Cross-Validation Selection of Regularisation Parameter(s) for Semiparametric Transformation Models

Senay Sokullu
Sami Stouli

Discussion Paper 16 / 672

21 March 2016

Revised 8 November 2017

Department of Economics
University of Bristol
Priory Road Complex
Bristol BS8 1TU
United Kingdom

Cross-Validation Selection of Regularisation Parameter(s) for Semiparametric Transformation Models*

Senay Sokullu† University of Bristol

Sami Stouli‡ University of Bristol

November 8, 2017

*We thank the editors Jean-Pierre Florens and Anna Simoni, two anonymous referees and Samuele Centorrino for helpful comments.

†senay.sokullu@bristol.ac.uk

‡s.stouli@bristol.ac.uk

Abstract

We propose cross-validation criteria for the selection of regularisation parameter(s) in the semiparametric instrumental variable transformation model proposed in Florens and Sokullu (2016). In the presence of an endogenous regressor, this model is characterized by the need to choose two regularisation parameters, one for the structural function and one for the transformation of the outcome. We consider two-step and simultaneous criteria, and analyze the finite-sample performance of the estimator using the corresponding regularisation parameters by means of several Monte-Carlo simulations. Our numerical experiments show that simultaneous selection of regularisation parameters provides significant improvements in the performance of the estimator. We also apply our methods to the choice of regularisation parameters in the estimation of two-sided network effects in the German magazine industry.

Keywords: Nonparametric IV Regression, Transformation models, Cross-Validation, Tikhonov Regularisation, Ill-posed inverse problems
JEL Classification: C14; C26; L14

1 Introduction

The semiparametric transformation model is a flexible specification of data generating mechanisms. Recent work by Florens and Sokullu (2016) proposes a nonparametric instrumental variables (NPIV) treatment of this model, allowing for a nonlinear relationship between the endogenous variable and an unrestricted (i.e. nonlinear and potentially non-monotone) transformation of the outcome of interest. This setup can be applied to a wide variety of economic problems, and recent applications include estimation of demand functions in two-sided markets (Sokullu, 2016b), estimation of demand in differentiated products markets (Berry and Haile, 2014), or the analysis of duration data (Abbring and van den Berg, 2003; Honore and Paula, 2010). In all of these economic problems, endogeneity is a crucial feature of the model, and allowing for nonlinear structural relationships is an essential component of reliable empirical econometric practice.

The main challenge in general NPIV procedures is addressing the so-called ill-posed inverse problem. This arises from expressing the unknowns of the model (the parameters) as the solution to a singular system of equations. In a finite-dimensional setting, this formulation requires the inversion of a nonsingular matrix in order to recover the parameters of interest. For infinite-dimensional settings one of the solutions adopted in the literature (see Newey and Powell, 2003; Darolles, Fan, Florens, and Renault, 2011; Horowitz, 2011) has been to regularize the problem by a method similar to ridge regression: the objective function should be penalized in order to circumvent ill-posedness.

Regularisation of ill-posed inverse problems necessitates selection of a tuning parameter which determines the degree of regularisation. It is of crucial importance since it balances the fitting and smoothing of the estimated functional parameters. Various selection methods have been considered in the literature, such as the discrepancy rule (Feve and Florens, 2010; Florens and Sokullu, 2016), truncation (Horowitz, 2011), and cross-validation (Centorinno, 2015). Except for Florens and Sokullu (2016), none of these papers consider a semiparametric transformation model, which is characterized by the need for selecting two different regularisation parameters: one for the transformation of the outcome, and one for the structural function. The main challenge in such a framework stems from the fact that these two different regularisation parameters should converge to zero at the same rate. Florens and Sokullu (2016) get over this challenge by assuming a constant ratio between the two parameters and constructing a two-step selection procedure.

In this paper, we first propose two selection methods based on cross-validation, then explore their relative performance compared to the two-step discrepancy rule method introduced in Florens and Sokullu (2016), as well as a simultaneous implementation of the

discrepancy rule method. The first method is a two-step approach which replaces the discrepancy rule criterion by a cross-validation criterion. Our second approach is a simultaneous cross-validation criterion which determines the values of the regularisation parameters of both the transformation and the structural function, in one step. We further describe an iterative procedure for the minimization of the simultaneous cross-validation criterion. We provide numerical evidence that cross-validation improves on the discrepancy rule (and its simultaneous implementation), and that simultaneous selection has good finite sample properties. This complements the recent contribution of Centorinno (2015). Although we show that one-step (simultaneous) selection provides the best finite sample properties among other selection methods, whether the parameters are converging to zero at the same rate is an open question that we leave for future work.

In the next section we describe the model analyzed in Florens and Sokullu (2016). In Section 3, we introduce our cross-validation criteria. In Section 4, we provide small-sample analysis by means of several Monte Carlo simulations. In Section 5, we apply our methods to the estimation of two-sided network effects in the German magazine industry. Section 6 concludes.

2 The Model

We consider the general model in Florens and Sokullu (2016). It is a semiparametric transformation model of the form:

$$H(Y) = \varphi(Z) + X\beta + U, \quad \mathbb{E}(U|X, W) = 0, \tag{1}$$

where Y and Z are endogenous variables, X is a vector of exogenous variables and W is a vector of instruments. $H(.)$ and $\varphi(.)$ are unknown functions to be estimated, along with the finite-dimensional parameter β. In this model, one element of the vector β needs to be normalized to 1 for identification purposes. The model can then be written as

$$H(Y) = \varphi(Z) + X_0 + X_1'\beta + U, \tag{2}$$

where $Y, Z, X_0, U \in \mathbb{R}$, $X = \{X_0, X_1\} \in \mathbb{R}^q$ and $W \in \mathbb{R}^p$. The variables Y, Z, X, W generate a random vector Λ with a cumulative distribution function F which is characterised by its square integrable density $f(y, z, x, w)$ with respect to Lebesgue measure. We denote by $L^2_F(Y), L^2_F(Z), L^2_F(X)$ and $L^2_F(W)$ the spaces of square integrable functions of Y, Z, X and W, respectively, with respect to the corresponding marginal of F. We assume that $L^2_F(Y), L^2_F(Z), L^2_F(X)$ and $L^2_F(W)$ are subspaces of a common Hilbert space denoted by L^2_F.

Florens and Sokullu (2016) show that the functions $H(.)$ and $\varphi(.)$ as well as the parameter vector β are identified. Below we present the assumptions needed for identification and state the theorem. For the proof, we refer the reader to Florens and Sokullu (2016).

Assumption 1 *There exist two square integrable functions H and φ such that:*

$$H(Y) = \varphi(Z) + X_0 + X_1'\beta + U$$

with

$$\mathbb{E}[U|X, W] = 0.$$

Assumption 2 ***Completeness.*** *The distribution of (Y, Z) given (X, W) is complete in the following sense:*

$$\forall m(Y, Z) \in L_F^2(Y \times Z), \quad \mathbb{E}[m(Y, Z)|X, W] = 0 \quad a.s. \quad \Rightarrow \quad m(Y, Z) = 0 \quad a.s.$$

Assumption 3 ***Conditional Additive Completeness.*** $\forall(m_1, m_2, \beta) \in L_F^2(Y) \times L_F^2(Z) \times \mathbb{R}^q \quad \mathbb{E}(m_1(Y) + m_2(Z) + X_1'\beta|X, W) = 0 \quad a.s. \Rightarrow m_1(Y) + m_2(Z) + X_1'\beta \quad a.s.$

Assumption 4 ***Separability.*** *Y and Z are measurably separable i.e., $\forall m(Y) \in L_F^2(Y)$ and $\forall l(Z) \in L_F^2(Z)$:*

$$m(Y) = l(Z) \Rightarrow m(.) = l(.) = constant.$$

Assumption 5 *(Y, Z) and X_1 are measurably separable:*

$$m(Y, Z) = l(X_1) \Rightarrow m(.) = l(.) = constant.$$

Assumption 6 ***Normalisation.*** *If $\varphi(Z)$ is constant a.s. then $\varphi(Z) = 0$ a.s. For simplicity, we will assume that $\varphi(.)$ is normalized by the condition $\mathbb{E}[\varphi(Z)] = 0$. We then consider as the parameter space:*

$$\mathcal{E}_0 = \{(H, \varphi) \in L_F^2(Y) \times L_F^2(Z) : \mathbb{E}[\varphi(Z)] = 0\}.$$

Assumption 7 *Let Σ_{X_1} denote the variance of X_1. Then, Σ_{X_1} is positive definite.*

Assumption 1 defines the model. One of the novelty of this model is that neither $H(Y)$ nor $\varphi(Z)$ needs to be monotone. In contrast to the previous literature on transformation models, this model allows the transformation to be nonmonotone. Assumptions 2 and 3 are completeness assumptions. The completeness assumption is standard in the NPIV literature and primitive conditions that lead to completeness have recently been analyzed in

D' Haultfoeuille (2011), Hu and Shiu (2011) and Andrews (2011). Intuitively, Assumption 2 means that (Y, Z) and (X, W) are sufficiently correlated while Assumption 3 means that (Y, Z, X_1) are sufficiently correlated with (X, W). Assumptions 4 and 5 are also standard in NPIV literature. Assumption 4 means that there is not an exact relationship between Y and Z, while Assumption 5 implies the absence of an exact relationship between (Y, Z) and X_1. Both assumptions are satisfied if $X_0 + U$ is not equal to a constant. A more detailed discussion of measurable separability can be found in Florens, Heckman, Meghir, and Vytlacil (2008). Assumption 6 is a normalisation assumption and Assumption 7 implies that the variance-covariance matrix of X_1 is positive definite.

Proposition 1 *(Theorem 8 in Florens and Sokullu (2016)) Under Assumptions 1-7, the functions $H(Y)$ and $\varphi(Z)$ and the parameter β are identified.*

For NPIV estimation of this model, define the operator:

$$T : \mathcal{E}_0 = \left\{ L_F^2(Y) \times \tilde{L}_F^2(Z) \right\} \mapsto L_F^2(X, W) : T(H, \varphi) = \mathbb{E}[H(Y) - \varphi(Z)|X, W],$$

where $\tilde{L}_F^2(Z) = \{\varphi \in L_F^2(Z) | \mathbb{E}(\varphi) = 0\}$. As Assumption 6 implies the normalisation $\mathbb{E}(\varphi(Z)) = 0$, the space $\tilde{L}_F^2(Z)$ only contains zero-mean functions. We also define the inner product

$$\langle (H_1(Y), \varphi_1(Z)), (H_2(Y), \varphi_2(Z)) \rangle_{L_F^2(Y) \times L_F^2(Z)} = \langle H_1(Y), H_2(Y) \rangle_{L_F^2(Y)} + \langle \varphi_1(Z), \varphi_2(Z) \rangle_{L_F^2(Z)},$$

where $\langle g(x), h(x) \rangle = \int_X g(x)h(x)f_X(x)dx$. The adjoint operator of T, T^*, satisfies

$$\langle T(H(Y), \varphi(Z)), \psi(X, W) \rangle_{L_F^2(X,W)} = \langle (H(Y), \varphi(Z)), T^*\psi(X, W) \rangle_{\mathcal{E}_0},$$

for any $(H, \varphi) \in \mathcal{E}_0$ where $\mathcal{E}_0 = \left\{ L_F^2(Y) \times \tilde{L}_F^2(Z) \right\}$ and $\psi \in L_F^2(X, W)$. This equality then gives the adjoint operator T^*:

$$T^*\psi = (\mathbb{E}[\psi(X, W)|Y], -\mathbb{P}\mathbb{E}[\psi(X, W)|Z]),$$

where $\mathbb{P}$ is the projection operator from $L_F^2(Z)$ onto $\tilde{L}_F^2(Z)$.[1] Further define the operator $T_X : \mathbb{R}^{q-1} \to L_F^2(X, W) : \beta \mapsto X_1'\beta$. Its adjoint is defined as $T_X^* : L_F^2(X, W) \to \mathbb{R}^{q-1} : g \mapsto \mathbb{E}[X_1 g(X, W)]$, following from the equality:

$$\langle T_X\beta, g(X, W) \rangle_{L_F^2(X,W)} = \langle \beta, T_X^* g(X, W) \rangle_{\mathbb{R}^{q-1}}.$$

[1] See Appendix A for the derivation of the adjoint operator T^*.

Given the above definitions, the unknowns (H, φ, β) solve a system of normal equations. Writing

$$T(H, \varphi) - T_X \beta = X_0, \tag{3}$$

the normal equations are

$$T^* T(H, \varphi) - T^* T_X \beta = T^* X_0 \tag{4}$$

$$T_X^* T(H, \varphi) - T_X^* T_X \beta = T_X^* X_0. \tag{5}$$

One can obtain $\beta = (T_X^* T_X)^{-1} T_X^* T(H, \varphi) - (T_X^* T_X)^{-1} T_X^* X_0$ from Equation (5). It can then be substituted into (4) to get:

$$(T^*(I - P_X)T)(H(Y), \varphi(Z)) = T^*(I - P_X)X_0, \tag{6}$$

where $P_X = T_X(T_X^* T_X)^{-1} T_X^*$ and I is the identity operator on $L_F^2(X, W)$. In order to obtain the functions $(H(Y), \varphi(Z))$, $(T^*(I - P_X)T)$ in (6) needs to be inverted. However, note that the operator T is infinite-dimensional and since $f(y, z, x, w)$ is square integrable, T is compact. Hence it has infinitely many eigenvalues in the neighbourhood of zero: the inverse of $(T^*(I - P_X)T)$ is discontinuous and causes an ill-posed inverse problem.[2,3] In order to be able to solve this ill-posed inverse problem we need to regularize it. In this paper, following Florens and Sokullu (2016), we adopt Tikhonov Regularisation. The functions (H, φ) are thus given by:

$$(H(Y), \varphi(Z)) = (\alpha I + T^*(I - P_X)T)^{-1} T^*(I - P_X)X_0, \tag{7}$$

where α is the regularisation parameter which is strictly positive and converges to zero at a suitable rate as the sample size increases. Equation (7) can then be rewritten as[4]:

$$\begin{pmatrix} \alpha_H H + \mathbb{E}\left[(I - P_X)\mathbb{E}(H|X, W)|Y\right] - \mathbb{E}\left[(I - P_X)\mathbb{E}(\varphi|X, W)|Y\right] \\ -\alpha_\varphi \varphi + \mathbb{PE}\left[(I - P_X)\mathbb{E}(H|X, W)|Z\right] - \mathbb{PE}\left[(I - P_X)\mathbb{E}(\varphi|X, W)|Z\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[(I - P_X)X_0|Y\right] \\ \mathbb{PE}\left[(I - P_X)X_0|Z\right] \end{pmatrix}. \tag{8}$$

The system of equations in (8) form the basis of our estimation strategy. In order to get estimates of H and φ, conditional expectations can be replaced by their empirical counterparts, i.e., by kernel estimators. In fact, the implementation of this method has already been discussed in detail in papers such as Darolles, Fan, Florens, and Renault (2011),

[2]Compact operators have infinitely many countable eigenvalues hence we can write the singular value decomposition of T, see Theorems 7.22 and 7.23 in Ryanne and Youngson (2008) and Theorem 2.31 in Carrasco, Florens, and Renault (2007).

[3]For more information on ill-posed inverse problem in the case of NPIV, see Darolles, Fan, Florens, and Renault (2011); Horowitz (2011).

[4]In Equation (8) we denote $H(Y)$ by H and $\varphi(Z)$ by φ for the sake of exposition.

Feve and Florens (2010) and Sokullu (2016b).

To explain the implementation of the defined method, consider an i.i.d. sample of $(y_i, z_i, x_i, w_i), i = 1, ..., N$. Let K_y, K_z, K_x and K_w be kernel functions chosen according to the dimension of Y, Z, X and W, respectively, and such that the technical conditions in Appendix B are satisfied, with associated bandwidth parameters h_y, h_z, h_x and h_w. Define the matrix $A_{xw}(w)$ whose (i,j)th element is:

$$A_{xw}(w)(i,j) = \frac{K_x\left(\frac{x_i - x_j}{h_x}\right) K_w\left(\frac{w - w_j}{h_w}\right)}{\sum_j K_x\left(\frac{x_i - x_j}{h_x}\right) K_w\left(\frac{w - w_j}{h_w}\right)}.$$

Moreover, let A_y and A_z be the matrices with (i,j)th elements:

$$A_y(i,j) = \frac{K_y\left(\frac{y_i - y_j}{h_y}\right)}{\sum_j K_y\left(\frac{y_i - y_j}{h_y}\right)} \quad \text{and} \quad A_z(i,j) = \frac{K_z\left(\frac{z_i - z_j}{h_z}\right)}{\sum_j K_z\left(\frac{z_i - z_j}{h_z}\right)}.$$

Let P be the matrix with $\frac{N-1}{N}$ on the diagonal and $-\frac{1}{N}$ elsewhere. Denoting by $\hat{P}_X$ the sample analog of P_X, the empirical counterpart of Equation (8) can be written as:

$$\begin{pmatrix} \alpha_H H + A_y(I - \hat{P}_X)A_{xw}H - A_y(I - \hat{P}_X)A_{xw}\varphi \\ -\alpha_\varphi \varphi + PA_z(I - \hat{P}_X)A_{xw}H - PA_z(I - \hat{P}_X)A_{xw}\varphi \end{pmatrix} = \begin{pmatrix} A_y(I - \hat{P}_X)X_0 \\ PA_z(I - \hat{P}_X)X_0 \end{pmatrix}. \tag{9}$$

For notational simplicity, we leave the dependence of the regularisation parameter $\alpha = (\alpha_H, \alpha_\varphi)$ on N implicit. The estimators $(\hat{H}, \hat{\varphi})$ are then given by:

$$\begin{pmatrix} \hat{H} \\ \hat{\varphi} \end{pmatrix} = \begin{pmatrix} \alpha_H I + A_y(I - \hat{P}_X)A_{xw} & -A_y(I - \hat{P}_X)A_{xw} \\ PA_z(I - \hat{P}_X)A_{xw} & -(\alpha_\varphi I + PA_z(I - \hat{P}_X)A_{xw}) \end{pmatrix}^{-1} \begin{pmatrix} A_y(I - \hat{P}_X)X_0 \\ PA_z(I - \hat{P}_X)X_0 \end{pmatrix}. \tag{10}$$

It should be noted that the estimates $(\hat{H}, \hat{\varphi})$ can also be obtained by using sieve approximation, see Horowitz (2011) and Blundell, Chen, and Kristensen (2007) among others. In this paper, we focus on kernel-based estimators.

Florens and Sokullu (2016) show that under some regularity conditions the estimators are consistent and $\hat{\beta}$ is asymptotically normal. We present the regularity conditions and the result of Florens and Sokullu (2016) in Appendix B and refer the reader to Florens and Sokullu (2016) for more details and for the proof.

3 Selection of Regularisation Parameter(s)

One of the key issues in the practical implementation of Tikhonov Regularised NPIV estimation is the selection of the regularisation parameter α. The regularisation parameter plays a very important role in the estimation as it balances the fitting and the smoothing. An arbitrary selection rule might result in highly oscillatory curves if it is picked too low, or it may result in very flat curves if it is picked too high.

Selection of regularisation parameter in NPIV problems has already been studied by Feve and Florens (2010); Darolles, Fan, Florens, and Renault (2011); Centorinno (2015) among others. Feve and Florens (2010) and Darolles, Fan, Florens, and Renault (2011) extend *the discrepancy rule* proposed by Morozov (1993) and Engl, Hanke, and Neubauer (1996) and suggest a data driven selection method. As explained in Engl, Hanke, and Neubauer (1996), the discrepancy principle is based on the comparison between the residual of the functional equation and the assumed bound for the noise level. Moreover, the regularisation parameter defined by this rule is shown to be convergent and of optimal order. Both Feve and Florens (2010) and Darolles, Fan, Florens, and Renault (2011) use the idea of minimizing a function of squared norm of residuals. The squared norm of residuals cannot be used directly and must be transformed as it reaches its minimum at $\alpha = 0$. Hence in these papers this function is constructed by taking the squared norm of residuals obtained from an estimation using iterated Tikhonov Regularisation of order 2 and then dividing this norm by α (Feve and Florens, 2010) or by α^2 (Darolles, Fan, Florens, and Renault, 2011).[5]

On the other hand, Centorinno (2015) uses cross-validation to select the regularisation parameter. He uses leave-one-out estimators to construct the cross-validation criterion function and his method does not require the use of iterated Tikhonov Regularisation of order 2, or division of the norm of residuals by any function of α. He shows that the regularisation parameter obtained with this method converges to zero at the optimal rate.

Deriving a selection rule for α in a semiparametric transformation model such as the one in Florens and Sokullu (2016), is not straightforward as one has to pick 2 regularisation parameters to estimate the model, see Equation (8). These two regularisation parameters for two unknown functions need not necessarily be the same, although they need to converge to zero at the same rate as the functions $H(Y)$ and $\varphi(Z)$ are estimated simultaneously. Florens and Sokullu (2016) show that $H(Y)$ and $\varphi(Z)$ can be estimated consistently and converge to their true values at the same rate. Since estimates of these functions depend on each other and since their rate of convergence is the same, we need α_H and α_φ to converge to zero at the

[5]The qualification of Tikhonov Regularisation is 2 and in case of estimation of a very regular function this prevents improvement of the convergence rate. The use of residuals obtained from a regression with iterated Tikhonov Regularisation of order 2 is especially done for cases where the function is very regular.

same rate, too. α's converging to zero at different rates would affect the rate of convergence of the functions. Florens and Sokullu (2016) proposed a method to get over this problem. They first assume that there is a constant ratio between two regularisation parameters, i.e. $\alpha_\varphi = c\alpha_H$ for $c > 0$, and then propose to choose optimal values for α_H and c in two steps.

The regularisation parameter α_H is chosen as if one is estimating a function of (Y, Z) instead of 2 separate functions $H(Y)$ and $\varphi(Z)$. It is then replaced in the original estimating equation in order to optimize over c. Let $G : L^2_F(Y, Z) \mapsto \mathbb{R}$ be the function defined as:

$$G(Y, Z) = H(Y) - \varphi(Z).$$

Define also the operators $T_G : L^2_F(Y, Z) \mapsto L^2_F(X, W) : T_G G = \mathbb{E}[G(Y, Z)|X, W]$. The adjoint T^*_G is defined as: $T^*_G : L^2_F(X, W) \mapsto L^2_F(Y, Z) : T^*_G \phi = \mathbb{E}[\phi(X, W)|Y, Z]$, with sample analog $\hat{T}_G$ and $\hat{T}^*_G$.

Since $H(Y) - \varphi(Z) = X_0 + X_1'\beta + U$, we can write:

$$T_G G(Y, Z) = X_0 + T_X \beta,$$

which leads to the normal equations:

$$T^*_G T_G G(Y, Z) = T^*_G X_0 + T^*_G T_X \beta \tag{11}$$

$$T^*_X T_G G(Y, Z) = T^*_X X_0 + T^*_X T_X \beta. \tag{12}$$

As is already done in the estimation of the model, Equation (12) is used to get an expression for β which is then substituted in Equation (11). From (12):

$$\beta = (T^*_X T_X)^{-1}(T^*_X T_G G(Y, Z) - T^*_X X_0),$$

then one obtains

$$T^*_G T_G G(Y, Z) = T^*_G X_0 + T^*_G T_X (T^*_X T_X)^{-1} T^*_X T_G G(Y, Z) - T^*_G T_X (T^*_X T_X)^{-1} T^*_X X_0.$$

Hence the estimate $\hat{G}^\alpha_{(1)}$ is given by:

$$\hat{G}^\alpha_{(1)} = (\alpha_H I + \hat{T}^*_G (I - \hat{P}_X)\hat{T}_G)^{-1} \hat{T}^*_G (I - \hat{P}_X) X_0, \tag{13}$$

with $\hat{P}_X = \hat{T}_X (\hat{T}^*_X \hat{T}_X)^{-1} \hat{T}^*_X$. The iterated Tikhonov regularized estimator of order 2 is given by:

$$\hat{G}^\alpha_{(2)} = (\alpha_H I + \hat{T}^*_G (I - \hat{P}_X)\hat{T}_G)^{-1} (\hat{T}^*_G (I - \hat{P}_X) X_0 + \alpha_H \hat{G}^\alpha_{(1)}),$$

whose vector of residuals can be written as:

$$\hat{u}^{\alpha}_{(2)} = \hat{T}^{*}_{G}(I - \hat{P}_X)X_0 - (\hat{T}^{*}_{G}(I - \hat{P}_X)\hat{T}_G)\hat{G}^{\alpha}_{(2)}.$$

The optimal α_H in Florens and Sokullu (2016) is then defined as

$$\alpha^{*}_{H} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{\alpha^2}\|\hat{u}^{\alpha}_{(2)}\|^2. \tag{14}$$

In the second step α^{*}_{H} is replaced in the original problem below and the optimal c is chosen as the minimizer of the squared norm of residuals.[6]

$$\begin{pmatrix} \alpha^{*}_{H}H + \mathbb{E}\left[(I - P_X)\mathbb{E}(H|X,W)|Y\right] - \mathbb{E}\left[(I - P_X)\mathbb{E}(\varphi|X,W)|Y\right] \\ -\alpha^{*}_{H}c\varphi + \mathbb{PE}\left[(I - P_X)\mathbb{E}(H|X,W)|Z\right] - \mathbb{PE}\left[(I - P_X)\mathbb{E}(\varphi|X,W)|Z\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[(I - P_X)X|Y\right] \\ \mathbb{PE}\left[(I - P_X)X|Z\right] \end{pmatrix}. \tag{15}$$

As already mentioned, in this paper we extend Centorinno (2015)'s cross-validation selection of regularisation parameter to the case of semi-parametric transformation models. We consider two main extensions: In the first one, we replicate the 2-step method proposed by Florens and Sokullu (2016), but we use a cross-validation criterion to select α^{*}_{H} and c. In the second one, we propose to select both α^{*}_{H} and α^{*}_{φ} simultaneously in one step by minimizing a cross-validation criterion obtained from the original problem. The performance of two additional variations are also studied numerically: 1. extension of the discrepancy rule to simultaneous selection of α_H and α_φ, and 2. iterative minimization of the one-step cross-validation criterion.

3.1 Two-step cross-validated selection of regularisation parameter

The two-step cross-validated selection of regularisation parameter follows closely the selection rule introduced in Florens and Sokullu (2016). The only difference is that we use leave-one-out estimators to construct the cross-validation criterion function. Note that the first step estimator of the function $G(Y,Z)$ is given by Equation (13). Define the leave-one-out matrices $A^{-i}_{xw}(w)$ and $A^{-i}_{yz}(z)$ with (i,j)th elements:

$$A^{-i}_{xw}(w)(i,j) = \frac{K_x\left(\frac{x_i - x_j}{h_x}\right)K_w\left(\frac{w - w_j}{h_w}\right)}{\sum_{j\neq i} K_x\left(\frac{x_i - x_j}{h_x}\right)K_w\left(\frac{w - w_j}{h_w}\right)}, \quad for \quad i \neq j$$

[6] In Equation (15) we denote $H(Y)$ by H and $\varphi(Z)$ by φ for the sake of exposition.

$$A_{yz}^{-i}(z)(i,j) = \frac{K_y\left(\frac{y_i - y_j}{h_y}\right) K_z\left(\frac{z - z_j}{h_z}\right)}{\sum_{j \neq i} K_y\left(\frac{y_i - y_j}{h_y}\right) K_z\left(\frac{z - z_j}{h_z}\right)}, \quad for \quad i \neq j$$

with diagonal elements set to zero. Hence when it is applied it gives exactly the same formula for leave-one-out estimator as in Li and Racine (2006) on page 69. For instance,

$$A_{yz}^{-i}(z)X_0 = \begin{pmatrix} 0 & a_{12} & a_{13} & ... & a_{1N} \\ a_{21} & 0 & a_{23} & ... & . \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ a_{N1} & a_{N2} & a_{N3} & ... & 0 \end{pmatrix} \begin{pmatrix} x_{10} \\ x_{20} \\ . \\ . \\ . \\ x_{N0} \end{pmatrix} = \begin{pmatrix} \sum_{j \neq 1} a_{1j} x_{j0} \\ \sum_{j \neq 2} a_{2j} x_{j0} \\ . \\ . \\ . \\ \sum_{j \neq N} a_{Nj} x_{j0} \end{pmatrix}.$$

Then the leave-one-out estimator of G in Equation (13) is given by:

$$\hat{G}_{-i}^{\alpha}(Y,Z) = (\alpha_H I + A_{yz}^{-i}(I - \hat{P}_X)A_{xw}^{-i})^{-1} A_{yz}^{-i}(I - \hat{P}_X)X_0,$$

which leads to the following cross-validation criterion function:

$$CV_1(\alpha) = \sum_{i=1}^{N} [\hat{T}_G^*(I - \hat{P}_X)\hat{T}_G \hat{G}_{-i}^{\alpha}(Y,Z) - \hat{T}_G^*(I - \hat{P}_X)X_0]^2. \tag{16}$$

The cross-validated α_H, α_{CV} is the minimizer of the cross-validation criterion in (16). As in Florens and Sokullu (2016) the second step consists in replacing the α_{CV} in the original problem below, and estimating H and φ by using leave-one-out operators for different values of c.

$$\begin{pmatrix} \hat{H}_{-i} \\ \hat{\varphi}_{-i} \end{pmatrix} = \begin{pmatrix} \alpha_{CV} I + A_y^{-i}(I - \hat{P}_X)A_{xw}^{-i} & -A_y^{-i}(I - \hat{P}_X)A_{xw}^{-i} \\ PA_z^{-i}(I - \hat{P}_X)A_{xw}^{-i} & -(c\alpha_{CV} I + PA_z^{-i}(I - \hat{P}_X)A_{xw}^{-i}) \end{pmatrix}^{-1} \begin{pmatrix} A_y^{-i}(I - \hat{P}_X)X_0 \\ PA_z^{-i}(I - \hat{P}_X)X_0 \end{pmatrix}. \tag{17}$$

The second step cross-validation function is given by:

$$CV_2(c) = \sum_{i=1}^{N} [(\hat{T}^*(I - \hat{P}_X)\hat{T})(\hat{H}_{-i}(Y), \hat{\varphi}_{-i}(Z)) - \hat{T}^*(I - \hat{P}_X)X_0]^2. \tag{18}$$

The cross-validated c, denoted c_{CV}, is then defined as the minimizer of Equation (18).

3.2 One-step cross-validated selection of regularisation parameters

One issue with having two regularisation parameters in the model we are considering is that although the two parameters may differ, they should converge to zero at the same rate. This is the reason why Florens and Sokullu (2016) propose to have a constant ratio between α_H and α_φ, although they do not consider cross-validation criteria. Following Centorinno (2015), we propose minimizing a cross-validation criterion over α_H and α_φ simultaneously, and compare the performance of this approach to the two-step selection methods described above. A formal proof that the cross-validated α's are converging to zero at the same rate is left for future work.

The leave-one-out estimation of (H, φ) is given by:

$$\begin{pmatrix} \hat{H}_{-i} \\ \hat{\varphi}_{-i} \end{pmatrix} = \begin{pmatrix} \alpha_H I + A_y^{-i}(I - \hat{P}_X)A_{xw}^{-i} & -A_y^{-i}(I - \hat{P}_X)A_{xw}^{-i} \\ PA_z^{-i}(I - \hat{P}_X)A_{xw}^{-i} & -(\alpha_\varphi I + PA_z^{-i}(I - \hat{P}_X)A_{xw}^{-i}) \end{pmatrix}^{-1} \begin{pmatrix} A_y^{-i}(I - \hat{P}_X)X_0 \\ PA_z^{-i}(I - \hat{P}_X)X_0 \end{pmatrix}, \tag{19}$$

which leads to the cross-validation criterion function:

$$CV(\alpha_H, \alpha_\varphi) = \sum_{i=1}^{N} [(\hat{T}^*(I - \hat{P}_X)\hat{T})(H_{-i}(Y), \varphi_{-i}(Z)) - \hat{T}^*(I - \hat{P}_X)X_0]^2. \tag{20}$$

The optimal values of α_H and α_φ are then those which minimize Equation (20):

$$(\alpha_H^*, \alpha_\varphi^*) = \underset{\alpha_H, \alpha_\varphi}{\operatorname{argmin}} \, CV(\alpha_H, \alpha_\varphi).$$

This is a two-dimensional minimization problem. We consider two methods in order to implement this minimization. First, we evaluate criterion (20) over a two-dimensional grid, and select the pair of parameter values which yields the smallest objective value. As a faster alternative, we also experiment with an iterative procedure which, at a given step m, proceeds by (i) evaluating criterion (20) over a one-dimensional grid for α_H given a value $\alpha_\varphi^{(m-1)}$, and selecting the optimal value $\alpha_H^{(m)}$, and (ii) evaluating criterion (20) over a one-dimensional grid for α_φ given $\alpha_H^{(m)}$, and selecting the optimal value $\alpha_\varphi^{(m)}$. We then iterate until convergence of the sum of squared differences $(\alpha_H^{(m)} - \alpha_H^{(m-1)})^2 + (\alpha_\varphi^{(m)} - \alpha_\varphi^{(m-1)})^2 \leq \tau$, for some specified tolerance τ.

The intuition behind our proposal is that the simultaneous criterion should be minimized by regularisation parameter values that are *jointly* optimal for both functions $H(.)$ and $\varphi(.)$. This may not be the case for the discrepancy rule or the two-step cross-validation criterion

which select a regularisation parameter which is optimal for the function $G(.)$. Moreover, we conjecture that regularisation parameters chosen jointly by simultaneous cross-validation converge to zero at the same rate, whereas simultaneous selection by the discrepancy rule might not preserve this property. In the next Section we provide numerical evidence that simultaneous cross-validation does perform better than other methods, and in particular than the discrepancy rule, as well as its simultaneous implementation: we implement the simultaneous discrepancy rule by mimimizing the sum of squared norms of residuals (scaled by $\alpha_H^2 + \alpha_\varphi^2$) formed with the estimates

$$\begin{pmatrix} \hat{H} \\ \hat{\varphi} \end{pmatrix} = \begin{pmatrix} \alpha_H I + A_y(I - \hat{P}_X)A_{xw} & -A_y(I - \hat{P}_X)A_{xw} \\ PA_z(I - \hat{P}_X)A_{xw} & -(\alpha_\varphi I + PA_z(I - \hat{P}_X)A_{xw}) \end{pmatrix}^{-1} \begin{pmatrix} A_y(I - \hat{P}_X)X_0 \\ PA_z(I - \hat{P}_X)X_0 \end{pmatrix}.$$

4 Numerical simulations

In this Section we study the sensitivity of the finite-sample performance of the estimator to the method chosen for selecting the regularisation parameters. We compare the two cross-validation methods introduced as well as the iterative variant of the one-step approach, to the method proposed in Florens and Sokullu (2016) based on the discrepancy principle. We also include a simultaneous implementation of the discrepancy principle, which minimizes the sum of squared norms of residuals of both steps of the original two-step implementation. The smoothing parameters are held fixed throughout the simulations, but we provide two further sets of simulation results in Appendix C.4 for different choices of bandwidth parameters.[7]

We study the performance of each selection method across two different data generating processes. For comparison purposes, the simulation data generating process of Florens and Sokullu (2016) is taken as our initial setup. We generate 499 samples of size $N = 100, 200$ and 400, from the semiparametric transformation model (**Design 1**):

$$\begin{aligned} \log\left(\frac{1-Y}{Y}\right) &= (Z^2 - \mathbb{E}(Z^2)) + X_0 + 0.3X_1 + U \\ Z &= 0.2W + \eta_W \\ \eta_W &= 0.5U + \varepsilon_W. \end{aligned}$$

The covariates X_0, X_1 and the instrumental variable W are drawn from a standard uniform distribution, and the disturbance U is taken normally distributed with mean 0 and variance $(X_0 + X_1 + W)/50$. In addition, ε_W is normally distributed with mean 0 and variance 0.4.

[7]The choice of bandwidth parameter values is fixed throughout simulations and chosen using Silverman's rule-of-thumb, see Florens and Sokullu (2016) for details about the bandwidth choice.

Florens and Sokullu (2016) provide additional discussion of this experimental design.

We also consider a second data generating process where only the specification of the $\varphi(.)$ function is altered. In this second design (**Design 2**), the specified model is:

$$\begin{aligned} \log\left(\frac{1-Y}{Y}\right) &= (\exp(-|Z|) - E(\exp(-|Z|))) + X_0 + 0.3X_1 + U \\ Z &= 0.2W + \eta_W \\ \eta_W &= 0.5U + \varepsilon_W. \end{aligned}$$

This model allows for studying the robustness of our numerical findings to the smoothness of $\varphi(.)$.

Following Centorinno (2015), the performance of each method is assessed relative to the performance of the estimator using the optimal α. Let $||\cdot||_p$ denote the L^p norm, where for $f : \mathbb{R} \to \mathbb{R}$, $||f||_p = \left\{\int_{\mathbb{R}} |f(s)|^p ds\right\}^{1/p}$. In addition, define the rectangular grid $\mathcal{G} = \mathcal{G}_1 \times \mathcal{G}_2$, with $\mathcal{G}_1 = \{0.0001, ..., 0.01\}$ and $\mathcal{G}_2 = \{0.001, ..., 0.1\}$, respectively, two logarithmically spaced grids of 20 elements. In the context of the semiparametric transformation model (2), we consider optimal values of α defined as the minimizer of the following simultaneous criteria

$$\alpha_p^* = \arg\min_{\alpha \in \mathcal{G}} ||\hat{\varphi}_\alpha - \varphi||_p^p + ||\hat{H}_\alpha - H||_p^p + ||\hat{\beta}_\alpha - \beta||^2, \quad p = 1, 2, \infty.$$

Thus the optimal regularisation parameter is chosen such that estimates of both the nonparametric and the parametric parts of the model are (jointly) optimal. We vary the choice of norm in the definition of α^* in order to assess the robustness of each method under different metrics.

We then compare the estimated parameters obtained by each of our five methods to those obtained using the optimal regularisation parameter α_p^* using the deviation statistics

$$DEV_p(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}}) = \frac{||\hat{\varphi}_{\hat{\alpha}} - \varphi||_p^p + ||\hat{H}_{\hat{\alpha}} - H||_p^p + ||\hat{\beta}_{\hat{\alpha}} - \beta||^2}{||\hat{\varphi}_{\alpha_p^*} - \varphi||_p^p + ||\hat{H}_{\alpha_p^*} - H||_p^p + ||\hat{\beta}_{\alpha_p^*} - \beta||^2}, \quad p = 1, 2, \infty,$$

where $\hat{\alpha}$ is the estimator of α obtained by the discrepancy principle (Disc.R), simultaneous implementation of the discrepancy principle (Disc.R 2), two-step cross-validation (CV1), simultaneous cross-validation (CV2), and iterated cross-validation (It.CV2).[8]

Our main simulation results for Design 1 are summarized in Tables 1–3. These tables show

[8]For both the discrepancy rule and CV1, we consider a grid of values $\mathcal{G}_\alpha = \mathcal{G}_1$ following the original implementation in Florens and Sokullu (2016). For the simultaneous cross-validation and simultaneous discrepancy rule procedures, we use $\mathcal{G}$ as our grid. Simulations with a finer and wider grid yield similar qualitative results.

the average and standard deviation of $DEV_p(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$, $p = 1, 2, \infty$, across simulations, for each sample sizes and each of the five selection methods.

The main feature of the results is that the simultaneous cross-validation method dominates the other methods both in terms of average performance relative to the optimal estimator as well as of precision, with smaller standard deviation of the deviation measure across simulations, across all sample sizes. Indeed, estimates from simultaneous cross-validation exhibit much less variability across simulations, with the standard error of $DEV_2(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ ranging from 0.34 to 0.51 across sample sizes under CV2 compared to 1.52 to 2.18 for the discrepancy rule-based estimates, and 0.75 to 1.16 for CV1 estimates.

The iterated implementation of simultaneous cross-validation is very competitive, especially for $N = 400$, and represents a useful, faster, alternative for large samples due to its implementation based on a sequence of minimizations over one-dimensional grids.[9] On the other hand, the performance of the simultaneous implementation of the discrepancy rule and the two-step cross-validation method appears to be weak compared to one-step cross-validation. In particular, as shown by Tables 1-3, their performance does not improve markedly as sample size increases. This reflects the fact that both methods choose a regularisation parameter for the G function, which need not be optimal for the target function H. Thus, the one-step criterion provides a principled approach to choosing regularisation parameters in transformation models.

The cross-validation method CV1 also exhibits good finite-sample performance relative to the estimator based on the discrepancy rule. Although their respective performance is comparable in terms of DEV_1, estimates based on CV1 appear more precise in terms of the DEV_2 and DEV_∞ metrics, indicating more stability across simulations of $CV1$ estimates.

Additional insights into the relative performance of each method can be gained by visual inspection of the simulation results shown in Figure 3-5 in Appendix C.1. As reflected by the standard deviation of DEV_p, estimates obtained using simultaneous cross-validation exhibit much less variability across simulations, and most of the gains arise from increased stability in the estimation of $\varphi(Z)$, across the support of Z. This is especially the case at the extremes of the support of Z where estimates using regularisation parameters chosen by the discrepancy rule and, to a lesser extent, by CV1 potentially diverge greatly from the true function, $\varphi(Z)$. Estimates obtained based on CV2, on the other hand, exihibit greater stability across the support of Z, and at the boundaries of the support as well.

These observations are confirmed by considering the ratios of mean square errors (MSE) for each function separately: Tables 4 and 5 present MSE ratios of CV2-based estimates over Disc.R-, CV1-, and It.CV2-based estimates of $H(Y)$ and $\varphi(Z)$. Table 4 shows that whereas

[9] We thank an anonymous referee for suggesting the study of iterative implementations of our proposals.

Table 1: **Summary statistics for $DEV_2(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	2.24	2.12	2.28	0.83	1.76	1.16	1.56	0.51	1.68	0.70
$N = 200$	1.86	2.18	2.00	0.51	1.51	0.96	1.34	0.35	1.35	0.36
$N = 400$	1.81	1.52	2.36	0.60	1.56	0.75	1.31	0.34	1.30	0.36

Table 2: **Summary statistics for $DEV_1(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.42	0.26	1.45	0.24	1.25	0.23	1.18	0.17	1.26	0.21
$N = 200$	1.33	0.19	1.34	0.19	1.19	0.17	1.13	0.14	1.17	0.15
$N = 400$	1.33	0.17	1.46	0.20	1.23	0.18	1.15	0.13	1.12	0.15

Table 3: **Summary statistics for $DEV_\infty(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.74	1.40	1.31	0.47	1.28	0.83	1.12	0.32	1.26	0.49
$N = 200$	1.22	0.83	1.13	0.31	1.08	0.38	1.04	0.20	1.02	0.24
$N = 400$	1.72	1.48	1.23	0.38	1.30	0.76	1.14	0.35	1.13	0.34

gains in MSE from using CV2 are non-negligible across sample sizes and methods for $H(Y)$, the relative performance of the estimator based on CV2 is much stronger for the structural function $\varphi(Z)$, gains in MSE ranging from 16% to 65%, with the noticeable exception of It.CV2 for $N = 400$ which outperforms CV2.

Table 4: **Ratio of Mean Square Errors for $\hat{H}$ - Design 1.**

Sample size	$\frac{MSE_{CV2}(\hat{H})}{MSE_{Disc.R}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{Disc.R2}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{CV1}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{It.CV2}(\hat{H})}$
$N = 100$	79.40	86.90	93.75	100.43
$N = 200$	73.58	87.96	94.29	99.59
$N = 400$	82.36	75.59	92.93	95.12

Table 5: **Ratio of Mean Square Errors for $\hat{\varphi}$ - Design 1.**

Sample size	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{Disc.R}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{Disc.R2}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{CV1}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{It.CV2}(\hat{\varphi})}$
$N = 100$	34.19	57.83	67.54	63.22
$N = 200$	42.61	70.52	76.65	82.58
$N = 400$	48.58	76.98	76.40	113.57

Sample size	Disc. R	Disc. R 2	CV 1	CV 2	It. CV 2
$N = 100$	0.25 (0.07)	0.20 (0.06)	0.23 (0.07)	0.23 (0.06)	0.24 (0.07)
$N = 200$	0.26 (0.05)	0.21 (0.04)	0.25 (0.05)	0.25 (0.05)	0.25 (0.05)
$N = 400$	0.27 (0.03)	0.22 (0.03)	0.26 (0.03)	0.26 (0.03)	0.26 (0.03)

(a) **Average and standard errors**.

Sample size	Disc. R	Disc. R 2	CV 1	CV 2	It. CV 2
$N = 100$	7.70	14.18	9.21	9.37	8.09
$N = 200$	3.22	8.90	4.58	4.55	3.90
$N = 400$	2.06	6.88	2.82	2.45	2.44

(b) **Mean Square Error** $\times 1000$.

Table 6: **Simulation results for estimation of β. Design 1.** (a) Average of $\hat{\beta}$ estimates across simulations and selection methods. Simulation standard errors in parenthesis; (b) Mean square error across simulations $\times 1000$.

It is interesting to note that estimates of the parametric component of the model, β, appear to be much less sensitive to the method of selection of the regularisation method. Table 6(a) summarizes the average and standard deviation of estimates of β across simulations, for each method, and Table 6(b) shows the mean square error. Except for Disc.R 2, all estimators perform similarly in terms of bias and standard errors, estimates based on CV2 even underperforming slightly in terms of MSE. These results seem to indicate that, perhaps unexpectedly, estimates of β are fairly robust to the method of choice of regularisation parameter.

Overall, our results are robust across metrics and sample sizes, although the contrast between CV2 and the other methods is especially stark under DEV_2 and DEV_∞. Thus CV2 yields the best overall performance, and It.CV2 provides a useful alternative as the sample size increases. The additional simulations provided in Appendix C.2-C.3 for Design 2 yield similar qualitative conclusions, and simulations with different bandwidth parameter values in Appendix C.4 show the robustness of our results to the choice of bandwidths.

5 Two-Sided Network Effects in the German Magazine Industry

In this section we estimate the demand system in German magazine industry studied in Sokullu (2016a). Magazines are two-sided platforms which serve to readers and advertisers on each side. Readers care about the amount of ad pages in the magazines and advertisers care about the circulation rate (number of readers) of the magazine which brings about two-sided network externalities, in other words, two-sided network effects.[10] These effects play a crucial role in the pricing strategy of the magazine. If, for instance, advertisers benefit more from the existence of readers of the magazine, this may lead to prices below marginal costs for readers and ad rates well above marginal costs for advertisers. Moreover, the benefit that the two sides gets may not be linear in two-sided network effects. Although the readers may enjoy seeing advertisements in the magazine, if the number of advertising pages increases too much, the benefit of readers may decrease as a consequence. In such a case, the pricing strategy of the magazine would change according to the number of advertising pages. Hence, an anti-trust economist may arrive at erroneous conclusions if he/she cannot estimate these two-sided network effects correctly.

Sokullu (2016b) shows that the two-sided network effects are nonlinear and nonmonotone on both sides in the local American newspaper industry while, using data from the German magazine industry, Sokullu (2016a) shows that the two-sided network effects are nonlinear and nonmonotone only on readers' side. We re-estimate the demand system in German magazine industry to demonstrate the effect of selection of regularisation parameter in empirical work.

In Sokullu (2016a) readers (r) and advertisers (a) are heterogenous in their net benefit of joining the platform and these benefits are drawn from a continuous distribution. The reader i buys the magazine if his net benefits b_i^r are higher than a threshold level, $\underline{b}^r$. Similarly, advertiser j advertises in the magazine if his net benefits b_j^a are higher than a threshold level, $\underline{b}^a$. The German magazine industry is composed of several segments and in each segment there are more than one magazine. Thus, threshold benefit levels on both sides, $\underline{b}^k$, $k \in \{a, r\}$, depend on relative magazine price for side k, P^k, the number of agents (market share of the magazine) on the other side of the platform, $N^{k'}$, and unobservable magazine characteristics U^k. All agents with net benefits higher than the threshold level join the platform. Then the probability of joining the platform, and hence the market share of the magazine on side k,

[10] Note that we do not refer to network formation literature in this section. Two-sided network effects mean the externality one side exerts on the other side on a two-sided platform.

is given by:

$$N^k = P(b_i^k \geq \underline{b}^k(N^{k'}, P^k, U^k)) = 1 - F^k(\underline{b}^k(N^{k'}, P^k, U^k)), \tag{21}$$

where F^k is the cumulative distribution function of the net benefits of agents on side k. Let $S^k(.) = 1 - F^k(.)$ be the survival function. Sokullu (2016a) assumes that the threshold benefit function is partially linear:

$$\underline{b}^k = \varphi^k(N^{k'}) + P^k + U^k,$$

which leads to the demand system of readers and advertisers:

$$N^r = S^r(\varphi^r(N^a) + P^r + U^r) \tag{22}$$

$$N^a = S^a(\varphi^a(N^r) + P^a + U^a). \tag{23}$$

Assuming that the survival functions $S^a(.)$ and $S^r(.)$ are strictly decreasing. The survival functions can be inverted to get the estimating equations:

$$H^r(N^r) = \varphi^r(N^a) + P^r + U^r \tag{24}$$

$$H^a(N^a) = \varphi^a(N^r) + P^a + U^a. \tag{25}$$

We estimate the system using the same data as in Sokullu (2016a), available online at *www.medialine.de*. It contains annual data on cover prices, ad prices, number of ad pages, number of content pages, and circulation numbers of German magazines for the year 2009. The sample consists of information on 171 magazines and there are 17 different group of magazines such as actuality, DIY, women's, sports, etc. Moreover, the magazines in the sample belong to 25 different publishers. Some of the publishers own magazines only in one group while some others are publishing in several different groups.

Prices and shares of agents on the other side in equations (24) and (25) are endogenous so that instruments are needed. We use the same instruments as in Sokullu (2016a). The cover price is instrumented with the average cover price of the publisher's other magazines and the ad rate is instrumented with the number of titles of the publisher. Moreover, the share of readers in the advertising demand equation is instrumented with the average circulation rate of the publisher's other magazines and the share of advertisers in the reader demand equation is instrumented with the average number of advertising pages of the publisher's other magazines.

We estimate the demand system (24) and (25) equation by equation using NPIV estimation for transformation models developed in Florens and Sokullu (2016) and extended in

Sokullu (2016b) for the case where all the right-hand side variables are endogenous. Moreover we estimate each equation using three different regularisation parameter selection criteria. Note that the model in this section is slightly different than the one given in (1). First neither (24) nor (25) includes any finite-dimensional parameter hence it is simpler than the model given in (1). Second, all the right-hand side variables in (24) and (25) are endogenous. Sokullu (2016b) shows that the model is identified under similar assumptions as in Florens and Sokullu (2016). Below we explain the estimation of this simple model briefly using the readers' demand equation (Equation 24) only.

Denote the instruments for reader demand equation by Z^r, so that $\mathbb{E}[U^r|Z^r] = 0$. Then one can write:

$$\mathbb{E}[H^r(N^r)|Z^r] = \mathbb{E}[\varphi^r(N^a)|Z^r] + \mathbb{E}[P^r|Z^r]$$

$$\mathbb{E}[H^r(N^r) - \varphi^r(N^a)|Z^r] = \mathbb{E}[P^r|Z^r]. \tag{26}$$

Define the operator T^r as:

$$T^r : \mathcal{E}^r = \left\{ L^2_{N^r} \times \tilde{L}^2_{N^a} \right\} \mapsto L^2_{Z^r} : T^r(H^r, \varphi) = \mathbb{E}[H^r(N^r) - \varphi(N^a)|Z^r]$$

where $\tilde{L}^2_{N^a} = \{\varphi \in L^2_{N^a} : \mathbb{E}(\varphi^r) = 0\}$. As in Section 2, $\tilde{L}^2_{N^a}$ is the space of functions where $\mathbb{E}[\varphi^r] = 0$. The adjoint operator of T^r, T^{r*} follows from Section 2, as well and it is equal to:

$$T^{r*}\xi = (\mathbb{E}[\xi|N^r], -\mathbb{P}\mathbb{E}[\xi|N^a])$$

where $\mathbb{P}$ is the projection operator from $L^2_{N^a}$ onto $\tilde{L}^2_{N^a}$. Then Equation (26) can be rewritten as:

$$T^r(H^r(N^r), \varphi^r(N^a)) = f^r \tag{27}$$

where $f^r = \mathbb{E}(P^r|Z^r)$. Note that in the model given in Section 2, there is an exogenous variable X instead of the P^r. The system given in (27) is also an ill-posed inverse problem and the regularized solution of $(H^r(N^r), \varphi^r(N^a))$ can be obtained by minimization of

$$\min_{H^r(N^r), \varphi^r(N^a)} \left\{ \|T^r(H^r(N^r), \varphi^r(N^a)) - f^r\|^2 + \alpha \|(H^r(N^r), \varphi^r(N^a))\|^2 \right\}.$$

The Tikhonov regularised solution is then given by:

$$(H^r(N^r), \varphi^r(N^a))' = (\gamma^r_n I + T^{r*}T^r)^{-1} T^{r*} f^r, \tag{28}$$

where I is the identity operator in $L^2_{N^r} \times L^2_{N^a}$. The counterpart of Equation (8) for this

endogenous simple model is given by

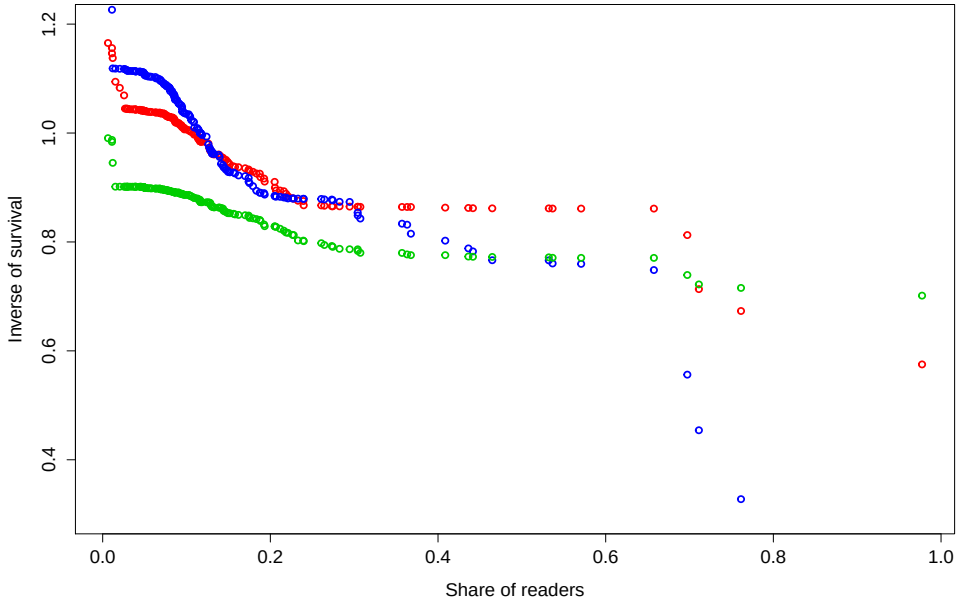
$$\begin{pmatrix} \gamma_n^r H^r + \mathbb{E}\left[\mathbb{E}(H^r|Z^r)|N^r\right] - \mathbb{E}\left[\mathbb{E}(\varphi|Z^r)|N^r\right] \\ \gamma_n^r \varphi^r - \mathbb{PE}\left[\mathbb{E}(H^r|Z^r)|N^a\right] + \mathbb{PE}\left[\mathbb{E}(\varphi|Z^r)|N^a\right] \end{pmatrix} = \begin{pmatrix} \mathbb{E}[\mathbb{E}(P^r|Z^r)|N^r] \\ -\mathbb{PE}[\mathbb{E}(P^r|Z^r)|N^a] \end{pmatrix} \tag{29}$$

The estimation strategy introduced in Section 2 can be used here too. Given an i.i.d. sample $(N_i^r, N_i^a, P_i^r, Z_i^r), i = 1, ..., N$, the expectations are replaced by their empirical counterparts and the system of equations is solved for $H^r(N_i^r)$ and $\varphi^r(N_i^a)$ for $i = 1, ..., N$. Details regarding the implementation of the model as well as its asymptotic properties can be found in Sokullu (2016b).

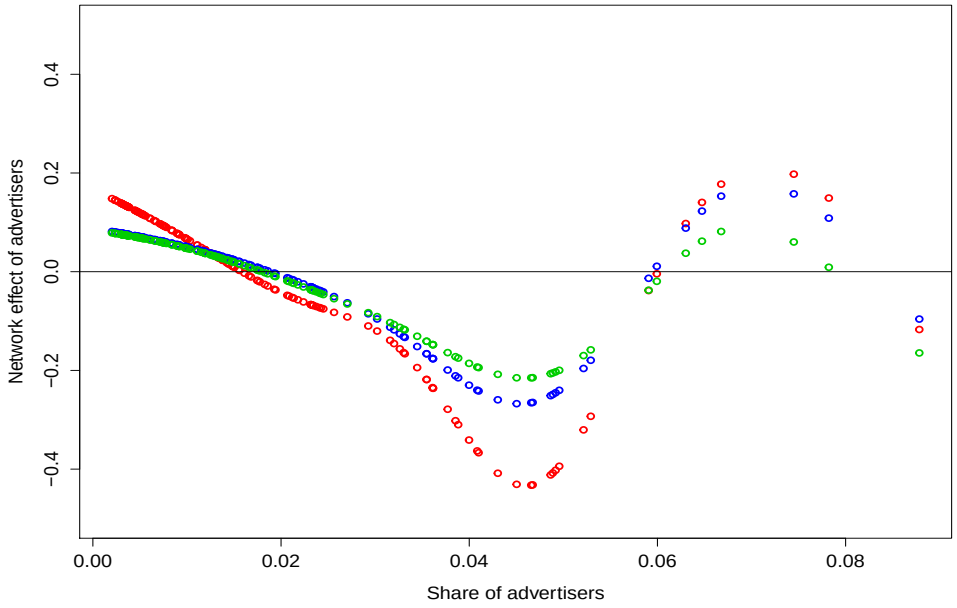
Results are presented in Figures 1 and 2. In all figures, red dots show the estimates obtained with discrepancy rule, green dots show those obtained with two-step cross-validation and finally blue ones show the estimates obtained with one-step cross-validation.

As can be seen from Figures 1a and 2a, all selection methods give similar rearranged downward sloping demand curves, especially inverse of advertiser demand is almost estimated to be the same with three methods. The more important issue in this empirical exercise is the estimation of the two-sided network effect functions. Sokullu (2016a) finds that two-sided network effects are nonlinear and nonmonotone on the readers' side (discrepancy rule), and our estimates obtained with cross-validation selection of regularisation parameter confirm this result. Indeed, Figure 1b shows that the network effects are estimated to be weaker compared to results of Sokullu (2016a) and the level of ads from which the readers start to get positive utility is a little bit higher. When we consider Figure 2b, it can be seen that the two-sided network effects are estimated to be monotonic with all methods although they are estimated to be stronger with cross-validation. In other words, benefits the advertisers get are estimated to be much higher with cross-validation selection rules. In Appendix D, Figure 9 includes estimates for reader's demand based on iterative cross-validation which are similar to those obtained with one-step cross-validation, and estimates based on simultaneous discrepancy rule, which are wiggly due to under-regularisation. Figure 10 shows that estimates for advertiser's demand based on the iterative method also differ. This is in line with results of Monte-Carlo simulations which show that for small sample sizes estimates based on CV2 and It.CV2 may differ.

To sum up, when the selected α's are not significantly different from each other, all selection rules perform similarly. Especially for an empirical application in which the nonparametric estimation is done to obtain information about the monotonicity of the function, all methods would give similar results. On the other hand, if we are more interested in the magnitude of an estimated effect, the use of a selection method with better small sample properties will be important.

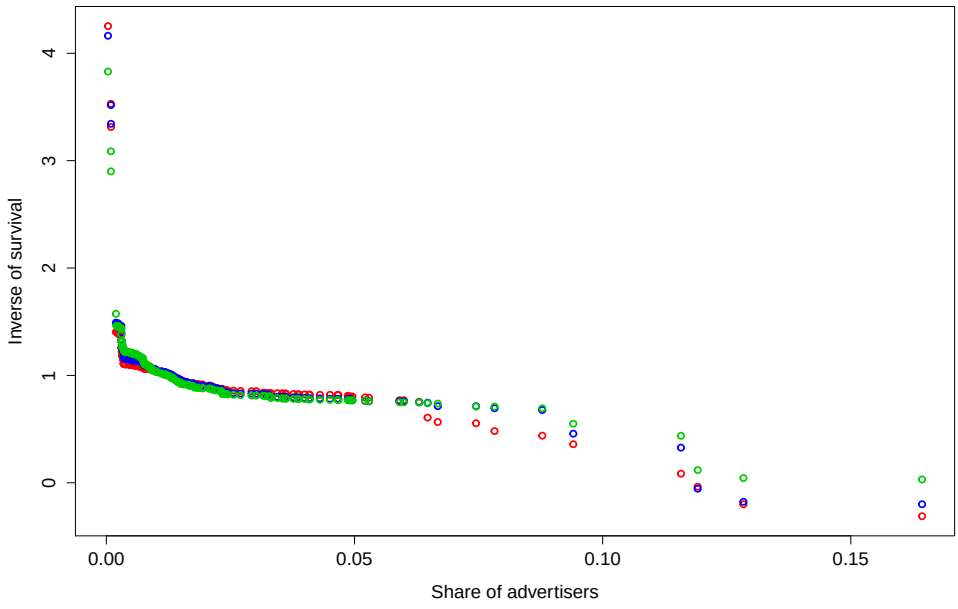


(a)

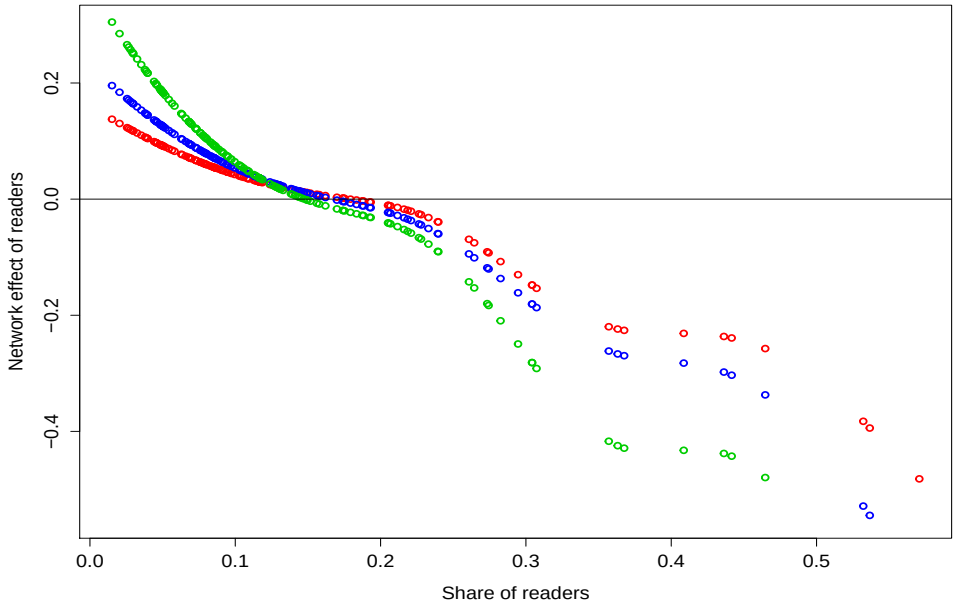


(b)

Figure 1: **Estimation of reader's demand**. (a) Inverse demand function, $H^r(N^r)$. (b) Network effect of advertisers on readers, $\varphi^r(N^a)$. Choice of regularisation parameter(s): Discrepancy rule (red), CV1 (green) and CV2 (blue).



(a)



(b)

Figure 2: **Estimation of advertiser's demand**. (a) Inverse demand function, $H^a(N^a)$. (b) Network effect of readers on advertisers, $\varphi^a(N^r)$. Choice of regularisation parameter(s): Discrepancy rule (red), CV1 (green) and CV2 (blue).

6 Conclusion

This paper proposes several criteria for the selection of regularisation parameters in semiparametric transformation models. We have provided extensive numerical simulations to study the finite sample behaviour of our various criteria.

In practice, we recommend using the one-step cross-validation criterion for choosing the regularisation parameters. In small samples, our preferred implementation requires performing the minimization over a two-dimensional grid. Our simulations show that this approach dominates alternative criteria and implementations. In large samples, our simulations show that the proposed iterative implementation provides a reliable alternative. This is useful numerically since minimization over a two-dimensional grid may be computationally demanding. The iterative approach is easy to implement, and not very sensitive to the convergence tolerance threshold.

Overall, this paper suggests that choosing regularisation parameters simultaneously in transformation models yields substantial improvements in the finite sample performance of the estimator introduced in Florens and Sokullu (2016). In future work, we will study the theoretical properties of the one-step criterion we have proposed.

References

Abbring, J. H., and G. J. van den Berg (2003): "The Nonparametric Identification of Treatment Effects in Duration Models," *Econometrica*, 71(5), 1491–1517.

Andrews, D. W. (2011): "Examples of L^2-Complete and Boundedly-Complete Distributions," Cowles Foundation Discussion Papers 1801, Cowles Foundation for Research in Economics, Yale University.

Berry, S. T., and P. A. Haile (2014): "Identification in Differentiated Products Markets Using Market Level Data," *Econometrica*, 82, 1749–1797.

Blundell, R., X. Chen, and D. Kristensen (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75(6), 1613–1669.

Carrasco, M., J.-P. Florens, and E. Renault (2007): "Linear Inverse Problems in Structural Econometrics Estimation Based on Spectral Decomposition and Regularization," in *Handbook of Econometrics*, ed. by J. Heckman, and E. Leamer, vol. 6 of *Handbook of Econometrics*, chap. 77. Elsevier.

Centorinno, S. (2015): "Data-Driven Selection of the Regularization Parameter in Additive Nonparametric Instrumental Regressions," Discussion paper, Stony Brook University.

Darolles, S., Y. Fan, J.-P. Florens, and E. Renault (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79(5), 1541–1565.

D' Haultfoeuille, X. (2011): "On The Completeness Condition In Nonparametric Instrumental Problems," *Econometric Theory*, 27(03), 460–471.

Engl, H. W., M. Hanke, and A. Neubauer (1996): *Regularization of Inverse Problems*. Kluwer Academic Publications, Dordrecht.

Feve, F., and J.-P. Florens (2010): "The Practice of Non Parametric Estimation by Solving Inverse Problems: The Example of Transformation Models," *Econometrics Journal*, 13.

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76(5), 1191–1206.

Florens, J.-P., and S. Sokullu (2016): "Nonparametric Estimation of Semiparametric Transformation Models," *Econometric Theory*, p. to appear.

Honore, B. E., and A. D. Paula (2010): "Interdependent Durations," *Review of Economic Studies*, 77(3), 1138–1163.

Horowitz, J. L. (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79, 347–394.

Hu, Y., and J.-L. Shiu (2011): "Nonparametric identification using instrumental variables: sufficient conditions for completeness," CeMMAP working papers CWP25/11, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Li, Q., and J. S. Racine (2006): *Nonparametric Econometrics: Theory and Practice*, vol. 1 of *Economics Books*. Princeton University Press.

Morozov, V. A. (1993): *Regularization Methods for Ill-Posed Problems*. CRC Press, Florida.

Newey, W. K., and J. L. Powell (2003): "Instrumental variable estimation of nonparametric models," *Econometrica*, 71(5), 1565–1578.

Ryanne, B. P., and M. A. Youngson (2008): *Linear Functional Analysis*. Springer-Verlag, London.

Sokullu, S. (2016a): "Network Effects in the German Magazine Industry," *Economics Letters*, 143, 77–79.

——— (2016b): "A Semi-Parametric Analysis of Two-Sided Markets: An Application to the Local Daily Newspapers in the USA," *Journal of Applied Econometrics*, 31, 843–864.

A Derivation of the Adjoint Operator

To compute the adjoint operator of T, T^*, let us first assume that $T : \mathcal{E} = \{L_F^2(Y) \times L_F^2(Z)\} \mapsto L_F^2(X,W)$, i.e we do not impose normalization. Then we can write:

$$\langle T(H(Y),\varphi(Z)),\psi(X,W)\rangle_{L_F^2(X,W)}$$

$$= \int \left[\int (H(y)-\varphi(Z))\frac{f(y,z,x,w)}{f_Y(y)f_Z(z)f_{XW}(x,w)} f_Y(y)f_Z(z)dydz\right]\psi(X,W)f_{XW}(x,w)dxdw$$

$$= \int \left[\int H(Y)\frac{f(y,z,x,w)}{f_Y(y)f_{XW}(x,w)} f_Y(y)dy\right]\psi(X,W)f_{XW}(x,w)dxdw$$

$$- \int \left[\int \varphi(Z)\frac{f(y,z,x,w)}{f_Z(z)f_{XW}(x,w)} f_Z(z)dz\right]\psi(X,W)f_{XW}(x,w)dxdw$$

$$= \int \underbrace{\left[\int \psi(X,W)\frac{f(y,z,x,w)}{f_Y(y)}\right]}_{\mathbb{E}[\psi(X,W)|Y]} H(Y)f_Y(y)dy - \int \underbrace{\left[\int \psi(X,W)\frac{f(y,z,x,w)}{f_Z(z)}\right]}_{-\mathbb{E}[\psi(X,W)|Z]} \varphi(Z)f_Z(z)dz.$$

Note however that our parameter space is $\mathcal{E}_0$. For this parameter space, following Lemma 3 in Florens and Sokullu (2016), T^* is given by:

$$T^*\psi = (\mathbb{E}[\psi(X,W)|Y], -\mathbb{PE}[\psi(X,W)|Z]).$$

B Consistency and Rate of Convergence

In this section we present the asymptotic properties of the estimators. These properties have already been shown in Florens and Sokullu (2016). Hence, here we present the needed assumptions and the main results and refer the reader to Florens and Sokullu (2016) for further details and proofs.

For the sake of exposition Florens and Sokullu (2016) show the asymptotic properties in two steps. In the first step, they assume that there is no finite dimensional parameter in the model and that $X \in \mathbb{R}$. Under this setting, it is easier to show the asymptotic properties of $\hat{H}(Y)$ and $\hat{\varphi}(Z)$. Then in the second step they assume that the model is the one given in Equation (1) and show the asymptotic properties of $\hat{\beta}$.

To present the results of the first step, assume that the model we consider is given by:

$$H(Y) = \varphi(Z) + X + U \tag{30}$$

$$\mathbb{E}[U|X,W] = 0,$$

where $Y, Z \in \mathbb{R}$ are endogenous variables, $X \in \mathbb{R}$ is an exogenous variable and $W \in \mathbb{R}^p$ is a vector of instruments. As in the general model introduced in Section 2, $U \in \mathbb{R}$ is an error term. Before turning to assumptions, let us introduce the following definitions:

Definition 1 *Let $\{\lambda_j, \phi_j, \psi_j\}$ be the singular system of the operator T such that:*

$$T\phi_j = \lambda_j \psi_j \quad and \quad T^*\psi_j = \lambda_j \phi_j,$$

where λ_j denote the sequence of nonzero singular values of the compact linear operator T, ϕ_j and ψ_j, for all $j \in \mathbb{N}$, are orthonormal sequences of functions in $\mathcal{E}_0$ and $L^2_F(X, W)$, respectively. We can moreover write the singular value decomposition for each $\varphi \in \mathcal{E}_0$:[11]

$$T\varphi = \sum_{j=1}^{\infty} \lambda_j \langle \varphi, \phi_j \rangle \psi_j$$

Definition 2 *If $K : \mathcal{E}_1 \mapsto \mathcal{E}_2$ is a linear operator between two normed spaces, then the operator norm of K is given by:*

$$\|K\| := sup\{\|K\phi\|_{\mathcal{E}_2} : \phi \in \mathcal{E}_1 \quad and \quad \|\phi\|_{\mathcal{E}_1} \leq 1\}$$

The following assumptions are needed for consistency:

Assumption 8 *Source Condition: There exists $\nu > 0$ such that:*

$$\sum_{j=1}^{\infty} \frac{\langle \Phi, \phi_j \rangle^2}{\lambda_j^{2\nu}} = \sum_{j=1}^{\infty} \frac{[\langle H, \phi_{1,j} \rangle + \langle \varphi, \phi_{2,j} \rangle]^2}{\lambda_j^{2\nu}} < \infty$$

where $\Phi = (H, \varphi)$.

Assumption 9 *There exists $s \geq 2$ such that:*

- $\left\| \hat{T} - T \right\|^2 = O_p\left(\frac{1}{N h_N^{p+2}} + h_N^{2s} \right)$
- $\left\| \hat{T}^* - T^* \right\|^2 = O_p\left(\frac{1}{N h_N^{p+2}} + h_N^{2s} \right)$

where s is the minimum between the order of the kernel and the order of the differentiability of f, p is the dimension of the instrument vector W and h_N is the bandwidth.

Assumption 10

$$\left\| \hat{T}^* X - \hat{T}^* \hat{T} \Phi \right\|^2 = O_p\left(\frac{1}{N} + h_N^{2s} \right)$$

[11] For more on singular value decomposition, see Carrasco, Florens, and Renault (2007).

Assumption 11 $\lim_{N\to\infty}\alpha_N = 0$, $\lim_{N\to\infty}\alpha_N^2 N \to \infty$, $\lim_{N\to\infty} N h_N^{p+2} \to \infty$, $\lim_{N\to\infty}\frac{h_N^{2s}}{\alpha_N^2} = 0$, $\lim_{N\to\infty}\alpha_N^{2-\nu} N h_N^{p+2} \to \infty \quad or \quad \nu \geq 2$.

Proposition 2 *(Theorem 5 in Florens and Sokullu (2016)) Let us define* $\Phi = (H(Y), \varphi(Z))$. *Let* s *be the minimum between the order of the kernel and the order of the differentiability of* f *and* ν *be the regularity of* Φ. *Under Assumptions 8 to 11:*

- $\left\|\hat{\Phi}_N^\alpha - \Phi\right\|^2 = O_p\left(\frac{1}{\alpha^2}\left(\frac{1}{N} + h_N^{2s}\right) + \frac{1}{\alpha^2}\left(\frac{1}{N h_N^{p+2}} + h_N^{2s}\right)\left(\alpha^{\min\{\nu,2\}}\right) + \alpha^{\min\{\nu,2\}}\right)$
- $\left\|\hat{\Phi}_N^\alpha - \Phi\right\| \to 0$ *in probability.*

For the second step, consider again the general model given in Assumption 1. Below we introduce the assumptions needed to show that $\hat{\beta}$ is consistent and asymptotically normal. Note that once $\sqrt{N}$-consistency for $\hat{\beta}$ is shown, consistency of $(\hat{H}, \hat{\varphi})$ follows from step 1 straightforwardly.

Let $\{\lambda_j, \phi_j, \psi_j\}$ for $j \geq 1$ be the singular system of the operator T as defined before and let $\{\mu_l, e_l, \tilde{\psi}_l\}$ for $l = 1, 2, .., q-1$ be the singular system of the operator T_X, such that for each $\beta \in \mathbb{R}^{q-1}$ we can write:

$$T_X \beta = \sum_{l=1}^{q-1} \mu_l \langle \beta, e_l \rangle \tilde{\psi}_l.$$

Assumption 12 *Source Condition: There exists* $\eta > 0$ *such that:*

$$\max_{l=1,\ldots,q-1} \sum_{j=1}^{\infty} \frac{\left\langle \tilde{\psi}_l, \psi_j \right\rangle^2}{\lambda_j^{2\eta}} < \infty.$$

Assumption 13 *Parameters given in the Source Conditions in Assumptions 8 and 12 are both greater than or equal to two, i.e.,* $\nu \geq 2$ *and* $\eta \geq 2$.

Assumption 14 $\lim_{N\to\infty} N\alpha \to 0$, $\lim_{N\to\infty} N\alpha_N h_N^{2s} \to 0$, $\lim_{N\to\infty}\frac{\alpha_N}{h_N^{p+q+1}} \to 0$.

We also modify Assumptions 9 and 11 to account for the change in the dimension of X.

Assumption 15 *There exists* $s \geq 2$ *such that:*

- $\left\|\hat{T} - T\right\|^2 = O_p\left(\frac{1}{N h_N^{p+q+1}} + h_N^{2s}\right)$
- $\left\|\hat{T}^* - T^*\right\|^2 = O_p\left(\frac{1}{N h_N^{p+q+1}} + h_N^{2s}\right)$

where s is the minimum between the order of the kernel and the order of the differentiability of f, p is the dimension of the instrument vector W, q is the dimension of X and h_N is the bandwidth.

Assumption 16 $\lim_{N\to\infty} \alpha_N \to 0$, $\lim_{N\to\infty} h_N^{2s} \to 0$, $\lim_{N\to\infty} N h_N^{p+q+1} \to \infty$.

Let us denote $\mathcal{R}(T)$ the range of T and $\mathcal{R}(T)^\perp$ its orthogonal space in $L_F^2(X,W)$. The null space of T^* is denoted by $\mathcal{N}(T^*)$. We assume that the set of instruments is sufficiently rich such that:

Assumption 17 $\mathcal{R}(T)^\perp = \mathcal{N}(T^*) \neq \{0\}$.

In practice, this assumption implies that there exists an element ψ_j defined by the SVD of T such that $\psi_j \in \mathcal{R}(T)^\perp$. For example, this condition is satisfied in the joint nondegenerate normal case, i.e, if (Y,Z,X,W) is jointly distributed as a nondegenerate normal distribution. In such a case, the null space of T^* is $\{0\}$ if the range of the covariance with (Y,Z) and (X,W) is equal to the dimension of (X,W). Note that this is impossible even if $X_0, X_1 \in \mathbb{R}$ and W has at least one element.

Assumption 18 *For $\delta > 0$, we have:*

- $\mathbb{E}[|U|^{2+\delta} |X,W] = c$, *for any* $c \in \mathbb{R}$
- $\mathbb{E}[|(I - P_{YZ})X_1|^{2+\delta}] < \infty$ *where* $P_{YZ} = T(T^*T)^{-1}T^*$

Proposition 3 *(Theorem 9 in Florens and Sokullu (2016)) Assume that $Var[U|X,W] = \sigma^2$. Moreover assume that Assumptions 8, 10 and 12-18 hold. Then:*

$$\sqrt{N}(\hat{\beta} - \beta) \to \mathcal{N}(0, V)$$

where $V = \sigma^2 M^{-1}[\sum_{j/\psi_j \in \mathcal{R}(T)^\perp} \mathbb{E}(X_1\psi_j)\mathbb{E}(X_1\psi_j)']M^{-1}$ and $M = T_X^* T(T^*T)^{-1}T^*T_X - T_X^*T_X$ and $\psi_j \in \mathcal{R}(T)^\perp$.

C Additional Simulations and Figures

C.1 Figures for Design 1

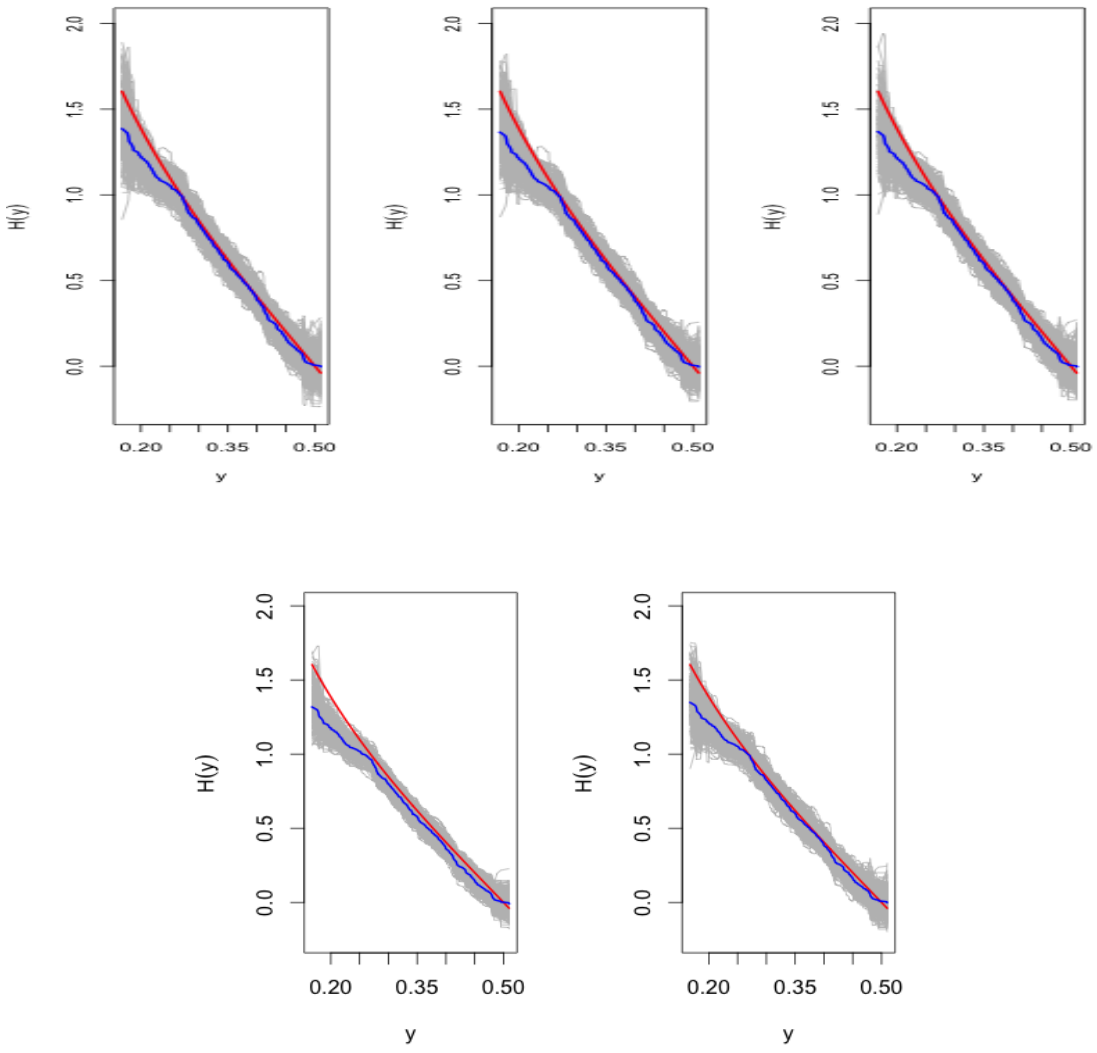


Figure 3: **Simulation results for estimation of $H(y)$.** $N = 400$. Simulation estimates (grey), average estimate (blue) and true function (red) across methods (TOP: left: Disc. R; middle: CV1; right: CV2. BOTTOM: left: Disc. R 2, right: It. CV 2).

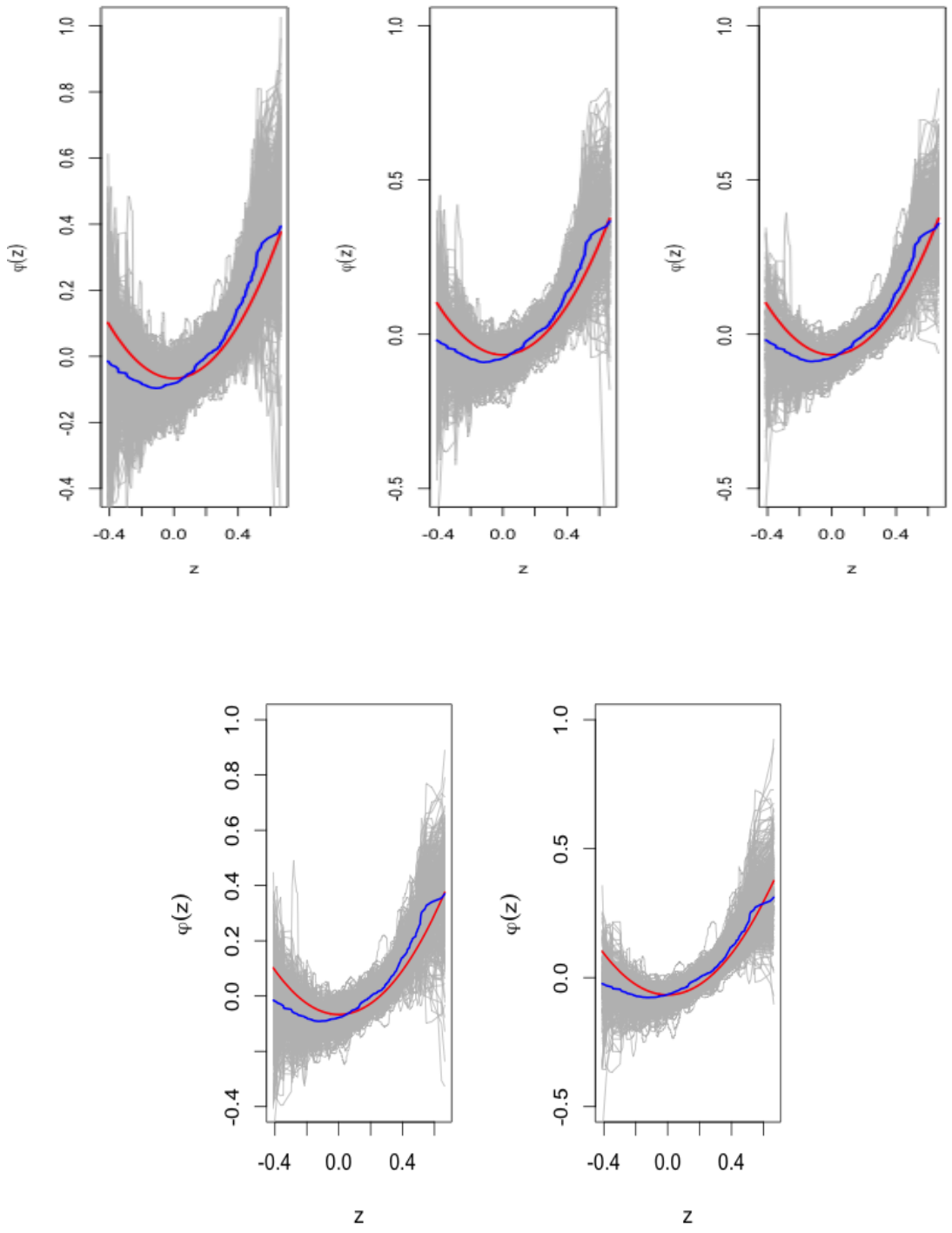


Figure 4: **Simulation results for estimation of $\varphi(z)$.** $N = 400$. Simulation estimates (grey), average estimate (blue) and true function (red) across methods (TOP: left: Disc. R; middle: CV1; right: CV2. BOTTOM: left: Disc. R 2, right: It. CV 2).

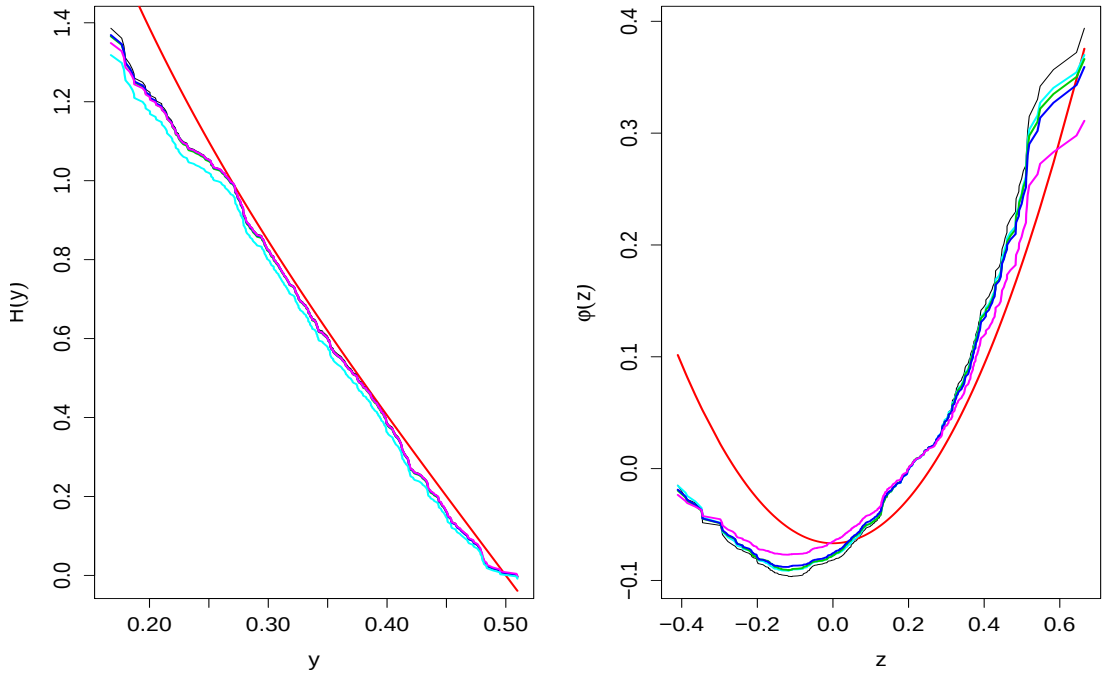


Figure 5: **Simulation results for estimation of $H(y)$ and $\varphi(z)$.** $N = 400$. Average simulation estimates of $H(y)$ (left) and $\varphi(z)$ (right) across methods: Disc. R (black), CV1 (green), CV2 (blue), Disc. R 2 (light blue), It. CV 2 (magenta), and true function (red).

C.2 Simulation results for Design 2

Table 7: **Summary statistics for $DEV_2(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 2.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.94	2.35	1.86	0.59	1.60	0.90	1.43	0.46	1.49	0.62
$N = 200$	1.62	0.99	1.97	0.47	1.47	0.59	1.39	0.41	1.34	0.39
$N = 400$	1.55	1.50	1.89	0.41	1.39	0.95	1.22	0.28	1.28	0.32

Table 8: **Summary statistics for $DEV_1(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 2.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.24	0.20	1.21	0.18	1.15	0.15	1.10	0.13	1.13	0.15
$N = 200$	1.18	0.13	1.29	0.14	1.14	0.14	1.11	0.12	1.10	0.11
$N = 400$	1.13	0.10	1.10	0.12	1.07	0.10	1.03	0.07	1.08	0.09

Table 9: **Summary statistics for $DEV_\infty(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 2.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.49	1.00	1.23	0.35	1.19	0.52	1.08	0.21	1.19	0.41
$N = 200$	1.61	1.03	1.29	0.39	1.23	0.47	1.13	0.29	1.18	0.35
$N = 400$	1.59	1.19	1.25	0.35	1.30	0.88	1.14	0.34	1.06	0.25

Table 10: **Ratio of Mean Square Errors for $\hat{H}$ - Design 2.**

Sample size	$\frac{MSE_{CV2}(\hat{H})}{MSE_{Disc.R}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{Disc.R2}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{CV1}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{It.CV2}(\hat{H})}$
$N = 100$	72.55	93.11	92.24	92.43
$N = 200$	93.40	82.05	99.08	103.32
$N = 400$	79.70	109.53	92.75	96.06

Table 11: **Ratio of Mean Square Errors for $\hat{\varphi}$ - Design 2.**

Sample size	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{Disc.R}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{Disc.R2}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{CV1}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{It.CV2}(\hat{\varphi})}$
$N = 100$	52.76	77.01	80.71	82.90
$N = 200$	57.96	79.28	83.94	91.37
$N = 400$	64.53	86.07	82.59	96.07

Sample size	**Disc. R**	**Disc. R 2**	**CV 1**	**CV 2**	**It. CV 2**
$N = 100$	0.24 (0.07)	0.20 (0.06)	0.23 (0.07)	0.22 (0.06)	0.24 (0.07)
$N = 200$	0.26 (0.05)	0.22 (0.04)	0.25 (0.05)	0.25 (0.05)	0.25 (0.05)
$N = 400$	0.27 (0.03)	0.22 (0.03)	0.26 (0.03)	0.26 (0.03)	0.26 (0.03)

(a) **Average and standard errors**.

Sample size	**Disc. R**	**Disc. R 2**	**CV 1**	**CV 2**	**It. CV 2**
$N = 100$	7.72	13.94	9.79	9.62	8.40
$N = 200$	3.90	8.90	5.04	5.24	4.46
$N = 400$	2.22	6.58	2.78	2.58	2.70

(b) **Mean Square Error** $\times 1000$.

Table 12: **Simulation results for estimation of β. Design 2** (a) Average of $\hat{\beta}$ estimates across simulations and selection methods. Simulation standard errors in parenthesis; (b) Mean square error across simulations $\times 1000$.

C.3 Figures for Design 2

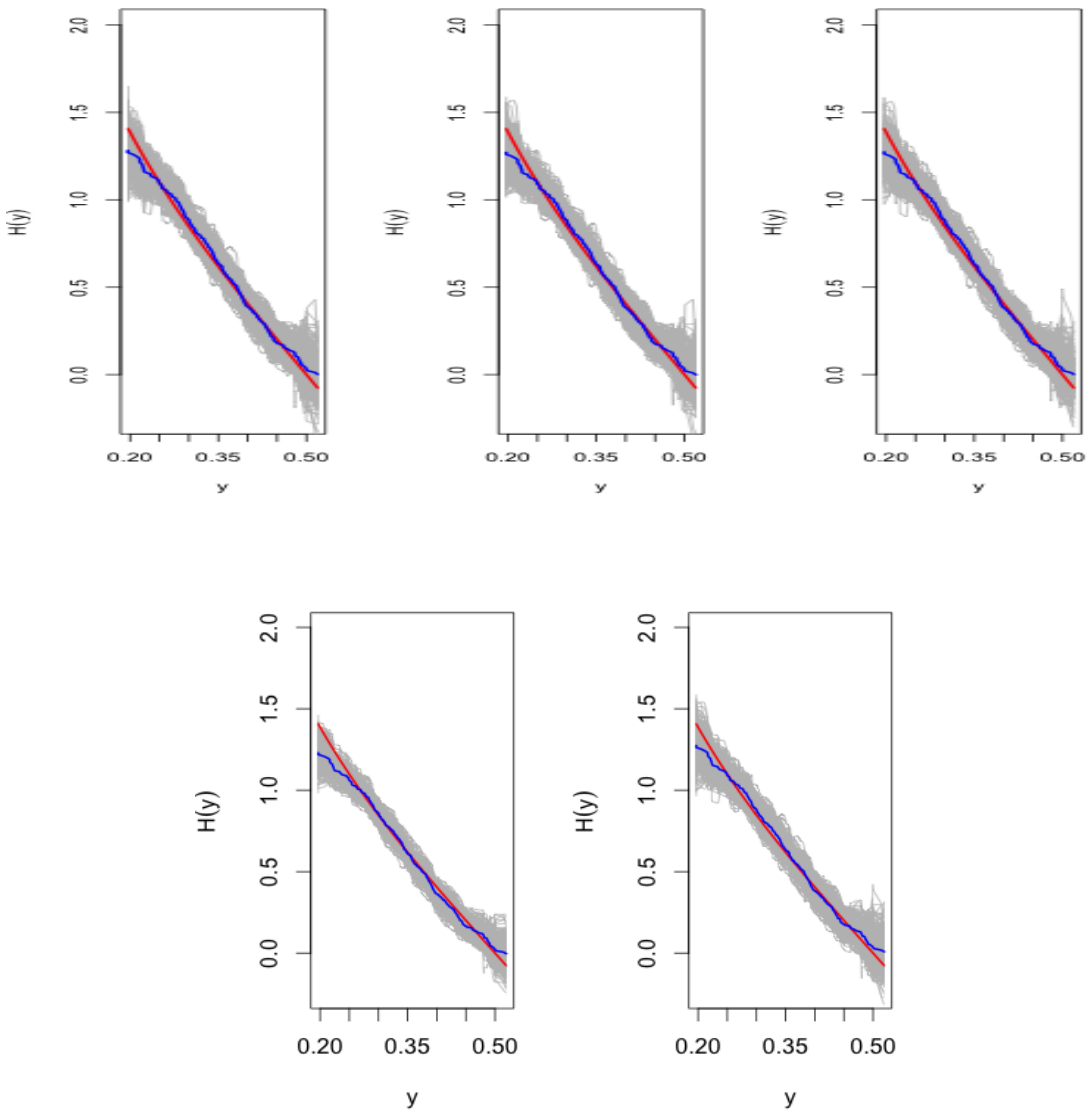


Figure 6: **Simulation results for estimation of $H(y)$.** $N = 400$. Simulation estimates (grey), average estimate (blue) and true function (red) across methods (TOP: left: Disc. R; middle: CV1; right: CV2. BOTTOM: left: Disc. R 2, right: It. CV 2).

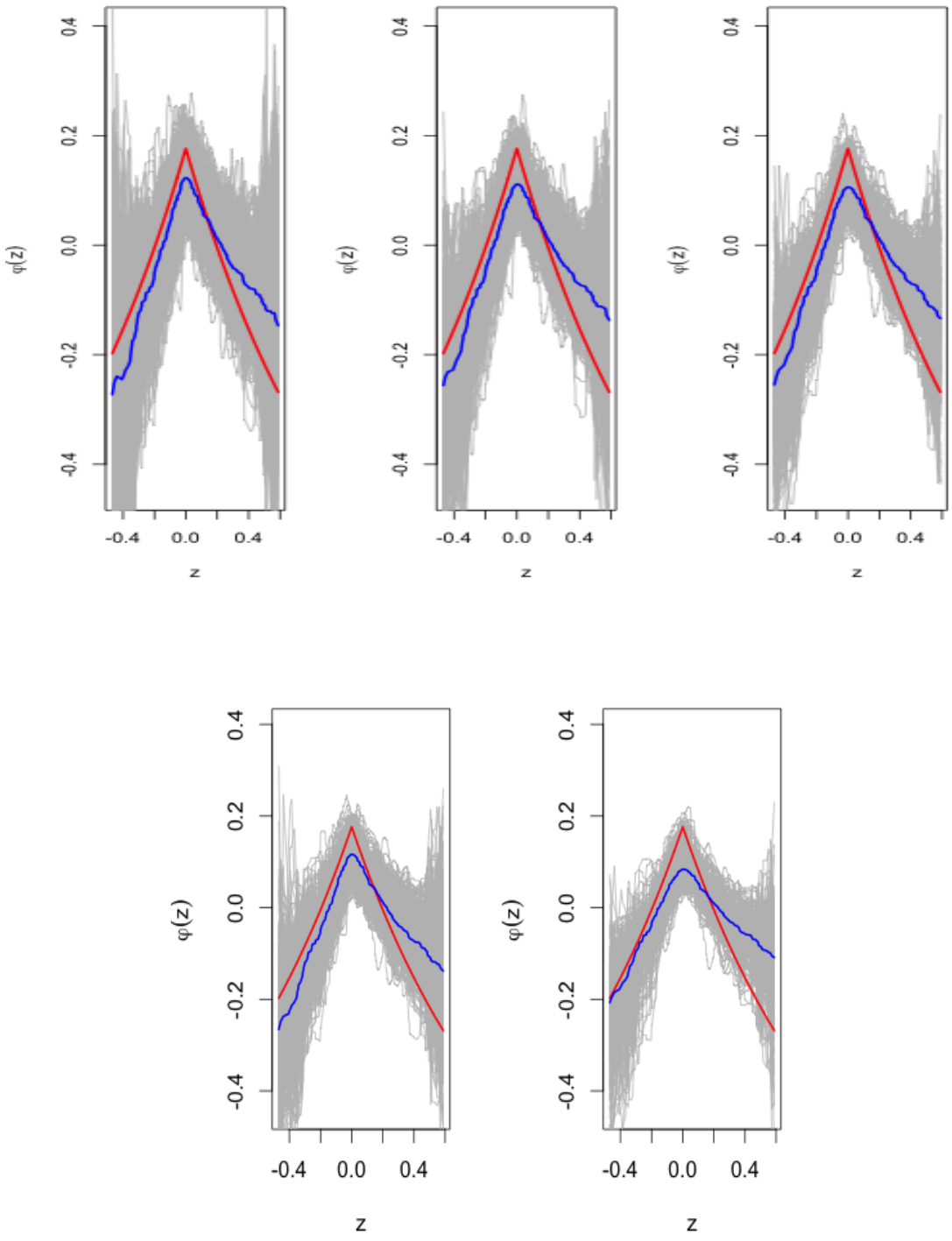


Figure 7: **Simulation results for estimation of $\varphi(z)$.** $N = 400$. Simulation estimates (grey), average estimate (blue) and true function (red) across methods (TOP: left: Disc. R; middle: CV1; right: CV2. BOTTOM: left: Disc. R 2, right: It. CV 2).

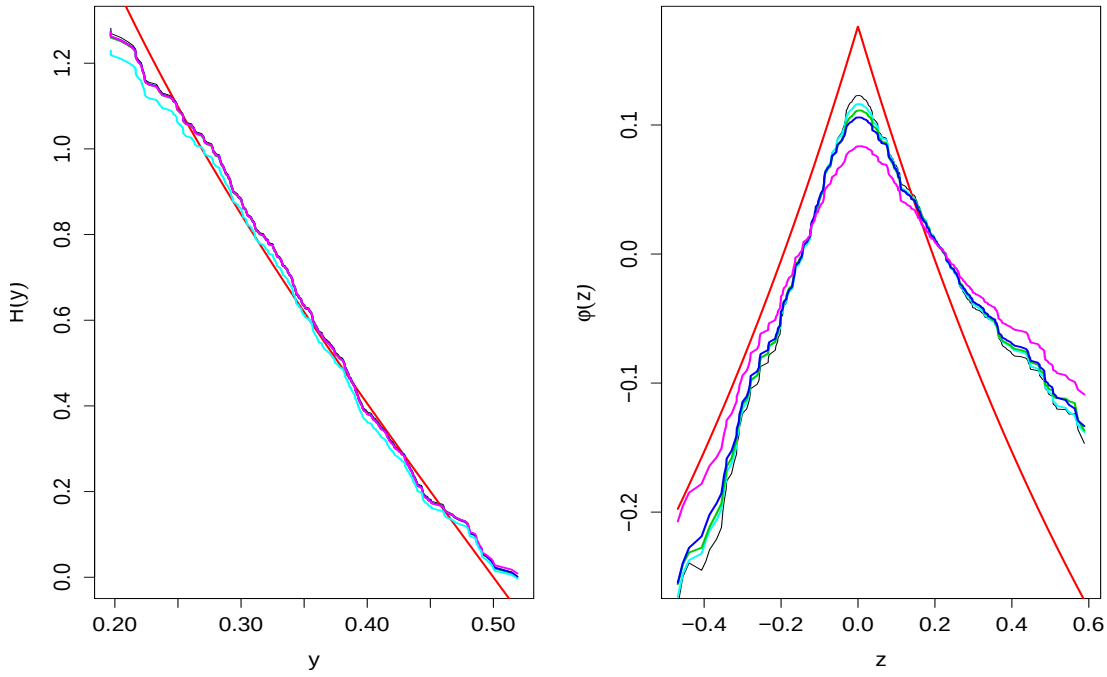


Figure 8: **Simulation results for estimation of $H(y)$ and $\varphi(z)$.** $N = 400$. Average simulation estimates of $H(y)$ (left) and $\varphi(z)$ (right) across methods: Disc. R (black), CV1 (green), CV2 (blue), Disc. R 2 (light blue), It. CV 2 (magenta), and true function (red).

C.4 Robustness to bandwidth choice

We show results of two Monte-Carlo simulations for Design 1 with two different bandwidth specifications: Silverman's rule-of-thumb divided by 2 (Tables 13-18), and Silverman's rule of thumb multiplied by 2 (Tables 19-24), for each of the bandwidth parameters h_y, h_z, h_x and h_w. Compared to simulations with bandwidths chosen by Silverman's rule-of-thumb, the main difference is that with larger bandwidths, Tables 13-15 show that the simultaneous implementation of the discrepancy rule performs relatively well in all metrics, while the best performing method is the iterative method It.CV2. As Table 16 indicates, the relative performance of CV2 mostly deteriorates due to estimation of H, since the relative MSE for $\hat{\varphi}$ remains favorable to CV2 (Table 17). Interestingly, It.CV2 dominates in terms of MSE of both $\hat{H}$ and $\hat{\varphi}$, suggesting that its performance may be more robust to bandwidth choice. With smaller bandwidths, CV2 and It.CV2 dominate. The performance of all methods deteriorates compared to results obtained with Silverman's rule-of-thumb. Overall the conclusions are qualitatively similar to those obtained with Silverman's rule of thumb.

Table 13: **Summary statistics for $DEV_2(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb $\times 2$.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	3.46	2.09	2.34	1.12	2.36	4.83	1.74	0.89	1.81	0.95
$N = 200$	3.50	2.01	2.29	0.94	3.18	2.75	1.81	0.97	1.89	1.15
$N = 400$	3.25	1.84	2.07	0.77	3.29	3.21	2.12	0.89	1.85	0.77

Table 14: **Summary statistics for $DEV_1(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb $\times 2$.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.82	0.49	1.53	0.35	1.43	0.46	1.29	0.31	1.31	0.32
$N = 200$	1.81	0.46	1.50	0.30	1.69	0.50	1.30	0.31	1.27	0.31
$N = 400$	1.82	0.43	1.44	0.29	1.84	0.49	1.48	0.33	1.35	0.28

Table 15: **Summary statistics for $DEV_\infty(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb $\times 2$.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	2.53	1.36	1.88	1.00	1.76	1.32	1.47	0.69	1.56	0.76
$N = 200$	2.80	1.75	1.96	1.16	2.54	1.83	1.60	0.80	1.71	0.97
$N = 400$	2.24	1.23	1.57	0.71	2.21	1.87	1.59	0.65	1.45	0.65

Table 16: **Ratio of Mean Square Errors for $\hat{H}$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb $\times 2$.**

Sample size	$\frac{MSE_{CV2}(\hat{H})}{MSE_{Disc.R}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{Disc.R2}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{CV1}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{It.CV2}(\hat{H})}$
$N = 100$	89.99	135.17	88.26	92.02
$N = 200$	96.32	157.37	75.30	91.68
$N = 400$	134.46	254.45	90.19	105.71

Table 17: **Ratio of Mean Square Errors for $\hat{\varphi}$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb $\times 2$.**

Sample size	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{Disc.R}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{Disc.R2}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{CV1}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{It.CV2}(\hat{\varphi})}$
$N = 100$	27.65	47.18	56.78	96.48
$N = 200$	25.48	47.54	33.84	117.17
$N = 400$	31.45	55.78	38.51	150.80

Sample size	Disc. R	Disc. R 2	CV 1	CV 2	It. CV 2
$N = 100$	0.23 (0.10)	0.20 (0.08)	0.25 (0.08)	0.25 (0.08)	0.26 (0.08)
$N = 200$	0.25 (0.07)	0.22 (0.05)	0.28 (0.06)	0.27 (0.06)	0.28 (0.06)
$N = 400$	0.27 (0.05)	0.24 (0.04)	0.30 (0.04)	0.30 (0.04)	0.30 (0.04)

(a) **Average and standard errors**.

Sample size	Disc. R	Disc. R 2	CV 1	CV 2	It. CV 2
$N = 100$	14.88	15.00	8.88	8.23	8.10
$N = 200$	7.58	9.03	4.29	4.01	3.96
$N = 400$	3.65	5.13	1.75	1.53	1.61

(b) **Mean Square Error** $\times 1000$.

Table 18: **Simulation results for estimation of β. Design 1. Bandwidths specification: Silverman's rule-of-thumb $\times 2$.** (a) Average of $\hat{\beta}$ estimates across simulations and selection methods. Simulation standard errors in parenthesis; (b) Mean square error across simulations $\times 1000$.

Table 19: **Summary statistics for $DEV_2(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb/2.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.98	1.59	1.94	0.49	1.58	0.48	1.62	0.43	1.64	0.53
$N = 200$	1.35	0.62	1.71	0.25	1.49	0.40	1.46	0.27	1.36	0.29
$N = 400$	1.42	0.98	1.82	0.23	1.51	0.39	1.48	0.31	1.34	0.32

Table 20: **Summary statistics for $DEV_1(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb/2.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.36	0.22	1.37	0.17	1.19	0.15	1.20	0.14	1.25	0.17
$N = 200$	1.16	0.09	1.24	0.10	1.13	0.10	1.12	0.09	1.13	0.09
$N = 400$	1.14	0.08	1.19	0.10	1.10	0.10	1.07	0.08	1.08	0.08

Table 21: **Summary statistics for $DEV_\infty(\hat{\varphi}_{\hat{\alpha}}, \hat{H}_{\hat{\alpha}}, \hat{\beta}_{\hat{\alpha}})$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb/2.**

	Disc. R		**Disc. R 2**		**CV 1**		**CV 2**		**It. CV 2**	
	Mean	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**	**Mean**	**S.d.**
$N = 100$	1.54	0.98	1.22	0.31	1.13	0.21	1.11	0.16	1.14	0.25
$N = 200$	1.36	0.78	1.16	0.27	1.09	0.28	1.06	0.12	1.09	0.21
$N = 400$	1.56	1.22	1.18	0.29	1.12	0.22	1.09	0.18	1.10	0.24

Table 22: **Ratio of Mean Square Errors for $\hat{H}$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb/2.**

Sample size	$\frac{MSE_{CV2}(\hat{H})}{MSE_{Disc.R}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{Disc.R2}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{CV1}(\hat{H})}$	$\frac{MSE_{CV2}(\hat{H})}{MSE_{It.CV2}(\hat{H})}$
$N = 100$	80.37	90.98	99.77	100.58
$N = 200$	100.67	94.02	97.82	105.33
$N = 400$	86.17	93.35	96.24	104.26

Table 23: **Ratio of Mean Square Errors for $\hat{\varphi}$ - Design 1. Bandwidths specification: Silverman's rule-of-thumb/2.**

Sample size	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{Disc.R}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{Disc.R2}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{CV1}(\hat{\varphi})}$	$\frac{MSE_{CV2}(\hat{\varphi})}{MSE_{It.CV2}(\hat{\varphi})}$
$N = 100$	25.70	55.19	90.47	61.49
$N = 200$	42.97	67.13	92.53	73.11
$N = 400$	40.96	72.27	92.75	78.22

Sample size	Disc. R	Disc. R 2	CV 1	CV 2	It. CV 2
$N = 100$	0.22 (0.08)	0.17 (0.07)	0.19 (0.07)	0.18 (0.06)	0.20 (0.07)
$N = 200$	0.24 (0.04)	0.19 (0.04)	0.20 (0.04)	0.20 (0.04)	0.22 (0.04)
$N = 400$	0.24 (0.03)	0.20 (0.03)	0.21 (0.03)	0.21 (0.03)	0.22 (0.03)

(a) **Average and standard errors**.

Sample size	Disc. R	Disc. R 2	CV 1	CV 2	It. CV 2
$N = 100$	12.02	20.31	16.30	17.22	15.68
$N = 200$	5.61	13.25	10.87	11.06	8.98
$N = 400$	4.35	11.22	8.65	8.71	6.89

(b) **Mean Square Error** $\times 1000$.

Table 24: **Simulation results for estimation of β. Design 1. Bandwidths specification: Silverman's rule-of-thumb/2.** (a) Average of $\hat{\beta}$ estimates across simulations and selection methods. Simulation standard errors in parenthesis; (b) Mean square error across simulations $\times 1000$.

D Empirical Application: Additional Figures

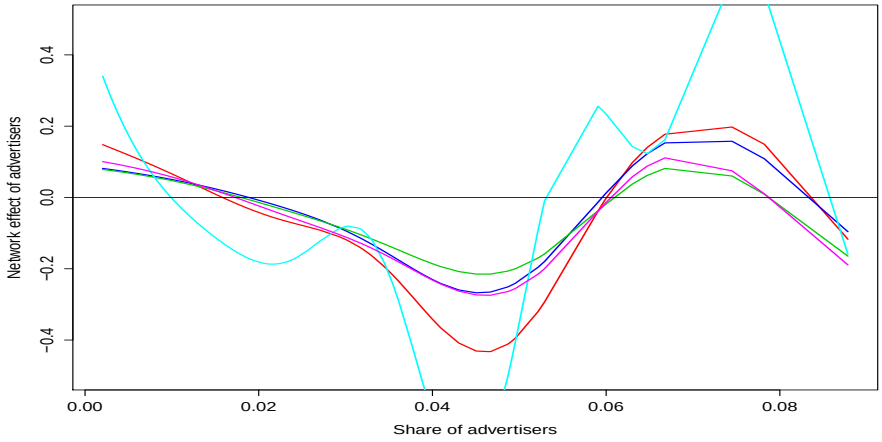


Figure 9: **Estimation of reader's demand**. (a) Inverse demand function, $H^r(N^r)$. (b) Network effect of advertisers on readers, $\varphi^r(N^a)$. Choice of regularisation parameter(s): Disc.R (red), Disc. R 2 (light blue), CV1 (green), CV2 (blue), It.CV2 (magenta).

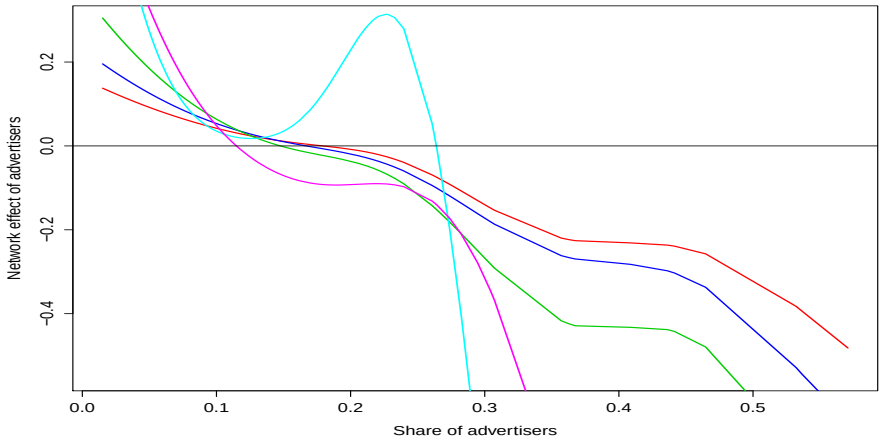


Figure 10: **Estimation of advertiser's demand**. (a) Inverse demand function, $H^a(N^a)$. (b) Network effect of readers on advertisers, $\varphi^a(N^r)$. Choice of regularisation parameter(s): Disc.R (red), Disc. R 2 (light blue), CV1 (green), CV2 (blue), It.CV2 (magenta).

E Illustration of cross-validation criteria

In this Section we illustrate the shape of the cross-validation criteria $CV_1(\alpha)$ and $CV(\alpha_H, \alpha_\varphi)$. We generate a random sample of size 400 according to Design 1 in the Monte-Carlo simulations and compute the optimal values of the regularisation parameters using the same grid that we used in the simulations. Then for illustrative purposes we recalculate the criteria's values in a much smaller area but with a larger grid (1500 grid points instead of 20). For $CV(\alpha_H, \alpha_\varphi)$, Figure 12 shows the value of the criterion fixing the value of α_φ to its optimal value. On the other hand, Figure 13 shows the value of the criterion fixing the value of α_H to its optimal value. These figures illustrate the presence of multiple minima which motivate our use of a grid in order to select the optimal values.

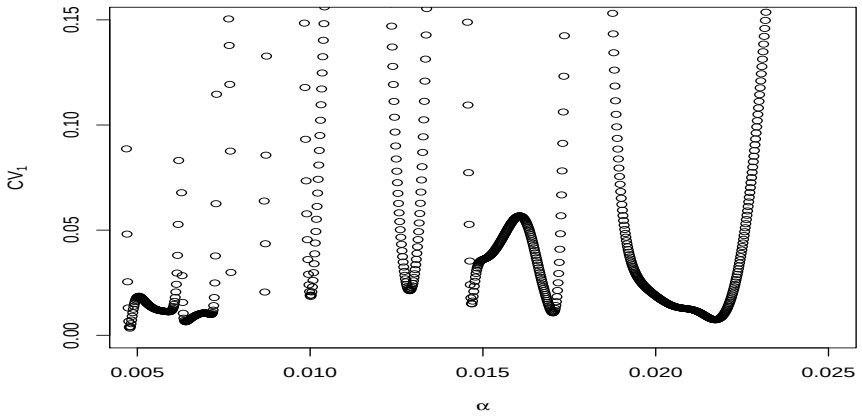


Figure 11: **Cross-validation objective** CV_1

.

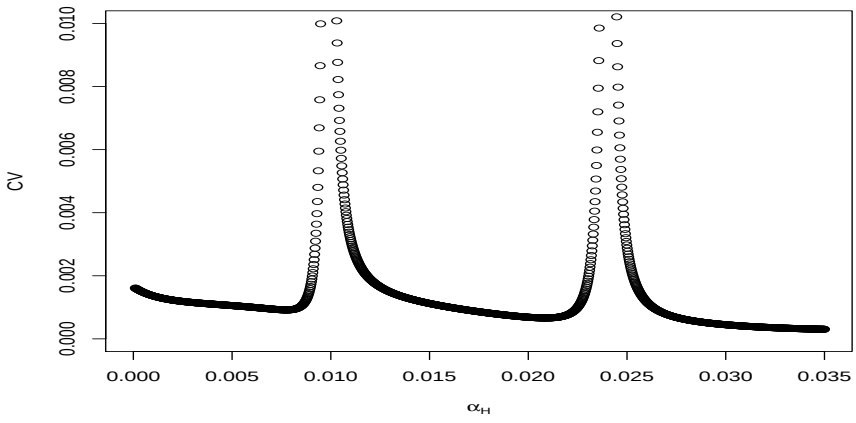


Figure 12: **Cross-validation objective** $CV(., \alpha^*_\varphi)$

.

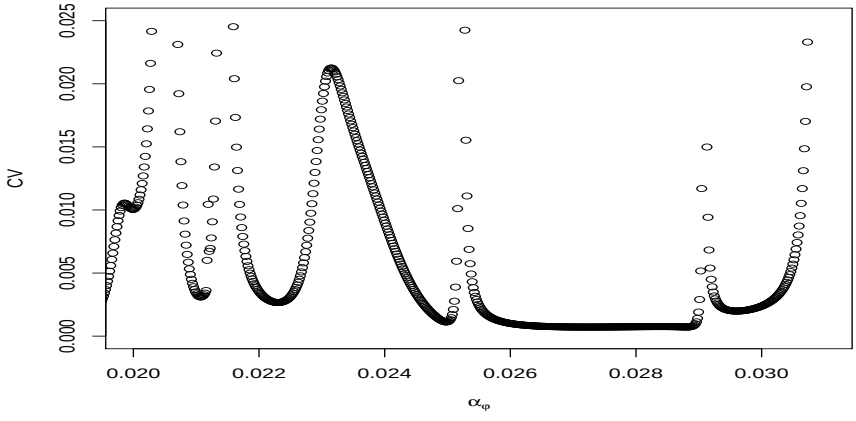


Figure 13: **Cross-validation objective** $CV(\alpha^*_H, .)$

.