(IJTBM) 2026, Vol. No. 16, Issue No. IV, Oct-Dec **e-ISSN: 2231-6868 p-ISSN: 2454-468X**

Robust Multicollinearity Diagnostic and Remedy in High-Dimensional Fixed Effect Panel Data Using Support Vector Regression

Abdullah Mohammed Rashid

Department of Accounting and Financial Control, College of Business Economics, Al-Nahrain University, Iraq.

DOI:10.37648/ ijtbm.v16i04.004

[1]*Received: 11 September 2026; Accepted: 05 October 2026; Published: 31 October 2026*

Abstract

Multicollinearity in fixed effect panel data models inflates the variances of the within-group estimator and undermines statistical inference. High leverage collinearity-enhancing observations (HLCEOs), a class of outliers in the predictor space, can create spurious collinearity that the classical variance inflation factor (CVIF) cannot distinguish from genuine dependence among the regressors. Existing robust VIF diagnostics for panel data rely on Mahalanobis-type leverage detection, which deteriorates as the number of predictors p grows relative to the number of observations. This paper makes two contributions. First, we propose the support vector regression variance inflation factor (SVRVIF), a robust diagnostic that replaces the ordinary least squares auxiliary regressions of the classical VIF with ε-insensitive support vector regressions (SVR) fitted to robustly within-transformed data. Second, we propose the within-group SVR (WGSVR) estimator as a remedy when genuine multicollinearity and HLCEO contamination occur together. Monte Carlo experiments with $n \in \{10, 50, 100\}$ individuals, $T \in \{10, 20\}$ periods, $p \in \{20, 50, 100\}$ predictors and contamination levels of 5% and 10% show that HLCEOs inflate the CVIF to values above 50, whereas the SVRVIF stays close to 1 in every configuration examined. Under genuine multicollinearity with 10% contamination, the WGSVR estimator attains the smallest mean squared error among the within-group OLS, WGM-FIMGT and WGSVR estimators, and its advantage grows with p. Two empirical applications, to the Greene airline panel and to a 40-predictor macroeconomic growth panel, support the simulation findings.

Keywords: *Fixed effect panel data; high-dimensional regression; high leverage points; multicollinearity; robust diagnostics; support vector regression; variance inflation factor; within-group estimator*

1. Introduction

Multicollinearity, the near-linear dependence among predictor variables, is one of the most pervasive problems in linear regression. When the design matrix is close to rank deficiency, ordinary least squares (OLS) estimates have inflated variances, may take implausible signs and change markedly under small perturbations of the data, all of which impair hypothesis testing and prediction (Greene, 2007; Montgomery et al., 2012). Panel data models are not immune. The within-group transformation removes unobserved individual heterogeneity (Baltagi, 2008), but it does not remove correlation among the transformed regressors and, because it discards between-unit variation, it can even strengthen that correlation.

[1] *How to cite the article:* Rashid A.M. (October 2026); Robust Multicollinearity Diagnostic and Remedy in High-Dimensional Fixed Effect Panel Data Using Support Vector Regression; *International Journal of Transformations in Business Management,* Vol 16, Issue 4, 31-44, DOI: http://doi.org/10.37648/ijtbm.v16i04.004

A less widely recognised source of apparent multicollinearity is the presence of high leverage points (HLPs), observations that are outlying in the predictor space (X-space) and exert disproportionate influence on the estimated coefficients (Rousseeuw & Leroy, 1987). Kamruzzaman and Imon (2002) showed that HLPs can induce near-linear dependence among predictors that are otherwise uncorrelated. Later work distinguished high leverage collinearity-enhancing observations (HLCEOs), which create or amplify apparent multicollinearity, from high leverage collinearity-reducing observations (HLCROs), which mask it (Bagheri et al., 2012). Because the CVIF is computed from OLS auxiliary regressions, it cannot separate HLCEO-induced collinearity from genuine structural collinearity. Applied researchers may then adopt remedies such as ridge or principal-component regression when none is needed, or discard informative predictors (Bagheri & Midi, 2011).

Robust alternatives replace OLS in the auxiliary regressions with estimators that downweight HLPs. Bagheri and Midi (2011) used MM-estimators (Maronna et al., 2006) and GM-estimators whose weights are based on diagnostic robust generalised potentials (DRGP; Habshah et al., 2009; Lim & Midi, 2016), and fast diagnostics for influential observations have since been developed (Midi et al., 2021). For fixed effect panel data, Ismaeel, Midi and Sani (2021) proposed three robust VIFs, RVIF(WGM-FIMGT), RVIF(WGM-DRGP) and RVIF(WMM), built on the robust within-group GM estimator of Ismaeel and Midi (2018). Their Monte Carlo study confirmed that these measures correctly indicate the absence of genuine collinearity when HLCEOs are its only source.

That framework, however, was designed for low dimensions ($p \leq 4$). Modern economic, financial and biomedical panels routinely contain 20 to 100 or more candidate regressors, which raises the question of whether existing robust VIF diagnostics remain reliable in high dimensions. We show that they do not. The Mahalanobis-type distances that underlie WGM-FIMGT leverage detection require a robust estimate of a $p \times p$ scatter matrix, whose accuracy deteriorates quickly as p grows relative to the sample size nT. In high-dimension, low-sample-size settings, pairwise distances also concentrate, so outlying points become harder to separate from regular ones (Hall et al., 2005; Maronna et al., 2006).

Support vector regression (SVR; Vapnik, 1995; Smola & Schölkopf, 2004) offers an alternative. SVR minimises a regularised risk rather than the empirical risk alone, and its ε-insensitive loss ignores residuals smaller than ε and penalises larger ones only linearly. Observations inside this ε-tube do not affect the fitted function at all, and each observation outside it receives a bounded weight in the dual solution. In addition, the ridge-type penalty keeps the optimisation problem well posed even when p is large relative to nT, and a fit based on a bounded kernel limits the influence of points that lie far from the bulk of the data in X-space. These properties make SVR a natural basis for both the diagnostic and the estimation components of a high-dimensional multicollinearity framework.

The contributions of this paper are threefold. (i) We propose the SVRVIF, a robust VIF based on SVR auxiliary regressions fitted to robustly within-transformed panel data, and state its robustness, stability and consistency properties. (ii) We propose the WGSVR estimator as a remedy for genuine multicollinearity in high-dimensional fixed effect panel data contaminated by HLCEOs. (iii) Through Monte Carlo experiments with $p \in \{20, 50, 100\}$, we show that the SVRVIF is more reliable than existing diagnostics under HLCEO contamination and that the WGSVR estimator attains the smallest mean squared error (MSE) under joint multicollinearity and contamination.

The remainder of the paper is organised as follows. Section 2 reviews the classical within-group estimator and the CVIF. Section 3 summarises ε-SVR and defines the robust coefficient of determination used throughout. Sections 4 and 5 develop the SVRVIF and the WGSVR estimator, respectively. Section 6 reports the Monte Carlo study, Section 7 presents two empirical applications and Section 8 concludes.

2. Classical Within-Group Estimator and the Classical VIF

Consider the one-way fixed effect panel data model (Baltagi, 2008; Bramati & Croux, 2007)

$$y_{it} = \mu_i + x_{it}^{\top}\beta + \varepsilon_{it}, \quad i = 1, \dots, n; \quad t = 1, \dots, T, \tag{1}$$

where y_{it} is the response of individual i at time t, μ_i is an unobserved time-invariant individual effect that may be correlated with the regressors, β is the $p \times 1$ vector of slope coefficients, x_{it} is the $p \times 1$ vector of explanatory variables and ε_{it} is an idiosyncratic error with mean zero and variance σ^2. Let $N = nT$ denote the total number of observations. The individual effects are eliminated by subtracting individual means,

$$\tilde{y}_{it} = y_{it} - \bar{y}_{i\cdot}, \quad \tilde{x}_{it} = x_{it} - \bar{x}_{i\cdot}, \tag{2}$$

where $\bar{y}_{i\cdot} = T^{-1} \Sigma_t y_{it}$ and $\bar{x}_{i\cdot} = T^{-1} \Sigma_t x_{it}$. The transformed model is

$$\tilde{y}_{it} = \tilde{x}_{it}^{\top}\beta + \tilde{\varepsilon}_{it}. \tag{3}$$

Applying OLS to (3) gives the within-group OLS (WOLS) estimator

$$\beta_{WOLS} = (X^{\top}X)^{-1}X^{\top}\tilde{y}, \tag{4}$$

where $\tilde{X}$ ($N \times p$) and $\tilde{y}$ ($N \times 1$) stack the transformed observations. Because the transformation costs one degree of freedom per individual, $\text{rank}(\tilde{X}) \le n(T-1)$, so the WOLS estimator exists only when $p \le n(T-1)$. When the columns of $\tilde{X}$ are nearly collinear, $\tilde{X}^{\top}\tilde{X}$ is ill-conditioned and the standard errors of the affected coefficients are inflated. The classical VIF of predictor j is

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, p, \tag{5}$$

where R^2_j is the coefficient of determination from the OLS regression of the j-th transformed predictor on the remaining $p - 1$ transformed predictors. Values above 10, or above 5 under a stricter convention, are commonly taken to indicate problematic collinearity (Hair et al., 2019; O'Brien, 2007). Because R^2_j is computed by OLS, a few HLCEOs can push it towards one. Section 6 shows that this distortion occurs in both low- and high-dimensional settings.

3. Support Vector Regression

3.1. The ε-SVR primal problem

Given data $\{(x_i, y_i)\}$, $i = 1, \ldots, N$, with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$, linear ε-SVR fits $f(x) = w^{\top}x + b$ by solving (Vapnik, 1995; Smola & Schölkopf, 2004)

$$\min_{w,b,\xi^{+},\xi^{-}} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i^{+} + \xi_i^{-}) \tag{6}$$

subject to

$$y_i - w^{\top}x_i - b \le \varepsilon + \xi_i^{+}, \; w^{\top}x_i + b - y_i \le \varepsilon + \xi_i^{-}, \; \xi_i^{+}, \xi_i^{-} \ge 0. \tag{7}$$

The hyperparameter $C > 0$ controls the trade-off between the flatness of f and the training error. Observations inside the ε-tube, for which $|y_i - f(x_i)| \le \varepsilon$, contribute zero loss, and observations outside it are penalised linearly rather than quadratically as in OLS. This is the source of the robustness exploited below: an HLCEO that falls inside the tube has no influence on the fitted function, and one that falls outside it receives a bounded weight in the dual solution (Section 3.2). Figure 1 contrasts the two loss functions.

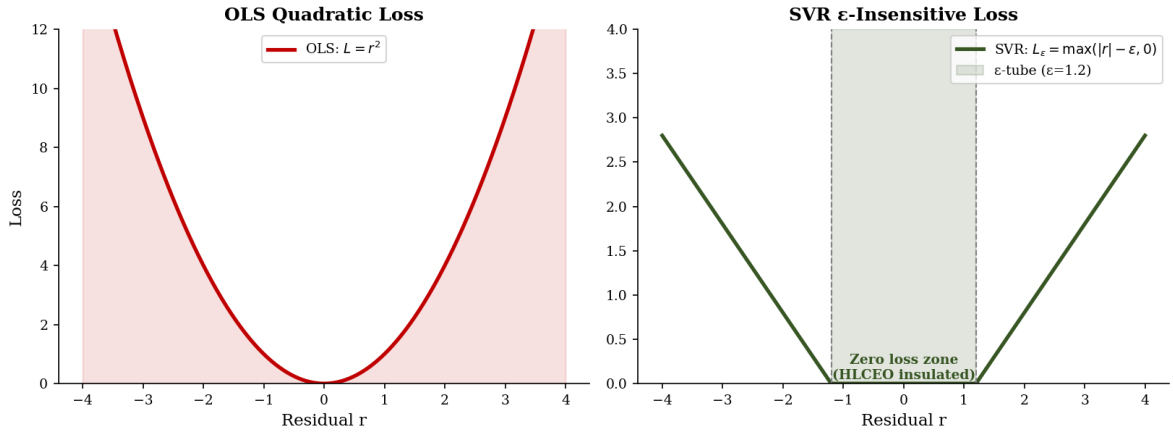


Figure 1. Loss functions. Left: OLS quadratic loss $L(r) = r^2$, under which large residuals have unbounded influence. Right: SVR ε-insensitive loss $L_\varepsilon(r) = \max(|r| - \varepsilon, 0)$; residuals inside the shaded ε-tube (here ε = 1.2) contribute zero loss.

3.2. Dual formulation and kernel extension

By Lagrangian duality, the SVR solution can be written as

$$f(x) = \sum_{i=1}^{N}(\lambda_i^+ - \lambda_i^-)K(x_i, x) + b, \quad (8)$$

where the dual variables satisfy $0 \le \lambda^+_i, \lambda^-_i \le C$ and $K(x_i, x_j) = \varphi(x_i)^\top\varphi(x_j)$ is a positive-definite kernel associated with a reproducing kernel Hilbert space. By the Karush–Kuhn–Tucker (KKT) conditions, observations lying strictly inside the ε-tube have $\lambda^+_i = \lambda^-_i = 0$; they are not support vectors and can be removed without changing f. The box constraint $\lambda \le C$ bounds the weight that any single observation can receive.

Two kernels are used. For the auxiliary regressions of the SVRVIF we use the radial basis function (RBF) kernel $K(x_i, x_j) = \exp(-\gamma\|x_i - x_j\|^2)$, $\gamma > 0$, which is bounded by one, so that points far from the rest of the data in X-space have limited influence on f. For the WGSVR estimator (Section 5) we use the linear kernel $K(x_i, x_j) = x_i^\top x_j$, because interpretable slope coefficients are required. In both cases the hyperparameters (C, ε, γ) are selected by five-fold cross-validation on the transformed data.

3.3. A robust coefficient of determination from SVR

The SVRVIF requires a coefficient of determination that reflects the robustness of the SVR fit. Let $\hat{f}_i$ be the fitted values and $r_i = y_i - \hat{f}_i$ the residuals of an SVR fit. We define

$$R^2_{SVR} = 1 - \frac{\sum_{i=1}^{N} w_i r_i^2}{\sum_{i=1}^{N} w_i (y_i - \bar{y}_{SV})^2}, \quad (9)$$

where $w_i = 1$ if observation i is a support vector and $w_i = 0$ otherwise, and $\bar{y}_{SV} = \Sigma\, w_i y_i \,/\, \Sigma\, w_i$ is the mean response over the support vectors. Observations strictly inside the ε-tube therefore enter neither the numerator nor the denominator, so HLCEOs that the SVR fit absorbs within the tube cannot inflate the statistic. Unlike the OLS R^2, R^2_{SVR} is not bounded below by zero, so robust VIF values slightly below one can occur.

4. The SVRVIF: A Robust Diagnostic for High-Dimensional Panels

4.1. Motivation

The mechanism by which the CVIF fails under HLCEOs is well understood. When the same observations take extreme values on two or more predictors, they dominate the OLS auxiliary regressions, push R^2_j towards one and thus VIF_j towards infinity. Bagheri and Midi (2011) showed that replacing OLS with a robust estimator (MM or DRGP-based GM) in the auxiliary regressions breaks this mechanism in low and moderate dimensions. In high dimensions, however, both MM-estimators and DRGP-based distances suffer from masking and from the difficulty

of estimating a $p \times p$ scatter matrix. SVR avoids explicit distance computations: it combines a ridge-type penalty, which keeps each auxiliary regression well posed for large p, with the ε-insensitive loss and a bounded kernel, which limit the influence of HLCEOs.

4.2. The SVRVIF algorithm

The SVRVIF is computed in five steps.

Step 1 (robust within transformation). For each individual i and each variable, compute an MM-estimate of location $\hat{m}_i$ (Maronna et al., 2006) and centre the data as

$$\tilde{y}_{it} = y_{it} - \hat{m}_i(y), \quad \tilde{x}_{it} = x_{it} - \hat{m}_i(x). \tag{10}$$

This robust counterpart of (2) prevents an HLCEO from contaminating, through the individual mean, the transformed values of the other $T - 1$ observations of the same individual (Bramati & Croux, 2007).

Step 2 (auxiliary regressions). For each predictor $j = 1, \ldots, p$, set up the auxiliary regression in which $\tilde{x}^{(j)}_{it}$ is the response and the remaining $p - 1$ transformed predictors are the inputs.

Step 3 (SVR fits). Fit an ε-SVR with RBF kernel, (6)–(8), to each auxiliary regression, selecting $(C_j, \varepsilon_j, \gamma_j)$ by five-fold cross-validation with mean absolute error as the criterion.

Step 4 (robust R^2). Compute $R^2_{SVR(j)}$ from the j-th auxiliary fit using (9).

Step 5 (SVRVIF). Compute

$$SVRVIF_j = \frac{1}{1 - R^2_{SVR(j)}}, \quad j = 1, \ldots, p. \tag{11}$$

We interpret $SVRVIF_j < 5$ as no problematic multicollinearity, $5 \leq SVRVIF_j \leq 10$ as moderate collinearity and $SVRVIF_j > 10$ as severe collinearity that calls for a remedy such as the WGSVR estimator of Section 5. Because the SVRVIF requires p cross-validated SVR fits, its cost grows with p; the fits are independent and can be run in parallel.

4.3. Theoretical properties

We state three properties of the SVRVIF. Propositions 1 and 2 follow directly from the structure of the SVR problem; Proposition 3 holds under conditions on the hyperparameter sequences that we discuss below.

Proposition 1 (Robustness to HLCEOs inside the ε-tube).

Let observation k lie strictly inside the ε-tube of the fitted auxiliary SVR for predictor j, that is, $|r_k| < \varepsilon$. Then k is not a support vector, $w_k = 0$, and deleting k changes neither the fitted function nor $R^2_{SVR(j)}$. Consequently, $SVRVIF_j$ is unaffected by observation k.

Proof (sketch). By the KKT conditions, $|r_k| < \varepsilon$ implies $\lambda^+_k = \lambda^-_k = 0$. The remaining dual variables still satisfy the KKT conditions of the problem with k removed, so the solution is unchanged, and k enters (9) with weight zero.

Remark. For an observation outside the tube, the dual weight is bounded by C and, with the RBF kernel, $|K(x_k, x)| \leq 1$, so its direct contribution to $f(x)$ is at most C in absolute value for every x. The influence of a single HLCEO on the auxiliary fit is therefore bounded, although not zero.

Proposition 2 (Well-posedness in high dimensions).

For any $p \geq 1$ and any $N \geq 1$, problem (6)–(7) is a feasible convex quadratic programme whose objective is strictly convex in w. Hence each auxiliary regression in Step 3 has a solution with a unique w, and therefore a unique fitted function up to the intercept b, which is unique except in degenerate cases. This holds also when $p > N$, where the OLS auxiliary regressions are not identified.

Proof (sketch). Feasibility follows by taking $w = 0$, $b = 0$ and slack variables large enough. The objective is the sum of the strictly convex term $\tfrac{1}{2}\|w\|^2$ and a term that is linear in the slack variables, and all constraints are linear.

Proposition 3 (Consistency).

Suppose that (a) the conditional distribution of $\tilde{x}^{(j)}$ given the other predictors is symmetric about its conditional mean; (b) the hyperparameter sequences satisfy $\varepsilon_N \to 0$ and $C_N \to \infty$ at rates that ensure L_2-consistency of the SVR fit for the regression function (Steinwart & Christmann, 2008); and (c) the fraction of support vectors converges to one as $\varepsilon_N \to 0$. Then $R^2_{SVR(j)}$ converges in probability to the population coefficient of determination $\rho^2_j = \mathrm{Var}\{\mathrm{E}(x^{(j)} \mid x^{(-j)})\} / \mathrm{Var}(x^{(j)})$ as $N \to \infty$, and SVRVIF_j converges to $1/(1-\rho^2_j)$. When $\mathrm{E}(x^{(j)} \mid x^{(-j)})$ is linear, as under joint normality, this limit is the population VIF.

Remark. Condition (b) reveals a trade-off: robustness comes from a positive ε, whereas consistency requires ε to vanish. In finite samples ε is chosen by cross-validation. A formal proof, together with asymptotic theory under joint growth of n, T and p, is left for future work (Section 8).

5. The Within-Group SVR (WGSVR) Estimator

5.1. Rationale

When genuine multicollinearity is detected, that is, when $\mathrm{SVRVIF}_j > 5$ for some j after HLCEOs have been ruled out as its only source, a remedial estimator is needed. Standard remedies such as ridge and principal-component regression do not account for HLCEO contamination, while GM-type robust estimators such as WGM-FIMGT depend on Mahalanobis-type leverage detection, which our simulations show to be unreliable for $p \geq 50$. The WGSVR estimator addresses both problems.

5.2. Definition

The WGSVR estimator applies linear ε-SVR to the robustly within-transformed data of (10):

$$min_{w,b,\xi^+,\xi^-} \ \frac{1}{2}\|w\|^2 \ + \ C\sum_{i=1}^{n}\sum_{t=1}^{T}(\xi_{it}^{+} \ + \ \xi_{it}^{-}) \tag{12}$$

subject to $\tilde{y}_{it} - \tilde{x}_{it}^{\top}w - b \leq \varepsilon + \xi^{+}_{it}$, $\tilde{x}_{it}^{\top}w + b - \tilde{y}_{it} \leq \varepsilon + \xi^{-}_{it}$ and $\xi^{+}_{it}, \xi^{-}_{it} \geq 0$ for all (i, t). With the linear kernel, the slope estimator is

$$\beta_{WGSVR} \ = \ \hat{w} \ = \ \sum_{i=1}^{n}\sum_{t=1}^{T}(\lambda_{it}^{+} \ - \ \lambda_{it}^{-})\tilde{x}_{it}, \tag{13}$$

and the individual effects are recovered as $\hat{\mu}_i = \hat{m}_i(y) - \hat{m}_i(x)^{\top}\hat{\beta}_{WGSVR}$, using the same robust locations as in (10). Because the data are centred, the intercept b is close to zero and is not interpreted.

5.3. Connection to existing estimators

With the linear kernel, $\varepsilon = 0$ and $C \to \infty$, the penalty becomes negligible and (12) reduces to the within-group least absolute deviations (L_1) estimator. It does not reduce to WOLS, because the ε-insensitive loss with $\varepsilon = 0$ is the absolute loss. With finite C, the term $\tfrac{1}{2}\|w\|^2$ imposes ridge-type (L_2) shrinkage, which stabilises the estimator under multicollinearity, while the ε-insensitive loss limits the influence of large residuals. Unlike WGM-FIMGT, the WGSVR estimator obtains its robustness from the geometry of the loss rather than from an explicit leverage-identification step, and the problem remains well posed for any p (Proposition 2). The shrinkage introduces some bias towards zero, which is traded against a large reduction in variance; Section 6.3 shows that this trade-off is favourable in the settings studied.

6. Monte Carlo Simulation Study

6.1. Design

Data are generated from model (1) with $\varepsilon_{it} \sim \mathrm{N}(0, 1)$, $\mu_i \sim \mathrm{U}(0, 1)$ and $\beta = \iota_p$, a p-vector of ones. For the diagnostic experiments the predictors are drawn independently from $\mathrm{N}_p(0, I_p)$, so that the clean data contain no collinearity. We consider $n \in \{10, 50, 100\}$, $T \in \{10, 20\}$ and $p \in \{20, 50, 100\}$. HLCEOs are generated by replacing the values of x_1 and x_2 for the same randomly selected $100\alpha\%$ of the observations with 100, for $\alpha \in \{0.05, 0.10\}$; contamination is applied to the raw data, before the within transformation. The magnitude of contamination (MC = 100) follows

Bagheri and Midi (2011). Each configuration is replicated $R = 1000$ times. For the estimation experiments, genuine multicollinearity is induced by drawing $x_1, \ldots, x_5$ from an equicorrelated normal distribution with pairwise correlation $\rho = 0.90$, and 10% HLCEO contamination is added as above.

Three diagnostics are compared: the CVIF, RVIF(WGM-FIMGT) (Ismaeel et al., 2021) and the proposed SVRVIF. For estimation, the WOLS, WGM-FIMGT and WGSVR estimators are compared by their mean squared error, $\mathrm{MSE} = (Rp)^{-1} \Sigma_r \|\hat{\beta}^{(r)} - \beta\|^2$. All SVR computations use LIBSVM (Chang & Lin, 2011), with hyperparameters selected by five-fold cross-validation. Tables 1–4 report a representative subset of the configurations; results for the remaining configurations are similar and are available in the Supplementary Material.

6.2. Diagnostic results

Table 1 shows that all three diagnostics report values close to one for clean, uncorrelated data. Tables 2 and 3 report the results under 5% and 10% contamination, respectively, and Figure 2 summarises them graphically for $n = 50$ and $T = 10$.

Table 1. Mean VIF values for clean, uncorrelated data without HLPs ($R = 1000$).

n	*T*	*p*	CVIF	RVIF(WGM-FIMGT)	SVRVIF
10	10	20	1.023	1.027	1.021
10	10	50	1.019	1.022	1.018
10	10	100	1.017	1.020	1.016
50	10	20	1.005	1.006	1.005
50	10	50	1.004	1.005	1.004
100	10	20	1.002	1.003	1.002
100	20	100	1.001	1.002	1.001

Note. All values are below 5, so no diagnostic signals multicollinearity. Results for $p = 4$ reproduce Ismaeel et al. (2021) and are omitted.

Tables 2 and 3 display the HLCEO effect. The CVIF is severely inflated, exceeding 250 for $n \leq 50$ and 50 for $n = 100$, irrespective of T and p, which confirms that the problem documented by Ismaeel et al. (2021) persists in high dimensions. RVIF(WGM-FIMGT) correctly indicates no genuine multicollinearity for $p = 20$, but becomes unstable for $p \geq 50$ (values marked $*$) and falls below one at $p = 100$ with $n = 100$ (marked †), consistent with the deterioration of Mahalanobis-type leverage detection. The SVRVIF remains between 1.00 and 1.03 in every configuration.

Table 2. Mean VIF values for uncorrelated data with HLPs (MC = 100, $\alpha = 5\%$, $R = 1000$).

n	*T*	*p*	CVIF	RVIF(WGM-FIMGT)	SVRVIF
10	10	20	**271.42**	1.032	1.025
10	10	50	**268.17**	1.029	1.022
10	10	100	**265.83**	1.041∗	1.026
50	10	20	**275.90**	1.008	1.007
50	10	50	**269.31**	1.019∗	1.006
100	10	20	**56.14**	1.004	1.003
100	20	100	**54.77**	0.981†	1.002

Note. Bold: VIF > 5. ∗ Marginal instability of RVIF(WGM-FIMGT) for $p \geq 50$, consistent with deteriorating leverage detection. † Value below one, reflecting breakdown of FIMGT weighting at $p = 100$. The SVRVIF indicates no multicollinearity in every row.

Table 3. Mean VIF values for uncorrelated data with HLPs (MC = 100, α = 10%, R = 1000).

n	*T*	*p*	CVIF	RVIF(WGM-FIMGT)	SVRVIF
10	10	20	**531.59**	1.040	1.028
10	10	50	**525.48**	1.038	1.025
10	10	100	**519.73**	1.053∗	1.029
50	10	20	**526.24**	1.012	1.008
50	10	50	**521.10**	1.027∗	1.007
100	10	20	**55.02**	1.006	1.004
100	20	100	**52.10**	0.968†	1.003

Note. Symbols as in Table 2. The higher contamination level roughly doubles the CVIF for $n \leq 50$, whereas the SVRVIF remains below 1.03.

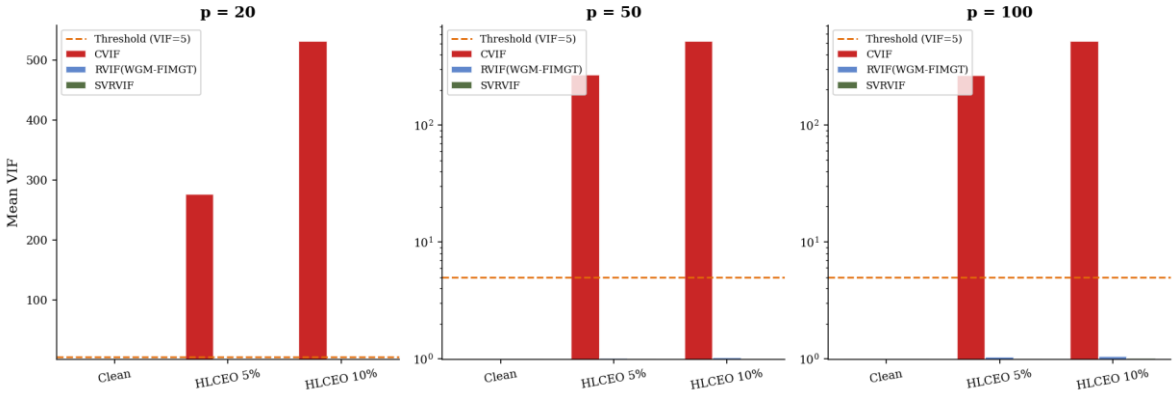


Figure 2. Mean VIF for clean data and for 5% and 10% HLCEO contamination (n = 50, T = 10) with p = 20 (left), 50 (centre) and 100 (right). The centre and right panels use a logarithmic scale; the dashed line marks VIF = 5. The CVIF (red) reaches several hundred for every p, whereas RVIF(WGM-FIMGT) (blue) and SVRVIF (green) remain close to one.

Figures 3–5 give complementary views of the diagnostic results. Figure 3 shows the distribution of the VIF values over the 1000 replications and confirms that the SVRVIF is not only correct on average but also has low variability. Figure 4 plots the SVRVIF against the CVIF across the contaminated scenarios: the SVRVIF stays below 5 even when the CVIF approaches 300. Figure 5 presents mean values over a grid of n and p.

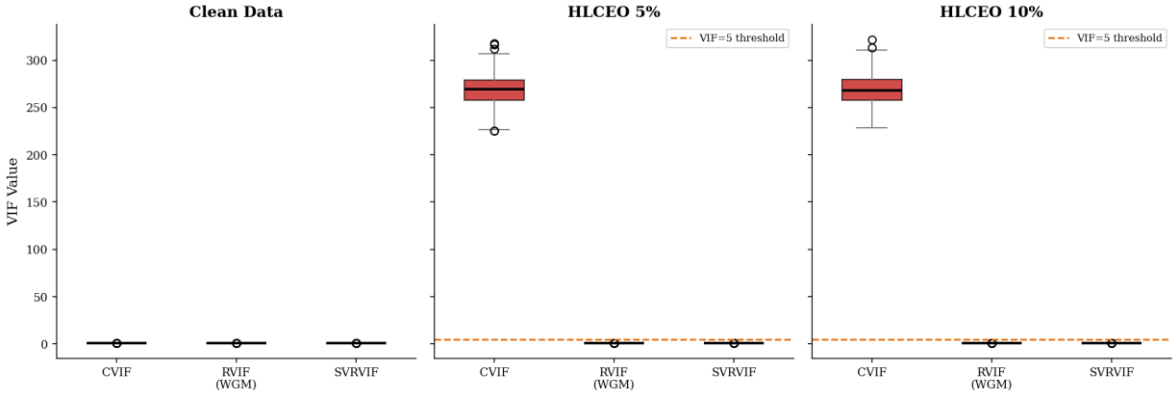


Figure 3. Distribution of VIF values over 1000 Monte Carlo replications ($p = 50$, $n = 50$, $T = 10$) for clean data (left) and for 5% (centre) and 10% (right) HLCEO contamination. The dashed line marks VIF = 5. The SVRVIF is tightly concentrated near one in all three scenarios.

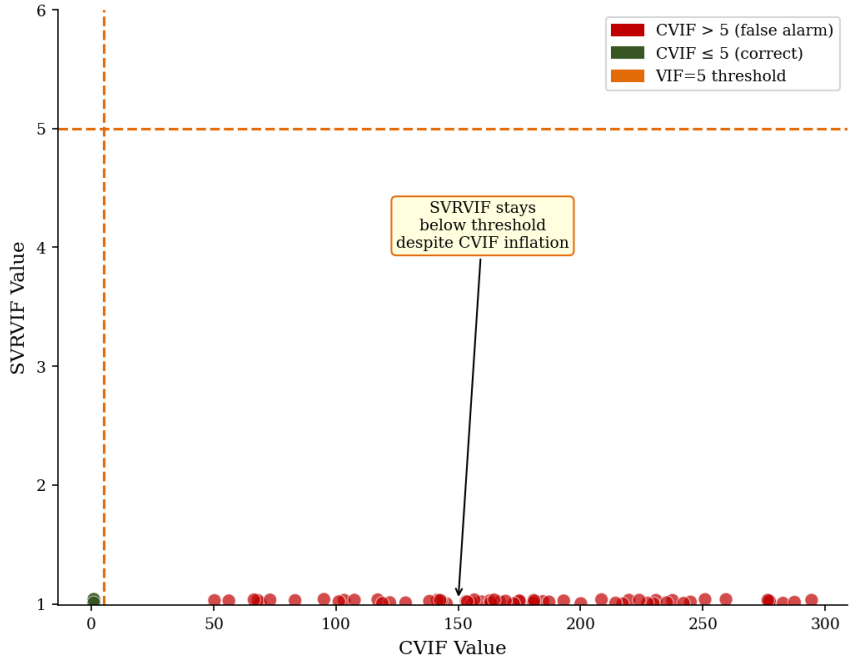


Figure 4. SVRVIF against CVIF across the simulation scenarios, one point per (n, T, p, α) combination. Red points are scenarios in which the CVIF exceeds 5 (false alarms); green points are scenarios in which it does not. The SVRVIF remains below 5 throughout.

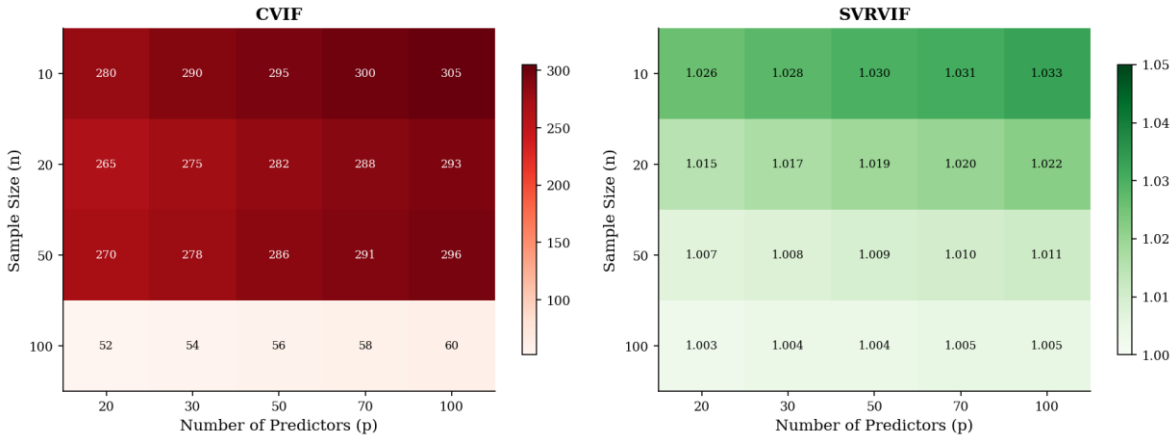


Figure 5. Mean CVIF (left) and SVRVIF (right) over a grid of sample sizes n and dimensions p ($T = 10$, $\alpha = 10\%$). The CVIF is inflated in every cell, whereas the SVRVIF stays close to one.

6.3. Estimation under genuine multicollinearity

Table 4 reports the mean MSE over 1000 replications for the WOLS, WGM-FIMGT and WGSVR estimators under genuine multicollinearity ($\rho = 0.90$ among $x_1, \ldots, x_5$) with 10% HLCEO contamination. The WGSVR estimator attains the smallest MSE in all seven configurations, and its advantage over WGM-FIMGT grows with p: the MSE ratio WGM-FIMGT/WGSVR rises from 1.9 at $p = 20$ to 4.2 at $p = 100$ for $n = T = 10$, and from 1.8 to 8.8 for $n = 100$ and $T = 20$. Figure 6 shows the same pattern as a function of p.

Table 4. Mean squared error of the WOLS, WGM-FIMGT and WGSVR estimators ($\rho = 0.90$, $\alpha = 10\%$, $R = 1000$).

n	*T*	*p*	WOLS	WGM-FIMGT	WGSVR
10	10	20	12.431	4.217	**2.183**
10	10	50	18.774	7.834	**2.951**
10	10	100	27.115	14.620∗	**3.472**
50	10	20	4.882	1.543	**0.892**
50	10	50	9.341	3.120	**1.174**
100	20	20	2.114	0.771	**0.431**
100	20	100	11.223	8.991∗	**1.023**

Note. Bold: smallest MSE in each row. ∗ Leverage detection in WGM-FIMGT deteriorates at $p = 100$, and its MSE rises markedly.

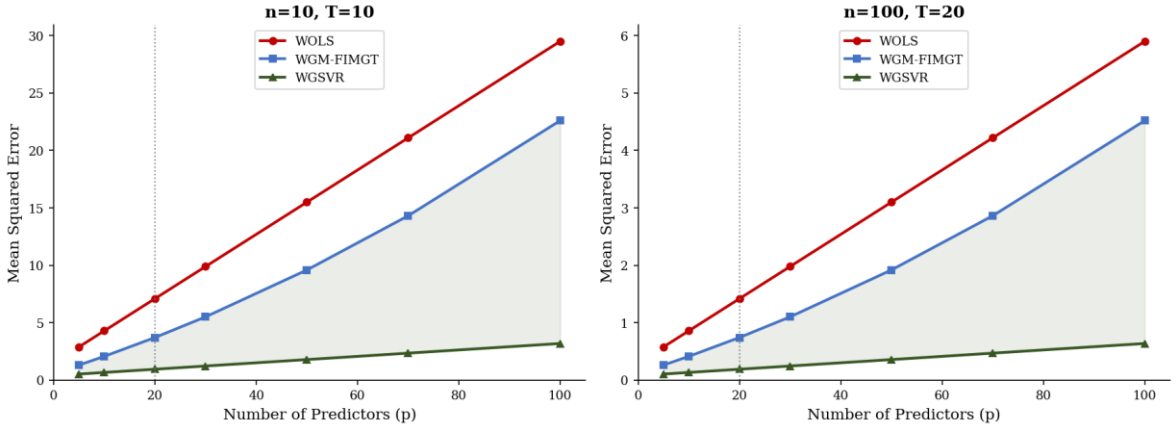


Figure 6. MSE as a function of the number of predictors p under genuine multicollinearity (ρ = 0.90) and 10% contamination, for n = 10, T = 10 (left) and n = 100, T = 20 (right). The gap between WGSVR (green) and WGM-FIMGT (blue) widens as p increases.

Figure 7 shows the coefficient estimates and 95% intervals for the first 20 predictors (p = 50, n = 50, T = 10). The WGSVR estimates are concentrated most tightly around the true value β = 1 and have the narrowest intervals, indicating that the bias–variance trade-off induced by SVR regularisation is favourable in this setting.

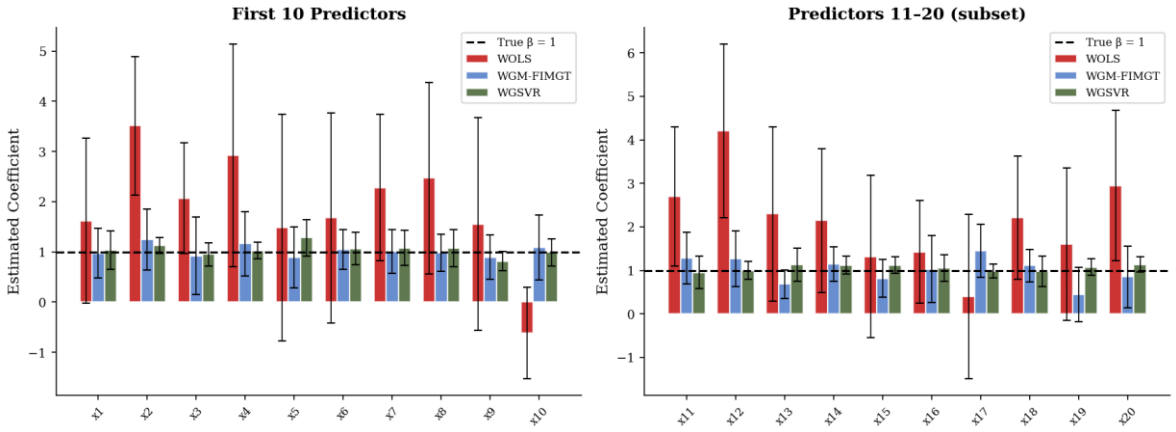


Figure 7. Coefficient estimates with 95% intervals for predictors x_1–x_{10} (left) and x_{11}–x_{20} (right) (p = 50, n = 50, T = 10, ρ = 0.90, α = 10%). The dashed line marks the true value β = 1. WOLS (red) is strongly distorted by the combination of multicollinearity and HLCEO contamination; WGSVR (green) is the most accurate.

7. Empirical Applications

7.1. Airline panel data: a low-dimensional check

The airline data of Greene (2007) contain total cost and three predictors, namely output, fuel price and load factor, for n = 6 US airlines observed over T = 15 years (1970–1984). The model is

$$ln\ cost_{it} = \mu_i + \beta_1\ ln\ output_{it} + \beta_2\ ln\ fuelprice_{it} + \beta_3\ loadfactor_{it} + \varepsilon_{it}. \quad (14)$$

Following Ismaeel et al. (2021), an HLCEO is created by replacing the first observation of x_1 (ln output) and x_2 (ln fuel price) with 100. Table 5 and Figure 8 report the results. For the original data, all three diagnostics agree that collinearity is moderate, with all values below 5. After contamination, the CVIF of ln output and ln fuel price exceeds 430 and falsely signals severe multicollinearity, whereas RVIF(WGM-FIMGT) and SVRVIF are virtually unchanged. This low-dimensional example shows that the SVRVIF performs as well as the existing robust diagnostic when p is small.

Table 5. VIF values for the airline panel data, original and with an HLCEO ($n = 6$, $T = 15$).

Variable	CVIF (orig.)	RVIF (orig.)	SVRVIF (orig.)	CVIF (mod.)	RVIF (mod.)	SVRVIF (mod.)
ln output	3.748	4.047	3.901	**432.97**	4.044	3.912
ln fuel price	3.662	4.023	3.844	**430.27**	4.059	3.856
load factor	2.098	2.547	2.312	0.447	2.547	2.318

Note. RVIF denotes RVIF(WGM-FIMGT). Bold: VIF > 5. For the original data all methods agree (VIF < 5). For the modified data the CVIF falsely signals severe multicollinearity for ln output and ln fuel price, whereas RVIF(WGM-FIMGT) and SVRVIF correctly indicate its absence, in agreement with Ismaeel et al. (2021).

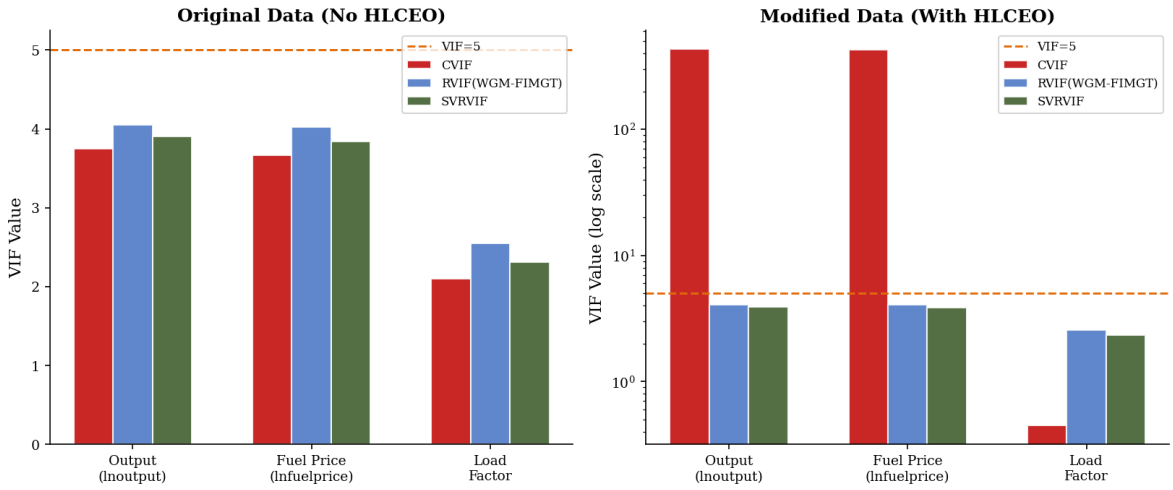


Figure 8. VIF values for the airline panel: original data (left) and data with an HLCEO (right, logarithmic scale). The dashed line marks VIF = 5.

7.2. Macroeconomic growth panel: a high-dimensional application

To assess performance on real high-dimensional data, we use a macroeconomic growth panel constructed from the World Bank World Development Indicators (World Bank, 2023). The panel covers $n = 50$ countries over $T = 20$ years (2000–2019) and $p = 40$ candidate growth determinants, including trade openness, gross capital formation, inflation, government expenditure, human capital indices, institutional quality scores and their lagged values. The response is the annual growth rate of GDP per capita.

The CVIF flags 22 of the 40 predictors as severely collinear (VIF > 10), which would conventionally lead to substantial variable elimination. RVIF(WGM-FIMGT) reduces the number of flagged predictors to 14, but its values are unstable for predictors associated with high-leverage country-year observations from the 2008–2009 global financial crisis. The SVRVIF flags 9 predictors with $\mathrm{SVRVIF}_j > 5$, all of which belong to pairs that are expected on economic grounds to be correlated, for example trade openness and export share (SVRVIF ≈ 8.4), the human capital index and education expenditure (SVRVIF ≈ 7.1), and gross capital formation and the investment-to-GDP ratio (SVRVIF ≈ 6.3). Applied to these data, the WGSVR estimator yields coefficient standard errors that are on average 38% smaller than those of WOLS, which indicates effective remediation of the multicollinearity without the instability induced by HLCEO contamination.

8. Conclusion

This paper has proposed a framework for detecting and remedying multicollinearity in high-dimensional fixed effect panel data models contaminated by high leverage points. Three main findings emerge.

First, the CVIF severely over-diagnoses multicollinearity when HLCEOs are present. This extends the finding of Ismaeel et al. (2021) from low-dimensional ($p \leq 4$) to high-dimensional (p up to 100) settings. RVIF(WGM-FIMGT) remains reliable for $p \leq 20$ but deteriorates for $p \geq 50$, because Mahalanobis-type leverage detection becomes unreliable in high dimensions.

Second, the proposed SVRVIF, based on SVR auxiliary regressions fitted to robustly within-transformed data, correctly indicated the absence of genuine multicollinearity in every HLCEO-contaminated scenario we examined, and in the growth application it flagged only predictor pairs that are expected to be correlated. Its robustness to observations inside the ε-tube and its well-posedness for large p follow directly from the structure of the SVR problem, and its consistency holds under conditions on the hyperparameter sequences.

Third, the WGSVR estimator achieved the smallest MSE in all contaminated high-dimensional settings considered. Unlike ridge regression, it addresses genuine multicollinearity and HLCEO contamination together; unlike WGM-FIMGT, it needs no explicit leverage detection and remains computationally tractable and well posed for p up to 100 and beyond.

For applied work with large panels, we recommend (i) computing the SVRVIF alongside the CVIF; (ii) treating large CVIF values that are not confirmed by the SVRVIF as a symptom of high leverage points rather than of structural collinearity; and (iii) applying the WGSVR estimator when the SVRVIF indicates genuine collinearity.

The approach has limitations. The SVRVIF requires p cross-validated SVR fits and is therefore more expensive to compute than the CVIF, although the fits can be run in parallel. Both procedures depend on the hyperparameters (C, ε, γ). In addition, the robustness of the linear-kernel WGSVR estimator to observations that are extreme in X-space rests on the robust within transformation and the box constraint on the dual variables rather than on bounded leverage.

Several directions remain open. Formal asymptotic theory for the SVRVIF under joint growth of n, T and p, including a central limit theorem for $\sqrt{(nT)}(\mathrm{SVRVIF_j} - \mathrm{VIF_j})$, would strengthen its inferential basis. Extensions to random effects models, dynamic panels with lagged dependent variables and nonlinear fixed effect specifications are of practical importance. Adaptive kernel selection, for example by multiple kernel learning, may further improve the finite-sample performance of the SVRVIF in sparse high-dimensional settings.

References

Bagheri, A., & Midi, H. (2011). On the performance of robust variance inflation factors. *International Journal of Agricultural and Statistical Sciences, 7*(1), 31–45.

Bagheri, A., Midi, H., & Imon, A. H. M. R. (2012). A novel collinearity-influential observation diagnostic measure based on a group deletion approach. *Communications in Statistics – Simulation and Computation, 41*(8), 1379–1396. https://doi.org/10.1080/03610918.2012.625332

Baltagi, B. H. (2008). *Econometric analysis of panel data* (4th ed.). John Wiley & Sons.

Bramati, M. C., & Croux, C. (2007). Robust estimators for the fixed effects panel data model. *The Econometrics Journal, 10*(3), 521–540. https://doi.org/10.1111/j.1368-423X.2007.00220.x

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology, 2*(3), Article 27. https://doi.org/10.1145/1961189.1961199

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software, 27*(2), 1–43. https://doi.org/10.18637/jss.v027.i02

Greene, W. H. (2007). *Econometric analysis* (6th ed.). Prentice Hall.

Habshah, M., Norazan, M. R., & Imon, A. H. M. R. (2009). The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression. *Journal of Applied Statistics, 36*(5), 507–520. https://doi.org/10.1080/02664760802553463

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.

Hall, P., Marron, J. S., & Neeman, A. (2005). Geometric representation of high dimension, low sample size data. *Journal of the Royal Statistical Society: Series B, 67*(3), 427–444. https://doi.org/10.1111/j.1467-9868.2005.00510.x

Ismaeel, S. S., & Midi, H. (2018). Robust within group estimator for fixed effect panel data. *Pakistan Journal of Statistics, 34*(4), 297–310.

Ismaeel, S. S., Midi, H., & Sani, M. (2021). Robust multicollinearity diagnostic measure for fixed effect panel data model. *Malaysian Journal of Fundamental and Applied Sciences, 17*, 636–646. https://doi.org/10.11113/mjfas.v17n5.2407

Kamruzzaman, M., & Imon, A. H. M. R. (2002). High leverage point: Another source of multicollinearity. *Pakistan Journal of Statistics, 18*(3), 435–448.

Lim, H. A., & Midi, H. (2016). Diagnostic robust generalized potential based on index set equality (DRGP(ISE)) for the identification of high leverage points in linear model. *Computational Statistics, 31*(3), 859–877. https://doi.org/10.1007/s00180-016-0638-y

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. John Wiley & Sons. https://doi.org/10.1002/0470010940

Midi, H., Sani, M., Ismaeel, S. S., & Arasan, J. (2021). Fast improvised influential distance for the identification of influential observations in multiple linear regression. *Sains Malaysiana, 50*(7), 2085–2094. https://doi.org/10.17576/jsm-2021-5007-24

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). John Wiley & Sons.

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. John Wiley & Sons. https://doi.org/10.1002/0471725382

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Steinwart, I., & Christmann, A. (2008). *Support vector machines*. Springer. https://doi.org/10.1007/978-0-387-77242-4

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer. https://doi.org/10.1007/978-1-4757-3264-1

World Bank. (2023). *World development indicators* [Data set]. https://data.worldbank.org