

Convex Support Vector Regression

Zhiqiang Liao^a, Sheng Dai^{a,*}, Timo Kuosmanen^b

^a *Aalto University School of Business, 02150 Espoo, Finland*

^b *Turku School of Economics, University of Turku, 20500 Turku, Finland*

September 2022

Abstract

Nonparametric regression subject to convexity or concavity constraints is increasingly popular in economics, finance, operations research, machine learning, and statistics. However, the conventional convex regression based on the least squares loss function often suffers from overfitting and outliers. This paper proposes to address these two issues by introducing the convex support vector regression (CSV) method, which effectively combines the key elements of convex regression and support vector regression. Numerical experiments demonstrate the performance of CSV in prediction accuracy and robustness that compares favorably with other state-of-the-art methods.

Keywords: Robustness and sensitivity analysis, Convex regression, Support vector regression, Overfitting, Regularization

*Corresponding author.

E-mail addresses: zhiqiang.liao@aalto.fi (Z. Liao), sheng.dai@aalto.fi (S. Dai), timo.kuosmanen@utu.fi (T. Kuosmanen).

1 Introduction

Convex regression (CR) is a classic approach to nonparametric regression that builds upon global concavity or convexity of the regression function (Hildreth, 1954). Since the explicit piecewise linear characterization of the multivariate model proposed by Kuosmanen (2008), CR has become an active research field with an increasing number of applications in economics, statistics, operational research and related fields (see, e.g., Guntuboyina & Sen, 2018; Johnson & Jiang, 2018). Recent methodological advances in CR include extensions to quantile-based approaches such as convex quantile regression (Wang et al., 2014; Kuosmanen et al., 2015) and convex expectile regression (Kuosmanen et al., 2020; Kuosmanen & Zhou, 2021). There has been significant development in the computational tools and algorithms (see, e.g., Lee et al., 2013; Mazumder et al., 2019; Dai, 2022; Lin et al., 2022).

Overfitting is a longstanding issue in nonparametric methods, including CR. The subgradients fitted by CR can be very large near the boundary of the convex hull of the design points (Seijo & Sen, 2011; Chen et al., 2020), which can seriously hamper the out-of-sample predictive power. To alleviate overfitting, (Lim 2014) has proposed to restrict the domain of the convex hull by imposing additional constraints on the subgradients of the regression function. Another approach bound subgradients is to impose regularization either in objective function or constraints, such as the L_2 -norm Lipschitz regularization (Mazumder et al., 2019) or the L_∞ -norm Lipschitz regularization (Balázs et al., 2015).

In the literature on machine learning, support vector regression (SVR) is a well-known approach firstly introduced by Vapnik (1999). SVR deviates from the linear regression in that it introduces an ε -insensitive loss function instead of the commonly used L_2 -norm loss function, which helps to improve its out-of-sample performance (Vapnik, 1999). Therefore, SVR has been considered as a robust alternative against outliers and to reduce overfitting in the context of linear regression.

Thus far, only few studies extend SVR to the context of shape-constrained regression or frontier estimation. The pioneering work by Wang & Ni (2012) was the first one to consider nonparametric convexity-constrained support vector regression (henceforth NCCSVR). In this approach, the Hessian matrix of a nonparametric representor function is constrained to be positive semidefinite in each observation. The authors transform the shape-constrained

SVR into a semidefinite programming problem, assuming the regression function to be continuous and twice differentiable throughout its domain. One notable limitation of NCCSVR is that it is not applicable for the univariate CR with a single regressor.

Recently, Valero-Carreras et al. (2021) and (Valero-Carreras et al. 2022) relax continuity and convexity assumptions, adapting SVR to the nonparametric estimation of frontier production and cost functions that envelop all observations. A notable limitation of their approach is the deterministic nature of the data generating process, which assumes away any noise in data. This observation motivates us to combine the key elements of both CR and SVR in a unified framework.

The main objective of the present paper is to improve the out-of-sample predictive power of CR by alleviating overfitting. To this end, we combine the key characteristics of both CR and SVR in a new approach referred to as convex support vector regression (CSVR). A notable difference between NCCSVR by Wang & Ni (2012) and the proposed CSVR approach concerns the implementation of the convexity constraints: NCCSVR imposes constraints on the Hessian matrix, whereas CSVR makes use of the inequality constraints known as the Afriat inequalities (see Kuosmanen, 2008). Using Monte Carlo simulations and two real-world examples, we show that the proposed CSVR approach yields a smaller mean squared error (MSE) than other state-of-the-art methods, including the NCCSVR method.

Our secondary objective is to outline how the proposed CSVR approach can be extended to facilitate the automatic variable selection in applications with high dimensionality. Inspired by such works as (Bradley & Mangasarian 1998), (Zhao et al. 2009) and (Negahban & Wainwright 2011), two alternative formulations of LASSO CSVR are considered (LASSO refers to the least absolute shrinkage and selection operator). This extension further enhances the linkages between SVR and LASSO that originated in the machine learning literature and CR that has emerged in econometrics and statistics.

The rest of this paper is organized as follows. Section 2 briefly reviews classical statistics and machine learning methods for regression problems. We then introduce the new shape-constrained SVR method, extend it to the Lasso version, and present a graphical illustration in section 3. Section 4 presents some evidence from Monte Carlo simulations. In the section 5 we experimentally compare CSVR against competing methods on two real-world datasets. Section 6 presents our concluding remarks.

2 Preliminaries on regression

2.1 Convex regression

Considering a general nonparametric regression model with a set of observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ satisfying

$$y_i = f(\mathbf{x}_i) + \varepsilon_i, \quad \text{for } i = 1, \dots, n, \quad (1)$$

where $\mathbf{x} \in \mathbb{R}^d$ is an observed vector of predictors, $y_i \in \mathbb{R}$ is the response variable, and ε_i is a random noise with zero mean. The regression function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ in Eq. (1) is unknown but satisfies certain shape restrictions such as monotonicity, concavity, and homogeneity (see, e.g., Kuosmanen & Johnson, 2010; Yagi et al., 2020). In this paper we focus exclusively on the class \mathcal{F} of concave function f , that is

$$\mathcal{F} := \left\{ f : \mathbb{R}^d \rightarrow \mathbb{R} \mid \forall \mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d, \tau f(\mathbf{x}_1) + (1 - \tau)f(\mathbf{x}_2) \leq f(\tau\mathbf{x}_1 + (1 - \tau)\mathbf{x}_2) \right\}.$$

The basic idea of CR is to find the best fitting function f from a family of continuous and concave functions \mathcal{F} by minimizing the sum of squares of the residuals

$$\begin{aligned} \min & \frac{1}{2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2 \\ \text{s.t.} & f \in \mathcal{F} \end{aligned} \quad (2)$$

While problem (2) is the infinite-dimensional multivariate convex regression problem, it can be equivalently represented by a finite-dimensional quadratic programming (QP) problem. Following Kuosmanen (2008), we consider the following least squares estimator as the operational multivariate convex regression model

$$\begin{aligned} \min_{\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\varepsilon}} & \frac{1}{2} \sum_{i=1}^n \varepsilon_i^2 \\ \text{s.t.} & y_i = \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i + \varepsilon_i \quad \forall i \\ & \alpha_i + \boldsymbol{\beta}'_i \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}'_h \mathbf{x}_i \quad \forall i, h \end{aligned} \quad (3)$$

where $\boldsymbol{\beta}'$ indicates the transpose of $\boldsymbol{\beta}$, and its subscript h is any index of data point not equal to i . The first constraint of (3) simply restates the regression equation (1) in terms of a piecewise linear approximation of the true but unknown regression function f , and the

second constraint enforces concavity of the piecewise linear regression function (reversing the sign of the inequality imposes convexity). Note that additional monotonicity constraints could be implemented by restricting the sign of $\boldsymbol{\beta}$ (e.g., $\boldsymbol{\beta} \geq 0$ for monotonic increasing and $\boldsymbol{\beta} \leq 0$ for monotonic decreasing functions). Further, imposing $\alpha = 0$ imposes linear homogeneity (constant returns to scale). See Kuosmanen et al. (2015) for more detailed discussion.

Given the optimal solutions $(\hat{\alpha}_i, \hat{\boldsymbol{\beta}}_i)$ to problem (3), we can reconstruct the explicit representor function $\hat{f}^{CR}(\mathbf{x})$ as (Kuosmanen, 2008)

$$\hat{f}^{CR}(\mathbf{x}) = \min_{i=1,\dots,n} \{ \hat{\alpha}_i + \hat{\boldsymbol{\beta}}_i' \mathbf{x} \} \quad (4)$$

However, the estimated coefficients $\hat{\boldsymbol{\beta}}_i$ could be arbitrarily large, particularly near the boundary of the convex hull of the covariate domain, due to the fact that the feasible set of problem (3) can be unbounded (see, e.g., Mazumder et al., 2019; Chen et al., 2020). This may lead to potential overfitting and deteriorate the out-of-sample performance of the estimated function. Even for the univariate case, the estimated $\hat{\beta}_i$ are also unbounded at the boundary (Ghosal & Sen, 2017).

To alleviate the overfitting problem in convex regression (3), (Mazumder et al. 2019) apply the L_2 -norm Lipschitz regularization on subgradients with a known bound $L > 0$ to reduce overfitting. For a prespecified $L > 0$, the class of F_L of concave functions with Lipschitz regularization

$$F_L := \{ f : \mathbb{R}^d \rightarrow \mathbb{R} \mid f \text{ is concave; } \sup_{\mathbf{x} \in \mathbb{R}^d} \|\partial f(\mathbf{x})\| \leq L \}.$$

where $\|\partial f(\mathbf{x})\|$ is the maximum of $\|\cdot\|_2$ -norm of vectors in $\partial f(\mathbf{x})$. The corresponding Lipschitz convex regression (LCR) is then formulated as

$$\begin{aligned} \min_{\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\varepsilon}} \quad & \frac{1}{2} \sum_{i=1}^n \varepsilon_i^2 & (5) \\ \text{s.t.} \quad & y_i = \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i + \varepsilon_i & \forall i \\ & \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}_h' \mathbf{x}_i & \forall i, h \\ & \|\boldsymbol{\beta}_i\|_2 \leq L & \forall i \end{aligned}$$

Note that the estimated function \hat{f}^{LCR} can be obtained by inserting the optimal $\hat{\alpha}_i$ and $\hat{\boldsymbol{\beta}}_i$ from (5) to the equation of explicit representor function (4). We can also resort to other

Lipschitz norms (e.g., $\|\cdot\|_\infty$ -norm) to address the overfitting problem in convex regression (Lim, 2014; Balázs et al., 2015). Following (Mazumder et al. 2019), $\|\cdot\|_2$ -norm is more effective than $\|\cdot\|_\infty$ -norm in avoiding overfitting. However, when the data with high noise, the outliers and the large variance in the error term could weaken the effectiveness of the Lipschitz norm approaches in reducing overfitting. Such a gap also motivates this paper to propose a new convex support vector regression approach and empirically compare the proposed method with other related approaches in terms of finite sample performance.

2.2 Support vector regression

Support vector regression (SVR) belongs to the class of support vector machines. As a regression method, it aims to estimate a function $f(x)$ that follows the structural risk minimization principle grounded on the statistical learning theory. It gives good generalization capacity by minimizing the upper bound of the risk (Vapnik, 1999).

SVR has excellent potential to reduce overfitting because the structural risk minimization principle makes a trade-off between the prediction accuracy and the complexity of the regression function. When some unwanted data exist, some parameters in the conventional regression models may become large to accommodate such outliers. Thus, these models may suffer from overfitting, while the SVR has good generalization performance as its optimization object function guarantees the flatness of the regression function. Flatness in the regression model means that one minimizes vector β . Taking into account data errors, one can introduce the slack variables ξ and ξ^* and penalty term C . The slack variables can be used to allow some errors that lie on the outside of the margin ε which refers to *soft margin* (Cortes & Vapnik, 1995). Hence we can obtain the regression function by solving the following optimization problem

$$\begin{aligned}
 \min_{\beta, \alpha, \xi_i, \xi_i^*} \quad & \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) & (6) \\
 \text{s.t.} \quad & y_i - \alpha - \beta' \mathbf{x}_i \leq \varepsilon + \xi_i & \forall i \\
 & \alpha + \beta' \mathbf{x}_i - y_i \leq \varepsilon + \xi_i^* & \forall i \\
 & \xi_i \geq 0, \xi_i^* \geq 0 & \forall i
 \end{aligned}$$

where C is a prespecified parameter that determines the trade-off between the complexity of regression function $f(x)$ and the prediction accuracy ε . This corresponds to the most

commonly adopted ε -insensitive loss function $|\xi|_\varepsilon$ described by

$$|\xi|_\varepsilon = \begin{cases} 0, & \text{if } |\xi| \leq \varepsilon; \\ |\xi| - \varepsilon, & \text{otherwise.} \end{cases} \quad (7)$$

for a user-determined nonnegative number ε . A potential benefit of using ε -insensitive loss function is robustness to outliers because it is less sensitive to noisy inputs. In practice, the penalty constant C and parameter ε can be chosen based on the user's experience. It can also be determined by cross-validation, a standard model selection technique in machine learning. Furthermore, the estimated support vector function is simply $\hat{f}^{SVR}(\mathbf{x}) = \alpha + \boldsymbol{\beta}' \mathbf{x}$.

3 Combining CR and SVR

3.1 Convex support vector regression (CSVR)

To address the overfitting problem and improve the model robustness, we blend the key elements of CR and SVR and propose the convex support vector regression (CSVR) approach. Consider the following quadratic programming problem

$$\begin{aligned} \min_{\boldsymbol{\beta}_i, \alpha, \xi, \xi^*} \quad & \frac{1}{2} \sum_{i=1}^n \|\boldsymbol{\beta}_i\|_2^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) & (8) \\ \text{s.t.} \quad & y_i - \alpha_i - \boldsymbol{\beta}_i' \mathbf{x}_i \leq \varepsilon + \xi_i & \forall i \\ & \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i - y_i \leq \varepsilon + \xi_i^* & \forall i \\ & \alpha_i + \boldsymbol{\beta}_i' \mathbf{x}_i \leq \alpha_h + \boldsymbol{\beta}_h' \mathbf{x}_i & \forall i, h \\ & \xi_i \geq 0, \xi_i^* \geq 0 & \forall i \end{aligned}$$

where the first two constraints restrict the error terms into a specified margin, noted as the maximum error (ε), and consider the possible outliers with the deviation from the margin as ξ_i . The third constraint, a system of Afriat inequalities, guarantees the concavity of the unknown function f . Compared to problem (6), problem (8) is a shape constrained extension by means of Afriat inequalities. Note that the coefficients $\boldsymbol{\beta}_i$ represent the subgradient of the concave function f at point \mathbf{x}_i .

To further investigate the relationship between the CSVR and regularized function esti-

mation, we rewrite problem (8) as the following equivalent formulation

$$\begin{aligned} \min_{\beta_i, \alpha} \quad & \sum_{i=1}^n |y_i - \alpha_i - \beta'_i \mathbf{x}_i|_\varepsilon + \frac{A}{2} \sum_{i=1}^n \|\beta_i\|_2^2 \\ \text{s.t.} \quad & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i \quad \forall i, h \end{aligned} \quad (9)$$

where A is the tuning parameter, playing a similar role as C . The function $|\cdot|_\varepsilon$ indicates the ε -insensitive loss function described by equation (7). Note that the objective function has a form of *loss+penalty*; hence, the parameter A controls the trade-off between loss and penalty. The standard cross-validation techniques can be used to determine the tuning parameter.

The penalty term in problem (9) is a L_2 -norm of the subgradient vector, the same as the penalized convex regression using the norm of the subgradients (Aybat & Wang, 2014; Lim, 2014; Dai et al., 2022). The ridge penalty shrinks the subgradients towards zero, which means the regression function will be as flat as possible. This shrinkage can control the variance of the subgradients, thus helping to alleviate the overfitting problem via the bias-variance trade-off, especially when many highly correlated variables exist. We will further study the effect of regularization in the following section.

Using the ε -insensitive loss function also enables a sparse set of support vectors to be obtained. That is, the number of support vectors increases more slowly than linearly. As an extended SVR model, the CSVr approach retains this advantage of sparsity. Moreover, the ε -insensitive loss function is less sensitive to outliers than the quadratic loss function used in problem (5). CSVr is more robust than conventional CR when there are outliers in the dataset, therefore, leading to a smaller prediction error than the usual CR; see Section 4 for simulation evidence. When $\varepsilon = 0$, we can obtain a special case of L_1 loss function in problem (9). Note that the L_1 loss function is relatively more robust to outliers than the quadratic loss function (Alquier et al., 2019).

To derive a similar version to problem (5), letting $\phi_\varepsilon(\cdot)$ be the ε -insensitive loss function, problem (9) can be rewritten as fully constrained optimization problem as follows

$$\begin{aligned} \min_{\beta_i, \alpha} \quad & \sum_{i=1}^n \phi_\varepsilon(y_i - f(\mathbf{x}_i)) \\ \text{s.t.} \quad & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i \quad \forall i, h \\ & \sum_{i=1}^n \|\beta_i\|_2 \leq C \end{aligned} \quad (10)$$

where $\sum_{i=1}^n \|\beta_i\|_2 \leq C$ is a L_2 -norm penalty, and C is a nonnegative tuning parameter. We note here that the L_2 -norm penalty used in this study differs from the L_2 -norm Lipschitz regularization in (Mazumder et al. 2019). (Mazumder et al. 2019) consider a least square estimator over a set of convex functions that are uniformly Lipschitz with a known bound (see problem (5)), whereas problem (10) is ridge regularized. Also, note that we use a different loss function, whose insensitivity may contribute to the performance of CSVr.

3.2 Lasso CSVr

In addition to L_2 -norm, we could introduce other regularization methods such as L_1 -norm to extend the present CSVr approach. The L_1 -norm support vector machine was first proposed by (Bradley & Mangasarian 1998) for solving classification problems. Inspired by this idea, we briefly describe an extension of CSVr: the Lasso CSVr model for the variable selection regression analysis. This extension provides an alternative path of extending CSVr to select variables automatically. The first version of Lasso CSVr replaces the L_2 -norm penalty in problem (9) with L_1 -norm penalty

$$\begin{aligned} \min_{\beta_i, \alpha} \quad & \sum_{i=1}^n |y_i - \alpha_i - \beta_i' \mathbf{x}_i|_\varepsilon + \frac{A}{2} \sum_{i=1}^n \|\beta_i\|_1 \\ \text{s.t.} \quad & \alpha_i + \beta_i' \mathbf{x}_i \leq \alpha_h + \beta_h' \mathbf{x}_i \quad \forall i, h \end{aligned} \quad (11)$$

Similar to the L_2 -norm penalty, the L_1 -norm penalty can control the variance of the estimation and improve prediction accuracy. Moreover, the Lasso performs automatic variable selection, which is not the case for the L_2 -norm penalty. Although the performance of Lasso does not uniformly dominate the ridge regression (Tibshirani, 1996), the L_1 -norm CSVr appears very promising because the variable selection is increasingly important in modern data science.

The variable selection aspect of the L_1 -norm CSVr approach is useful for regression analysis in the case that there exist no highly correlated variables (see the limitations of L_1 -norm penalty in Zou & Hastie, 2005). While the L_1 -norm penalty can shrink the subgradients of the function and make the subgradients of irrelevant variables small, it cannot reduce them to zero exactly. The reason is that selecting variables by regularizing the subgradient β_i in problem (11) with a group sparsity penalty is not an effective way due to the existing of Afriat inequalities (Xu et al., 2016; Dai, 2022). That is, the small changes to β_i in each

Afraid inequality $\alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i$ may not violate the concavity assumption. Therefore, the L_1 -norm CSVR approach can make certain coefficients very small but not be zero, and thus does not necessarily make the representor function (4) sparse.

This motivates us to consider a L_∞ -regularized Lasso CSVR. As shown in (Zhao et al. 2009) and (Negahban & Wainwright 2011), the L_∞ -norm taking the maximum encourages all d components of the subgradient β_i to be zero simultaneously or to be nonzero simultaneously. The second version of Lasso CSVR can be formulated as the following optimization problem

$$\begin{aligned} \min_{\beta_i, \alpha} \quad & \sum_{i=1}^n |y_i - \alpha_i - \beta'_i \mathbf{x}_i|_\varepsilon + \frac{A}{2} \sum_{i=1}^n \|\beta_i\|_\infty \\ \text{s.t.} \quad & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i \quad \forall i, h \end{aligned} \quad (12)$$

Compared to the L_1 -norm penalty, the L_∞ -norm penalty restricts each subgradient $\beta_{j,i}$ with a fixed bound instead of all d components of the subgradient β_i . Furthermore, the L_∞ -regularized Lasso approach is a special case of the block L_1/L_∞ -regularization when the number of blocks is n in (Negahban & Wainwright 2011).

3.3 Illustrative example

We proceed to demonstrate how the fitted functions estimated by CSVR and convex regression are different and illustrate the potential advantages of CSVR in reducing overfitting via an artificial example. In so doing, we generate 50 observations with $y_i = 3 + \ln(x_i) + \epsilon_i$, where x_i is randomly drawn from $U(1, 10)$ and the error term ϵ_i is generated independently from $N(0, 0.7^2)$. To search the optimal hyperparameters, we resort to the standard fivefold cross-validation approach (see, e.g., Mazumder et al., 2019; Dai, 2022).

Fig. 1 depicts the fitted functions estimated by CSVR and convex regression. We observe that both shape-constrained approaches yield piecewise concave lines and can capture the shape of data points. The fitted convex regression function appears to be more sharper, while the fitted CSVR functions seem to be relatively smooth, implying the difference in the loss function. When comparing the approximation errors, CR has a smaller in-sample error (i.e., $MSE = 0.44$) in comparison with CSVR. However, a smaller training error may lead to a larger test error, suggesting that CR is hampered by overfitting, and CSVR can effectively avoid this problem.

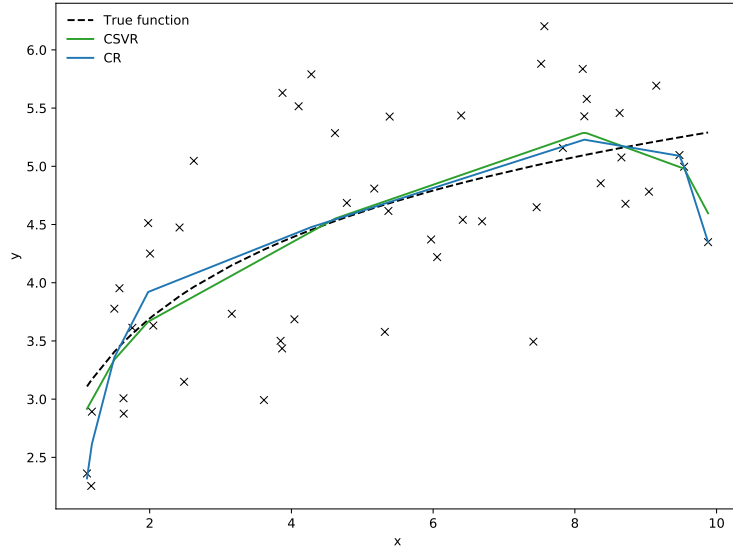


Fig. 1. Illustration of the fitted functions estimated by CSV and convex regression.

We then illustrate the robustness of the estimated CSV function to the choice of the tuning parameter C . In Fig. 2, the hyperparameter ε is fixed at 0.1 and C is tuned over 5 values from the set $\{1, 2, 4, 6, 10\}$. Note that the optimal hyperparameter $C^* = 6$ is determined by the standard cross-validation approach. It is evidently from Fig. 2 that the parameter C can reshape the estimated CSV functions but produce very similar estimated piecewise-linear curves.

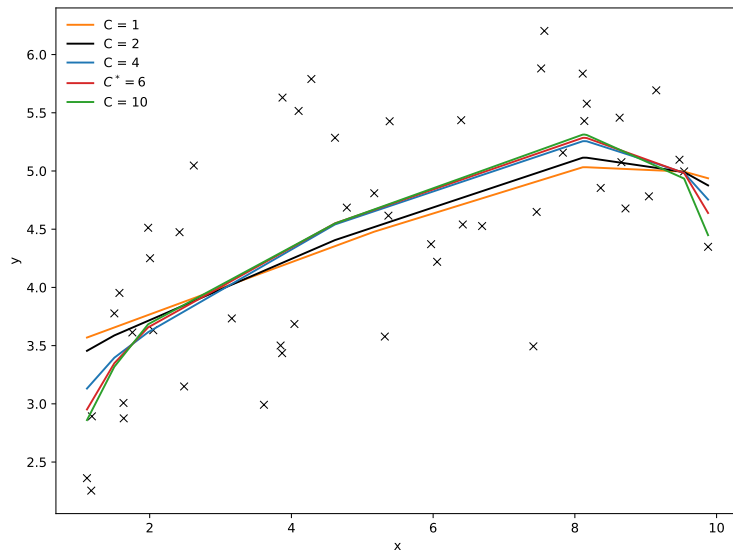


Fig. 2. Illustration of the estimated CSV functions with different values of C .

4 Monte Carlo study

Having illustrated the estimated CSVN function, we proceed to investigate the finite sample performance of the CSVN, CR, SVR with radial basis function kernel, and LCR approaches in the controlled environment of Monte Carlo simulations. The main objective of our simulations is to examine whether the proposed CSVN approach can better fit the true function by addressing the overfitting problem.

4.1 Setup

Consider the following data generating processes (DGP) (see, e.g., Valero-Carreras et al., 2021)

- 1) DGP I: $y = 3 + x_1^{0.5} + \epsilon$
- 2) DGP II: $y = 3 + x_1^{0.2} + x_2^{0.3} + \epsilon$
- 3) DGP III: $y = 3 + x_1^{0.05} + x_2^{0.15} + x_3^{0.3} + \epsilon$

where x_1 , x_2 , and x_3 are independently and randomly sampled from the uniform distribution $U[1, 10]$ and the error term ϵ is drawn from $N(0, \sigma^2)$. For each DGP, we consider 12 different scenarios with different number of observations ($n = 50, 100, 200, 500$) and the levels of noise ($\sigma = 0.5, 1, 2$). Each scenario is replicated 50 times to calculate the in-sample and out-of-sample performances with the mean squared error (MSE) statistic.

Regarding the tuning parameter selection, the optimal hyperparameters (i.e., C and ϵ) are determined in the SVR and CSVN approaches by using the fivefold cross-validation method. Following (Valero-Carreras et al. 2021), the tuning values C and ϵ are varied from the multiplier sets $\{0.1, 0.5, 1, 2, 5\}$ and $\{0, 0.001, 0.01, 0.1, 0.2\}$, respectively. For the LCR approach, as in Mazumder et al. (2019), we employ the one standard error rule in cross-validation to search the optimal Lipschitz parameter L .

In the following experiments, we resort to the pyStoNED package (Dai et al., 2021) with the standard solver Mosek (9.2.44). All computations are performed on Aalto University's high-performance computing cluster Triton with Xeon @2.8 GHz processors, 1 CPU, and 8 GB RAM per task. The simulation code and data are available at the GitHub repository (<https://github.com/ds2010/CSVN>).

4.2 In-sample performance

We first compare the in-sample performance of CSVr with alternatives. Table 1 reports the MSE statistic of each approaches with $n \in \{50, 100, 200, 500\}$, $d \in \{1, 2, 3\}$, and $\sigma = 1$. Note that the MSEs of NCCSVr in the univariate cases are missing due to the fact that the NCCSVr approach requires the multidimensional data space (i.e., $d \geq 2$) (Wang & Ni, 2012). The optimal hyperparameters in LCR, CSVr, and NCCSVr are prespecified via the standard cross-validation technique.

Table 1 indicates that CSVr exhibits the lowest values of MSE in almost all scenarios. Compared to CSVr, LCR has competitive performance in the univariate case (i.e., DGP I), but its performance deteriorates in multivariate cases.

Compared to the regularized approaches, the traditional CR approach performs poorly in the multivariate cases (i.e., DGP II and III), even though it is actually quite competitive in the univariate setting. After restricting the subgradients in CR, the performance in fitting the true function will be improved as the optimal subgradients cannot take any values for given feasibility. As expected, the performance of each approach deteriorates as more input variables or smaller sample sizes are introduced.

Table 1. MSE comparison of different approaches with $\sigma = 1$.

DGP	n	CSVr	SVr	CR	LCR	NCCSVr
I	50	0.0644	0.1099	0.0838	0.0673	–
	100	0.0384	0.0694	0.0427	0.0377	–
	200	0.0209	0.0403	0.0252	0.0192	–
	500	0.0086	0.0152	0.0107	0.0080	–
II	50	0.0771	0.1125	0.2193	0.1674	0.1085
	100	0.0386	0.0696	0.1270	0.0904	0.0596
	200	0.0271	0.0400	0.0829	0.0536	0.0300
	500	0.0083	0.0188	0.0341	0.0222	0.0123
III	50	0.0758	0.0994	0.4189	0.3198	0.1543
	100	0.0556	0.0788	0.2765	0.2124	0.0883
	200	0.0365	0.0484	0.1855	0.1276	0.0453
	500	0.0184	0.0274	0.0974	0.0599	0.0188

To assess the robustness of the CSVr approach, we next consider scenarios with different levels of error variance (i.e., σ) using a fixed sample size $n = 500$ (see Table 2). As expected, we observe that the MSE values increase for all methods as the data noise variance increases.

However, the proposed CSVN approach has the smallest MSEs in all cases and is robust to the increasing noise. Note that those methods that use the regularization techniques are more robust to noise than the original CR approach. Overall, CSVN maintains its good performance in different levels of data noise and dimensions.

Table 2. MSE comparison of different approaches with $n = 500$.

DGP	σ	CSVN	SVN	CR	LCR	NCCSVN
I	0.5	0.0030	0.0043	0.0031	0.0025	–
	1	0.0086	0.0152	0.0107	0.0080	–
	2	0.0257	0.0522	0.0392	0.0271	–
II	0.5	0.0038	0.0076	0.0090	0.0066	0.0043
	1	0.0083	0.0188	0.0341	0.0222	0.0123
	2	0.0214	0.0482	0.1341	0.0755	0.0494
III	0.5	0.0077	0.0114	0.0248	0.0185	0.0070
	1	0.0184	0.0274	0.0974	0.0599	0.0188
	2	0.0493	0.0642	0.3874	0.1862	0.0671

Compared to other regularized approaches, CSVN is relatively more robust when the data noise varies from 0.5 to 2 (see Table 2). While the LCR approach can benefit from the Lipschitz regularization and avoid overfitting, it still uses the squared L_2 norm loss function and thus tends to be sensitive to outliers and heteroscedasticity. However, the CSVN approach introduces the ϵ -insensitive loss function inherited from SVN to increase robustness. For smaller σ values, as shown in Table 2, it seems that NCCSVN may outperform the proposed CSVN method (i.e. DGP III, $\sigma = 0.5$).

Recall that (Wang & Ni 2012) have proposed a similar NCCSVN approach, where they introduce convexity/concavity into SVN by using the Hessian matrix and then transform it to a semidefinite programming problem. In contrast, our CSVN approach resorts to the system of Afriat inequalities to ensure the fitted function to be convex/concave. Moreover, the NCCSVN approach can be applied to the multivariate cases only (i.e., $d \geq 2$) and requires one more tuning parameter. To further understand the performance difference between these two similar approaches, we implement the following additional experiment to report the prediction accuracy.

In previous experiments, we have observed that the variation of the noise σ has a significant impact on the MSE values, especially in the cases with small σ . Here we investigate the

impact of noise variation on these two models' performance. Fig. 3 depicts the performance of CSVr and NCCSVr when $n = 500$, $d = \{2, 3\}$, and σ varies from 0.2 to 3. Fig. 3 a) indicates that the performance of CSVr dominates NCCSVr, whereas Fig. 3 b) shows that there are tiny differences in the finite sample performances between NCCSVr and CSVr, but increasing larger differences occur as σ grows. This might suggest that imposing the Hessian matrix is not as efficient as the system of Afriat inequalities in terms of overfitting reduction.

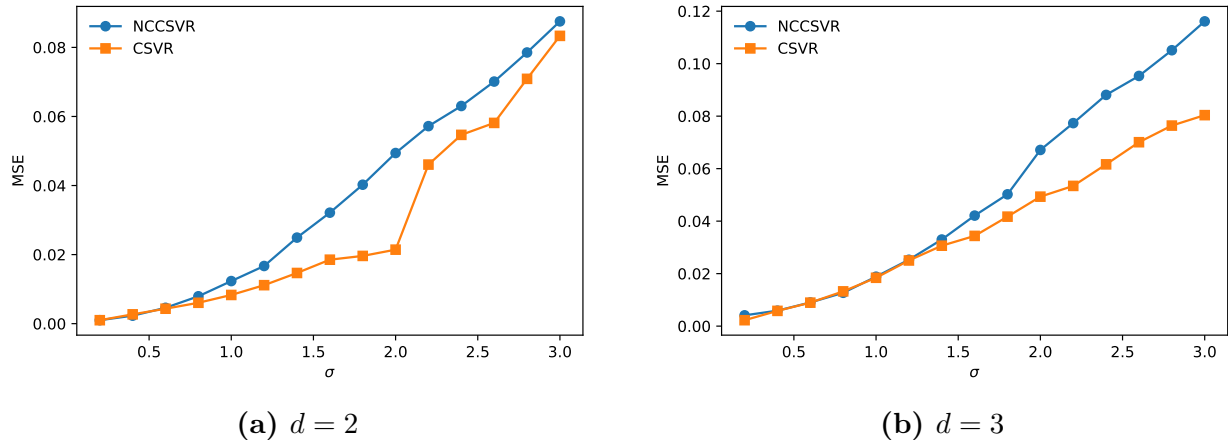


Fig. 3. Illustration of the impacts of noise variation on MSE.

We further investigate the impact of outliers. The additional five outliers are drawn from the uniform distribution $U[90, 100]$, and the other normal observations are also drawn from $U[1, 10]$. For the sake of illustration, we simply consider the instances with $n \in \{50, 100\}$, $d \in \{2, 3\}$, and $\sigma = 1$. The in-sample MSEs of each scenario are averaged in a total of 50 replications. As expected, Table 3 shows that the CSVr approach performs best among all compared approaches in terms of prediction accuracy. The Lipschitz norm convex regression approach described in Mazumder et al. (2019) can also control the impact of outliers and, in this case, outperforms the conventional SVR approach, which would have better performance if there were no additional outliers. Furthermore, compared to results reported in Table 1, the results shown in Table 3 imply that the additional outliers can lead to worse accuracy for all methods. We also observe that the performance of NCCSVr varies sharply with the parameters, but this does not happen in CSVr. Moreover, the choice of its kernel parameter is highly dependent on the values of input data, which might lead to a limitation

in practical applications. Thus it could be hurt for practitioners to decide the range of the three parameters used in NCCSVR.

Table 3. MSE comparison of different approaches with additional five outliers.

DGP	n	CSVR	SVR	CR	LCR	NCCSVR
II	50	0.0979	0.1900	0.1948	0.1477	0.1096
	100	0.0563	0.1175	0.1154	0.0879	0.0595
III	50	0.0996	0.2251	0.4093	0.1034	0.1556
	100	0.0658	0.1115	0.2074	0.0897	0.1065

4.3 Out-of-sample performance

We proceed to investigate out-of-sample performance by considering six scenarios with $d \in \{2, 3\}$, $\sigma \in \{0.5, 1, 2\}$, and $n = 500$ for the training set and another 1000 hold-out observations for the test set. We then replicate each scenario 50 times to obtain an empirical distribution of the out-of-sample MSE statistic. Note that the in-sample overfitting may result in low prediction accuracy of the out-of-sample model, and if an estimator is likely to overfit in multidimensional data space, then it will have a smaller in-sample MSE and a larger out-of-sample MSE.

The boxplots in Figs. 4 and 5 illustrate the distributions of out-of-sample MSE. When comparing traditional CR with regularized alternatives, we observe that CR has a relatively large out-of-sample MSE, which is far more than that of the other regularized approaches. For instance, the values of out-of-sample MSE for CR are $0.121(\pm 0.342)$, $0.454(\pm 1.192)$, and $1.705(\pm 4.300)$ respectively with $d = 2$ and $\sigma \in \{0.5, 1, 2\}$. To facilitate a comparison of the most competitive alternatives, we exclude the MSE results for CR from Figs. 4 and 5.

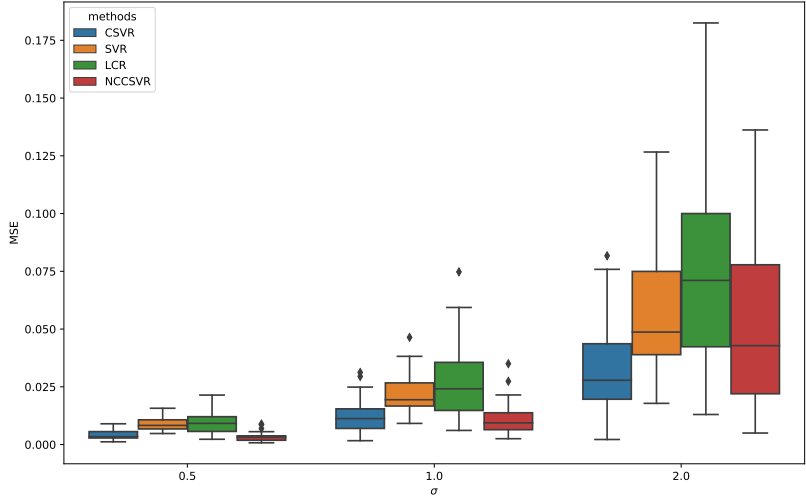


Fig. 4. Out-of-sample MSE of the four methods with $d = 2$ and $\sigma \in \{0.5, 1, 2\}$.

CSVR performs better than other methods in alleviating the overfitting problem. Compared to the LCR, the three SVR-based approaches seem to perform better in alleviating the overfitting problem. The ϵ -insensitive loss function used in SVR-based approaches is more robust to outliers and large errors than the least squares loss function, which, in turn, helps to build a robust prediction model as demonstrated in Fig. 5. In the experiments, we also find that the out-sample performance of LCR deteriorates rapidly as the data noise increases.

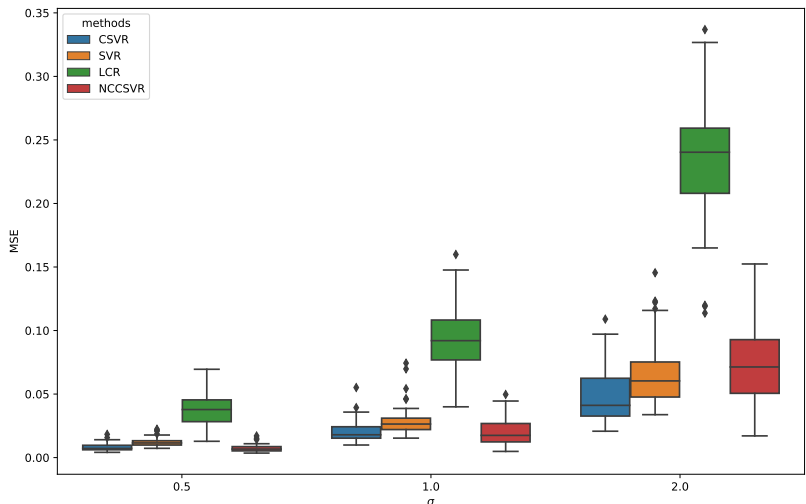


Fig. 5. Out-of-sample MSE of the four methods with $d = 3$ and $\sigma \in \{0.5, 1, 2\}$.

5 Experiments with real data

In this section, we apply the proposed approach to two real-world datasets: the Boston housing data and the NBER-CES manufacturing industry data. Those real datasets have been used extensively in various fields of economics, econometrics, statistics, and machine learning for different purposes such as new models test or algorithms benchmarking (see, e.g., Lin et al., 2011; Wang & Wang, 2013). In both examples below, all hyperparameters are tuned over from 50 candidate values via the fivefold cross-validation technique, and the in-sample MSE and out-of-sample MSE for each approach are calculated to compare the effectiveness in reducing overfitting.

5.1 Boston housing data

This dataset contains housing price information in the Boston area collected from the StatLib archive.¹ The data includes 13 variables with 506 observations.² Following the commonly used setting, the variable *MEDV* is taken as the response variable of interest, and others are the explanatory variables. The descriptive statistics for all variables are presented in Table A1.

Table 4 reports the descriptive statistics for the estimated coefficients (i.e., $\hat{\alpha}_i$ and $\hat{\beta}_i$) of CSVR, LCR, and CR. We omit SVR and NCCSVR here due to the incomparable dual variable (i.e., $\hat{\beta}_i$) (see Smola & Schölkopf, 2004). As shown in Table 4, all values of β for CSVR and LCR lie roughly between -0.5 and 0.5, whereas those values for CR lie in a larger interval, which can be a symptom of overfitting. We also note that although both CSVR and LCR have small β , LCR obtains coefficients even a bit smaller than CSVR. Furthermore, LCR produces a more flat function f , but that does not mean a better performance automatically.

To evaluate the prediction accuracy, we divide the 506 samples randomly into 405 training samples and 101 test samples and then compute the in-sample and out-of-sample prediction error in terms of MSE. The calculated in-sample and out-of-sample MSEs and estimated standard deviations are demonstrated in Table 5. Overall, as expected, we observe that

¹StatLib-Datasets Archive: <http://lib.stat.cmu.edu/datasets/boston>.

²Note that the dummy variable (i.e., *CHAS*) is excluded from the dataset for the sake of simplicity. Note that CR can handle contextual variables, but this falls beyond the scope of this paper; cf. Johnson & Kuosmanen (2012), for further details.

the CSVR method achieves the best prediction performance. The results show that the restriction on the subgradients or regularization described in Section 3 leads to a flatter estimated function that can overcome overfitting. Although LCR is also developed for solving overfitting problems, our approach still outperforms the LCR method. A possible explanation is that owing to the structure of the SVR-based approaches, CSVR can achieve a good bias-variance trade-off. Compared to the non-regularized method (i.e., CR in Table 5), all methods benefit from additional regularization in terms of the out-of-sample prediction accuracy. However, note that CR yields the lowest in-sample MSE in this empirical data.

Table 4. Descriptive statistics for estimates of all explanatory variables.

	CSV		LCR		CR	
	Mean	Std.Dev.	Mean	Std.Dev.	Mean	Std.Dev.
$\hat{\beta}_{CRIM}$	0.08	0.17	-0.01	0.08	2.66	10.47
$\hat{\beta}_{ZN}$	0.12	0.15	0.07	0.06	0.80	9.22
$\hat{\beta}_{INDUS}$	-0.02	0.22	-0.03	0.11	2.59	12.29
$\hat{\beta}_{NOX}$	0.00	0.04	0.00	0.02	147.61	963.24
$\hat{\beta}_{RM}$	0.19	0.34	0.05	0.10	11.10	64.63
$\hat{\beta}_{AGE}$	-0.02	0.19	-0.01	0.08	-0.39	2.87
$\hat{\beta}_{DIS}$	-0.17	0.34	-0.04	0.10	-0.60	48.50
$\hat{\beta}_{RAD}$	0.09	0.26	0.06	0.10	-5.01	20.08
$\hat{\beta}_{TAX}$	-0.01	0.05	-0.01	0.02	0.09	0.85
$\hat{\beta}_{PTRATIO}$	-0.18	0.33	-0.06	0.11	1.89	42.85
$\hat{\beta}_B$	-0.21	0.26	-0.08	0.10	-0.25	0.88
$\hat{\beta}_{LSTAT}$	-0.42	0.29	-0.28	0.12	0.24	7.68
$\hat{\alpha}$	132.61	103.68	69.44	39.31	276.42	1264.22

Table 5. In-sample and out-of-sample MSEs: Boston housing data.

MSE	CSV	SVR	CR	LCR	NCCSVR
Out-of-sample MSE	38.44(5.04)	64.00(14.28)	2334.17(933.04)	43.72(6.47)	42.72(9.28)
In-sample MSE	21.43(0.86)	62.57(2.90)	1.41(0.33)	35.05(1.52)	40.17(2.05)

Note: standard deviation in parentheses.

5.2 NBER-CES manufacturing industry data

The data consist of 473 manufacturing industries (the 1997 6-digit NAICS codes) and are collected from the NBER-CES Manufacturing Industry Database.³ In this application, we apply the present approaches to estimate the production function, where, as in Wang & Ni (2012), the input includes the capital (INVEST), labor (PAY), and raw materials (MAT-COST), and the output is the value added (VADD). Furthermore, the training and test set are 376 and 97 observations, respectively.

CSVR’s in-sample and out-of-sample MSEs are 0.53×10^7 and 0.99×10^7 (see Table 6). While the in-sample MSE value of CSVR is not the lowest in comparison with other shape-constrained approaches, the out-of-sample MSE remains the lowest. This is perhaps due to the fact that there is a good trade-off between model performance and overfitting alleviation.

Table 6. In-sample and out-of-sample MSEs ($\times 10^7$): NBER-CES manufacturing industry data.

MSE	CSVR	SVR	CR	LCR	NCCSVR
Out-of-sample MSE	0.99	3.96	74.78	1.09	1.83
In-sample MSE	0.53	3.94	0.37	0.91	1.01

It is worth noting that CR has the best in-sample fit in the applications but not in the MC simulations. It is because the MSE is measured differently. In the simulations, we measure the deviation between the estimated function f and the true function F . In the applications, the true function f is unknown, and we thus measure the deviation of the predictions from the observed y . This is not the same MSE because $y = f(x) + \epsilon$ also includes the noise. CR will always minimize the MSE wrt y , but due to overfitting not wrt $f(x)$.

In conclusion, both simulations and real-world applications demonstrate that all regularized shape-constrained methods have a superior ability to control overfitting, but CSVR would be more appealing than other regularized shape-constrained methods because of its simplicity, capacity for univariate regression, and robust performance.

³NBER database: <https://www.nber.org/research/data/nber-ces-manufacturing-industry-database>.

6 Conclusions

Overfitting is a commonly seen phenomenon in nonparametric regression. To mitigate the effects of overfitting, we have introduced a new approach called convex support vector regression, which effectively combines the key elements of support vector regression and convex regression. The paper investigates the finite sample performance of the developed CSVN approach in contrast to other state-of-the-art regression methods through Monte Carlo simulations. Additional two real-world datasets are also used to test and compare the performance of these approaches. We hope that the proposed approach can help to further bridge the gaps between the data-driven estimation approaches known in econometrics and statistics, machine learning, and operations research and management science.

The evidence from the simulations indicates that CSVN performs at least as well as LCR and much better than SVR and traditional convex regression. Two real-world applications also show that our approach outperforms other state-of-the-art regression methods. The regularized convex regression model can help to alleviate the overfitting problem, also owing to its insensitive loss function and robustness in the presence of outliers.

In this paper, we have restricted attention to regularizations known in the literature, but there could be more efficient ways to restrict the domain of subgradients (e.g., weight-restricted regression). Another promising future research direction is developing the simple and fast algorithmic framework of convex support vector regression. The main bottleneck of CSVN is that the full problem (8) has $n(n-1)$ constraints and thus becomes computationally inefficient for more than a few thousand observations. Furthermore, we have deliberately kept away from statistical inferences, and further work in this direction, e.g., exploring the asymptotic property of CSVN, would be needed.

Acknowledgments

We acknowledge the computational resources provided by the Aalto Science-IT project. Zhiqiang Liao gratefully acknowledges financial support from the Foundation for Economic Education (Liikesivistysrahasto) [grant no. 210038]. Sheng Dai gratefully acknowledges financial support from the Foundation for Economic Education (Liikesivistysrahasto) [grants no. 180019, 190073, 210075] and the HSE Support Foundation [grant no. 11-2290].

References

- Alquier, P., Cottet, V. & Lecué, G. (2019), Estimation bounds and sharp oracle inequalities of regularized procedures with lipschitz loss functions, *The Annals of Statistics* 47, 2117–2144.
- Aybat, N. S. & Wang, Z. (2014), A parallel method for large scale convex regression problems, *Proceedings of the IEEE Conference on Decision and Control*, pp. 5710–5717.
- Balázs, G., György, A. & Szepesvári, C. (2015), Near-optimal max-affine estimators for convex regression, *18th Artificial Intelligence and Statistics*, PMLR, pp. 38:56–64.
- Bradley, P. S. & Mangasarian, O. L. (1998), Feature selection via concave minimization and support vector machines., *Proceedings of the 15th International Conference on Machine Learning*, pp. 98:82–90.
- Chen, X., Lin, Q. & Sen, B. (2020), On degrees of freedom of projection estimators with applications to multivariate nonparametric regression, *Journal of the American Statistical Association* 115, 173–186.
- Cortes, C. & Vapnik, V. (1995), Support-vector networks, *Machine Learning* 20, 273–297.
- Dai, S. (2022), Variable selection in convex quantile regression: L_1 -norm or L_0 -norm regularization?, *European Journal of Operational Research* .
- Dai, S., Fang, Y. H., Lee, C. Y. & Kuosmanen, T. (2021), pyStoNED: A Python package for convex regression and frontier estimation, *arXiv preprint arXiv:2109.12962*.
- Dai, S., Kuosmanen, T. & Zhou, X. (2022), Non-crossing convex quantile regression, *arXiv preprint arXiv:2204.01371*.
- Ghosal, P. & Sen, B. (2017), On univariate convex regression, *Sankhya A* 79, 215–253.
- Guntuboyina, A. & Sen, B. (2018), Nonparametric shape-restricted regression, *Statistical Science* 33, 568–594.
- Hildreth, C. (1954), Point estimates of ordinates of concave functions, *Journal of the American Statistical Association* 49, 598–619.
- Johnson, A. L. & Jiang, D. R. (2018), Shape constraints in economics and operations research, *Statistical Science* 33, 527–546.
- Johnson, A. L. & Kuosmanen, T. (2012), One-stage and two-stage DEA estimation of the effects of contextual variables, *European Journal of Operational Research* 220, 559–570.
- Kuosmanen, T. (2008), Representation theorem for convex nonparametric least squares, *Econometrics Journal* 11, 308–325.
- Kuosmanen, T. & Johnson, A. L. (2010), Data envelopment analysis as nonparametric least-squares regression, *Operations Research* 58, 149–160.
- Kuosmanen, T., Johnson, A. L. & Saastamoinen, A. (2015), Stochastic nonparametric approach to efficiency analysis: A unified framework, in J. Zhu (ed.), *Data Envelopment Analysis*, Springer, Boston, MA, chapter 7, pp. 191–244.

- Kuosmanen, T. & Zhou, X. (2021), Shadow prices and marginal abatement costs: Convex quantile regression approach, *European Journal of Operational Research* 289, 666–675.
- Kuosmanen, T., Zhou, X. & Dai, S. (2020), How much climate policy has cost for oecd countries?, *World Development* 125, 104681.
- Lee, C. Y., Johnson, A. L., Moreno-Centeno, E. & Kuosmanen, T. (2013), A more efficient algorithm for convex nonparametric least squares, *European Journal of Operational Research* 227, 391–400.
- Lim, E. (2014), On convergence rates of convex regression in multiple dimensions, *INFORMS Journal on Computing* 26, 616–628.
- Lin, D., Foster, D. P. & Ungar, L. H. (2011), VIF regression: A fast regression algorithm for large data, *Journal of the American Statistical Association* 106, 232–247.
- Lin, M., Sun, D. & Toh, K. C. (2022), An augmented Lagrangian method with constraint generation for shape-constrained convex regression problems, *Mathematical Programming Computation* 14, 223–270.
- Mazumder, R., Choudhury, A., Iyengar, G. & Sen, B. (2019), A computational framework for multivariate convex regression and its variants, *Journal of the American Statistical Association* 114, 318–331.
- Negahban, S. N. & Wainwright, M. J. (2011), Simultaneous support recovery in high dimensions: Benefits and perils of block ℓ_1/ℓ_∞ -regularization, *IEEE Transactions on Information Theory* 57, 3841–3863.
- Seijo, E. & Sen, B. (2011), Nonparametric least squares estimation of a multivariate convex regression function, *The Annals of Statistics* 39, 1633–1657.
- Smola, A. J. & Schölkopf, B. (2004), A tutorial on support vector regression, *Statistics and computing* 14, 199–222.
- Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society Series B* 58, 267–288.
- Valero-Carreras, D., Aparicio, J. & Guerrero, N. M. (2021), Support vector frontiers: A new approach for estimating production functions through support vector machines, *Omega* 104, 102490.
- Valero-Carreras, D., Aparicio, J. & Guerrero, N. M. (2022), Multi-output support vector frontiers, *Computers & Operations Research* 143, 105765.
- Vapnik, V. (1999), *The nature of statistical learning theory*, Springer, New York.
- Wang, Y. & Ni, H. (2012), Multivariate convex support vector regression with semidefinite programming, *Knowledge-Based Systems* 30, 87–94.
- Wang, Y. & Wang, S. (2013), Estimating α -frontier technical efficiency with shape-restricted kernel quantile regression, *Neurocomputing* 101, 243–251.
- Wang, Y., Wang, S., Dang, C. & Ge, W. (2014), Nonparametric quantile frontier estimation under shape restriction, *European Journal of Operational Research* 232, 671–678.

- Xu, M., Chen, M. & Lafferty, J. (2016), Faithful variable screening for high-dimensional convex regression, *The Annals of Statistics* 44, 2624–2660.
- Yagi, D., Chen, Y., Johnson, A. L. & Kuosmanen, T. (2020), Shape-constrained kernel-weighted least squares: Estimating production functions for Chilean manufacturing industries, *Journal of Business & Economic Statistics* 38, 43–54.
- Zhao, P., Rocha, G. & Yu, B. (2009), The composite absolute penalties family for grouped and hierarchical variable selection, *The Annals of Statistics* 37, 3468–3497.
- Zou, H. & Hastie, T. (2005), Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society Series B* 67, 301–320.

Appendix

Table A1. Descriptive statistics of all variables: the Boston housing data.

Variable	Description	Mean	Std. Dev.	Min.	Max.
CRIM	per capita crime rate by town	3.61	8.60	0.01	88.98
ZN	proportion of residential land zoned	11.36	23.32	0.00	100.00
INDUS	proportion of non-retail business acres per town	11.14	6.86	0.46	27.74
NOX	nitric oxides concentration	0.55	0.11	0.36	0.87
RM	average number of rooms per dwelling	6.28	0.70	3.56	8.78
AGE	proportion of owner built prior to 1940	68.57	28.15	2.90	100.00
DIS	weighted distances to city centres	3.80	2.11	1.13	12.13
RAD	index of accessibility to radial highways	9.55	8.71	1.00	24.00
TAX	full-value property-tax rate per \$10,000	408.24	168.54	187.00	711.00
PTRATIO	pupil-teacher ratio by town	18.46	2.16	12.60	22.00
B	black proportion of population	356.67	91.29	0.32	396.90
LSTAT	proportion of population that is lower status	12.65	7.14	1.73	37.97
MEDV	median value of owner-occupied homes	22.53	9.20	5.00	50.00