
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Liao, Zhiqiang; Dai, Sheng; Kuosmanen, Timo

Convex support vector regression

Published in:
European Journal of Operational Research

DOI:
[10.1016/j.ejor.2023.05.009](https://doi.org/10.1016/j.ejor.2023.05.009)

E-pub ahead of print: 16/03/2024

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Liao, Z., Dai, S., & Kuosmanen, T. (2024). Convex support vector regression. *European Journal of Operational Research*, 313(3), 858-870. Advance online publication. <https://doi.org/10.1016/j.ejor.2023.05.009>

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.



Continuous Optimization

Convex support vector regression

Zhiqiang Liao^a, Sheng Dai^{b,*}, Timo Kuosmanen^b^a Department of Information and Service Management, Aalto University School of Business, Finland^b Department of Economics, Turku School of Economics, University of Turku, Finland

ARTICLE INFO

Article history:

Received 26 September 2022

Accepted 4 May 2023

Keywords:

Robustness and sensitivity analysis

Convex regression

Support vector regression

Overfitting

Regularization

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 (CSVR) method, which effectively combines the key elements of convex regression and support vector regression. Numerical experiments demonstrate the performance of CSVR in prediction accuracy and robustness that compares favorably with other state-of-the-art methods.

© 2023 The Author(s). Published by Elsevier B.V.

This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>)

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 (Kuosmanen et al., 2015; Wang et al., 2014) and convex expectile regression (Kuosmanen & Zhou, 2021; Kuosmanen et al., 2020). There has been significant development in the computational tools and algorithms (see, e.g., Lee et al., 2013; Lin et al., 2022; Mazumder et al., 2019).

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 (Chen et al., 2020; Seijo & Sen, 2011), 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 with bounded 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) is 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 provides us with the theoretical motivation to combine the key elements of both CR and SVR in a unified framework.

* Corresponding author.

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

Our empirical motivation to combine CR and SVR comes from the benchmark regulation of electricity distribution networks, which is one of the most significant real-world applications of CR. For example, the Finnish Energy Authority systematically applies CR to implement incentive regulation for Finnish electricity distribution firms since 2012 (Kuosmanen, 2012; Kuosmanen & Johnson, 2020). The key advantage of CR is that it enables the regulator to ensure that the cost function used as a benchmark satisfies the necessary monotonicity, convexity, and linear homogeneity conditions that are essential to creating the right incentives for regulated firms. However, the incentives for future time periods involve predictive analytics based on historical data, and hence performance of the nonparametric CR may suffer from overfitting. The proposed combination of CR and SVR can provide a useful tool to combine flexible nonparametric modeling of appropriate shape constraints that allows one to alleviate overfitting to improve the accuracy of the out-of-sample predictions.

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 four 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 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 Section 5, we experimentally compare CSVr against competing methods on four 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

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

where $\mathbf{x}_i \in \mathbb{R}^d$ 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

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

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

$$\min_{f \in \mathcal{F}} \frac{1}{2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2. \tag{2}$$

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

$$\min_{\beta, \alpha, \varepsilon} \frac{1}{2} \sum_{i=1}^n \varepsilon_i^2 \tag{3}$$

$$\text{s.t. } y_i = \alpha_i + \beta_i' \mathbf{x}_i + \varepsilon_i \quad \forall i,$$

$$\alpha_i + \beta_i' \mathbf{x}_i \leq \alpha_h + \beta_h' \mathbf{x}_i \quad \forall i, h,$$

where β' indicates the transpose of β , and its subscript h is any index of data point not equal to i . The first constraint of (3) simply restates the regression Eq. (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 β (e.g., $\beta \geq 0$ for monotonic increasing and $\beta \leq 0$ for monotonic decreasing functions). Further, imposing $\alpha = 0$ imposes linear homogeneity (constant returns to scale). See Kuosmanen et al. (2015) for a more detailed discussion.

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

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

However, the estimated coefficients $\hat{\beta}_i$ 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., Chen et al., 2020; Mazumder et al., 2019). 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 \mathcal{F}_L of concave functions with Lipschitz regularization

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

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

$$\min_{f \in \mathcal{F}_L} \frac{1}{2} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2. \tag{5}$$

Similarly, such an infinite-dimensional problem can be solved by the following regularized optimization problem

$$\min_{\beta, \alpha, \varepsilon} \frac{1}{2} \sum_{i=1}^n \varepsilon_i^2 \tag{6}$$

$$\begin{aligned}
 \text{s.t. } & y_i = \alpha_i + \beta'_i \mathbf{x}_i + \varepsilon_i && \forall i, \\
 & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i && \forall i, h, \\
 & \|\beta_i\|_2 \leq L && \forall i.
 \end{aligned}$$

Note that the estimated function f^{LCR} can be obtained by inserting the optimal $\hat{\alpha}_i$ and $\hat{\beta}_i$ from (6) to the equation of explicit representer function (4). We can also resort to other Lipschitz norms (e.g., $\|\cdot\|_\infty$ -norm) to address the overfitting problem in convex regression (Balázs et al., 2015; Lim, 2014). 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), thereby reducing overfitting and contributing to the small generalization error.¹

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 the unwanted data exit, 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

$$\begin{aligned}
 \min_{\beta, \alpha, \xi, \xi^*} & \frac{1}{2} \|\beta\|_2^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) && (7) \\
 \text{s.t. } & y_i - \alpha - \beta'_i \mathbf{x}_i \leq \varepsilon + \xi_i && \forall i, \\
 & \alpha + \beta'_i \mathbf{x}_i - y_i \leq \varepsilon + \xi_i^* && \forall i, \\
 & \xi_i \geq 0, \xi_i^* \geq 0 && \forall i,
 \end{aligned}$$

where C is a prespecified parameter determining the trade-off between the complexity of regression function $f(x)$ and the prediction accuracy. This corresponds to the most commonly adopted ε -insensitive loss function $|\xi|_\varepsilon$ described by

$$|\xi|_\varepsilon = \begin{cases} 0, & \text{if } |\xi| \leq \varepsilon; \\ |\xi| - \varepsilon, & \text{otherwise,} \end{cases} \quad (8)$$

for a user-determined nonnegative number ε . A potential benefit of using the ε -insensitive loss function is robustness to outliers

¹ Note that by restricting the capacity of the function, we can limit the complexity of the model and hence improve the generalization performance of the model while also making it more interpretable, faster to train, and more computationally resource-efficient (Goodfellow et al., 2016).

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 $f^{SVR}(\mathbf{x}) = \alpha + \beta'_i \mathbf{x}$.

The SVR approach (7) can be extended to integrate the kernel techniques such as radial basis function (RBF) kernel and polynomial (poly) kernel. By using such a kernel trick, we can model data with nonlinear relationships in a linear way and fit the function in the corresponding feature space. For more detailed theoretical properties and development of SVR and kernel SVR, we refer the reader to Vapnik (1999).

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

$$\begin{aligned}
 \min_{\beta_i, \alpha, \xi, \xi^*} & \frac{1}{2} \sum_{i=1}^n \|\beta_i\|_2^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) && (9) \\
 \text{s.t. } & y_i - \alpha_i - \beta'_i \mathbf{x}_i \leq \varepsilon + \xi_i && \forall i, \\
 & \alpha_i + \beta'_i \mathbf{x}_i - y_i \leq \varepsilon + \xi_i^* && \forall i, \\
 & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i && \forall i, h, \\
 & \xi_i \geq 0, \xi_i^* \geq 0 && \forall i,
 \end{aligned}$$

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 (7), problem (9) is a shape-constrained extension by means of Afriat inequalities. Note that the coefficients β_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 estimation, we rewrite problem (9) as the following equivalent formulation

$$\begin{aligned}
 \min_{\beta_i, \alpha} & \sum_{i=1}^n |y_i - \alpha_i - \beta'_i \mathbf{x}_i|_\varepsilon + \frac{A}{2} \sum_{i=1}^n \|\beta_i\|_2^2 && (10) \\
 \text{s.t. } & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i \quad \forall i, h,
 \end{aligned}$$

where A is the tuning parameter, playing a similar role as C . The function $|\cdot|_\varepsilon$ indicates the ε -insensitive loss function described by Eq. (8). Note that the objective function has a form of *loss+penalty*; hence, 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 (10) is an L_2 -norm of the subgradient vector, the same as the penalized convex regression using the norm of the subgradients (Aybat & Wang, 2014; Dai et al., 2022; Lim, 2014). The ridge penalty can control the magnitude of estimated subgradients, indicating that the fitted regression function can 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 (6). CSVR is more robust than conventional CR when there are outliers in the dataset, thereby 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 (10). 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 (6), letting $\phi_\varepsilon(\cdot)$ be the ε -insensitive loss function, problem (10) can be rewritten as a fully constrained optimization problem as follows

$$\begin{aligned} \min_{\beta_i, \alpha} \quad & \sum_{i=1}^n \phi_\varepsilon(y_i - f(\mathbf{x}_i)) \quad (11) \\ \text{s.t.} \quad & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i \quad \forall i, h, \\ & \sum_{i=1}^n \|\beta_i\|_2^2 \leq C, \end{aligned}$$

where $\sum_{i=1}^n \|\beta_i\|_2^2 \leq C$ is an 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 (6)), whereas problem (11) is ridge regularized. Also, note that we use a different loss function, whose insensitivity may contribute to the performance of CSVR.

The proposed CSVR approach does not need to find the appropriate kernel function like the kernel SVR approach. Instead, CSVR directly fits the optimal concave function in the input space by imposing multiple shape constraints. Nevertheless, learning a concave function in the feature space is an interesting and promising avenue. Most of the existing work on function estimation in the high-dimensional feature space is for linear functions—imposing shape constraints in the feature space is a challenging problem. A few attempts have been made to apply the kernel techniques to handle shape constraints in nonparametric regression (see, e.g., Du et al., 2013; Wang & Ni, 2012). We leave further development of the kernel CSVR and its comparisons with CVSR for future research.

3.2. Lasso CSVR

In addition to L_2 -norm, we could introduce other regularization methods (e.g., L_1 -norm) to extend the present CSVR approach. The L_1 -norm support vector machine is 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 (10) with an L_1 -norm penalty

$$\min_{f \in \mathcal{F}} \sum_{i=1}^n \phi_\varepsilon(y_i - f(\mathbf{x}_i)) + A \sum_{i=1}^n \|\beta_i\|_1, \quad (12)$$

where $\phi_\varepsilon(\cdot)$ is the ε -insensitive loss function, and A is the tuning parameter. Problem (12) can be transformed into the following equivalent finite-dimensional optimization problem

$$\begin{aligned} \min_{\beta_i, \alpha} \quad & A \sum_{i=1}^n \|\beta_i\|_1 + \frac{1}{n} \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (13) \\ \text{s.t.} \quad & y_i - \alpha_i - \beta'_i \mathbf{x}_i \leq \varepsilon + \xi_i \quad \forall i, \\ & \alpha_i + \beta'_i \mathbf{x}_i - y_i \leq \varepsilon + \xi_i^* \quad \forall i, \\ & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i \quad \forall i, h, \\ & \xi_i \geq 0, \xi_i^* \geq 0 \quad \forall i. \end{aligned}$$

$$\begin{aligned} \alpha_i + \beta'_i \mathbf{x}_i - y_i &\leq \varepsilon + \xi_i^* && \forall i, \\ \alpha_i + \beta'_i \mathbf{x}_i &\leq \alpha_h + \beta'_h \mathbf{x}_i && \forall i, h, \\ \xi_i \geq 0, \xi_i^* &\geq 0 && \forall i. \end{aligned}$$

Similar to the L_2 -norm penalty, the L_1 -norm penalty also 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 (13) with a group sparsity penalty is not an effective way due to the existing of Afriat inequalities (Dai, 2023; Xu et al., 2016). That is, the small changes to β_i in each Afriat 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 completely, and thus does not necessarily make the representor function (4) sparse.

This motivates us to consider a $L_{\infty/1}$ -regularized Lasso CSVR. As shown in Zhao et al. (2009) and Negahban & Wainwright (2011), the $L_{\infty/1}$ -norm taking the maximum encourages all n components of the subgradient $\beta_{j,i}$ to be zero simultaneously or to be nonzero simultaneously. The second version of Lasso CSVR can be formulated as the following optimization problem

$$\min_{f \in \mathcal{F}} \sum_{i=1}^n \phi_\varepsilon(y_i - f(\mathbf{x}_i)) + A \|\beta\|_{L_{\infty/1}}, \quad (14)$$

where $\|\beta\|_{L_{\infty/1}}$ indicates the $L_{\infty/1}$ -norm and it can be determined by

$$\|\beta\|_{\infty/1} = \sum_{k=1}^d \|(\beta_1^k, \beta_2^k, \dots, \beta_n^k)\|_{\infty},$$

where $k = 1, \dots, d$ is the dimension of β_i . We can also convert problem (14) to the following equivalent and tractable optimization problem

$$\begin{aligned} \min_{\beta_i, \alpha} \quad & A \|\beta\|_{\infty/1} + \frac{1}{n} \sum_{i=1}^n (\xi_i + \xi_i^*) \quad (15) \\ \text{s.t.} \quad & y_i - \alpha_i - \beta'_i \mathbf{x}_i \leq \varepsilon + \xi_i \quad \forall i, \\ & \alpha_i + \beta'_i \mathbf{x}_i - y_i \leq \varepsilon + \xi_i^* \quad \forall i, \\ & \alpha_i + \beta'_i \mathbf{x}_i \leq \alpha_h + \beta'_h \mathbf{x}_i \quad \forall i, h, \\ & \xi_i \geq 0, \xi_i^* \geq 0 \quad \forall i. \end{aligned}$$

Compared to the L_1 -norm penalty, the $L_{\infty/1}$ -norm penalty restricts each subgradient $\beta_{j,i}$ with a fixed bound instead of all d components of the subgradient β_i . Furthermore, the $L_{\infty/1}$ -regularized Lasso approach is a special case of the block $L_{\infty/1}$ -regularization when the number of blocks is n (Negahban & Wainwright, 2011). In practice, the L_1 -norm Lasso CSVR (13) and $L_{\infty/1}$ -regularized Lasso CSVR (15) can be solved directly by the off-the-shelf commercial solvers (e.g., Mosek and Cplex).

3.3. Illustrative example

We proceed to demonstrate how the fitted functions estimated by CSVR and convex regression are different and illustrate the

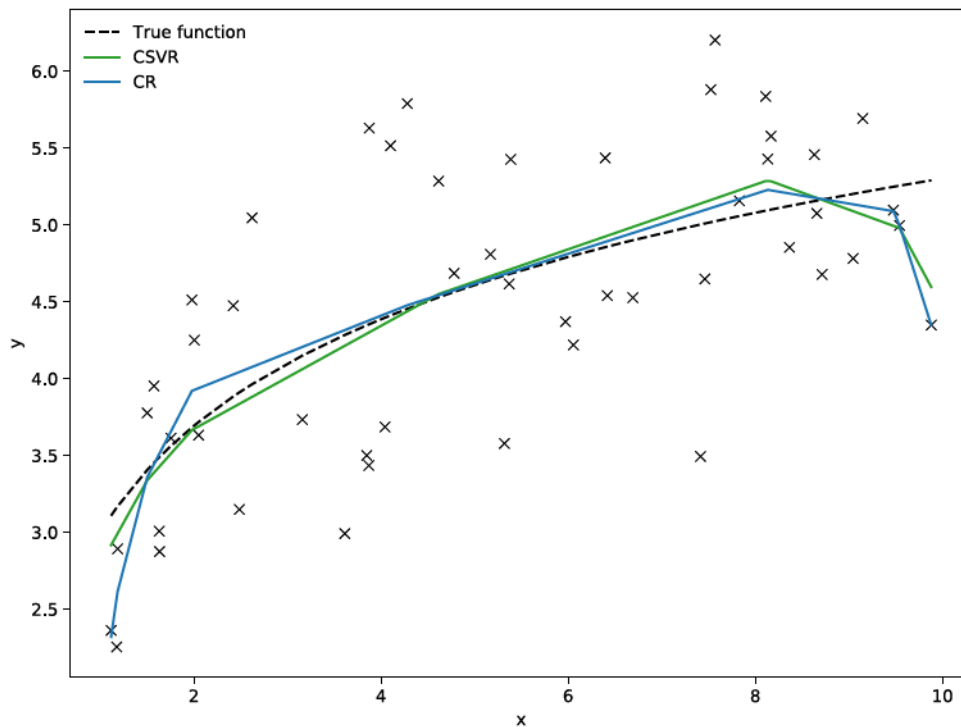


Fig. 1. Illustration of the fitted functions estimated by CSVR and convex regression.

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., Dai, 2023; Mazumder et al., 2019).

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 sharper, while the fitted CSVR functions seem relatively smooth, implying the difference in the loss function. To quantitatively illustrate the potential overfitting in CR, we split the data into a training set (80% of observations) and a testing set (20% of observations). When comparing the approximation errors, CR has a smaller in-sample MSE ($MSE = 0.121$) than CSVR ($MSE = 0.130$). By contrast, the out-of-sample MSE (i.e., the generalization error on the test data) of CR ($MSE = 0.372$) is higher than that of CSVR ($MSE = 0.206$). The results suggest that CR is hampered by overfitting, and CSVR can effectively avoid this problem.

We then illustrate the robustness of the estimated CSVR 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 CSVR functions but produce very similar estimated piecewise-linear curves.

4. Monte Carlo study

Having illustrated the estimated CSVR function, we proceed to investigate the finite sample performance of the CSVR, CR, SVR with RBF and polynomial kernels, and LCR approaches in the controlled environment of Monte Carlo simulations. The main objective of our simulations is to examine whether the proposed CSVR

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., Dai, 2023; 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: $y = 3 + x_1^{0.05} + x_2^{0.15} + x_3^{0.3} + \epsilon$,
- (4) DGP IV: $y = \prod_{d=1}^D x_d^{\frac{0.8}{D}} + \epsilon$,

where x_1, x_2, x_3 , and x_d 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 numbers 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 MSE statistic.

Regarding the tuning parameter selection, the optimal hyperparameters (i.e., C and ϵ) are determined in the SVR and CSVR 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 @2.8GHz processors, 1 CPU, and 8 GB RAM per task. The simulation code and data are available at the GitHub repository (<https://github.com/ds2010/CSVR>).

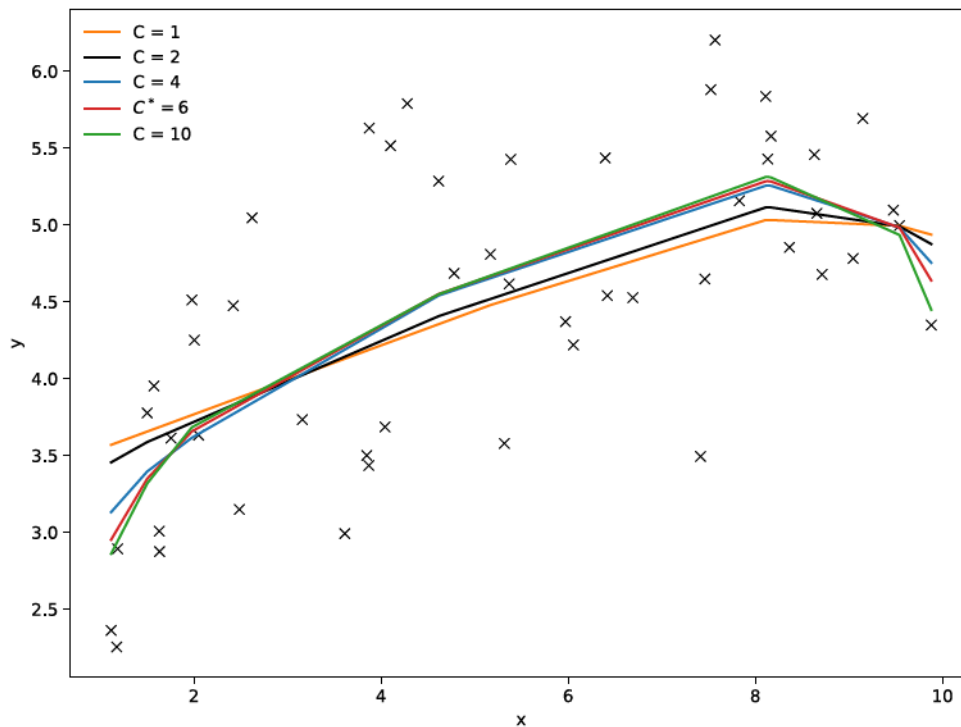


Fig. 2. Illustration of the estimated CSVR functions with different values of C.

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 deteriorates in multivariate cases.

Compared to the regularized approaches, the traditional CR approach performs poorly in 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 the given feasibility. As expected, the performance of each approach deteriorates as more input variables or smaller sample sizes are introduced.

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 CSVR approach has the smallest MSEs in most 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, CSVR maintains its good performance in different levels of data noise and dimensions.

Compared to other regularized approaches, CSVR 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 CSVR approach introduces the ϵ -insensitive loss function inherited from SVR to increase robustness. For smaller σ values, as shown in Table 2, it seems that NCCSVR may outperform the proposed CSVR method (i.e. DGP III, $\sigma = 0.5$).

Recall that Wang & Ni (2012) have proposed a similar NCCSVR approach, where they introduce convexity/concavity into SVR by using the Hessian matrix and then transform it to a semidefi-

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

DGP	n	CSVR	SVR(RBF)	SVR(poly)	CR	LCR	NCCSVR
I	50	0.0644	0.1133	0.1273	0.0783	0.0705	–
	100	0.0384	0.0719	0.1043	0.0485	0.0422	–
	200	0.0201	0.0359	0.0902	0.0271	0.0214	–
	500	0.0086	0.0152	0.0824	0.0107	0.0081	–
II	50	0.0771	0.1085	0.1493	0.2214	0.1468	0.1054
	100	0.0386	0.0638	0.0810	0.1368	0.0920	0.0505
	200	0.0271	0.0411	0.0543	0.0766	0.0525	0.0320
III	500	0.0083	0.0219	0.0334	0.0383	0.0296	0.0120
	50	0.0758	0.1058	0.1922	0.4208	0.1282	0.1374
	100	0.0556	0.0652	0.1154	0.2884	0.0991	0.0746
	200	0.0365	0.0432	0.0752	0.1864	0.0713	0.0398
	500	0.0184	0.0283	0.0406	0.0974	0.0599	0.0188

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

DGP	σ	CSVR	SVR(RBF)	SVR(poly)	CR	LCR	NCCSVR
I	0.5	0.0030	0.0043	0.0078	0.0031	0.0025	–
	1	0.0086	0.0152	0.0824	0.0107	0.0081	–
	2	0.0257	0.0522	0.1012	0.0392	0.0271	–
II	0.5	0.0038	0.0076	0.0223	0.0090	0.0066	0.0043
	1	0.0083	0.0219	0.0334	0.0383	0.0296	0.0120
	2	0.0214	0.0482	0.0771	0.1341	0.0755	0.0494
III	0.5	0.0085	0.0131	0.0203	0.0248	0.0185	0.0072
	1	0.0184	0.0283	0.0406	0.0974	0.0599	0.0188
	2	0.0493	0.0658	0.1200	0.3874	0.1862	0.0671

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

DGP	n	CSVR	SVR(RBF)	SVR(poly)	CR	LCR	NCCSVR
II	50	0.0897	0.1940	0.1389	0.2185	0.1277	0.1096
	100	0.0763	0.0977	0.1250	0.1107	0.0863	0.0595
III	50	0.0806	0.1269	0.1418	0.3652	0.0891	0.1556
	100	0.0693	0.1103	0.1120	0.2377	0.0815	0.1065

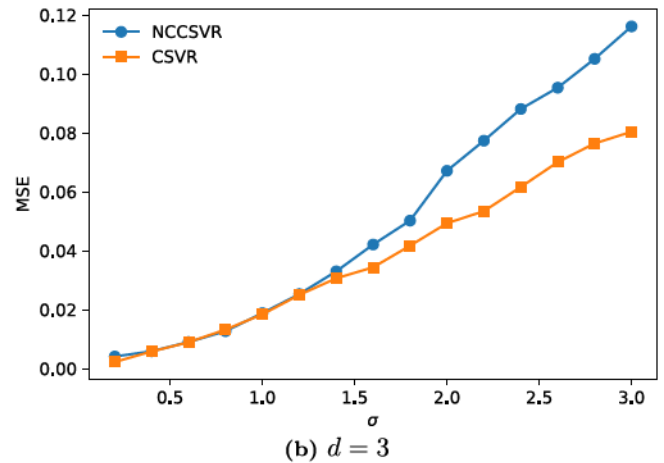
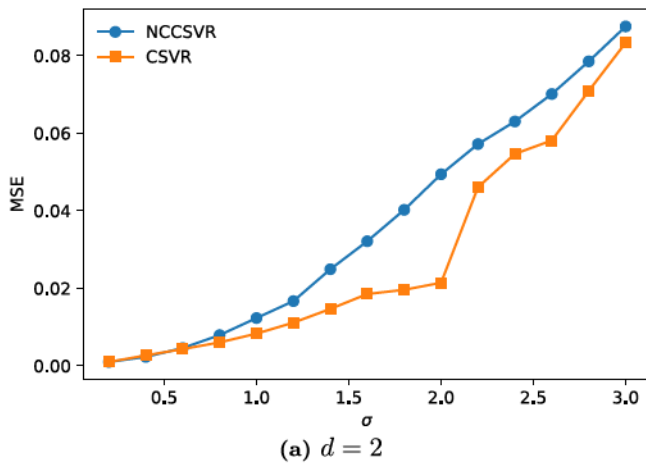


Fig. 3. Illustration of the impacts of noise variation on MSE.

nite programming problem. In contrast, our CSVR approach resorts to the system of Afriat inequalities to ensure the fitted function to be convex/concave. Moreover, the NCCSVR approach can only be applied to multivariate cases (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.

We next 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.

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 Lips-

chitz 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.

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.

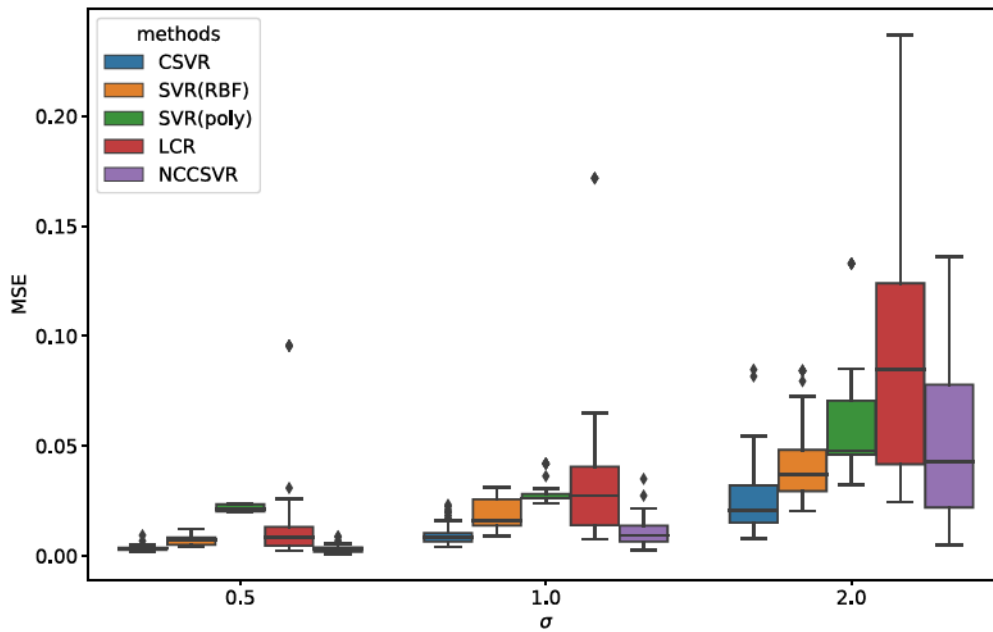


Fig. 4. Out-of-sample MSE of five methods with $d = 2$ and $\sigma \in \{0.5, 1, 2\}$.

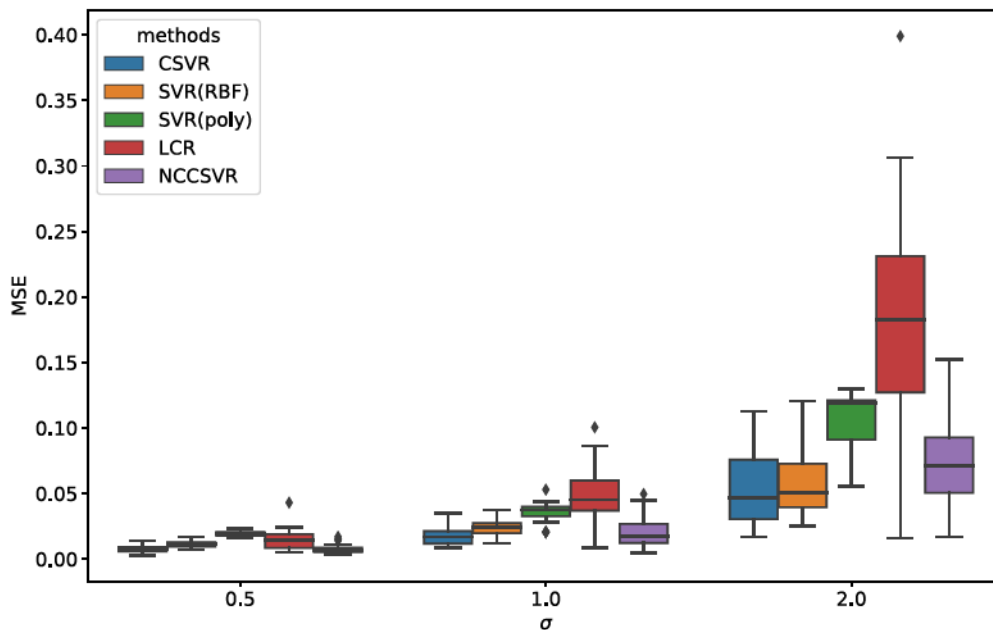


Fig. 5. Out-of-sample MSE of five methods with $d = 3$ and $\sigma \in \{0.5, 1, 2\}$.

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.

CSVR performs better than other methods in alleviating the overfitting problem. Compared to the LCR approach, the four 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-of-sample performance of LCR deteriorates rapidly as the data noise increases. Additional experiments on the mean absolute error (MAE) statistic also can support those conclusions drawn from the MSE comparisons (see Section A in Appendix).

4.4. High-dimensional performance

To investigate the performance of Lasso CSVR methods in high-dimensional variable selection, we consider DGP IV with $d \in \{5, 10, 20, 30\}$, $\sigma \in \{0.1, 0.5, 1, 1.5, 2\}$, and $n \in \{50, 100, 200\}$ and focus on the out-of-sample performance and the number of se-

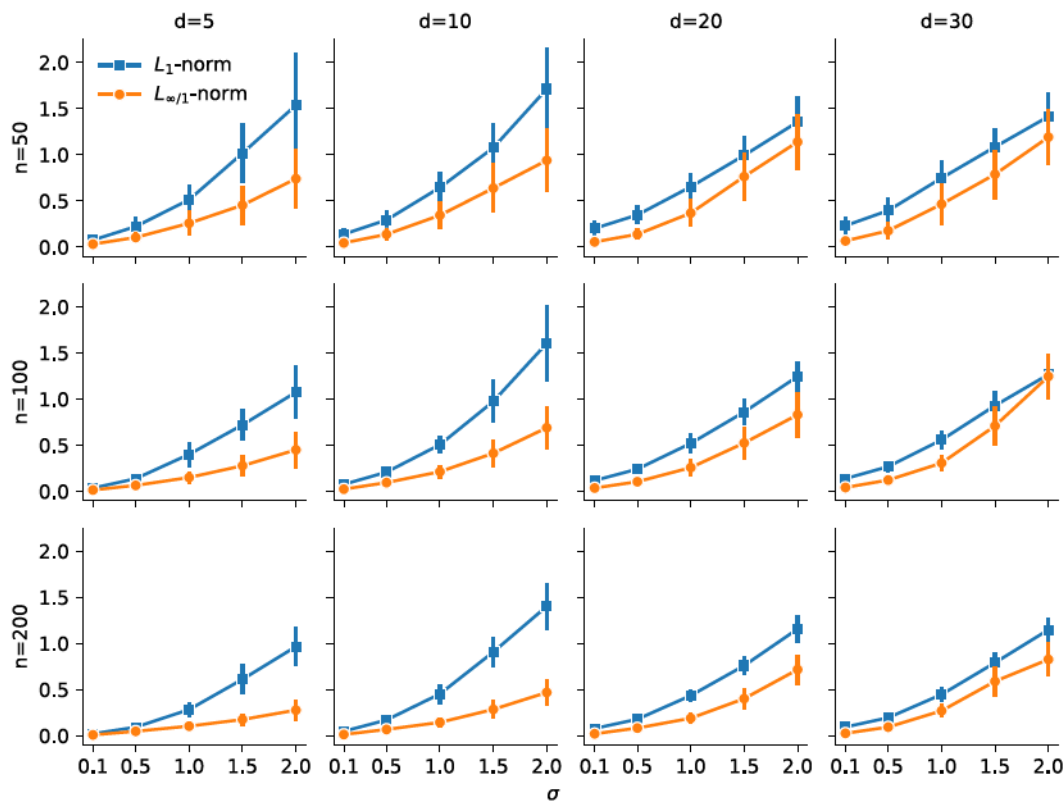


Fig. 6. Out-of-sample MSE of the $L_{\infty/1}$ and L_1 -regularized Lasso CSVR.

lected variables. Fig. 6 shows the out-of-sample MSEs of two Lasso CSVR approaches in different scenarios. We observe that the $L_{\infty/1}$ -regularized Lasso CSVR exhibits superior out-of-sample performance over the L_1 -regularized Lasso CSVR. While the two methods have a close MSE value under lower noise levels (e.g., $\sigma = 0.1$), the Lasso CSVR regularized by $L_{\infty/1}$ -norm is better than L_1 -norm as σ gets larger. Furthermore, the large sample size can decrease the values of MSE in both methods, which confirms the same conclusion from, e.g., Chen et al. (2020) and Dai (2023).

Fig. 7 demonstrates the number of selected variables with varying data noise, sample sizes, and dimensions. Overall, the $L_{\infty/1}$ -regularized Lasso CSVR performs better in variable selection than the L_1 -regularized Lasso CSVR, even with high statistical noise in the input data. Considering that variable selection in convex regression is a notoriously difficult problem (see, e.g., Bertsimas & Mundru, 2021; Xu et al., 2016), it is no surprise to see that both approaches cannot reduce the dimensionality in some scenarios (e.g., $d = 5$ and $n = 200$). Furthermore, while the $L_{\infty/1}$ -regularized Lasso CSVR can make all elements in β_i become zero simultaneously, it might still be difficult to reduce dimension when the sample size becomes large (e.g., $n = 200$).

5. Experiments with real data

In this section, we apply the proposed approach to real-world datasets: two housing datasets (Boston and California) and two manufacturing industry datasets (NBER-CES and METI). 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., Arreola et al., 2020; Lin et al., 2011; Wang & Wang, 2013). In the examples below, all hyperparameters are tuned over from 50 candidate values via the fivefold cross-validation technique, and

the in-sample and out-of-sample MSE and MAE for each approach are calculated to compare the effectiveness in reducing overfitting.

5.1. Housing data

The Boston housing dataset contains housing price information in the Boston area collected from the StatLib archive.² The data includes 13 variables with 506 observations. Following the commonly used setting, the variable *MEDV* is taken as the response variable of interest, and others are the explanatory variables. Note that the dummy variable (i.e., *CHAS*) is excluded from the dataset for the sake of simplicity.³ The California housing dataset with the Bay Area consists of 9 variables and 2290 observations (Pace & Barry, 1997).⁴ Similarly, *MEDV* is denoted as the response variable, and the other six variables—*MedInc*, *HouseAge*, *AveRooms*, *AveBedrms*, *Population*, *AveOccup*—are the explanatory variables. Note that the longitude and latitude variables are eliminated from the models. The definitions and corresponding descriptive statistics for all variables are presented in Tables B.1 and B.2 in Appendix.

Table 4 reports the mean and the standard deviation of the estimated coefficients (i.e., $\hat{\alpha}_i$ and $\hat{\beta}_i$) of CSVR, LCR, and CR. Unlike in linear regression, these coefficients are not constants but differ across observations.⁵ We omit SVR and NCCSVR here due to the

² StatLib–Datasets Archive: <http://lib.stat.cmu.edu/datasets/boston>.

³ While CR can handle contextual variables, this falls beyond the scope of this paper; cf. Johnson & Kuosmanen (2012), for further details.

⁴ Luís Torgo: https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html.

⁵ Statistical testing and confidence intervals can be implemented separately for each hyperplane following Deng et al. (2022). It is straightforward to the present results of such statistical inferences in the univariate setting, but in the present case, we have 12 regressors and a large number of hyperplanes. The question of how to present the results of such statistical inferences in practice is left as an interesting operational challenge for future research.

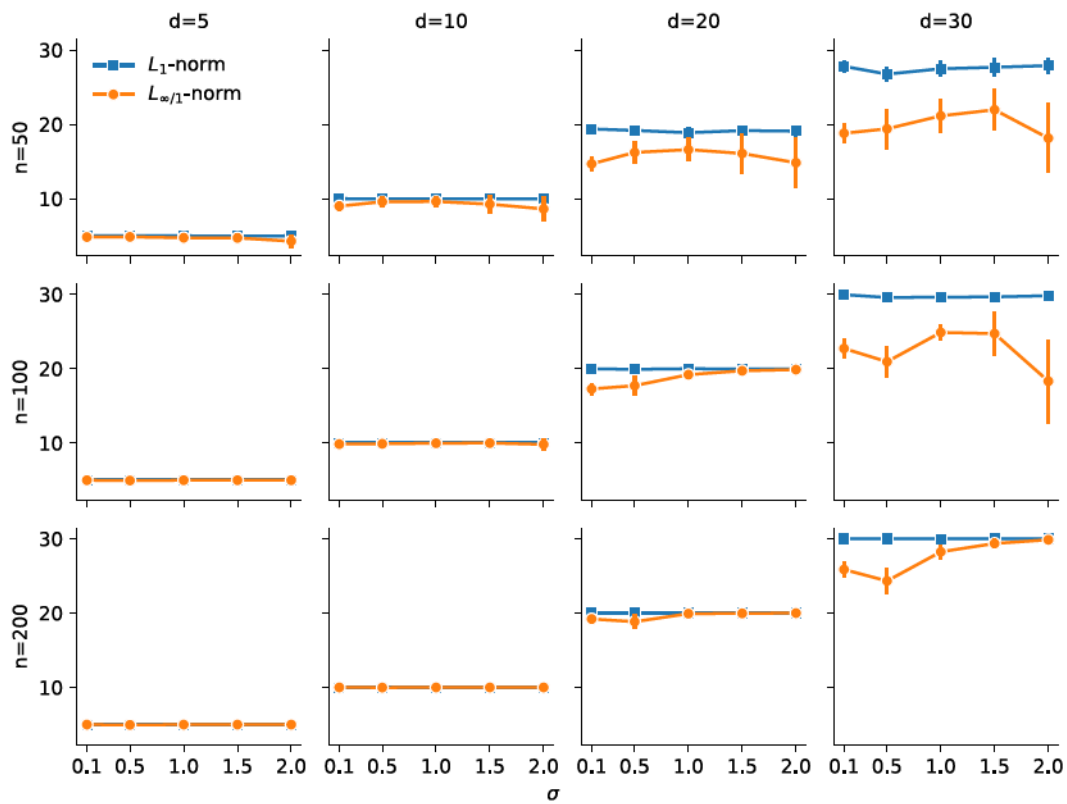


Fig. 7. Number of selected variables of the $L_{\infty/1}$ and L_1 regularized Lasso CSVR.

Table 4
Descriptive statistics for estimates of all explanatory variables.

	CSVR		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
$\hat{\beta}_{NGX}$	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

incomparable dual variable (i.e., $\hat{\beta}_i$) (see Smola & Schölkopf, 2004). As shown in Table 4, all values of $\hat{\beta}$ 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 $\hat{\beta}$, LCR obtains coefficients even a bit smaller than CSVR. Furthermore, LCR produces a flatter function f , but that does not mean a better performance automatically.

To evaluate the prediction accuracy, we perform fivefold cross-validation on the housing data and then compute the in-sample and out-of-sample prediction error in terms of MSE and MAE. The calculated average in-sample and out-of-sample MSEs and MAEs are demonstrated in Tables 5 and 6. Overall, as expected, we observe that the CSVR method achieves the best predic-

tion performance in both housing datasets. 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 MSE accuracy. However, note that CR yields the lowest in-sample MSE and MAE in both datasets.

5.2. Manufacturing industry data

The NBER-CES manufacturing industry dataset consists of 473 manufacturing industries (the 1997 6-digit NAICS codes) and is collected from the NBER-CES Manufacturing Industry Database.⁶ 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).

Japan's Ministry of Economy, Trade and Industry (METI) provides a series of open-access manufacturing industry data for the years 2010–2014.⁷ Such a Census of Manufacture database has been applied in Arreola et al. (2020) and Tanaka & Managi (2021). Following Arreola et al. (2020), we set Labor and Capital as input variables and Value Added as the output variable. The descriptive

⁶ NBER database: <https://www.nber.org/research/data/nber-ces-manufacturing-industry-database>.

⁷ Census of Manufacture (METI): <https://www.meti.go.jp/english/statistics/tyo/kougyo/index.html>.

Table 5
Performance comparisons: Boston housing data.

	CSVR	SVR(RBF)	SVR(poly)	CR	LCR	NCCSVR
MSE(out)	38.44(5.04)	64.00(14.28)	62.04(14.28)	2334.17(933.04)	43.72(6.47)	42.72(9.28)
MSE(in)	21.43(0.86)	62.57(2.90)	62.57(2.90)	1.41(0.33)	35.05(1.52)	40.17(2.05)
MAE(out)	3.94(0.43)	5.14(0.47)	4.89(0.54)	23.35(3.98)	4.24(0.41)	4.39(0.25)
MAE(in)	2.45(0.08)	5.00(0.12)	4.77(0.13)	0.33(0.05)	3.16(0.10)	3.98(0.11)

Note: standard deviation in parentheses ; MSE(out): out-of-sample MSE. MSE(in): in-sample MSE.

Table 6
Performance comparisons: California housing data.

	CSVR	SVR(RBF)	SVR(poly)	CR	LCR	NCCSVR
MSE(out)	0.81(0.20)	1.52(0.15)	1.52(0.16)	10.72(10.58)	1.22(0.14)	1.07(0.32)
MSE(in)	0.61(0.02)	1.50(0.04)	1.51(0.03)	0.45(0.02)	1.15(0.03)	0.90(0.07)
MAE(out)	0.70(0.06)	0.99(0.06)	0.99(0.06)	1.22(0.37)	0.91(0.05)	0.77(0.05)
MAE(in)	0.63(0.01)	0.98(0.02)	0.99(0.01)	0.49(0.01)	0.89(0.01)	0.68(0.01)

Note: standard deviation in parentheses.

Table 7
Performance comparisons: NBER-CES manufacturing industry data.

	CSVR	SVR(RBF)	SVR(poly)	CR	LCR	NCCSVR	SFA
MSE(out)	3.16	7.34	22.86	4.51	4.11	3.45	3.17
MSE(in)	4.88	18.9	17.34	4.03	4.08	6.72	5.09
MAE(out)	0.71	1.02	2.18	0.88	0.85	0.73	0.78
MAE(in)	0.70	1.28	2.10	0.78	0.79	0.82	0.85

Table 8
Performance comparisons: METI manufacturing industry data.

	CSVR	SVR(RBF)	SVR(poly)	CR	LCR	NCCSVR	SFA
MSE(out)×10 ³	13.90	186.94	63.00	25.38	31.15	27.17	19.99
MSE(in)×10 ³	25.40	241.21	114.67	17.86	60.16	59.15	30.12
MAE(out)	44.14	89.85	123.23	60.56	75.02	58.76	54.38
MAE(in)	47.54	101.17	133.51	58.29	86.05	67.32	59.22

statistics for both manufacturing industry data are reported in Tables B.3 and B.4 in Appendix.

For each case, we randomly split the dataset into new training and test sets. All the experiment results are the averaged values over 30 random splits. In estimating the production function, we also include stochastic frontier analysis (SFA) as one of the benchmark methods. For the MSE comparison, the in-sample MSE value of CSVR is not the lowest in comparison with other shape-constrained approaches (e.g., CR), but the out-of-sample MSE remains the lowest (see Table 7). This is perhaps due to the fact that there is a good trade-off between model performance and overfitting alleviation. SFA's out-of-sample MSE is very close to that of the CSVR model. This is because the limited complexity of the SFA model can also reduce the risk of overfitting. For the MAE comparison, CSVR performs best among all the benchmark methods in both in-sample and out-of-sample sets. Table 8 also demonstrates the superiority of CSVR in terms of the lowest out-of-sample MSE and MAE values with the larger METI manufacturing industry dataset.

To statistically compare the performance difference between CSVR and other methods demonstrated in Tables 7 and 8, we perform a resampled paired *t*-test by following the procedure in Dieterich (1998), where the null hypothesis states that model A and model B have equal performance. Here we focus on the comparison of predictive performance. Table 9 reports the *t*-statistic obtained by the resampled paired *t*-test and its corresponding significance level. While CSVR is better than SFA in the NBER-CES dataset (see Table 7), the difference is not significant in terms of out-of-sample MSE. However, for all other cases, CSVR has the low-

Table 9
Results of resampled paired *t*-test, showing the significance of differences in out-of-sample performance metrics of CSVR vs. other methods.

CSVR vs.	NBER-CES		METI	
	MSE(out)	MAE(out)	MSE(out)	MAE(out)
SVR(RBF)	-6.87**	-13.04**	-6.83**	-15.60**
SVR(poly)	-5.75**	-25.62**	-13.77**	-65.77**
CR	-4.90**	-8.33**	-2.78**	-18.92**
LCR	-6.12**	-9.91**	-9.13**	-37.30**
NCCSVR	-2.51**	-2.10*	-8.25**	-21.28**
SFA	-0.02	-8.04**	-8.58**	-18.42**

Note: * *p* < 0.05, ** *p* < 0.01.

est out-of-sample MSE and MAE, and its performance metrics significantly differ from those of other methods.

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 with respect to *y*, but due to overfitting not with respect to *f*(*x*).

In conclusion, both simulations and real-world applications demonstrate that the regularized shape-constrained methods have a superior ability to control overfitting, but CSVR would be more appealing than other regularized shape-constrained methods be-

cause of its simplicity, capacity for univariate regression, and robust performance.

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 CSVR approach in contrast to other state-of-the-art regression methods through Monte Carlo simulations. Additional four 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 CSVR performs at least as well as LCR and much better than SVR and traditional convex regression. Four 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 a simple and fast algorithmic framework of convex support vector regression. The main bottleneck of CSVR is that the full problem (9) 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 CSVR, would be needed.

Acknowledgments

The authors would like to thank the three anonymous reviewers for their helpful comments. 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] and the Jenny and Antti Wihuri Foundation [grant no. 00220201]. Sheng Dai gratefully acknowledges financial support from the Foundation for Economic Education (Liikesivistysrahasto) [grant no. 220074] and the OP Group Research Foundation [grant no. 20230008].

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2023.05.009.

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.
- Arreola, J. L. P., Johnson, A. L., Chen, X., & Morita, H. (2020). Estimating stochastic production frontiers: A one-stage multivariate semiparametric Bayesian concave regression method. *European Journal of Operational Research*, 287, 699–711.
- Aybat, N. S., & Wang, Z. (2014). A parallel method for large scale convex regression problems. In *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. In *18th artificial intelligence and statistics* (pp. 38:56–64). PMLR.
- Bertsimas, D., & Mundru, N. (2021). Sparse convex regression. *INFORMS Journal on Computing*, 33, 262–279.
- Bradley, P. S., & Mangasarian, O. L. (1998). Feature selection via concave minimization and support vector machines. In *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. (2023). Variable selection in convex quantile regression: L_1 -norm or L_0 -norm regularization? *European Journal of Operational Research*, 305, 338–355.
- Dai, S., Fang, Y. H., Lee, C. Y., & Kuosmanen, T. (2021). pyStoNED: A Python package for convex regression and frontier estimation. arXiv:2109.12962.
- Dai, S., Kuosmanen, T., & Zhou, X. (2022). Non-crossing convex quantile regression. arXiv:2204.01371.
- Deng, H., Han, Q., & Sen, B. (2022). Inference for local parameters in convexity constrained models. *Journal of the American Statistical Association*. <https://doi.org/10.1080/01621459.2022.2071721>.
- Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10, 1895–1923.
- Du, P., Parmeter, C. F., & Racine, J. S. (2013). Nonparametric kernel regression with multiple predictors and multiple shape constraints. *Statistica Sinica*, 23, 1347–1371.
- Ghosal, P., & Sen, B. (2017). On univariate convex regression. *Sankhya A*, 79, 215–253.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.
- 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. (2012). Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the stoned method in the finnish regulatory model. *Energy Economics*, 34, 2189–2199.
- Kuosmanen, T., & Johnson, A. L. (2010). Data envelopment analysis as nonparametric least-squares regression. *Operations Research*, 58, 149–160.
- Kuosmanen, T., & Johnson, A. L. (2020). Conditional yardstick competition in energy regulation. *The Energy Journal*, 41, 67–92.
- Kuosmanen, T., Johnson, A. L., & Saastamoinen, A. (2015). Stochastic nonparametric approach to efficiency analysis: A unified framework. In J. Zhu (Ed.), *Data envelopment analysis: chapter 7* (pp. 191–244). Boston, MA: Springer.
- 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.
- Pace, R. K., & Barry, R. (1997). Sparse spatial autoregressions. *Statistics and Probability Letters*, 33, 291–297.
- 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.
- Tanaka, K., & Managi, S. (2021). Industrial agglomeration effect for energy efficiency in Japanese production plants. *Energy Policy*, 156, 112442.
- 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 and Operations Research*, 143, 105765.
- Vapnik, V. (1999). *The nature of statistical learning theory*. New York: Springer.

- 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 and 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.