targeted UWLS.

(ii) TCC: A targeted convex combination between OLS and a properly chosen UWLS.

(iii) TGMM: A targeted matrix combination between OLS and a properly chosen UWLS. We show that it is equivalent to efficient GMM based on all the moment conditions associated to the two estimators.

Even though these three estimators deliver striking improvements in estimation of parameters of interest by comparison with the extant contenders with a little less cost of implementation, we acknowledge that researchers ready to combine even more UWLS estimators could get even more accurate estimators. More precisely, our current paper paves the way for at least three new research agendas.

First, how to elicit optimally a set of UWLS estimators to combine them for efficient estimation of a given target. As discussed in Section 5, the computation of an efficient TGMM may be flawed in finite samples in case of near collinearity between moment conditions corresponding to different values of the heteroskedasticity parameters γ. Considering the combination of more than two sets of moment conditions, we will obviously face a trade-off between the minimization of variance of the estimator resulting from many moment conditions (corresponding to an infinite number of values of γ) and the instability that their near-multicollinearity may cause. As a result, we expect to have to develop regularization methods beyond the simplest binary case considered in Section 5 above.

Second, beyond the infinite number of UWLS estimators provided by a user-specified parametric model for heteroskedasticity, one may even consider nonparametric families of skedastic functions. We have shown that our (parametrically) targeted estimators are good contenders with respect to machine learning approaches. Of course, it should be possible to generate even more accurate estimators by taking advantage of machine learning, but within the methodology of targeting a parameter of interest. While this implies computational issues due to non-convexity of the corresponding optimization programs, iterative approaches should be devised to keep the simplicity of TCC estimators at each step of the recursion.

Third, as already explained, the possibility of devising an infinite set of valid UWLS estimators is a benefit of the maintained assumption of a regression model that is defined by a conditional expectation. In case of a regression equation that is only defined by linear projection, there is no reason to believe that moment conditions that are reweighted by nonlinear functions of exogenous variables are still valid. However, it may intuitively be informative to keep using the UWLS estimators, even though they are biased, even asymptotically. In this case, optimality must be defined by some minimization of (asymptotic) mean squared error instead of asymptotic variance. This extension seems important for many applications. A well-documented example is estimation of weak ARMA-GARCH models (see Francq and Zakoian (2000) and the references therein) for which strategies to improve upon OLS are still largely missing in the

literature. More generally, there is room for extension of our targeted approaches to approximately linear models in the sense of Sacks and Ylvisaker (1978).

References

Angrist, J. D. and Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspective*, 24: 3–30.

Breusch, T., Hailong, Q., Schmidt, P., and Wyhowski, D. (1999). Redundancy of moment conditions. *Journal of Econometrics*, 91: 89–111.

Breusch, T. S. and Pagan, A. R. (1979). A Simple Test for Heteroskedasticity and Random Coefficient Variation. *Econometrica*, 47:1287–1294.

Cao, W., Tsiatis, A., and Davidian, M. (2009). Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. *Biometrika*, 96:723–734.

Chen, X., Jacho-Chavez, D. T., and Linton, O. (2016). Averaging of an increasing number of moment condition estimators. *Econometric Theory*, 32: 30–70.

Chernozhukov, V., Escanciano, J.-C., Ichimura, H., Newey, W., and Robins, J. (2022). Locally Robust Ssemiparametric Estimation. *Econometrica*, 90: 1501–1535.

Cragg, J. G. (1992). Quasi-aitken estimation for heteroskedasticty of unknown form. *Journal of Econometrics*, 54: 179–201.

DiCiccio, C. J., Romano, J. P., and Wolf, M. (2019). Improving weighted least squares inference. *Econometrics and Statistics*, 10:96–119.

Dovonon, P., Atchade, Y., and Tchatoka, F. (2022). Efficiency bounds for moment condition models with mixed identification strength. Concordia University Working Papers.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regressions. *Biometrika*, 85: 645–660.

Francq, C. and Zakoian, J. (2000). Estimating Weak GARCH Representations. *Econometric Theory*, 16: 692–728.

Gonzales Coya, E. and Perron, P. (2022). Estimation in the Presence of Heteroskedasticity of Unknown Form: A Lasso-based Approach. Boston University Working Paper.

Gourieroux, C., Monfort, A., and Renault, E. (1996). Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form. *Journal of Statistical Planning and Inference*, 50: 37–63.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2: 849–879.

Koenker, R. (1981). A Note on Studentizing a Test for Heteroscedasticity. *Journal of Econometrics*, 17: 107–112.

Leamer, E. E. (2010). Tantalus on the Road to Asymptotia. *Journal of Economic Perspective*, 24: 31–46.

Lu, C. and Wooldridge, J. M. (2020). A GMM estimator asymptotically more efficient than OLS and WLS in the presence of heteroskedasticity of unknown form. *Applied Economics Letters*, 27: 997–1001.

Miller, S. and Startz, R. (2019). Feasible Generalized Least Sqaures Using Machine Learning. Forthcoming: Economics Letters.

Papadopoulosa, A. and Tsionas, M. G. (2021). Efficiency gains in least squares estimation: A new approach. Forthcoming: Econometric Reviews.

Rilstone, P. (1991). Some Monte Carlo Evidence on the Relative Efficiency of Parametric and Semiparametric EGLS Estimators. *Journal of Business and Economic Statistics*, 9:179–187.

Robinson, P. M. (1987). Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form. *Econometrica*, 55: 875–891.

Romano, J. P. and Wolf, M. . (2017). Resurrecting Weighted Least Squares. *Journal of Econometrics*, 197: 1–19.

Sacks, J. and Ylvisaker, D. (1978). Linear Estimation for Approximately Linear Models. *The Annals of Statistics*, 6: 1122–1137.

Spady, R. and Stouli, S. (2019). Simultaneous Mean-Variance Regression. Working paper.

Stock, J. H. and Watson, M. W. (2011). *Introduction to Econometrics*. Pearson, 3 edition.

van der Laan, M. J. and Rubin, D. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, 2:https://doi.org/10.2202/1557–4679.1043.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heterogeneity. *Econometrica*, 48:817–838.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, J. M. (2012). *Introductory Econometrics*. South-Western, Mason, Ohio.

A Appendix A: Simulation evidence and Empirical illustration

We will explore the small-sample performance of the proposed targeted estimators TWLS, TCC and TGMM using simulation experiments based on 10000 Monte Carlo trials. Results will be reported for the eight estimators: OLS, WLS, ALS, MIN, TWLS, CC, TCC, GMM and TGMM.[1]

We omit estimators such as Cragg's estimator, GMM estimators where γ is estimated by maximum likelihood, etc. because in results unreported here we find their performance to be relatively poor.

The main message of the numerical results here is that if the user's model $\omega^2(x;\gamma)$ for $V(u|x)$ allows for improvement in precision over the existing estimators then the proposed targeted estimators achieve it. Like Romano and Wolf (2017), we report the improvement in the empirical mean squared error (MSE), and find that its reduction by the proposed estimators can be huge by any conceivable standard. Under all cases there does not seem to be any major cost in terms of empirical bias, size, etc. to using the proposed estimators. Comparison among the proposed estimators TWLS, TCC and TGMM does not however give a clear winner. Based on these observations and the simplicity of the estimators we recommend all three proposed estimators in practice.

A.1 Simulations under the design in Romano and Wolf (2017):

Romano and Wolf (2017) take the linear model $y = x_{(1)}\beta_1 + x_{(2)}\beta_2 + u$ with $x_{(1)} = 1, x_{(2)} \sim U(1,4)$, $x = (x_{(1)}, x_{(2)})'$; $\beta = (\beta_1, \beta_2)'$, $\beta^0 = (0,0)'$; $u = s(x)z$ where $z \sim N(0,1)$ is independent of $x_{(2)}$ and thus $E[u|x] = 0$ and $V(u|x) = s^2(x)$. They consider 10 cases for the skedastic function:

Case 1: (a) $s^2(x) = 1$; (b) $s^2(x) = x_{(2)}$; (c) $s^2(x) = x_{(2)}^2$; (d) $s^2(x) = x_{(2)}^4$.

Case 2: (a) $s^2(x) = \left(\log(x_{(2)})\right)^2$; (b) $s^2(x) = \left(\log(x_{(2)})\right)^4$.

Case 3: (a) $s^2(x) = \exp\left(.1(x_{(2)} + x_{(2)}^2)\right)$; (b) $s^2(x) = \exp\left(.15(x_{(2)} + x_{(2)}^2)\right)$.

Case 4: (a) $s^2(x) = \begin{cases} 1 & \text{if } x_{(2)} < 2 \\ 2 & \text{if } 2 \le x_{(2)} < 3 \\ 3 & \text{if } x_{(2)} \ge 3 \end{cases}$; (b) $s^2(x) = \begin{cases} 1 & \text{if } x_{(2)} < 2 \\ 2^2 & \text{if } 2 \le x_{(2)} < 3 \\ 3^2 & \text{if } x_{(2)} \ge 3 \end{cases}$.

To emphasize the gain in precision, we will add a Case 2(c) with $s^2(x) = \left(\log(x_{(2)})\right)^6$.

Romano and Wolf (2017) consider two parametric models $\omega^2(x;\gamma)$ — Model 1: $\omega^2(x;\gamma) := \exp(\gamma_1 + \gamma_2 \log(x_{(2)}))$ and Model 2: $\omega^2(x;\gamma) := \exp(\gamma_1 + \gamma_2 x_{(2)})$ — and like them our results here are also very similar for both models. However, since there is slightly more action in terms of improved precision in case of estimators based on Model 2, for brevity we report here the results based on Model 2 only.[2]

[1] The extensive simulation study here, of which only a subset of results is presented while the rest are available from us, complements Rilstone (1991)'s early simulations that focused on OLS, WLS and its semiparametric versions.

[2] Model 1 is correct for $V(u|x)$ under Cases 1(a)-1(d) with $\gamma_2^0 = 0, 1, 2, 4$ respectively. Model 2 is correct for $V(u|x)$ only under Case 1(a) with $\gamma_2^0 = 0$. So, all estimators are asymptotically efficient under Case 1(a), and all estimators other than OLS are asymptotically efficient under Cases 1(b)-1(d) when using Model 1.

Following Romano and Wolf (2017) we will report for $h(\beta) = \beta_2$ the empirical MSE's (their ratios) of estimators and empirical coverage probability of 95% confidence intervals (1 - empirical size of 5% t tests). We will consider sample sizes $n = 50, 100, 200, 400$.

Table 3 presents the ratio of the empirical MSE of each estimator for $h(\beta) = \beta_2$ with respect to that of OLS. Besides Case 1(a) (conditional homoskedasticity), the other estimators lead to smaller, sometimes much smaller, MSE. (To compare any two non-OLS estimators, say A with respect to B, divide the ratio under A with that under B.) Importantly, the proposed estimators either performs very similar to the other estimators that they are supposed to improve upon or leads to really big gain in precision as in Cases 2 (a), (b) and (c).

Table 4 presents the empirical size (empirical rejection probability of the truth) of 5% Wald tests based on each estimator for $h(\beta) = \beta_2$. The results look reasonable. While the size-corrected empirical power is not reported here for brevity (but is available from us), we note that the proposed estimators always have either the same or substantially greater (in Cases 2) empirical size-corrected power than its competitors.

Table 5 presents the average length of each of the non-OLS confidence intervals for $h(\beta) = \beta_2$ with respect that of the OLS intervals. For brevity we report this for Case 2 only where, as noted above, the benefit of the proposed estimators' precision is most prominently evident. These are indeed big gains in precision of confidence intervals by any standard.

A.2 Empirically relevant simulations in Romano and Wolf (2017):

Romano and Wolf (2017)'s simulation based on a real-life example revisits the well-known cross-sectional data set from 1970 containing $n = 506$ observations from communities in the Boston area (see, Wooldridge (2012)). They consider a linear regression with:

$$E[y|x] = x'\beta = x_{(1)}\beta_1 + x_{(2)}\beta_2 + x_{(3)}\beta_3 + x_{(4)}\beta_4 + x_{(5)}\beta_5$$

where y is the log of the median housing price in a community, $x_{(1)} = 1$, $x_{(2)}$ is the log of nitrogen oxide in the air (in parts per million), $x_{(3)}$ is the log of weighted distance from five employment centers (in miles), $x_{(4)}$ is the average number of rooms per house, and $x_{(5)}$ is the average student–teacher ratio in the community's schools.

To mimic the true conditional heteroskedasticity in this data, Romano and Wolf (2017): (i) obtain $\widehat{e}_i = (y_i - x'\widehat{\beta}_{OLS})/\sqrt{1 - q_{i,i}}$ for $i = 1, \ldots, n$ where $q_{i,i} = x_i'(\sum_j x_j x_j')^{-1} x_i$ is i-th diagonal element of the hat-matrix; (ii) generate artificial data (y_i^*, x_i^*) for $i = 1, \ldots, n$ where $x_i^* = x_i$ and $y_i^* = x_i'\widehat{\beta}_{OLS} + \widehat{e}_i v_i$ where $v_i \sim N(0, 1)$ independently of the system. Thus, the true β in this artificial data is $\widehat{\beta}_{OLS}$. Romano and Wolf (2017) then report for each element of β the empirical MSE's (their ratios) of estimators, empirical coverage probability of 95% confidence intervals (1 - empirical size of 5% t tests) and ratios of the average length of these intervals.

We will do the same, and since the improvement shown by Romano and Wolf (2017) is noticeably better with their Model 1, i.e., $\omega^2(x;\gamma) = \exp(\gamma_1 + \sum_{k=2}^{5} \log(x_{(k)}))$, we will for brevity only report the further improvement provided by our proposed estimators based on Model 1. These are reported in Tables 6 and 7 respectively for the ratio of the empirical MSE's with respect to OLS and the empirical size of 5% Wald test with, within parentheses, the ratio of the average length of confidence intervals based on other estimators to that based on OLS. As is clearly evident, the proposed estimators deliver noticeably big further gains over its competitors.

A.3 Simulations under the design in Lu and Wooldridge (2020):

Lu and Wooldridge (2020) take the linear model $y = x_{(1)}\beta_1 + x_{(2)}\beta_2 + x_{(3)}\beta_3 + x_{(3)}\beta_4 + u$ with $x_{(1)} = 1, x_{(2)} \sim N(1,1)$, $x_{(3)} = .8 + .2x_{(2)} + e_1$, $x_{(4)} = 1(x_{(5)} > x_{(3)})$, $u = s(x)e_3$ where e_1, e_2, e_3 are independent $N(0,1)$, and $x_{(5)} = .3 + .1x_{(2)} + .1x_{(3)} + e_2$. They take $x = (x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)})'$, e_3 as independent of x (giving $E[u|x] = 0$ and $V(u|x) = s^2(x)$), and $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)'$ with $\beta^0 = (.5, 1, 1, 1)'$. They consider 4 cases for the skedastic function:

Case 1: $s^2(x) = (\beta_1 + \beta_2 x_{(2)} + \beta_3 x_{(3)} - 3\beta_4 x_{(4)} + .1x_{(2)}(x_{(3)} + x_{(4)}) - .1x_{(3)}x_{(4)} - .05x_{(2)}^2 + .05x_{(3)}^2))^2$

Case 2: $s^2(x) = (\beta_1 + \beta_2|x_{(2)}| + \beta_3 x_{(3)}^2 + \beta_4 x_{(4)})^2$.

Case 3: $s^2(x) = \exp(\beta_1 + \beta_2|x_{(2)}| + \beta_4 x_{(4)})$.

Case 4: $s^2(x) = \exp(\beta_1 + \beta_2 x_{(2)} + \beta_3 x_{(3)} + \beta_4 x_{(4)})$.

They consider the parametric model $\omega^2(x;\gamma) = \exp(x'\gamma)$, which is correct for $V(u|x)$ with $\gamma^0 = \beta^0$ in Case 4.

We take $h(\beta) = \beta_1, \beta_2, \beta_3, \beta_4$ respectively and sample size $n = 1000, 5000$. Lu and Wooldridge (2020) take $n = 1000, 10000$ and report Monte Carlo mean and standard deviations in their Table 1. Comparison of standard deviations of course works in favor of our proposed estimators. Therefore, to put our proposed estimators to stricter tests, we will instead continue with reporting the empirical MSE's of estimators and empirical size of 5% Wald tests.

Table 8 presents the ratio of the empirical MSE of each estimator with respect to that of OLS.[3] It is of interest to note that in our implementation of Cases 1 and 2, WLS based on an incorrect model $\omega^2(x;\gamma)$ can be much less precise than OLS, which is a possibility that DiCiccio et al. (2019) (p.2, paragraph 7) noted as motivation to their MIN and CC estimators but conjectured as "rare". ALS also suffers from the same problem in this case since ALS and WLS are very similar (almost identical) here because of high level of heteroskedasticity of u.

On the other hand, the MIN, CC and MCC estimators deliver big gains in precision over OLS. Additionally, when the parametric model $\omega^2(x;\gamma)$ is far from correct for $V(u|x)$, i.e., Cases 1 and 2, we see

[3]Our results for WLS and GMM are not the same as that in Lu and Wooldridge (2020) because they use Gamma QMLE for γ whereas we use least squares estimators of γ to maintain uniformity with the rest of the simulations.

that our proposed estimators deliver further substantial gains in precision. However, when $\omega^2(x;\gamma)$ is correct for $V(u|x)$, i.e., in Case 4, there is no room for improvement since all non-OLS estimators are then asymptotically efficient (not considering the information that β's appear in both $E[y|x]$ and $V(y|x)$). Then our proposed estimators are less precise than their non-OLS competitors. This problem however diminishes with larger sample size $n = 5000$.

Table 9 presents the empirical size of 5% Wald tests based on each estimator. The results for the proposed estimators look reasonable except for our proposed estimators of β_2 and β_3 in the case of correct specification, i.e., Case 4, and for TGMM of β_2 and β_3 in Case 3. As with the case of MSE, this problem diminishes with larger sample size $n = 5000$ and ultimately disappears when, like Lu and Wooldridge (2020), $n = 10000$ (not reported).

A.4 Empirical illustration in Lu and Wooldridge (2020):

Lu and Wooldridge (2020) use a subset of the well-known cross-sectional individual-level data set ' 401ksubs' (see Wooldridge (2012)) to estimate a linear regression with:

$$E[y|x] = x'\beta = \sum_{k=1}^{10} x_{(k)}\beta_k$$

where y is net total financial assets (in \$ 1000) and is denoted by "nettfa"; $x_{(1)} = 1$ and is denoted by "constant"; $x_{(2)}$ is annual income (in \$1000) in excess of population (data) average and is denoted by "inc_0"; $x_{(3)} = x_{(2)}^2$ and is denoted by "inc_0^2"; $x_{(4)}$ is age in excess of population (data) average and is denoted by "age_0"; $x_{(5)} = x_{(4)}^2$ and is denoted by "age_0^2"; $x_{(6)} = x_{(2)} \times x_{(4)}$ and is denoted by "$\text{inc}_0.\text{age}_0$"; $x_{(7)}$ is a dummy variable for eligibility for a 401k plan and is denoted by "e401k"; $x_{(8)}$ is a dummy variable for male and is denoted by "male"; $x_{(9)} = x_{(7)} \times x_{(2)}$ and is denoted by "e401k.inc_0"; and $x_{(10)} = x_{(7)} \times x_{(4)}$ and is denoted by "e401k.age_0".

We use the same data set, matching the descriptive statistics and OLS coefficients in Lu and Wooldridge (2020)'s Table 2 and 3 respectively (the OLS standard errors don't match because we report the HC3 version). We report in Table 10 the various estimates and standard errors (in parentheses) for the coefficients of this regression model. We use Lu and Wooldridge (2020)'s parametric model $\omega^2(x;\gamma) = \exp(x'\gamma)$. Lu and Wooldridge (2020) showed big gains in precision by WLS over OLS, and then further improvement over WLS by their GMM estimator. Our results in Table 10 of course confirm these findings of Lu and Wooldridge (2020). Additionally, our results also demonstrate that even further gains, and often substantial ones, in precision over all those estimators can be obtained by our proposed estimators.

$V(u\|x)$	n	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
	50	1.037	1.037	1.021	1.064	1.017	1.078	1.081	1.108
Case	100	1.023	1.023	1.016	1.036	1.012	1.057	1.045	1.067
(1a)	200	1.014	1.014	1.008	1.016	1.007	1.034	1.026	1.043
	400	1.004	1.004	1.004	1.009	1.003	1.014	1.015	1.023
	50	.954	.974	.958	.964	.946	.996	.972	1.001
Case	100	.934	.943	.942	.938	.934	.967	.955	.980
(1b)	200	.921	.922	.926	.922	.922	.934	.930	.951
	400	.909	.909	.911	.908	.909	.916	.914	.928
	50	.756	.764	.768	.762	.760	.771	.791	.825
Case	100	.729	.729	.733	.731	.732	.739	.744	.768
(1c)	200	.729	.729	.730	.730	.730	.729	.733	.746
	400	.715	.715	.715	.715	.715	.710	.711	.717
	50	.430	.430	.431	.426	.434	.414	.444	.459
Case	100	.379	.379	.379	.374	.380	.362	.382	.387
(1d)	200	.360	.360	.360	.360	.359	.348	.358	.357
	400	.339	.339	.339	.339	.338	.328	.335	.328
	50	.618	.618	.622	.615	.620	.547	.605	.584
Case	100	.594	.594	.594	.579	.594	.483	.554	.479
(2a)	200	.589	.589	.589	.588	.588	.491	.552	.450
	400	.569	.569	.569	.563	.567	.461	.524	.418
	50	.409	.409	.410	.417	.416	.335	.413	.380
Case	100	.357	.357	.357	.336	.360	.249	.340	.277
(2b)	200	.333	.333	.333	.300	.334	.203	.310	.218
	400	.315	.315	.315	.282	.315	.201	.286	.187
	50	.274	.274	.274	.101	.278	.060	.257	.227
Case	100	.197	.197	.197	.060	.198	.036	.176	.141
(2c)	200	.170	.170	.170	.048	.170	.026	.147	.097
	400	.153	.153	.153	.048	.153	.028	.125	.070
	50	.860	.890	.872	.877	.866	.896	.896	.942
Case	100	.848	.856	.860	.854	.853	.886	.880	.912
(3a)	200	.827	.828	.829	.830	.829	.840	.844	.859
	400	.815	.815	.816	.815	.817	.818	.818	.826
	50	.682	.694	.693	.687	.690	.702	.720	.759
Case	100	.673	.675	.674	.672	.677	.682	.688	.711
(3b)	200	.651	.651	.651	.648	.653	.652	.651	.667
	400	.634	.634	.634	.632	.635	.634	.633	.638
	50	.964	.968	.968	.978	.954	1.009	.979	1.014
Case	100	.959	.959	.967	.958	.953	.996	.970	.989
(4a)	200	.923	.923	.933	.924	.922	.940	.932	.948
	400	.922	.922	.926	.921	.920	.924	.926	.932
	50	.795	.808	.819	.813	.794	.819	.827	.860
Case	100	.772	.773	.784	.775	.768	.781	.781	.802
(4b)	200	.762	.762	.765	.761	.757	.765	.763	.778
	400	.736	.736	.736	.734	.731	.732	.731	.736

Table 3: Ratio of MSE of estimators with respect to MSE of OLS estimator of $h(\beta) := \beta_2$ based on 10000 Monte Carlo trials under the design of Romano and Wolf (2017) and using their Model 2.

$V(u\|x)$	n	OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
	50	5.0	5.4	5.4	5.5	6.5	5.5	7.2	7.0	8.5
Case	100	5.3	5.5	5.5	5.6	6.0	5.6	6.5	6.4	7.1
(1a)	200	4.9	5.1	5.1	5.0	5.2	5.0	5.6	5.5	6.1
	400	4.8	5.1	5.1	5.1	5.2	5.1	5.4	5.3	5.6
	50	4.7	5.1	5.4	5.3	5.8	5.3	7.0	6.1	6.8
Case	100	4.8	5.0	5.1	5.1	5.3	5.0	6.4	5.6	6.5
(1b)	200	5.2	5.2	5.2	5.3	5.3	5.3	5.7	5.6	6.4
	400	4.8	4.8	4.8	4.9	4.9	4.9	5.2	5.2	5.4
	50	4.9	5.0	5.2	5.2	5.6	5.3	6.5	6.1	7.5
Case	100	5.0	5.0	5.0	5.0	5.4	5.3	5.9	5.8	6.7
(1c)	200	5.2	5.1	5.1	5.1	5.4	5.1	5.6	5.4	6.1
	400	5.2	5.2	5.2	5.2	5.2	5.3	5.4	5.3	5.5
	50	5.3	5.2	5.2	5.2	5.9	5.7	6.5	6.0	7.2
Case	100	5.5	5.2	5.2	5.2	5.4	5.3	5.8	5.3	5.9
(1d)	200	5.0	4.7	4.7	4.7	5.0	4.8	5.2	4.7	5.3
	400	5.1	4.9	4.9	4.9	4.8	4.9	5.3	5.0	5.3
	50	4.9	4.9	4.9	5.0	5.5	5.2	5.9	6.0	6.4
Case	100	5.1	5.1	5.1	5.1	5.5	5.3	5.5	5.6	5.3
(2a)	200	5.0	5.0	5.0	5.0	5.2	5.1	5.5	5.3	5.0
	400	4.7	4.9	4.9	4.9	5.0	5.0	5.0	5.0	4.6
	50	5.3	5.3	5.3	5.4	8.0	5.7	9.4	6.1	6.0
Case	100	5.1	5.4	5.4	5.4	6.7	5.6	7.0	5.3	4.6
(2b)	200	4.8	4.9	4.9	4.9	5.9	5.1	5.8	5.0	4.0
	400	5.1	4.7	4.7	4.7	6.0	4.9	5.6	4.6	3.7
	50	5.7	5.8	5.8	5.8	6.0	6.0	7.1	5.0	4.6
Case	100	5.0	5.2	5.2	5.2	5.0	5.2	4.8	3.9	3.0
(2c)	200	5.2	4.9	4.9	4.9	4.6	4.9	4.5	3.9	2.2
	400	5.3	5.2	5.2	5.2	5.1	5.2	4.9	3.8	2.0
	50	4.8	5.2	5.5	5.4	5.9	5.4	6.6	6.2	7.2
Case	100	5.0	5.4	5.6	5.6	5.9	5.7	6.7	6.3	7.0
(3a)	200	4.9	4.8	4.8	4.9	5.0	4.9	5.3	5.3	5.7
	400	4.5	4.7	4.7	4.7	4.7	4.7	4.8	4.9	5.2
	50	5.4	5.4	5.6	5.5	6.0	5.6	6.6	6.6	7.8
Case	100	4.9	5.6	5.6	5.6	5.8	5.7	6.0	6.3	7.1
(3b)	200	5.2	5.1	5.1	5.1	5.2	5.2	5.4	5.5	6.0
	400	5.3	5.4	5.4	5.4	5.5	5.4	5.6	5.6	5.6
	50	4.9	5.3	5.4	5.5	5.8	5.6	6.9	6.5	7.6
Case	100	5.3	5.5	5.5	5.8	6.1	5.8	7.0	6.5	7.3
(4a)	200	4.6	4.7	4.7	4.9	4.8	4.8	5.3	5.2	5.7
	400	5.1	5.2	5.2	5.3	5.3	5.3	5.4	5.4	5.8
	50	5.1	4.9	5.0	5.5	5.8	5.5	6.5	6.3	7.6
Case	100	5.3	5.5	5.5	5.7	6.0	5.7	6.4	6.5	7.3
(4b)	200	5.3	5.2	5.2	5.2	5.2	5.3	5.7	5.6	6.4
	400	5.3	5.0	5.0	5.0	5.3	5.2	5.3	5.3	5.6

Table 4: Empirical size (in %) of 5% Wald test for $h(\beta) := \beta_2$ based on 10000 Monte Carlo trials under the simulation design of Romano and Wolf (2017) and using their Model 2.

$V(u\|x)$	n	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
	50	.77	.77	.77	.75	.76	.70	.74	.71
Case	100	.76	.76	.76	.74	.75	.68	.72	.68
(2a)	200	.76	.76	.76	.76	.76	.69	.73	.67
	400	.75	.75	.75	.75	.75	.68	.72	.65
	50	.63	.63	.63	.58	.62	.50	.62	.58
Case	100	.58	.58	.58	.53	.58	.46	.57	.52
(2b)	200	.57	.57	.57	.52	.57	.43	.55	.49
	400	.56	.56	.56	.51	.55	.43	.53	.45
	50	.50	.50	.50	.30	.50	.23	.50	.46
Case	100	.43	.43	.43	.24	.43	.19	.44	.40
(2c)	200	.41	.41	.41	.22	.41	.16	.40	.34
	400	.39	.39	.39	.22	.39	.17	.37	.30

Table 5: Ratio of the average length of confidence intervals of $h(\beta) := \beta_2$ using each estimators with respect to the average length of confidence intervals using OLS. Results are based on 10000 Monte Carlo trials under the design of Romano and Wolf (2017) and using their Model 2.

$h(\beta)$	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
β_1	.607		.607	.502	.607	.505	.487	.468
β_2	.667	same	.668	.554	.667	.553	.555	.479
β_3	.506	as	.506	.341	.506	.343	.373	.332
β_4	.496	WLS	.497	.350	.497	.355	.340	.319
β_5	.930		.920	.888	.909	.904	.820	.773

Table 6: Ratio of MSE of estimators with respect to MSE of OLS estimator of coefficients based on 10000 Monte Carlo trials under the empirical design of Romano and Wolf (2017) [c.f. their Table C7] and using their Model 1 that, in their Table C7, performed noticeably better than Model 2.

$h(\beta)$	OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
β_1	4.7 (1)	5.1 (.778)		5.1 (.778)	7.2 (.655)	5.1 (.778)	7.4 (.655)	5.4 (.690)	7.2 (.636)
β_2	4.7 (1)	4.8 (.813)	same	4.8 (.813)	5.9 (.721)	4.8 (.812)	5.8 (.720)	5.3 (.736)	5.8 (.664)
β_3	5.0 (1)	4.9 (.713)	as	4.9 (.713)	6.3 (.564)	4.9 (.713)	6.3 (.565)	5.3 (.609)	6.4 (.555)
β_4	4.2 (1)	4.7 (.707)	WLS	4.7 (.707)	8.2 (.538)	4.7 (.707)	8.9 (.533)	5.0 (.585)	7.5 (.524)
β_5	4.9 (1)	5.3 (.952)		5.4 (.944)	5.6 (.920)	5.4 (.939)	5.9 (.927)	5.6 (.882)	6.1 (.838)

Table 7: Empirical size (in %) of 5% Wald test for coefficients based on 10000 Monte Carlo trials under the empirical design of Romano and Wolf (2017) [c.f. their Table C8] and using their Model 1 that, in their Table C8, performed noticeably better than Model 2.Withiin parenthesis is presented the ratio of the average length of confidence interval for each $h(\beta)$ using each estimators with respect to the average length of confidence interval of that $h(\beta)$ using OLS from this same Monte Carlo experiment.

$V(u\|x)$	$h(\beta)$	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
	β_1	.812	.812	.796	.559	.787	.512	.637	.433
Case	β_2	.978	.978	.923	.809	.904	.889	.847	.674
(1)	β_3	.871	.872	.841	.661	.829	.697	.718	.553
	β_4	1.551	1.551	1.001	.895	1.001	.989	.845	.790
	β_1	1.453	1.453	.894	.547	.790	.578	.418	.402
Case	β_2	1.389	1.389	.954	.730	.849	.746	.807	.796
(2)	β_3	2.240	2.240	.914	.553	.800	.608	.439	.432
	β_4	1.329	1.329	.891	.664	.785	.665	.621	.558
	β_1	.867	.867	.867	.872	.866	.872	.827	.830
Case	β_2	.796	.796	.792	.773	.789	.794	.763	.716
(3)	β_3	.807	.807	.809	.804	.806	.861	.793	.880
	β_4	.963	.963	.963	.950	.959	.964	.875	.799
	β_1	.169	.169	.169	.179	.169	.179	.177	.211
Case	β_2	.073	.073	.073	.095	.073	.094	.089	.135
(4)	β_3	.073	.073	.073	.099	.073	.098	.089	.136
	β_4	.120	.120	.120	.137	.120	.140	.132	.160
	β_1	.795	.795	.787	.622	.784	.564	.650	.465
Case	β_2	.975	.975	.948	.799	.923	.777	.881	.588
(1)	β_3	.860	.860	.850	.673	.837	.669	.755	.508
	β_4	1.583	1.583	1.000	.900	1.000	.999	.850	.768
	β_1	1.713	1.713	.986	.569	.853	.589	.438	.412
Case	β_2	1.507	1.507	.999	.742	.890	.739	.866	.710
(2)	β_3	2.880	2.880	.992	.588	.881	.610	.447	.407
	β_4	1.570	1.570	.979	.685	.832	.685	.631	.559
	β_1	.875	.875	.875	.873	.874	.875	.831	.784
Case	β_2	.793	.793	.792	.784	.793	.783	.781	.612
(3)	β_3	.807	.807	.807	.804	.807	.804	.800	.822
	β_4	.959	.959	.959	.955	.958	.957	.874	.784
	β_1	.156	.156	.156	.158	.156	.158	.158	.164
Case	β_2	.064	.064	.064	.075	.064	.074	.070	.096
(4)	β_3	.066	.066	.066	.077	.066	.075	.072	.088
	β_4	.111	.111	.111	.115	.112	.116	.115	.121

Table 8: Ratio of MSE of estimators with respect to MSE of OLS estimator of various $h(\beta)$'s based on 10000 Monte Carlo trials under the design of Lu and Wooldridge (2020). The top panel (above the horizontal line) corresponds to sample size $n = 1000$, and the bottom panel to $n = 5000$. The parametric model $\omega^2(x;\gamma)$ is correctly specified for $V(u|x)$ under Case 4.

$V(u\|x)$	$h(\beta)$	OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
	β_1	4.7	5.0	5.0	5.0	6.3	5.0	5.8	5.3	7.6
Case	β_2	5.1	5.4	5.4	5.7	7.3	5.6	9.5	5.9	11.8
(1)	β_3	4.9	5.2	5.2	5.2	6.6	5.2	8.4	5.8	10.4
	β_4	5.2	4.9	4.9	5.2	5.4	5.3	5.4	5.4	6.4
	β_1	4.9	5.2	5.2	5.9	6.8	5.4	6.9	5.2	6.7
Case	β_2	4.7	4.9	4.9	5.2	6.2	5.2	6.7	5.2	11.0
(2)	β_3	5.4	5.2	5.2	6.1	6.8	5.8	7.5	5.2	8.6
	β_4	5.1	4.5	4.5	4.7	5.0	5.1	5.0	5.1	5.4
	β_1	5.1	5.0	5.0	5.0	5.6	5.0	6.1	5.1	8.0
Case	β_2	5.0	5.1	5.1	5.2	5.7	5.3	6.3	5.5	9.8
(3)	β_3	5.2	5.0	5.0	5.0	5.5	5.0	6.8	5.1	11.1
	β_4	5.1	5.0	5.0	5.1	5.2	5.1	5.6	5.3	5.6
	β_1	4.5	5.2	5.2	5.2	6.8	5.2	6.8	3.8	6.7
Case	β_2	4.8	5.0	5.0	5.0	12.9	5.0	13.2	3.8	12.2
(4)	β_3	5.2	5.4	5.4	5.4	13.0	5.4	13.3	3.6	11.6
	β_4	4.8	5.3	5.3	5.3	9.1	5.3	9.7	3.5	6.8
	β_1	4.9	5.0	5.0	5.0	5.6	5.0	5.4	4.9	6.0
Case	β_2	5.0	4.9	4.9	5.1	5.3	5.2	5.8	5.3	7.3
(1)	β_3	4.9	4.8	4.8	4.8	4.9	4.8	5.4	4.9	6.4
	β_4	5.2	4.9	4.9	5.2	5.0	5.2	5.2	5.1	5.3
	β_1	5.0	4.8	4.8	5.5	5.5	5.2	5.7	5.3	5.4
Case	β_2	5.0	4.3	4.3	5.1	5.5	4.9	5.5	5.1	6.5
(2)	β_3	5.4	4.4	4.4	5.4	5.1	5.1	5.3	5.0	5.3
	β_4	5.2	4.9	4.9	5.3	5.4	5.5	5.4	5.3	5.2
	β_1	5.2	5.0	5.0	5.0	5.2	5.1	5.4	5.2	5.7
Case	β_2	5.4	5.5	5.5	5.5	5.5	5.6	5.5	5.6	6.1
(3)	β_3	4.9	4.9	4.9	4.9	5.1	4.9	5.1	5.0	6.7
	β_4	5.0	5.2	5.2	5.2	5.3	5.2	5.3	5.3	5.1
	β_1	5.3	5.4	5.4	5.4	5.7	5.4	5.8	4.6	5.3
Case	β_2	4.9	5.3	5.3	5.3	9.1	5.3	9.2	4.2	8.8
(4)	β_3	5.6	5.0	5.0	5.0	8.6	5.0	8.5	3.7	7.3
	β_4	4.9	5.1	5.1	5.1	6.0	5.1	6.2	4.0	5.1

Table 9: Empirical size (in %) of 5% Wald test for $h(\beta) := \beta_j$ for $j = 1, \ldots, 4$ based on 10000 Monte Carlo trials under the simulation design of Lu and Wooldridge (2020). The top panel (above the horizontal line) corresponds to sample size $n = 1000$, and the bottom panel to $n = 5000$. The parametric model $\omega^2(x; \gamma)$ is correctly specified for $V(u|x)$ under Case 4.

$h(\beta)$	OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
constant	5.905	6.393			6.176	6.350	6.074	6.615	6.205
	(2.115)	(.978)			(.911)	(.961)	(.910)	(.922)	(.889)
inc_0	.633	.463			.472	.482	.473	.502	.458
	(.152)	(.063)			(.055)	(.061)	(.055)	(.056)	(.050)
inc_0^2	.000	.003			.002	.003	.002	.002	.002
	(.005)	(.002)			(.002)	(.002)	(.002)	(.002)	(.002)
age_0	.704	.605			.581	.608	.581	.676	.595
	(.141)	(.087)			(.076)	(.087)	(.076)	(.075)	(.068)
age_0^2	.031	.011			.004	.011	.006	.013	.009
	(.014)	(.005)	same		(.004)	(.005)	(.004)	(.004)	(.004)
$\text{inc}_0.\text{age}_0$	.044	.026	as		.027	.027	.028	.031	.029
	(.013)	(.006)			(.005)	(.006)	(.005)	(.005)	(.004)
e401k	6.346	6.770	WLS		5.758	6.647	5.758	7.400	4.762
	(2.022)	(1.844)			(1.432)	(1.807)	(1.432)	(1.540)	(1.039)
male	1.799	1.505			1.558	1.517	1.580	1.656	1.131
	(1.959)	(.756)			(.531)	(.752)	(.523)	(.740)	(.558)
e401k.inc_0	.307	.258			.216	.265	.226	.309	.206
	(.216)	(.128)			(.089)	(.125)	(.087)	(.112)	(.081)
e401k.age_0	.154	.160			.228	.160	.228	.161	.161
	(.262)	(.120)			(.102)	(.118)	(.102)	(.116)	(.101)

Table 10: Estimates and standard errors (in parentheses) of regression coefficients in the financial wealth equation in Lu and Wooldridge (2020)'s empirical application [c.f. their Table 3].

B Appendix B: Technical Proofs

B.1 Proof of Propositions 1, 2, and Remarks 1 to 5 in Section 3.2-3.3:

For the sake of notational simplicity, all computations are made in the joint asymptotic Gaussian distribution of estimators of β, without making explicit any notation of limit in distribution. We are interested in CC and MCC estimators:

$$\begin{aligned}\hat{h}_\lambda &= h\left((1-\lambda)\,\hat{\beta}_1 + \lambda\hat{\beta}_2\right),\\ \hat{h}_A &= h\left((I_p - A)\,\hat{\beta}_1 + A\hat{\beta}_2\right).\end{aligned}$$

After asymptotic expansion:

$$\begin{aligned}\hat{h}_\lambda &= \delta' U(\lambda), \quad U(\lambda) = \hat{\beta}_1 - \lambda\left(\hat{\beta}_1 - \hat{\beta}_2\right)\\ \hat{h}_A &= \delta' U(A), \quad U(A) = \hat{\beta}_1 - A\left(\hat{\beta}_1 - \hat{\beta}_2\right).\end{aligned}$$

If we find A^* such that:

$$Var\left(U\left(A^*\right)\right) \ll Var\left(U\left(A\right)\right) \ \forall A$$

with inequalities in the sense of positive semi-definite matrices, we can be sure to have defined a minimum for:

$$Var\left(\hat{h}_A\right) = \delta' Var\left(U\left(A\right)\right)\delta.$$

Hence, we have an optimal MCC estimator from multivariate regression coefficients:

$$A^* = Cov\left(\hat{\beta}_1, \hat{\beta}_1 - \hat{\beta}_2\right)\left[Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)\right]^{-1}, \quad \hat{h}_{A^*} = \delta'\left[\hat{\beta}_1 - A^*\left(\hat{\beta}_1 - \hat{\beta}_2\right)\right].$$

By contrast, there does not exist in general a real number λ^* such that:

$$Var\left[U\left(\lambda^*\right)\right] \ll Var\left[U\left(\lambda\right)\right] \ \forall\lambda. \tag{27}$$

The optimal CC is defined from an optimal number λ^* that depends on the target δ:

$$\lambda^* = \arg\min_{\lambda\in\mathbb{R}} \delta' Var\left[U\left(\lambda\right)\right]\delta = \arg\min_{\lambda\in\mathbb{R}}\left[\delta'\hat{\beta}_1 - \lambda\delta'\left(\hat{\beta}_1 - \hat{\beta}_2\right)\right].$$

Hence, we have an optimal CC estimator from univariate regression coefficient of $\delta'\hat{\beta}_1$ on $\left[\delta'\hat{\beta}_1 - \delta'\hat{\beta}_2\right]$:

$$\begin{aligned}\lambda^* &= Cov\left(\delta'\hat{\beta}_1, \delta'\hat{\beta}_1 - \delta'\hat{\beta}_2\right)\left[Var\left(\delta'\hat{\beta}_1 - \delta'\hat{\beta}_2\right)\right]^{-1}\\ \hat{h}_{\lambda^*} &= \delta'\hat{\beta}_1 - \lambda^*\delta'\left(\hat{\beta}_1 - \hat{\beta}_2\right).\end{aligned}$$

While λ^* does not solve (27) in general, it does solve if:

$$\delta' A^* = \lambda^* \delta'$$

meaning that δ is an eigenvector of $A^{*\prime}$ with eigenvalue λ^*. Otherwise, we have in general:

$$Var\left(\hat{h}_{A^*}\right) < Var\left(\hat{h}_{\lambda^*}\right).$$

The optimal TMCC estimator is strictly more accurate than the optimal TCC.

However, if for all $\delta \in \mathbb{R}^p$, no TCC estimator based on CC of the two estimators $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ can improve upon $h\left(\hat{\beta}_1\right)$, it is also the case for TMCC estimators. To see that, note that if $\hat{h}_{A^*}$ stands for our optimal TMCC and for all eigenvectors δ of $A^{*\prime}$ if $\lambda^*(\delta)$ stands for the corresponding eigenvalue, then:

$$\begin{aligned} Var\left(\hat{h}_{\lambda^*(\delta)}\right) &= \delta' Var\left(U\left(A^*\right)\right)\delta = \delta'\left\{Var\left(\hat{\beta}_1\right) - Var\left(A^*\left(\hat{\beta}_1 - \hat{\beta}_2\right)\right)\right\}\delta \\ &= \delta' Var\left(\hat{\beta}_1\right)\delta - \delta' A^* Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)A^{*\prime}\delta \\ &= \delta' Var\left(\hat{\beta}_1\right)\delta - \lambda^*\delta' Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)\lambda^*\delta. \end{aligned}$$

The fact that no TCC estimator based on CC of the two estimators $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ can improve upon $h\left(\hat{\beta}_1\right)$ means that:

$$Var\left(\hat{h}_{\lambda^*(\delta)}\right) = \delta' Var\left(\hat{\beta}_1\right)\delta \implies \lambda^*\delta' Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)\lambda^*\delta = 0 \implies \lambda^*\delta = 0 \implies A^{*\prime}\delta = 0.$$

Since this must be true for all eigenvectors δ of $A^{*\prime}$, we conclude:

$$A^* = 0.$$

No TMCC can improve upon $h\left(\hat{\beta}_1\right)$. In other words, the optimal TCC is $h\left(\hat{\beta}_1\right)$ for all possible target $h(\beta)$ if and only if:

$$Cov\left(\hat{\beta}_1, \hat{\beta}_1 - \hat{\beta}_2\right) = 0.$$

Since (see Section 3.1.) estimators are one-to-one linear functions of moment conditions, this can be rewritten with obvious simplified notations:

$$Cov\left(g_1, G_1^{-1}g_1 - G_2^{-1}g_2\right) = 0.$$

This can be rewritten:

$$\Omega_{11}G_1^{-1\prime} = \Omega_{12}G_2^{-1\prime} \iff G_2^{-1}\Omega_{21} = G_1^{-1}\Omega_{11} \iff \Omega_{21} = G_2G_1^{-1}\Omega_{11} \iff G_2 = \Omega_{21}\Omega_{11}^{-1}G_1.$$

B.2 Proof of Proposition 3:

We first compute the matrix of multivariate regression coefficients:

$$A^* \;=\; Cov\left(\hat{\beta}_1, \hat{\beta}_1 - \hat{\beta}_2\right)\left[Var\left(\hat{\beta}_1 - \hat{\beta}_2\right)\right]^{-1} \;=\; \left[\Sigma_{11} - \Sigma_{12}\right]\left[\Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{21}\right]^{-1}$$

where Σ stands for the joint asymptotic variance matrix of the couple $\left(\hat{\beta}_1', \hat{\beta}_2'\right)'$ of estimators:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

For efficient GMM, with weighting matrix $W = \left[\Omega\left(\beta^0\right)\right]^{-1}$, we have shown that it is an MCC of the two estimators with a matrix of weights:

$$\begin{aligned} \Delta &= \left[G'WG\right]^{-1}\left[G_1'W_{12}G_2 + G_2'W_{22}G_2\right] \\ &= \left[\Sigma^{-1}\right]^{-1}\left[\Sigma^{12} + \Sigma^{22}\right] \\ &= \left[\Sigma^{11} + \Sigma^{21} + \Sigma^{12} + \Sigma^{22}\right]^{-1}\left[\Sigma^{12} + \Sigma^{22}\right] \end{aligned}$$

with the notations:

$$\Sigma^{-1} \;=\; \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} = G'WG \;=\; G_1'W_{11}G_1 + G_1'W_{12}G_2 + G_2'W_{21}G_1 + G_2'W_{22}G_2.$$

Hence, to prove Proposition 3, we need to check that, when $W = \left[\Omega\left(\beta^0\right)\right]^{-1}$:

$$\left[\Sigma_{11} - \Sigma_{12}\right]\left[\Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{21}\right]^{-1} \;=\; \left[\Sigma^{11} + \Sigma^{21} + \Sigma^{12} + \Sigma^{22}\right]^{-1}\left[\Sigma^{12} + \Sigma^{22}\right].$$

We know (see discussion in Section 3.2.) that we can assume without loss of generality that $\sqrt{n}\bar{g}_{1,n}$ and $\sqrt{n}\bar{g}_{2,n}$ are asymptotically independent, or equivalently the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are asymptotically independent. In this case, matrices Σ and Σ^{-1} are block diagonal and, to prove Proposition 3, we only need to check that:

$$\Sigma_{11}\left[\Sigma_{11} + \Sigma_{22}\right]^{-1} = \left[\Sigma^{11} + \Sigma^{22}\right]^{-1}\Sigma^{22}$$

that is:

$$\left[\Sigma^{11} + \Sigma^{22}\right]\Sigma_{11} = \Sigma^{22}\left[\Sigma_{11} + \Sigma_{22}\right]$$

which is obvious, since in case of block-diagonality:

$$\Sigma^{11} = \left(\Sigma_{11}\right)^{-1},\; \Sigma^{22} = \left(\Sigma_{22}\right)^{-1}.$$