Information Sharing for Robust and Stable Cross-Validation

David Kepplinger*
and
Siqi Wei
Department of Statistics, George Mason University, Fairfax, VA, USA

Abstract

Robust estimators for linear regression require non-convex objective functions to shield against adverse effects of contamination, including outliers. This non-convexity brings challenges, particularly when combined with penalization in high-dimensional settings. A crucial challenge is selecting hyperparameters for the penalty based on a finite sample. In practice, cross-validation (CV) is the prevalent strategy with good performance for convex estimators. Applied with robust estimators, however, CV often gives subpar results due to the interplay between multiple local minima and the penalty. The best local minimum attained on the full training data may not be the minimum with the desired statistical properties. Furthermore, there may be a mismatch between this minimum and the minima attained in the CV folds which are used for evaluating the prediction error. This paper introduces a novel adaptive CV strategy that tracks multiple minima for each combination of hyperparameters and subsets of the data. A matching scheme is presented for correctly evaluating minima computed on the full training data using the best-matching minima from the CV folds. We show that the proposed strategy reduces the variability of the estimated performance metric, leads to smoother CV curves, and therefore substantially increases the reliability and utility of robust penalized estimators.

Keywords: Robust regression, hyperparameter selection, cross-validation, non-convexity, penalized estimation.

*This project was supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Award Number 2018631).

1 Introduction

In this paper, we revisit a critical issue for applying robust penalized estimators: how to reliably select hyperparameters of the penalty function. In practice, cross-validation (CV) is by far the most prevalent strategy used for this hyperparameters selection. Besides adjustments of the CV sampling scheme (e.g., stratified CV), performing multiple replications of CV or using different evaluation metrics, the general procedure is almost always the same. Although computations can be burdensome, CV has become a ubiquitous tool in any statistical learning framework. Recent advances in asymptotic results for K-fold CV (e.g., Austern and Zhou 2020; Bates et al. 2024; Li 2023) further underline the advantages of CV that were previously noticed primarily empirically. While these theoretical guarantees for CV do not apply to robust estimators, some empirical studies suggest that good out-of-sample accuracy and variable selection can be achieved using CV with robust measures of prediction accuracy (Amato et al. 2021; Cohen Freue et al. 2019; Filzmoser and Nordhausen 2021; Khan et al. 2007; Loh 2021; Maronna 2011; Monti and Filzmoser 2021; Ronchetti et al. 1997; Sun et al. 2019). However, reproducing these benefits in practical applications has proven difficult because CV for robust penalized estimators tends to be highly unstable (Kepplinger and Cohen Freue 2023; She et al. 2021), particularly in the presence of contamination in the response and/or the predictors, such as outliers. As mentioned in Datta and Zou (2019), even in the simpler measurement error model, leave-one-out CV with variants of the (convex) LASSO estimator fails. When dealing with arbitrary contamination, and when using non-convex estimators, these issues tend to be more severe. A more reliable CV method is therefore desperately needed for these cases.

Practically, the issues with CV manifest in very different solutions for different random CV splits of the data. These differences are often substantial and affect both the estimate of the prediction accuracy and hyperparameter selection, leading to questionable results. In

the following, we illuminate the issues underlying the instability of CV for robust penalized estimators. We propose a novel CV strategy called Robust Information Sharing (RIS) CV, which provides a more reliable and stable estimation of prediction accuracy and thus hyperparameter selection in the presence of contamination.

1.1 Background

We focus on robust penalized estimators for the linear regression model with response vector $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$, p-dimensional predictors $\mathbf{x}_i \in \mathbb{R}^p$, $i \in \mathcal{T} = \{1, \ldots, n\}$, true coefficients $\boldsymbol{\beta}_0$, and i.i.d, errors ε_i from an arbitrary symmetric distribution. We focus on studies where the primary goal is to predict out-of-sample responses for a new observation $\mathbf{x}^*$ and the number of predictors, p, is potentially greater than n. Importantly, up to but less than half of the observations may deviate from the model.

In this setting, we consider hyperparameter selection for robust estimators defined as the minimizer of an objective function that can be decomposed into a loss term, ℓ, a penalty term, P, and the penalization level $\lambda \geq 0$, i.e.,

$$\begin{aligned} \mathcal{O}(\boldsymbol{\beta}; \mathscr{T}, \lambda) &:= \ell(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}), \\ \text{where} \quad \mathbf{y} &= (y_i)_{i \in \mathscr{T}}^\intercal, \mathbf{X} = (x_{ij})_{i \in \mathscr{T}, j = 1, \ldots, p}. \end{aligned} \tag{1}$$

The loss depends on the parameters only through the residuals, and we assume that it can be recast as a weighted least-squares loss: $\ell(\mathbf{r}) = \sum_{i \in \mathscr{T}} w_i(\mathbf{r}) r_i^2$, where the weights depend on the residuals $\mathbf{r} = (r_i)_{i \in \mathscr{T}}$. The penalty may involve additional hyperparameters, like the elastic net's mixing parameter, but we focus on the penalization level.

The majority of popular robust regression methods for high-dimensional data fall into this framework, for example SparseLTS (Alfons et al. 2013), MM-LASSO (Smucler and Yohai 2017), MM-Bridge (Arslan 2016), Tukey-Lasso (Chang et al. 2018), PENSE/PENSEM

(Cohen Freue et al. 2019), the penalized τ estimator (Mozafari-Majd and Koivunen 2025) and the adaptive PENSE/PENSEM (Kepplinger and Cohen Freue 2023). Some of these estimators, such as MM-LASSO or adaptive PENSE, are two-step estimators where both stages seek a minimizer of (1), but employ different loss functions and/or different penalties.

To apply these estimators successfully in practice, the appropriate value for the hyper-parameter λ, which governs the strength of the penalty, must be chosen in a data-driven fashion. The solution path, that is, the minimizers of (1) for a decreasing sequence of penalization levels, is usually expected to be smooth. In other words, a slight relaxation of the penalization level is expected to lead to only a small change in the minimizer. Unfortunately, this expectation often does not align with reality when employing robust estimators, as detailed in Section 2. This is the main reason why common strategies to select λ often fail for robust estimators and why our focus in this work is on selecting λ. In cases where the penalty term depends on additional hyperparameters, an appropriate choice for those most often hinges on a reliable and stable selection of λ.

Stability and reliability of the regression estimator in (1) amid contamination is achieved through a robust loss function, ℓ. A common choice is the S-loss function given by $\ell_S(\mathbf{r}) = \frac{1}{2}\sigma_M^2(\mathbf{r})$, with $\sigma_M^2(\mathbf{r})$ defined implicitly by

$$\delta = \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{r_i}{\sigma_M^2(\mathbf{r})}\right). \tag{2}$$

Here, $\delta \in (0, 0.5)$ is a tuning constant that determines the finite sample breakdown point of the S-loss (Cohen Freue et al. 2019; Kepplinger 2023), i.e., the proportion of observations that can be arbitrarily contaminated without leading to an infinitely biased estimate. The S-loss is used by PENSE and S-Ridge (Maronna 2011) which also forms the first stage of MM-LASSO, adaptive PENSE and adaptive PENSEM. The M-loss is another popular robust loss function, $\ell_M(\mathbf{r}) = \frac{1}{2n}\sum_{i=1}^{n} \rho(r_i/s)$, and requires a predetermined scale, $s > 0$.

It is utilized in the second stages of MM-LASSO, PENSEM and adaptive PENSEM where the first stage is used to estimate the residual scale s and the penalty loadings for the adaptive EN penalty. For the M- and S-loss functions, a bounded and therefore non-convex ρ function is necessary to achieve high robustness towards arbitrarily contaminated data points. Common ρ functions behave like the square function around 0 and smoothly transition to a constant beyond a certain cutoff value. Typical examples for ρ are Tukey's bisquare or the linear-quadratic-quadratic (LQQ) function (Koller and Stahel 2011).

Another choice for ℓ is the τ-loss, an attempt to improve the efficiency of S-estimators. It is based on two bounded functions, ρ and ρ_1, and is defined as $\ell_\tau(\mathbf{r}) = \frac{\sigma_M^2(\mathbf{r})}{n\delta_1} \sum_{i=1}^n \rho_1(r_i/\sigma_M(\mathbf{r}))$. The SparseLTS estimator, on the other hand, is a trimmed least squares (LTS) estimator that considers only the h smallest residuals, $n/2 < h < n$, with $\ell_{\text{LTS}}(\mathbf{r}) = \sum_{i=1}^h |r|_{(i)}^2$, where $|r|_{(i)}$ are the ordered absolute residuals.

A common feature of all these robust loss functions is their non-convexity, which is necessary to achieve high robustness against arbitrary contamination. This non-convexity leads to computational challenges which have been widely discussed in the literature. However, in combination with the penalization, the non-convexity also poses a substantial issue for hyperparameter selection. To our knowledge, this has not yet been formally acknowledged or discussed. Yet, as we will show in Section 2, this issue often restricts the practical utility of robust penalized regression estimators and hinders their adaptation.

A substantial body of literature on penalized regression estimators is concerned with fit-based criteria to select λ, such as AIC and BIC. These criteria have also been adapted for robust estimation, e.g., the robust BIC criterion (Alfons et al. 2013) or the predictive information criterion (She et al. 2021). In practice, however, CV is the dominating strategy for multiple reasons: it allows comparisons between different estimators and does not rely on a consistent and robust estimate of the scale, which itself is a very challenging and unsolved problem in high dimensions (Dicker 2014; Fan et al. 2012; Loh 2021; Maronna

and Yohai 2010; Reid et al. 2016).

The primary objective of this paper is to select the overall penalization level, λ, for robust penalized estimators (1). The main contribution is twofold. First, in Section 2, we illuminate and investigate the major drivers causing issues in applying CV to robust penalized estimators, i.e., the combination of a non-convex objective function with local minima driven by contaminated observations and the penalty. Second, we propose a new, robust, and reliable CV strategy, called Robust Information Sharing CV (RIS-CV), to mitigate these issues in practical applications. In Sections 4 and 5 we demonstrate that RIS-CV can be applied with different robust penalized estimators and leads to more reliable and stable hyperparameter selection than standard CV in a variety of settings.

1.2 Notation

The index set of the complete training data for the n observations is denoted as $\mathscr{T} = \{1, \ldots, n\}$. The subsets of the training data used to estimate the parameters in the K CV folds are denoted by $\mathscr{T}_1, \ldots, \mathscr{T}_K$, $\mathscr{T}_k \subset \mathscr{T}$. The set of minima for a given index set $\mathscr{T}$ and penalty parameter λ is denoted by $\mathcal{B}_\lambda^{\mathscr{T}} = \left\{\hat{\boldsymbol{\beta}} \colon \nabla_{\boldsymbol{\beta}} \mathcal{O}(\hat{\boldsymbol{\beta}}; \mathscr{T}, \lambda) = \mathbf{0}\right\}$. We assume that only the $Z \geq 1$ best minima are retained and that the minima are ordered by the value of the objective function, i.e., in a set of minima $\mathcal{B}_\lambda^{\mathscr{T}} = \{\hat{\boldsymbol{\beta}}_1, \ldots, \hat{\boldsymbol{\beta}}_Z\}$, $\mathcal{O}(\hat{\boldsymbol{\beta}}_1; \mathscr{T}, \lambda) \leq \mathcal{O}(\hat{\boldsymbol{\beta}}_2; \mathscr{T}, \lambda) \leq \cdots \leq \mathcal{O}(\hat{\boldsymbol{\beta}}_Z; \mathscr{T}, \lambda)$. If the index set on which the objective function is evaluated is obvious from the context, $\mathscr{T}$ will be omitted from the notation.

None of the non-convex optimization routines employed in this paper can guarantee to find the actual global minimum. Any notion of a "global" minimum is therefore to be understood as the local minimum with the smallest objective function value among all local minima uncovered by the non-convex optimization routine.

2 Cross-validation for Robust Penalized Estimators

Before we shed light on why standard or "naïve" CV (N-CV) fails for robust penalized estimators, we give a brief review of N-CV commonly used for robust and non-robust penalized regression estimators.

For N-CV, we first compute the global minimizers of (1) on a grid of λ values, $\mathcal{L} = \{\lambda_1, \ldots, \lambda_q\}$, $\lambda_1 > \lambda_2 > \cdots > \lambda_q > 0$, using the complete training data $\{(y_i, \mathbf{x}_i) \colon i \in \mathscr{T}\}$. The number of values, q, typically ranges from 50–100, with larger q giving a "finer resolution" and hence better chances for finding a good solution, but at higher computational costs. To estimate the prediction accuracy of these minimizers, N-CV randomly splits the training indices $\mathscr{T}$ into K approximately equally sized subsets, or folds, $\mathscr{F}_1, \ldots, \mathscr{F}_K$ with $\mathscr{F}_k \subset \mathscr{T}$ such that $\bigcup_k \mathscr{F}_k = \mathscr{T}$ and $\mathscr{F}_k \cap \mathscr{F}_{k'} = \emptyset$ for all $k \neq k'$. In each of these CV folds, the global minimizers of (1) are computed using only the observations in $\mathscr{T}_k = \mathscr{T} \setminus \mathscr{F}_k$ over the same penalty grid $\mathcal{L}$ as for the complete training data. Denoting these global minimizers by $\hat{\boldsymbol{\beta}}_{k,\lambda}$, we compute the prediction errors on the left-out observations as $e_{i,\lambda} = y_i - \mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}}_{k,\lambda}, i \in \mathscr{F}_k$. The prediction accuracy of the estimate $\hat{\boldsymbol{\beta}}_\lambda$ is then estimated for each $\lambda \in \mathcal{L}$ by summarizing these prediction errors, usually using a measure of their scale. We will denote this measure of prediction accuracy as $E(\lambda)$. The prevalent choice for $E(\lambda)$ is the root mean squared prediction error (RMSPE),

$$\widehat{\mathrm{RMSPE}}(\lambda) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} e_{i,\lambda}^2}.$$

In the potential presence of contamination, however, it is commonly argued (Cohen Freue et al. 2019; Kepplinger 2023; She et al. 2021; Smucler and Yohai 2017) that robust estimates of the prediction accuracy should be used since contamination in the response is not expected to be well predicted by the model. Robust choices are the mean absolute prediction error

(MAPE) or the τ-size (Yohai and Zamar 1988) of the prediction errors:

$$\hat{\tau}(\lambda) = \left(\underset{i=1,\ldots,n}{\text{Med}} \, |e_{i,\lambda}| \right) \sqrt{\frac{1}{n} \sum_{i=1}^{n} \min \left(c_\tau, \frac{|e_{i,\lambda}|}{\text{Med}_{i=1,\ldots,n} \, |e_{i,\lambda}|} \right)^2}, \quad c_\tau > 0.$$

To reduce the Monte Carlo error incurred by a single CV split and to obtain an estimate of the variance of the error measure, CV can be repeated with different random splits. With R replications of K-fold N-CV, the estimated prediction accuracy using metric E is

$$\hat{E}(\lambda) = \frac{1}{R} \sum_{r=1}^{R} \hat{E}^{(r)}(\lambda), \qquad \text{SD}\left(\hat{E}(\lambda)\right) = \sqrt{\frac{1}{R} \sum_{r=1}^{R} \left(\hat{E}^{(r)}(\lambda) - \hat{E}(\lambda) \right)^2}.$$

The penalty level λ is then either chosen as the one that leads to the smallest measure of the prediction error, $\hat{\lambda} = \min_{\lambda \in \mathcal{L}} \hat{E}(\lambda)$, or, to mitigate overfitting, as the one whose prediction error is within one standard deviation of the minimum, i.e., the "1-SE rule:"

$$\hat{\lambda}_{\text{1-SE}} = \max \left\{ \lambda \geq \hat{\lambda} \colon \hat{E}(\lambda) \leq \hat{E}(\hat{\lambda}) + \text{SD}\left(\hat{E}(\hat{\lambda})\right) \right\}. \tag{3}$$

In practice, great utility lies in a plot of the prediction accuracy against the penalization strength. This "CV curve" plots the estimated prediction accuracy and the estimated standard error against the penalization strength or the L_1 norm of the estimates at different $\lambda \in \mathcal{L}$. The practitioner can then choose the λ leading to the best prediction accuracy or may select a value that better balances model complexity with prediction accuracy.

Figure 1 shows the CV curve for a classical EN estimator (least squares; left) and the robust adaptive PENSE estimator (right) in the biomarker discovery study from Section 4. It is obvious that the CV curve provides valuable insight into the effects of penalization and the overall prediction accuracy of the models on the penalization path, with two striking observations. First, the classical adaptive EN estimator does not seem to perform

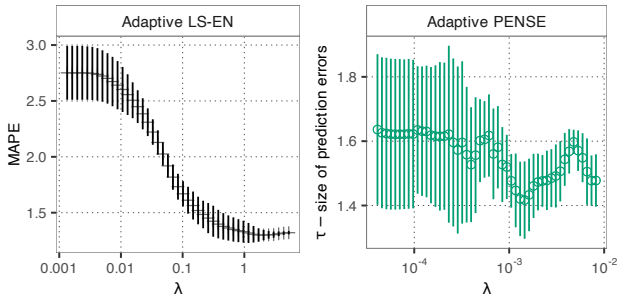

Figure 1: Estimated prediction performance of the non-robust adaptive EN estimator (left) and the robust adaptive PENSE (right) in the CAV study from Section 4 as estimated by N-CV.

particularly well in this example, with the intercept-only model (at the far right with the highest considered penalization strength) yielding almost as good a prediction accuracy as the less sparse estimates. Clearly, a practitioner may question the applicability of the EN estimator or the linear regression model in this case. Second, the CV curve for adaptive PENSE is highly non-smooth. On the other hand, adaptive PENSE seems to find models with better prediction accuracy than the intercept-only model, but the standard errors are large, and the highly irregular shape of the CV curve undermines confidence in those estimates. In Sections 4 and 5 we demonstrate that the poor performance of the classical EN estimator is likely due to contamination. But the other question is why small changes in λ lead to such drastic changes in the estimated prediction accuracy for adaptive PENSE.

*Remark 1.*A slight variation of N-CV described above, which is sometimes used in practice, is to compute the global minimizers only at the $\hat{\lambda}$ chosen by N-CV, rather than for every value in $\mathcal{L}$. While this does not affect the choice of $\hat{\lambda}$, it can lead to faster computations in some settings. However, since (1) is usually solved by an iterative algorithm that can be started with the solution(s) from the previous slightly higher penalty level, the computational overhead of computing the entire regularization path is typically small.

Remark 2. The non-convexity of (1) requires a careful choice of starting values for the iterative optimization algorithm. The most prevalent strategy to obtain starting values for

robust estimators is to compute non-robust penalized estimates on "outlier-free" subsets of the data. A common approach is to consider many random subsets of size less than $n - o > 1$, where o is a conservative upper bound on the number of contaminated observations, in the hope that at least some of those subsets are free of contamination and hence give appropriate starting values close to the global minimum (e.g., Alfons et al. 2013). Another option is a guided search for outlier-free subsets, e.g., the extension of the Peña-Yohai method (Peña and Yohai 1999) to penalized estimation (Cohen Freue et al. 2019; Maronna 2011). A detailed discussion on finding appropriate starting values for robust regression can be found in Cohen Freue et al. (2019) and Maronna, Martin, et al. (2019). For cross-validation, obtaining good starting values in each CV fold is computationally very taxing. A common computational shortcut (e.g., Khan et al. 2010) circumventing this problem is to use the minimum from the complete training data as the starting point in the CV folds.

2.1 Failings of N-CV for Penalized Robust Estimators

In short, the blame for the non-smoothness of the N-CV curve is on the non-convexity of the objective function combined with the presence of contamination. In the following, we highlight the issues that the non-convexity and contamination create for CV of robust penalized estimators. Together, these issues may lead to an undesirable coefficient estimate and a CV prediction error that is substantially biased, has high variance, or both.

Non-smoothness of the penalization path. Due to the non-convexity of the loss function, (1) can have more than one minimum. While robust penalized estimators are usually defined as the global minimizer of (1), the "global" designation assigned to the minimum with the smallest objective function value can be misleading. Although the objective function (1) is smooth in λ, the path of the global minimum is not necessarily smooth. In Proposition 1 in the supplementary material, we show that in the simple

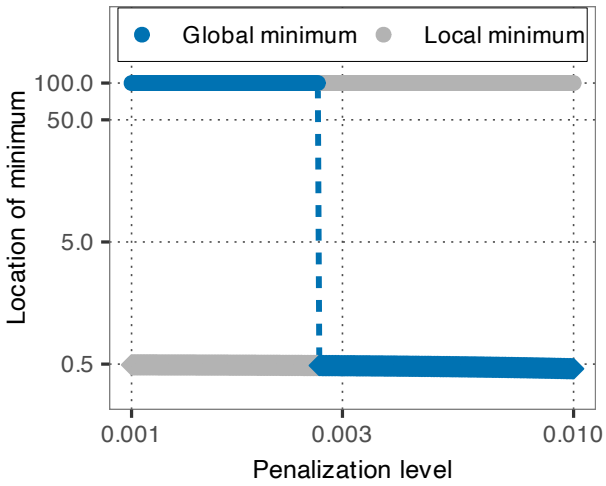

Figure 2: Demonstration of a non-smooth regularization path for a penalized M-estimator of regression in a simulation setting detailed in Section S.1.1 of the supplementary materials. The dots represent local minima at each penalization level, while blue dots indicate the minimum that is designated as the "global" minimum, i.e., has the smallest value of the objective function at that particular penalization level λ.

univariate setting and under certain conditions, we can find at least one λ where the path of the global minimum of a penalized M-estimator, denoted by $\hat{\beta}^*(\lambda)$, has a discontinuity, i.e., $\lim_{\delta \to 0} |\hat{\beta}^*(\lambda - \delta) - \hat{\beta}^*(\lambda + \delta)| > 0$. We further provide a simple example setting where these conditions are satisfied with positive probability. Figure 2 shows an instance where the penalized M-estimator has two minima for each λ in $\mathcal{L}$, one at $\beta \approx 0.5$ and one at $\beta \approx 100$. The minima that are designated as the "global" minima, however, have a discontinuity around $\lambda = 0.003$, jumping from the local minimum at $\beta \approx 100$ to the minimum at $\beta \approx 0.5$. In high-dimensional problems, the non-smoothness of the global minimum is even more severe, as robust loss functions tend to have more local minima as the dimension of the covariates increases.

The primary reason to focus on the global minimum of (1) is that it may possess desirable statistical properties. The global minimum of the objective function of PENSE and adaptive PENSE, for instance, is root-n consistent, finite-sample robust, and for adaptive PENSE it possesses the oracle property (Kepplinger 2023). These properties, however, only pertain to the global minimum at a properly chosen penalty parameter λ^*. In practice,

when many different penalization levels must be tried, considering only the global minimum is fallacious. In the example scenario from Figure 2, the majority of the data follow a linear regression model with $\beta_\star = 100$, and therefore an estimate close to that would lead to a reasonably small prediction error for all λ. However, if we were to consider only the global minima (blue points), a stark difference would be seen between prediction accuracy with small versus large λ values. In situations where a larger λ value and therefore a sparser solution may be preferable, restricting attention to only the global minimum would prevent the selection of a good estimate.

Mismatch between the global minima and the N-CV solutions. A related problem arises when estimating the prediction accuracy of the minima of the objective function using N-CV. The objective function evaluated on the randomly chosen CV training indices, $\mathscr{T}_k$, also possesses multiple minima. Whether the global minimum in $\mathscr{T}_k$ describes a similar signal as the global minimum on the full training data, $\mathscr{T}$, however, is unknown. Referring again to the example in Figure 2, if the global minimum on $\mathscr{T}_k$ for a small $\lambda = 0.001$ is around 0.5, it would contain little information for estimating the prediction accuracy of the global minimum on $\mathscr{T}$, which is around 100. As the number of local minima increases, so too does the probability that the global minimum on $\mathscr{T}_k$ is unrelated to the global minimum on $\mathscr{T}$. Among the K CV folds, some may yield a global minimum related to that of the full dataset, while others do not. With N-CV, however, the prediction accuracy is estimated from both the related and unrelated minima, potentially introducing substantial bias and hence leading to nonsensical results.

Robust measures of prediction accuracy for N-CV, like the MAPE or the τ-size, cannot solve this problem either. The unrelated minima could give prediction errors that appear as outliers. Combined with the truly contaminated observations in the data, they may outnumber the useful prediction errors and hence break the robust measures and lead

to unbounded bias. Repeating N-CV many times helps to mitigate these instabilities, but depending on the severity of the mismatches, a large number of replications may be necessary. This creates a computational bottleneck and can also lead to an overestimation of the variance of the prediction accuracy.

3 Robust Information Sharing CV

We now present Robust Information Sharing CV (RIS-CV). The central component of RIS-CV is the tracking of multiple minima. This requires a strategy to match minima on the complete training data with the minima from the CV folds.

Tracking multiple minima. The main component of RIS-CV is to keep track of multiple minima for the complete training data, $\mathscr{T}$, and all the CV folds, $\mathscr{F}_k$, $k = 1, \ldots, K$. For every $\lambda \in \mathcal{L}$ we retain $Z_\lambda^{(\mathscr{T})} \geq 1$ unique minima, denoted by $\mathcal{B}_\lambda^{\mathscr{T}}$. To keep computational complexity at bay, we limit the maximum number of unique minima to Z, ensuring that $Z_\lambda^{(\mathscr{T})} \leq Z$. In the simulation study in Section 5, for example, we set $Z = 10$.

We have shown in Section 2 that the regularization path may not be smooth if $Z = 1$. It is also easy to see that relaxing the penalty increases the number of local minima monotonically. Moreover, if $\hat{\boldsymbol{\beta}}(\lambda) \in \mathcal{B}_\lambda^{\mathscr{T}}$ is a minimum of the objective function for λ, then for any $\delta > 0$ there exists an $\epsilon > 0$ such that there is a minimum $\hat{\boldsymbol{\beta}}(\lambda - \epsilon)$ with $\|\hat{\boldsymbol{\beta}}(\lambda - \epsilon) - \hat{\boldsymbol{\beta}}(\lambda)\| < \delta$. Although this may not be the global minimum of the objective function at $\lambda - \epsilon$, tracking multiple minima increases the chances that $\hat{\boldsymbol{\beta}}(\lambda - \epsilon) \in \mathcal{B}_{\lambda-\epsilon}^{\mathscr{T}}$. Instead of a single, non-smooth regularization path, RIS-CV therefore captures multiple smooth regularization paths.

Remark 3. Common software implementations for estimators based on (1) employ an "exploration-concentration" strategy, i.e., explore all the provided starting values for a

few iterations, and iterate only the "most promising" solutions until convergence. These implementations nevertheless return and utilize only the global minima for estimation. In RIS-CV, on the other hand, we harness these multiple smooth regularization paths to improve the reliability of the regularization path and the estimated CV curve. In fact, too aggressive screening of minima in the exploration stage is harmful to RIS-CV, as it depletes the final set of minima of diverse solutions. To apply RIS-CV successfully in practice, implementations must therefore be adjusted to screen out only likely duplicates, without eliminating solutions based on their objective function value.

Matching minima based on similarity. To estimate the prediction error of the minima in $\mathcal{B}_\lambda^{\mathscr{T}}$, they must be matched with the corresponding minima in each CV fold, $\mathcal{B}_\lambda^{\mathscr{T}_k}$. We propose to measure the similarity of the minima based on the robustness weights associated with these minima, leveraging the fact that the loss function in (1) can be recast as a weighted least-squares loss. The weight for the S-loss function, for example, is given by

$$w_i(\mathbf{r}) = \frac{\rho'(\tilde{r}_i)/\tilde{r}_i}{\sum_{k\in\mathscr{D}} \rho'(\tilde{r}_k)\tilde{r}_k}, \quad i \in \mathscr{T}, \tag{4}$$

where $\tilde{r}_i = r_i/\hat{\sigma}_M(\mathbf{r})$ is the residual scaled by the M-scale estimate.

The weights encode the "inlyingness" of an observation relative to the regression hyperplane, indicating how well each observation conforms to the estimated model. An observation close to the hyperplane receives a larger weight, while an observation far away and hence deemed outlying receives a weight of 0. For RIS-CV we define the similarity between two coefficient vectors, $\omega(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \mathscr{I})$, as the correlation between the corresponding weight vectors $\mathbf{w}(\mathbf{r}_1)$ and $\mathbf{w}(\mathbf{r}_2)$ over index set $\mathscr{I}$,

$$\omega(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \mathscr{I}) = \frac{\frac{1}{|\mathscr{I}|}\sum_{i\in\mathscr{I}} w_i(\mathbf{r}_1)w_i(\mathbf{r}_2) - \overline{w}(\boldsymbol{\beta}_1)\overline{w}(\boldsymbol{\beta}_2)}{\sqrt{\left[\frac{1}{|\mathscr{I}|}\sum_{i\in\mathscr{I}} w_i(\boldsymbol{\beta}_1)^2 - \overline{w}(\boldsymbol{\beta}_1)^2\right]\left[\frac{1}{|\mathscr{I}|}\sum_{i\in\mathscr{I}} w_i(\boldsymbol{\beta}_2)^2 - \overline{w}(\boldsymbol{\beta}_2)^2\right]}}, \tag{5}$$

with $r_{li} = y_i - \mathbf{x}_i^\intercal \boldsymbol{\beta}_l$, $l \in \{1, 2\}, i \in \mathscr{I}$ and $\overline{w}(\mathbf{r}_l) = \frac{1}{|\mathscr{I}|} \sum_{i \in \mathscr{I}} w_i(\mathbf{r}_l)$.

We utilize the weight-similarity $\omega(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \mathscr{I})$ to match the minima in $\mathcal{B}_\lambda^{\mathscr{T}}$ with their closest counterpart in each CV fold, $\mathcal{B}_\lambda^{\mathscr{T}_k}$. Specifically, for a set of minima $\mathcal{B}_\lambda^{\mathscr{T}}$ we define its collection of CV-surrogates from fold $k = 1, \ldots, K$ as

$$\check{\mathcal{B}}_\lambda^{\mathscr{T}_k} = \left\{ \underset{\boldsymbol{\beta}^* \in \mathcal{B}_\lambda^{\mathscr{T}_k}}{\arg\min}\, \omega(\boldsymbol{\beta}^*, \boldsymbol{\beta}_j; \mathscr{T}_k) \colon j = 1, \ldots, Z_\lambda^{\mathscr{T}}, \boldsymbol{\beta}_j \in \mathcal{B}_\lambda^{\mathscr{T}} \right\}. \tag{6}$$

Hence, the q-th element in $\check{\mathcal{B}}_\lambda^{\mathscr{T}_k}$ is the minimum from CV fold $\mathscr{T}_k$ most similar to the q-th element in $\mathcal{B}_\lambda^{\mathscr{T}}$. There can be duplicates in $\check{\mathcal{B}}_\lambda^{\mathscr{T}_k}$.

Matching minima from the complete training data to their surrogates in each CV fold allows us to more reliably estimate the prediction accuracy of each minimum than relying solely on the ordering of the minima based on their objective function value. With a non-convex loss function, the minimum with the lowest objective value in $\mathcal{B}_\lambda^{\mathscr{T}}$ may capture a very different signal than the minimum with the lowest objective value in $\mathcal{B}_\lambda^{\mathscr{T}_k}$. In contrast, our strategy matches each minimum in $\mathcal{B}_\lambda^{\mathscr{T}}$ with a minimum in $\mathcal{B}_\lambda^{\mathscr{T}_k}$ that best agrees with the outlyingness of the observations in CV fold $\mathscr{T}_k$.

The weight-based similarity has several advantages over distances between coefficient estimates or residuals. First, it is dimensionless and independent of the number of covariates, covariance structure, or response scale. Second, observations that are deemed contaminated do not affect the weight-similarity, while measures using residuals can be arbitrarily affected. Third, contaminated observations often cause local minima; therefore, minima that agree on the outlyingness of observations thus likely describe a similar signal.

The approximation for CV estimates proposed by Khan et al. (2010) can be considered a strategy to keep the minima in $\mathcal{B}_\lambda^{\mathscr{T}_k}$ close to the minima in $\mathcal{B}_\lambda^{\mathscr{T}}$. Using $\mathcal{B}_\lambda^{\mathscr{T}}$ as starting points and limiting the algorithm to a few iterations to compute $\mathcal{B}_\lambda^{\mathscr{T}_k}$, it is likely that the resulting minima are close to the starting points. However, for penalized robust estimators

and high-dimensional data, it is in general difficult to choose a fixed number of iterations that yields a good balance between adapting to the data in the CV fold and staying close to the minima in $\mathcal{B}_\lambda^{\mathscr{J}}$. Our matching scheme, on the other hand, does not require such tuning and hence leads to more reliable estimates of the prediction accuracy, albeit at the cost of more expensive computations.

Once the CV-surrogates from each CV fold $\mathscr{T}_k$ and each penalty parameter $\lambda \in \mathcal{L}$ are determined, RIS-CV utilizes the robustness weights to quantify the prediction accuracy of every minimum $\hat{\boldsymbol{\beta}}_q \in \mathcal{B}_\lambda^{\mathscr{J}}$. The prediction accuracy is estimated by a weighted standard deviation of the CV prediction errors, where the weights reflect the outlyingness of each observation as estimated on the complete training data:

$$\hat{E}_q^{(2)}(\lambda) = \sqrt{\frac{1}{\sum_{i \in \mathscr{J}} w_i(\mathbf{r}_q)} \sum_{k=1}^{K} \sum_{i \in \mathscr{F}_k} w_i(\mathbf{r}_q) \left(y_i - \mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}}_q^k \right)^2}, \quad q = 1, \ldots, |\mathcal{B}_\lambda^{\mathcal{T}}|, \tag{7}$$

where $\mathbf{r}_k = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_k$ are the residuals from the complete training data and $\hat{\boldsymbol{\beta}}_q^k \in \check{\mathcal{B}}_\lambda^{\mathscr{T}_k}$ is the CV-surrogate of the q-th minimum. This effectively ignores the prediction error of observations that are deemed contaminated by that particular $\hat{\boldsymbol{\beta}}_q$. For heavy-tailed error distributions, a weighted mean absolute error may be more appropriate,

$$\hat{E}_q^{(1)}(\lambda) = \frac{1}{\sum_{i \in \mathscr{J}} w_i(\mathbf{r}_q)} \sum_{k=1}^{K} \sum_{i \in \mathscr{F}_k} w_i(\mathbf{r}_q) \left| y_i - \mathbf{x}_i^\intercal \hat{\boldsymbol{\beta}}_q^k \right|, \quad q = 1, \ldots, |\mathcal{B}_\lambda^{\mathcal{T}}|. \tag{8}$$

Contaminated observations cannot be expected to be predicted well by the model, and their prediction errors should therefore not affect the overall assessment of an estimate's prediction accuracy. Robust estimators assign a zero weight to a bounded number of observations (e.g., S-estimators assign zero weights to fewer than $\lfloor \delta n \rfloor$ observations). Therefore, many of the good, inlying observations will have a weight greater than 0, even if $\hat{\boldsymbol{\beta}}_q$ describes an illicit signal driven by contamination. Consequently, the weighted metric will propagate

the high prediction errors for these observations and the illicit signal will thus have poor prediction accuracy.

Compared to the usual robustification of N-CV through robust measures of the prediction error, our approach connects the estimated prediction error more closely to the estimated outlyingness, reducing the risk of misrepresenting an estimate's prediction accuracy. While robust measures such as the τ-size guard against the effects of arbitrarily large prediction errors, these measures do not discriminate whether the large prediction errors are due to observations that are deemed contaminated or not. Therefore, even inliers are allowed to have very large prediction errors. This disconnect between the outlyingness of the observations and their effect on the estimated prediction error can lead to a high variance in the estimated prediction accuracy.

Remark 4. As suggested by an anonymous reviewer, an alternative is to use the average weights from the CV fits instead of the weights from the fit to the full training data. Specifically, one could replace $w_i(\mathbf{r}_q)$ in equations 7 and 8 by $w^*_{iq} = \frac{1}{K-1} \sum_{k:i \in \mathscr{T}_k} w_i \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^k_q \right)$. These weights reduce the dependence of the estimated prediction error on the full fit and hence can alleviate the potential underestimation of the actual prediction error caused by overfitting the complete training data. The R package in the supplementary materials supports this alternative choice of weights, although our numerical experiments in Section 5 suggest that this approach often leads to sparser solutions with worse prediction accuracy.

Similar to N-CV, the prediction accuracy estimated by RIS-CV depends on the random CV splits and is thus a stochastic quantity. RIS-CV should therefore be repeated several times to assess the variability of the estimate, but the number of replications can usually be much smaller than the number of replications needed for N-CV. We generally suggest repeating RIS-CV 5 to 20 times, depending on the complexity of the problem. For

more complex problems, higher variability in the estimated prediction accuracy may occur, requiring additional RIS-CV replications. The RIS-CV strategy is detailed in Algorithm 1.

For faster computations, the R package in the supplementary materials performs step 1 in Algorithm 1 for all $\lambda \in \mathcal{L}$ before the replicated RIS-CV (steps 2 through 10) is applied. This is done to efficiently utilize the set of minima from the previous penalization level $\lambda' > \lambda$, $\mathcal{B}^{\mathcal{T}}_{\lambda'}$ as starting points.

The RIS-CV procedure yields estimates of the prediction accuracy for up to Z minima at each level of the penalty parameter, and the "optimal" minimum can be selected in several ways. We simply select the minimum with the best prediction accuracy for each $\lambda \in \mathcal{L}$, $q^*_\lambda = \arg\min_{q=1,\ldots,|\mathcal{B}^{\mathcal{T}}_\lambda|} \hat{E}_q(\lambda)$. These estimates together with the associated standard errors can be plotted (i.e., the RIS-CV curve) to judge the model's suitability for the problem at hand and to select the desired penalty parameter. Standard errors could also be considered when choosing q^*_λ, but the numerical experiments below do not suggest an improvement over the simpler strategy applied here. The R package in the supplementary materials returns the estimated prediction accuracy and its standard error for all minima and therefore allows the user to utilize more sophisticated strategies or to fine-tune RIS-CV.

Remark 5. In practical applications, we suggest first running a small number of RIS-CV replications with a large number of minima, e.g., $Z = 50$, and plotting the CV curve. The R package in the supplementary material allows the user to construct CV curves for different Z and different CV measures (the weighted RMSPE 7, the weighted MAPE 7 or their variants using the average weights from the CV fits as in Remark 4). If a smaller Z also yields a smooth CV curve, further CV replications can be added with this smaller Z to save computation time. On the other hand, if the regularization path and/or the RIS-CV curve are still highly irregular, the user can try increasing Z and investigating different measures for the prediction error.

Remark 6. Computation speed of RIS-CV can be improved by using only the minima from the fit to the complete training data as starting points for the CV folds, similar to the shortcut used in Khan et al. (2010). For RIS-CV, all the local minima from the complete training data, not only the global minimum, can be leveraged as starting points. This restricts the search space around local minima of interest. However, in some situations, the CV surrogates identified by this strategy are inferior to the CV surrogates obtained by re-computing the starting points. Section S3.3 in the supplementary materials shows the differences between re-computing the starting points for each CV fold and using only the minima from the complete training data for RIS-CV in the simulation study of Section 5. In high-dimensional problems, this computational shortcut may be necessary when re-computing starting points for each CV fold becomes infeasible. The R package in the supplementary materials supports this shortcut to ensure applicability of RIS-CV to a wide range of applications. In one of the applications of RIS-CV shown in the supplementary materials (Section S2.1), for instance, this computational shortcut is used to reduce the computation time to a reasonable level without sacrificing prediction accuracy.

4 Biomarker Discovery Study

We show the benefits of RIS-CV over naïve CV for developing a biomarker for cardiac allograft vasculopathy (CAV) based on protein expression levels. In this application, we use adaptive PENSE (Kepplinger 2023), a two-stage estimator where both stages use the S-loss in (1), ℓ_S, but with different penalty terms. The first stage is an S-Ridge (Maronna 2011) estimator, i.e., $P(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2$. The second stage uses an adaptive EN penalty, $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\check{\beta}_j|^{-1} \left(\frac{1-\alpha}{2}\beta_j^2 + \alpha|\beta_j|\right)$, where $\check{\boldsymbol{\beta}}$ is the estimate from the first stage. We apply N-CV and RIS-CV in both stages of this estimator.

CAV is a life-threatening complication after receiving a cardiac transplant characterized

Algorithm 1 Robust Information Sharing CV

Input: Standardized data set, $\mathcal{D} = \{(y_i, \mathbf{x}_i) \colon i \in \mathscr{T}\}$, fixed hyperparameter, λ, the number of folds, K, the maximum number of minima retained, Z, and the number of cross-validation replications, R.

1: Compute up to Z unique local minima of (1) using all observations $\mathscr{T} = \{1, \ldots, n\}$ of $\mathcal{D}$ in the non-convex optimization algorithm. Denote this set by by $\mathcal{B}_\lambda^{\mathscr{T}}$.
2: **for** $r = 1, \ldots, R$ **do**
3: Split the data into K cross-validation folds, denoted by $\mathscr{F}_1, \ldots, \mathscr{F}_K$, such that $\mathscr{F}_k \cap \mathscr{F}_{k'} = \emptyset$, for $k \neq k'$, and $\bigcup_{k=1}^K \mathscr{F}_k = \mathscr{T}$.
4: **for** $k = 1, \ldots, K$ **do**
5: Compute up to Z unique local minima of (1) using only the observations in $\mathscr{T}_k = \mathscr{T} \setminus \mathscr{F}_k$, denoted by $\mathcal{B}_\lambda^{\mathscr{T}_k}$.
6: From $\mathcal{B}_\lambda^{\mathscr{T}_k}$ determine the CV-surrogates $\check{\mathcal{B}}_\lambda^{\mathscr{T}_k}$ according to (6).
7: **end for**
8: Estimate the robust weighted RMSPE (7) or MAPE (8) for each minimum $\hat{\boldsymbol{\beta}}_q \in \mathcal{B}_\lambda^{\mathscr{T}}$, $q = 1, \ldots, |\mathcal{B}_\lambda^{\mathscr{T}}|$, denoted by $\hat{E}_q^{(r)}(\lambda)$.
9: **end for**
10: Compute the average robust weighted prediction errors and their standard errors

$$\hat{E}_q(\lambda) = \frac{1}{R} \sum_{r=1}^{R} \hat{E}_q^{(r)}(\lambda), \quad \widehat{\mathrm{SD}}_q(\lambda) = \sqrt{\frac{1}{R-1} \sum_{r=1}^{R} \left(\hat{E}_q^{(r)}(\lambda) - \hat{E}_q(\lambda) \right)^2}.$$

by narrowing of vessels that supply oxygenated blood to the heart. The usual clinical biomarker for CAV is the percentage of diameter stenosis of the left anterior descending artery. The data is obtained from Kepplinger and Cohen Freue (2023), who use a synthetic replicate of the restricted original data. The goal is to predict the stenosis of the artery using the expression levels of the 81 protein groups available for a total of $N = 37$ patients. The adaptive PENSE estimator is tuned to a breakdown point of 20% and uses an EN penalty with $\alpha = 0.8$, following the analysis in Kepplinger and Cohen Freue (2023). For RIS-CV, up to $Z = 30$ local minima are tracked.

Figure 3 shows the CV curves and the regularization path of adaptive PENSE with 5-fold CV. The top parts of panels (a) and (b) show two independent CV runs for N-CV and RIS-CV, respectively. It is clear that the individual CV curves from RIS-CV are more consistent in their shape than those from N-CV.

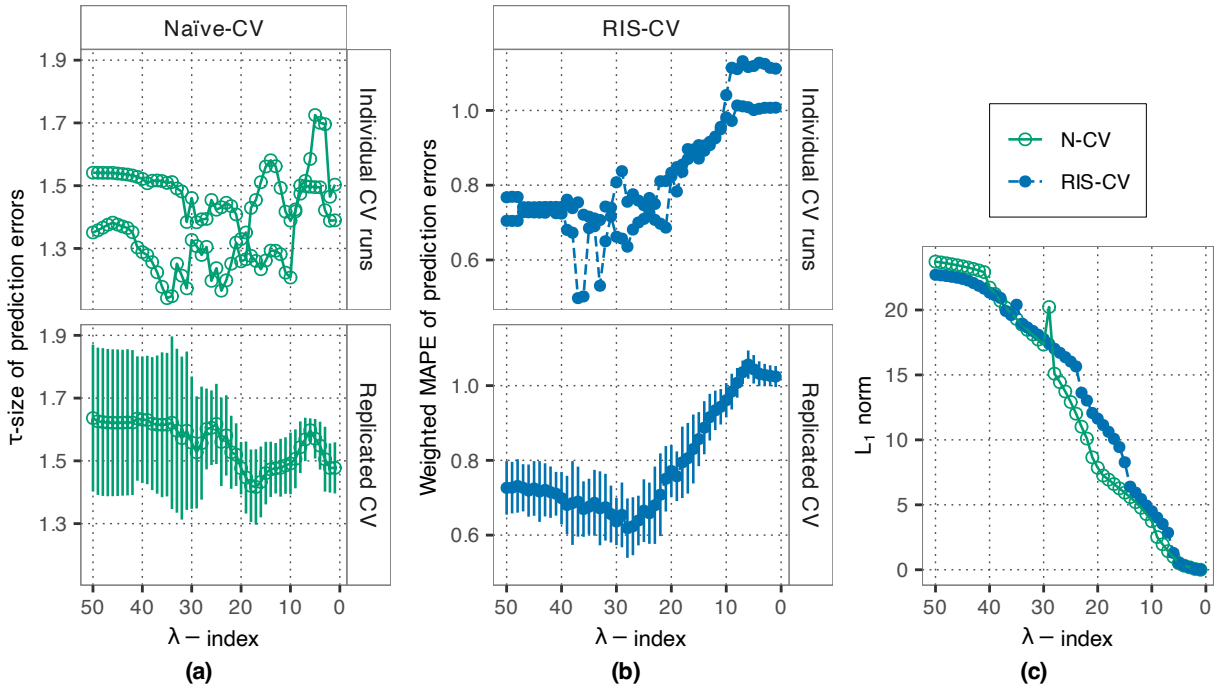

Figure 3: (a) Individual CV curves from 5-fold N-CV (top) and the average ±1 SD over 10 replications (bottom). (b) Individual CV curves from 5-fold RIS-CV (top) and the average ±1 SD over 10 replications (bottom). (c) L_1 norm (excluding the intercept) of the selected minimum of the adaptive PENSE objective function.

The benefits of RIS-CV become more obvious when averaging 10 replications of 5-fold CV, as shown in the bottom parts of Figures 3(a) and (b). Even after averaging, it is difficult to identify an appropriate penalization level with N-CV. The intercept-only model and the model at λ-index $= 17$ have similar prediction accuracy, and the 1-SE rule would select a penalization level in between (λ-index$= 9$), even though these models are very different (0 vs. 10 vs. 4 non-zero slope coefficients). RIS-CV, on the other hand, identifies a tight range of penalization levels around λ-index $= 29$ that seem to yield the best prediction accuracy in this data set, and the 1-SE rule would select λ-index $= 18$. The smoothness and stability of the RIS-CV curve are clearly advantageous in identifying a good penalization level in this application.

We further analyze the smoothness of the regularization path, in terms of the L_1 norm of the slope coefficients for the minima selected by N-CV and RIS-CV (Figure 3c). N-CV selects the global minimum at each penalization level, while RIS-CV selects the minimum

with the best estimated prediction accuracy. RIS-CV's flexibility in choosing a non-global minimum results in a smoother CV curve and regularization path.

In Section S2 of the supplementary materials, we present two additional applications with similar conclusions as in this CAV study. In these applications, we further estimate the out-of-sample (OOS) prediction error on an independent test set. The effects of the local minima in these applications are again noticeable, but less pronounced than in the CAV study. Nevertheless, RIS-CV leads to smoother CV curves and regularization paths, as well as better OOS prediction accuracy.

5 Simulation Study

In the biomarker discovery study and the applications in the supplementary materials, we demonstrate that RIS-CV leads to smoother CV curves and in turn to better selection of the penalty parameters. In this simulation study, we illustrate that the minimum selected by RIS-CV, which is not necessarily a global minimum for the selected penalization level, can lead to a better out-of-sample prediction. Throughout this study, we compare RIS-CV with N-CV for the PENSE estimator. We consider a data-generating process (DGP) similar to Kepplinger (2023). The data is generated according to the linear model $y_i = \mathbf{x}_i^\intercal \boldsymbol{\beta}^0 + \varepsilon_i$, $i = 1, \dots, n$, where $\mathbf{x}_i$ is the p-dimensional covariate vector, $\boldsymbol{\beta}^0 = (1, \dots, 1, 0, \dots 0)^\intercal$ is the true coefficient vector with the first $s = \lfloor \log(n) \rfloor$ entries equal to 1 and the others are all 0. The covariates $\mathbf{x}_i$ follow a multivariate t_4 distribution and AR(1) correlation structure, $\mathrm{Cor}(X_j, X_{j'}) = 0.5^{|j-j'|}$, $j, j' = 1, \dots, p$. The i.i.d. errors, ε_i, follow a symmetric distribution, F, with a scale chosen such that $\boldsymbol{\beta}^0$ explains about 50% of the variation in y_i (i.e., SNR ≈ 1). For this, the empirical variance in $\boldsymbol{\varepsilon}$ is measured by the empirical standard deviation if F is Gaussian and by the τ-size for other error distributions.

We consider different scenarios for the number of observations, $n \in (100, 200)$, the

number of available predictors, $p \in (50, 100, 200)$, and error distribution, F, (Gaussian, Laplace, Symmetric Stable with stability parameter $\alpha = 1.5$). Good leverage points are introduced by multiplying the $(p-s)/2$ largest covariate values for 25% of observations by 8. These values are introduced in covariates with a true coefficient of 0 and hence should not affect the estimators. Furthermore, 25% of the observations come from a different model with three distinct contamination signals. We would expect that the objective function has at least one local minimum close to each of them. The contamination signals all follow the linear model but with a different $\boldsymbol{\beta}^0$ and further introduce leverage points in the truly relevant covariates. The details are described in Section S3.1 of the supplementary materials.

For each of the 18 settings, we repeat the simulation 100 times and compare the prediction performance of the solutions/penalty levels selected by N-CV and RIS-CV with $K = 5$ folds and $R = 5$ replications. We select the penalization level according to the 1-SE rule (3). Across all simulations, $Z = 10$ solutions are retained for RIS-CV, and PENSE is tuned to a breakdown point of 33% and α is set to 0.5.

Figure 4 summarizes the simulation results in terms of the prediction accuracy. The boxplots show the relative prediction error attained with RIS-CV versus the prediction error attained with N-CV for replication $r = 1, \ldots, 100$ of setting $s = 1, \ldots, 18$,

$$\text{RelPredErr}^{(r,s)} = \begin{cases} \frac{\sqrt{\mathbb{E}_{y,\mathbf{x}}\left[(y-\mathbf{x}^\intercal \hat{\boldsymbol{\beta}}^{(r,s)}_{\text{RIS-CV}})^2\right]}}{\sqrt{\mathbb{E}_{y,\mathbf{x}}\left[(y-\mathbf{x}^\intercal \hat{\boldsymbol{\beta}}^{(r,s)}_{\text{N-CV}})^2\right]}} - 1 & \text{if } s \text{ uses Gaussian errors,} \\ \frac{\mathbb{E}_{y,\mathbf{x}}\left[\left|y-\mathbf{x}^\intercal \hat{\boldsymbol{\beta}}^{(r,s)}_{\text{RIS-CV}}\right|\right]}{\mathbb{E}_{y,\mathbf{x}}\left[\left|y-\mathbf{x}^\intercal \hat{\boldsymbol{\beta}}^{(r,s)}_{\text{N-CV}}\right|\right]} - 1 & \text{if } s \text{ uses any other error distribution,} \end{cases} \tag{9}$$

where $\hat{\boldsymbol{\beta}}_{\text{RIS-CV}}$ and $\hat{\boldsymbol{\beta}}_{\text{N-CV}}$ are the estimates chosen by RIS-CV and N-CV. The expectations are computed via a Monte Carlo approximation using 10,000 draws.

In most simulation runs, RIS-CV selects solutions with better prediction accuracy than N-CV. Section S3.2 of the supplementary materials includes additional results. Common

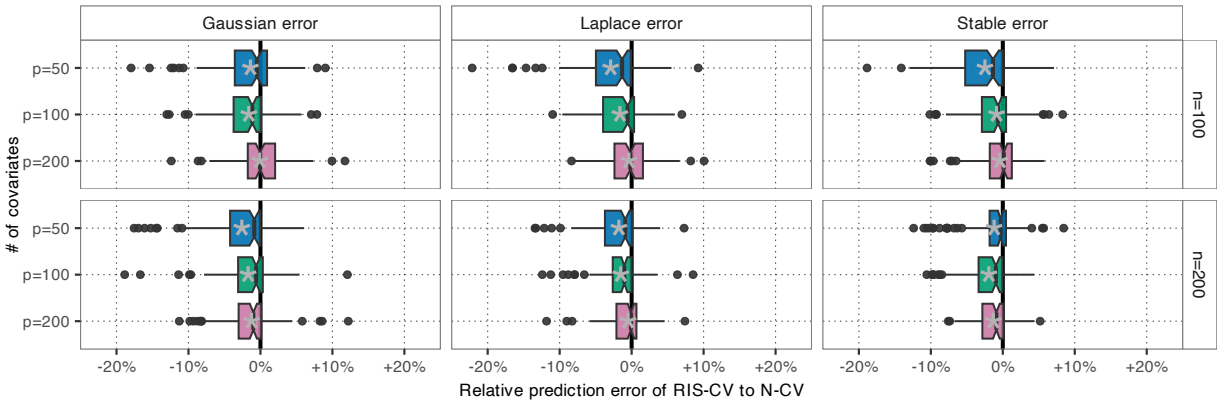

Figure 4: Relative prediction error attained with RIS-CV vs. N-CV according to (9) in 100 simulation replications across 18 different settings. The penalty parameter is chosen by the 1-SE rule (3). The gray asterisks depict the mean. Negative differences mean the solution chosen by RIS-CV leads to better prediction accuracy than N-CV.

measures of smoothness indicate that RIS-CV produces a smoother CV curve than N-CV, offering better guidance on the optimal penalization level. The supplementary materials also compare RIS-CV with the two-step approximation of Khan et al. (2010) for N-CV, showcasing improvements by considering more local minima.

6 Conclusion

Cross-validation is the prevalent data-driven method for selecting models with penalized estimators, including robust penalized estimators. We show that standard CV (N-CV) is unstable when applied to non-convex robust penalized regression estimators defined as the minimizer of (1). Our theory and empirical results reveal that multiple local minima, caused by contamination, can create highly non-smooth penalization paths and CV curves. We further highlight that the local minimum with the smallest value of (1) at a given λ, i.e., the "global" minimum at λ, is not necessarily the minimum closest to the true signal due to the interaction between the robust loss and the penalty term. This nonsmoothness can in turn lead to low-quality estimates of the prediction accuracy, and thus an ill-guided selection of the penalization level. Such issues become more pronounced as the complexity

of the data analysis task increases. The instability of N-CV has been a major obstacle for the utility and adoption of robust penalized estimators.

In this paper, we therefore propose a novel strategy, RIS-CV, where we retain all local minima uncovered by the numerical algorithm to optimize (1) and share contamination information from these minima on the complete training data with the individual cross-validation folds. Our results show that leveraging the robustness weights associated with each local minimum makes it possible to (a) determine which minima in the CV folds correspond most closely with the minima on the full data set and (b) estimate the prediction accuracy of all of those local minima, not only the global minimum. This allows RIS-CV to select the local minimum with the best prediction accuracy, which is not necessarily the global minimum as discussed in Section 2.1, and thus often yielding a smoother CV curve. A single replication of RIS-CV is in general computationally more expensive than N-CV due to the requirement to track many more minima. However, the overall computation time is most often comparable to N-CV because fewer CV replications are necessary to obtain a smooth and informative CV curve.

The proposed matching scheme currently does not differentiate between good and poor matches. The most similar CV solution is chosen as a surrogate, regardless of the actual similarity. Since the metric is a unitless correlation coefficient, thresholding rules could be developed in the future to avoid using unrelated minima in RIS-CV. For example, one could require CV surrogates to have a similarity of at least 0.75. Further research would be necessary, however, to devise appropriate strategies to handle situations where some CV folds do not yield a CV surrogate and how to properly choose the threshold. Moreover, the smoothness of the regularization path and CV curve can be made a more explicit metric in choosing CV surrogates. Similarly, better graphs could be devised to not only show the CV curve for a single solution at each λ, but rather plot multiple smooth CV curves to give the user more control over which solution to choose.

The ideas developed here are not only applicable to robust penalized linear regression, but can easily be extended to robust estimators for the generalized linear model (e.g., Avella-Medina and Ronchetti 2017; Bianco et al. 2022) or to smoothing parameter selection in robust additive regression models (e.g., Kalogridis and Van Aelst 2023; Tharmaratnam et al. 2010).

RIS-CV is a much-needed tool to improve the practicality, utility, and acceptance of robust penalized estimators. Our numerical studies reveal that RIS-CV leads to smoother CV curves, more reliable selection of the penalty parameter, and identification of the most promising minimum of the objective function at the chosen penalization level. We show that improved smoothness and identification of useful minima lead to better out-of-sample prediction accuracy in a large-scale simulation study and in several applications. RIS-CV thus improves the reliability of the robust model selection process and thereby instills more trustworthiness in the results.

Supplementary Materials

The file *supplemental.pdf* contains further theoretical insights into the failures of naïve cross-validation (Section S1), two more empirical studies in Section S2, and details about the simulation settings along with additional results (Section S3).

The file *codes.zip* contains all codes and data sets to reproduce the results and figures in this paper, alongside a copy of the `pense` R package (version 2.5.0-01), which can also be found on CRAN (https://cran.r-project.org/package=pense). All codes are also available on GitHub: https://github.com/dakep/riscv-empirical-studies.

Acknowledgments

The authors are grateful for the detailed and constructive feedback from the Editor, Associate Editor, and two anonymous referees, which greatly improved the manuscript. The authors further thank the George Mason University School of Computing for providing SW with the scholarship to work on this project.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This project was supported by resources provided by the Office of Research Computing at George Mason University (https://orc.gmu.edu) which is funded in part by grants from the National Science Foundation (Award Number 2018631).

References

Alfons, A., Croux, C., and Gelper, S. (2013). "Sparse Least Trimmed Squares Regression for Analyzing High-Dimensional Large Data Sets". *The Annals of Applied Statistics*, *7*(1), 226–248. https://doi.org/10.1214/12-AOAS575

Amato, U., Antoniadis, A., De Feis, I., and Gijbels, I. (2021). "Penalised Robust Estimators for Sparse and High-Dimensional Linear Models". *Statistical Methods & Applications*, *30*(1), 1–48. https://doi.org/10.1007/s10260-020-00511-z

Arslan, O. (2016). "Penalized MM Regression Estimation With L_γ Penalty: A Robust Version of Bridge Regression". *Statistics*, *50*(6), 1236–1260. https://doi.org/10.1080/02331888.2016.1159682

Austern, M., and Zhou, W. (2020). "Asymptotics of Cross-Validation". *arXiv.org*. https://doi.org/10.48550/arXiv.2001.11111

Avella-Medina, M., and Ronchetti, E. (2017). "Robust and Consistent Variable Selection in High-Dimensional Generalized Linear Models". *Biometrika*, *105*(1), 31–44. https://doi.org/10.1093/biomet/asx070

Bates, S., Hastie, T., and Tibshirani, R. (2024). "Cross-Validation: What Does It Estimate and How Well Does It do It?" *Journal of the American Statistical Association*, *119*(546), 1434–1445. https://doi.org/10.1080/01621459.2023.2197686

Bianco, A. M., Boente, G., and Chebi, G. (2022). "Penalized Robust Estimators in Sparse Logistic Regression". *TEST*, *31*(3), 563–594. https://doi.org/10.1007/s11749-021-00792-w

Chang, L., Roberts, S., and and, A. W. (2018). "Robust Lasso Regression Using Tukey's Biweight Criterion". *Technometrics*, *60*(1), 36–47. https://doi.org/10.1080/00401706.2017.1305299

Cohen Freue, G. V., Kepplinger, D., Salibián-Barrera, M., and Smucler, E. (2019). "Robust Elastic Net Estimators for Variable Selection and Identification of Proteomic Biomarkers". *Annals of Applied Statistics*, *13*(4), 2065–2090. https://doi.org/10.1214/19-AOAS1269

Datta, A., and Zou, H. (2019). "A Note on Cross-Validation for Lasso Under Measurement Errors". *Technometrics*, *62*(4), 549–556. https://doi.org/10.1080/00401706.2019.1668856

Dicker, L. H. (2014). "Variance Estimation in High-Dimensional Linear Models". *Biometrika*, *101*(2), 269–284. https://doi.org/10.1093/biomet/ast065

Fan, J., Guo, S., and Hao, N. (2012). "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(1), 37–65. https://doi.org/10.1111/j.1467-9868.2011.01005.x

Filzmoser, P., and Nordhausen, K. (2021). "Robust Linear Regression for High-Dimensional Data: an Overview". *WIREs Computational Statistics*, *13*(4), e1524. https://doi.org/10.1002/wics.1524

Kalogridis, I., and Van Aelst, S. (2023). "Robust Penalized Estimators for Functional Linear Regression". *Journal of Multivariate Analysis*, *194*, 105104. https://doi.org/10.1016/j.jmva.2022.105104

Kepplinger, D. (2023). "Robust Variable Selection and Estimation via Adaptive Elastic Net S-Estimators for Linear Regression". *Computational Statistics & Data Analysis*, *183*, 107730. https://doi.org/10.1016/j.csda.2023.107730

Kepplinger, D., and Cohen Freue, G. V. (2023). "Robust Prediction and Protein Selection With Adaptive PENSE". In T. Burger (Ed.), *Statistical analysis of proteomic data: Methods and tools* (pp. 315–331). Springer US. https://doi.org/10.1007/978-1-0716-1967-4_14

Khan, J. A., Van Aelst, S., and Zamar, R. H. (2007). "Robust Linear Model Selection Based on Least Angle Regression". *Journal of the American Statistical Association*, *102*(480), 1289–1299. https://doi.org/10.1198/016214507000000950

Khan, J. A., Van Aelst, S., and Zamar, R. H. (2010). "Fast Robust Estimation of Prediction Error Based on Resampling". *Computational Statistics & Data Analysis*, *54*(12), 3121–3130. https://doi.org/10.1016/j.csda.2010.01.031

Koller, M., and Stahel, W. A. (2011). "Sharpening Wald-Type Inference in Robust Regression for Small Samples". *Computational Statistics & Data Analysis*, *55*(8), 2504–2515. https://doi.org/10.1016/j.csda.2011.02.014

Li, J. (2023). "Asymptotics of K-Fold Cross Validation". *Journal of Artificial Intelligence Research*, *78*, 491–526. https://doi.org/10.1613/jair.1.13974

Loh, P.-L. (2021). "Scale Calibration for High-Dimensional Robust Regression". *Electronic Journal of Statistics*, *15*(2), 5933–5994. https://doi.org/10.1214/21-EJS1936

Maronna, R. A. (2011). "Robust Ridge Regression for High-Dimensional Data". *Technometrics*, *53*(1), 44–53. https://doi.org/10.1198/TECH.2010.09114

Maronna, R. A., Martin, D. R., Yohai, V. J., and Salibián-Barrera, M. (2019). *Robust Statistics: Theory and Methods (With R)*. John Wiley & Sons, Inc.

Maronna, R. A., and Yohai, V. J. (2010). "Correcting MM Estimates for "Fat" Data Sets". *Computational Statistics & Data Analysis*, *54*(12), 3168–3173. https://doi.org/10.1016/j.csda.2009.09.015

Monti, G. S., and Filzmoser, P. (2021). "Sparse Least Trimmed Squares Regression With Compositional Covariates for High-Dimensional Data". *Bioinformatics*, *37*(21), 3805–3814. https://doi.org/10.1093/bioinformatics/btab572

Mozafari-Majd, E., and Koivunen, V. (2025). "The Adaptive τ-Lasso: Robustness and Oracle Properties". *IEEE Transactions on Signal Processing*, 1–16. https://doi.org/10.1109/TSP.2025.3563225

Peña, D., and Yohai, V. J. (1999). "A Fast Procedure for Outlier Diagnostics in Large Regression Problems". *Journal of the American Statistical Association*, *94*(446), 434–445. https://doi.org/10.1080/01621459.1999.10474138

Reid, S., Tibshirani, R., and Friedman, J. (2016). "A Study of Error Variance Estimation in Lasso Regression". *Statistica Sinica*, *26*, 35–67. https://doi.org/10.5705/ss.2014.042

Ronchetti, E., Field, C., and Blanchard, W. (1997). "Robust Linear Model Selection by Cross-Validation". *Journal of the American Statistical Association*, *92*(439), 1017–1023. https://doi.org/10.1080/01621459.1997.10474057

She, Y., Wang, Z., and Shen, J. (2021). "Gaining Outlier Resistance With Progressive Quantiles: Fast Algorithms and Theoretical Studies". *Journal of the American Statistical Association*, *117*(539), 1282–1295. https://doi.org/10.1080/01621459.2020.1850460

Smucler, E., and Yohai, V. J. (2017). "Robust and Sparse Estimators for Linear Regression Models". *Computational Statistics & Data Analysis*, *111*, 116–130. https://doi.org/10.1016/j.csda.2017.02.002

Sun, Q., Zhou, W.-X., and Fan, J. (2019). "Adaptive Huber Regression". *Journal of the American Statistical Association*, *115*(529), 254–265. https://doi.org/10.1080/01621459.2018.1543124

Tharmaratnam, K., Claeskens, G., Croux, C., and Salibián-Barrera, M. (2010). "S-Estimation for Penalized Regression Splines". *Journal of Computational and Graphical Statistics*, *19*(3), 609–625. https://doi.org/10.1198/jcgs.2010.08149

Yohai, V. J., and Zamar, R. H. (1988). "High Breakdown-Point Estimates of Regression by Means of the Minimization of an Efficient Scale". *Journal of the American Statistical Association*, *83*(402), 406–413. https://doi.org/10.1080/01621459.1988.10478611

Supplemental to "Information Sharing for Robust and Stable Cross-Validation"

David Kepplinger* and Siqi Wei†
Department of Statistics, George Mason University, Fairfax, VA, USA

*dkeppling@gmu.edu; ORCID: 0000-0002-3776-4819
†swei3@gmu.edu; ORCID: 0000-0002-2758-9922

S1 Failures of Naïve Cross-Validation

S1.1 Non-smooth path of global minima

Here we demonstrate that the chances of the global minimum "jumping" between local minima when the penalization level changes is non-negligible. The following proposition shows that if there are two local minima (the "good" and the "bad" minimum), with the bad minimum being much closer to the origin than the good minimum, the bad minimum will take over from the good minimum as the global minimum when the penalization level is increased. Under the stated conditions, the proposition is entirely deterministic. We will show later that there are indeed situations in which the conditions for the proposition are satisfied with non-zero probability.

Proposition 1. *Consider a bounded robust loss function ρ with $\psi(x) = \rho'(x) = 0$ for $|x| > c_1 > 0$ and $\psi'(x) \geq 0$ for $|x| < c_2 < c_1$ and $c_2 > 0$. Assume that $\hat{\beta}_c$ and $\hat{\beta}_\star$ are the only two minima of the objective function at a penalization level λ, that $\hat{\beta}_c > 0$, $\hat{\beta}_\star - \hat{\beta}_c \gg 2c_1/x_{min}$, where $x_{min} = \min\{|x_i| \colon x_i \neq 0\}$, and $\mathcal{O}(\hat{\beta}_c; \lambda) = \mathcal{O}(\hat{\beta}_\star; \lambda)$. Assume further that there exists a subset of the n observations, $\mathcal{C} \subset \{1, \ldots, n\}$, such that $|y_i - \hat{\beta}_c x_i| < c_1$ for all $i \in \mathcal{C}$, $|y_i - \hat{\beta}_\star x_i| < c_1$ for all $i \notin \mathcal{C}$, and that the cardinality of the set $\{i \colon x_i = 0\}$ is less than bn.*

Then, for any δ with $|\delta|$ small enough, $\tilde{\beta}_c = \hat{\beta}_c - \delta/S_c$ with

$$S_c = \frac{1}{n} \sum_{i \in \mathcal{C}} \psi' \left(y_i - (\hat{\beta}_c + \nu_c) x_i \right) x_i^2$$

and $\nu_c \in (0, \delta)$ is a minimum of the objective function for penalization level $\lambda + \delta$. Similarly, with $S_\star = \frac{1}{n} \sum_{i \notin \mathcal{C}} \psi' \left(y_i - (\hat{\beta}_\star + \nu_\star) x_i \right) x_i^2$ and $\nu_\star \in (0, \delta)$, $\tilde{\beta}_\star = \hat{\beta}_\star - \delta/S_\star$ is another

minimum of the objective function. Furthermore,

$$\mathcal{O}(\hat{\beta}_c + |\delta|/S_c; \lambda - |\delta|) > \mathcal{O}(\hat{\beta}_\star + |\delta|/S_\star; \lambda - |\delta|), \text{ and}$$
$$\mathcal{O}(\hat{\beta}_c - |\delta|/S_c; \lambda + |\delta|) < \mathcal{O}(\hat{\beta}_\star - |\delta|/S_\star; \lambda + |\delta|).$$

Therefore,

$$\lim_{\tilde{\delta}\to 0} |\hat{\beta}(\lambda - \tilde{\delta}) - \hat{\beta}(\lambda + \tilde{\delta})| > 0.$$

In other words, under the above assumptions, the regularization path for parameter β has a discontinuity at λ and hence is non-smooth.

Proof. The first step is to show that $\tilde{\beta}_c$ and $\tilde{\beta}_\star$ are minima of $\mathcal{O}(\beta; \lambda + \delta)$. From the assumptions, we see that for all $i \in \mathcal{C}$, $|y_i - x_i\hat{\beta}_c| < c_1$ and either $x_i = 0$ or $|y_i - x_i\hat{\beta}_\star| \geq c_1$. For all $i \notin \mathcal{C}$, on the other hand, $|y_i - x_i\hat{\beta}_\star| < c_1$ and either $|y_i - x_i\hat{\beta}_c| \geq c_1$ or $x_i = 0$. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\rho(y_i - x_i\beta)\Big|_{\beta=\hat{\beta}_c+\eta} = \begin{cases} \psi\left(y_i - x_i(\hat{\beta}_c + \eta)\right) x_i \neq 0 & \text{if } i \in \mathcal{C}, \\ 0 & \text{if } i \notin \mathcal{C}, \end{cases}$$

for any small enough perturbation η. This shows that the derivative at $\hat{\beta}_c$ is determined solely by observations in $\mathcal{C}$ and, using the same arguments, we can show that the derivative at $\hat{\beta}_\star$ is determined solely by observations not in $\mathcal{C}$. From this, we can further show that

$$
\begin{aligned}
n(\lambda+\delta) &= \sum_{i\in\mathcal{C}} \psi(y_i - (\hat{\beta}_c+\eta)x_i)x_i \\
\Leftrightarrow n(\lambda+\delta) &= \sum_{i\in\mathcal{C}} \psi(y_i - \hat{\beta}_c x_i)x_i - \eta\psi'(y_i-(\hat{\beta}_c+\nu)x_i)x_i^2 \\
\Leftrightarrow n(\lambda+\delta) &= n\lambda - \eta\sum_{i\in\mathcal{C}} \psi'(y_i-(\hat{\beta}_c+\nu)x_i)x_i^2 \\
\Leftrightarrow \eta &= -\frac{\delta}{\frac{1}{n}\sum_{i\in\mathcal{C}}\psi'(y_i-(\hat{\beta}_c+\nu)x_i)x_i^2},
\end{aligned}
$$

where $\nu \in (0,\eta)$. Therefore, we have $\tilde{\beta}_c = \hat{\beta}_c - \frac{\delta}{S_c}$ and if δ is small enough, $S_c > 0$. Similarly, $\tilde{\beta}_\star = \hat{\beta}_\star - \frac{\delta}{S_\star}$ with $S_\star > 0$.

Next we need to show that $\mathcal{O}(\tilde{\beta}_c;\lambda+\delta) < \mathcal{O}(\tilde{\beta}_\star;\lambda+\delta)$ A Taylor series expansion of $\mathcal{O}(\hat{\beta}_c+\delta/S_c;\lambda+\delta)$ gives

$$
\mathcal{O}(\hat{\beta}_c+\delta/S_c;\lambda+\delta) = \mathcal{O}(\hat{\beta}_c;\lambda) + \delta\hat{\beta}_c - \frac{\delta^2}{S_c} + \frac{2\delta^2}{S_c^2}\tilde{S}_c,
$$

where $\tilde{S}_c = \sum_{i\in\mathcal{C}}\psi'(y_i - x_i\xi)x_i^2 > 0$ and $\xi \in (0,\delta/S_c)$. Applying a similar expansion for $\mathcal{O}(\hat{\beta}_\star+\delta/S_\star;\lambda+\delta)$ and noting that $\mathcal{O}(\hat{\beta}_c;\lambda) = \mathcal{O}(\hat{\beta}_\star;\lambda)$ we get

$$
\begin{aligned}
\mathcal{O}(\hat{\beta}_\star+\frac{\delta}{S_\star};\lambda+\delta) - \mathcal{O}(\hat{\beta}_c+\frac{\delta}{S_c};\lambda+\delta) &= \delta\hat{\beta}_\star - \frac{\delta^2}{S_\star} + \frac{2\delta^2}{S_\star^2}\tilde{S}_\star - \delta\hat{\beta}_c + \frac{\delta^2}{S_c} - \frac{2\delta^2}{S_c^2}\tilde{S}_c \\
&= \delta(\hat{\beta}_\star - \hat{\beta}_c) + \delta^2\left[\frac{1}{S_c} - \frac{1}{S_\star} + \frac{2\tilde{S}_\star}{S_\star^2} - \frac{2\tilde{S}_c}{S_c^2}\right].
\end{aligned} \tag{S1}
$$

Since ρ is quadratic around 0 and bounded, and because less than bn observations have $x_i = 0$, $S_c, S_\star, \tilde{S}_c, \tilde{S}_\star$ are all bounded and $S_c, S_\star$ are greater than 0 for $|\delta|$ small enough. Therefore, $0 \leq \left|\frac{1}{S_c} - \frac{1}{S_\star} + \frac{2\tilde{S}_\star}{S_\star^2} - \frac{2\tilde{S}_c}{S_c^2}\right| < \infty$. In turn, for $|\delta|$ small enough, (S1) is strictly greater than 0 for positive δ and strictly less than 0 for negative δ.

□

S1.1.1 Example scenario

Here we consider a simple case where the conditions required for Proposition 1 hold with high probability. Consider a bounded robust loss function ρ such that $\rho(x) = \rho_\infty$ for all $|x| > c_1$ and ρ is quadratic for $|x| \leq c_1$. Examples of such loss functions are Hampel's loss (Hampel et al. 1986) or the GGW and LQQ loss functions (Koller and Stahel 2011)). Other loss functions which are at least approximately quadratic in an open neighborhood of 0 would have similar behavior.

When we have n independent realizations from the simple model

$$y_i = \begin{cases} x_i\beta_c + \gamma_{c,i} & i \in \mathscr{C} \\ x_i\beta_\star + \gamma_{\star,i} & i \notin \mathscr{C}, \end{cases}$$

where $\beta_c > 0$, $\beta_\star - \beta_c$ large enough (see below), x_i i.i.d. $N(0,1)$, $\gamma_{c,i}$ and $\gamma_{\star,i}$ i.i.d. $N(0,\sigma_c^2)$ and $N(0,\sigma_\star^2)$, respectively, and $\mathscr{C} \subset \{1,\ldots,n\}$ such that $|\mathscr{C}| = bn < n/2$. For simplicity we assume that bn and hence $(1-b)n$ are integer. In this setting we can choose the parameters such that the conditions on $\hat{\beta}_c$ and $\hat{\beta}_\star$ for Proposition 1 are satisfied with arbitrarily high probability.

In the following we will consider a "bad" minimum $\hat{\beta}_c \in \mathcal{B}_c = [\beta_c - \frac{\lambda n}{(bn-2)} - \delta, \beta_c - \frac{\lambda n}{(bn-2)} + \delta]$ and a "good" minimum $\hat{\beta}_\star \in \mathcal{B}_\star = [\beta_\star - \frac{\lambda n}{(n(1-b)-2)} - \delta, \beta_\star - \frac{\lambda n}{(n(1-b)-2)} + \delta]$. We assume that β_c and $\beta_\star$ are far enough apart such that the objective function $\mathcal{O}(\hat{\beta}_c;\lambda)$ depends only on observations $i \in \mathscr{C}$ and $\mathcal{O}(\hat{\beta}_\star;\lambda)$ depends only on observations $i \notin \mathscr{C}$ with probability at least $1-\kappa$. In other words, the residuals for observations in $\mathcal{C}$ are within the quadratic part of the loss function for any $\hat{\beta}_c$ and in the bounded region for any $\hat{\beta}_\star$, and vice versa for observations not in $\mathcal{C}$.

Proof. Since we assume that the robust loss function ρ is bounded, we need to show that for all $\hat{\beta}_c \in \mathcal{B}_c$ all residuals for observations in $\rfloor$ are less than c_2 in absolute value with high probability. Writing $r_{c,i} = y_i - \hat{\beta}_c x_i$ we have that $r_{c,i} = \gamma_{c,i} + \frac{\lambda n}{(bn-2)} x_i + \delta_c$ for all $i \in \mathcal{C}$ and hence

$$\begin{aligned}
\mathbb{P}\left\{|r_{c,i}| < c_2\right\} &= \mathbb{P}\left\{|\gamma_{ci} + \frac{\lambda n}{(bn-2)} x_i + \delta_c| < c_2\right\} \\
&= 1 - 2\Phi\left(-\frac{c_2}{\sqrt{\sigma_c^2 + \frac{\lambda^2 n^2}{(bn-2)^2} + \delta_c^2}}\right) \\
&= p_{c,\mathcal{C}} > 0.
\end{aligned}$$

Similarly, for $i \notin \mathcal{C}$,

$$\begin{aligned}
\mathbb{P}\left\{|r_{c,i}| > c_1\right\} &= \mathbb{P}\left\{|\gamma_{ci} + (\beta_\star - \beta_c) x_i + \frac{\lambda n}{(bn-2)} x_i + \delta_c| > c_1\right\} \\
&= 2\Phi\left(-\frac{c_1}{\sqrt{\left(\sigma_c^2 + \frac{\lambda^2 n^2}{(bn-2)^2} + \delta_c^2\right) + (\beta_\star - \beta_c)^2}}\right) \\
&= p_{c,\mathcal{C}^c} \gg 1 - p_{c,\mathcal{C}}.
\end{aligned}$$

The probability $p_{c,\mathcal{C}^c}$ can be made arbitrarily large by moving β_c and $\beta_\star$ arbitrarily far apart. Since r_i are i.i.d., $\mathbb{P}\left\{\forall i \in \mathcal{C} \colon |r_{c,i}| \leq c_2 \wedge \forall i \notin \mathcal{C} \colon |r_{c,i}| > c_1\right\} = p_{c,\mathcal{C}}^{bn} p_{c,\mathcal{C}^c}^{(1-b)n} \gg 0$.

The same calculations can be done for $\hat{\beta}_\star \in \mathcal{B}_\star$, and hence the objective function value at $\hat{\beta}_\star$ and $\hat{\beta}_c$ do not depend on the same observations with arbitrarily high probability. □

Conditioned on the partitioning of the observations from above, the objective function has at least one minimum in each of $\mathcal{B}_c$ and $\mathcal{B}_\star$ with probability at least $1 - \kappa$, i.e.,

$$\mathbb{P}\left\{\exists(\hat{\beta}_c, \hat{\beta}_\star) \in \mathcal{B}_c \otimes \mathcal{B}_\star \colon \mathcal{O}'(\hat{\beta}_c; \lambda) = \mathcal{O}'(\hat{\beta}_\star; \lambda) = 0\right\} > 1 - \kappa. \tag{S2}$$

Proof. Consider $\hat{\beta}_c = \beta_c - \frac{\lambda n}{(bn-2)} + \Delta$. For $\hat{\beta}_c$ to be a minimum, the derivative of the penalized loss must be 0. As we assume that $\hat{\beta}_c > 0$, this is equivalent to:

$$n\lambda = \sum_{i \in \mathcal{C}} \psi \left(y_i - \beta_c x_i - \frac{\lambda n}{(bn-2)} x_i + \Delta x_i \right) x_i. \tag{S3}$$

Since ρ is quadratic at 0 ψ is linear and hence, for σ_c^2 and λ small enough, we can re-write (S3) as

$$\begin{aligned} n\lambda &= \sum_{i \in \mathcal{C}} \gamma_i x_i + \frac{\lambda n}{(bn-2)} \sum_{i \in \mathcal{C}} x_i^2 + \Delta \sum_{i \in \mathcal{C}} x_i^2 \\ \Leftrightarrow \Delta &= \frac{n\lambda(-1 + \frac{1}{(bn-2)} \sum_{i \in \mathcal{C}} x_i^2) + \sum_{i \in \mathcal{C}} \gamma_i x_i}{\sum_{i \in \mathcal{C}} x_i^2}. \end{aligned}$$

Therefore, $\mathbb{E}[\Delta] = 0$ and $\mathrm{Var}[\Delta] \leq \frac{1}{(bn-2)} \left[\sigma_c^2 + \frac{2\lambda^2 n^2}{(bn-4)(bn-2)} \right]$. Similar calculations can be carried out for $\hat{\beta}_\star$. For any given $\epsilon, \kappa > 0$ we can therefore find suitable σ_c^2, $\sigma_\star^2$ and n to satisfy (S2). □

Furthermore, there exists a sequence λ_n such that the objective function values at $\hat{\beta}_c$ and $\hat{\beta}_\star$ are within an ϵ neighborhood with arbitrarily high probability $1 - \kappa$, i.e.,

$$\lim_{n \to \infty} \mathbb{P} \left\{ \mathcal{O}(\hat{\beta}_\star; \lambda_n) = \mathcal{O}(\hat{\beta}_c; \lambda_n) \right\} > 1 - \kappa. \tag{S4}$$

Proof. Define $D = \mathcal{O}(\hat{\beta}_\star; \lambda_n) - \mathcal{O}(\hat{\beta}_c; \lambda_n)$. Since the loss function is quadratic in a neighborhood around 0, we can write D as

$$
\begin{aligned}
D = & \rho_\infty(b-(1-b))+ \\
& \frac{1}{n}\left(\sum_{i\notin\mathcal{C}}\gamma_i^2 - \sum_{i\in\mathcal{C}}\gamma_i^2\right)+ \\
& \lambda_n^2\left(\frac{n}{(n(1-b)-2)^2}\sum_{i\notin\mathcal{C}}x_i^2 - \frac{n}{(nb-2)^2}\sum_{i\in\mathcal{C}}x_i^2\right)+ \\
& 2\lambda_n\left(\frac{1}{(n(1-b)-2)}\sum_{i\notin\mathcal{C}}\gamma_i x_i - \frac{1}{(nb-2)}\sum_{i\in\mathcal{C}}\gamma_i x_i\right)+ \\
& 2\lambda_n\left(\frac{\Delta_\star}{(n(1-b)-2)}\sum_{i\notin\mathcal{C}}x_i - \frac{\Delta_c}{(nb-2)}\sum_{i\in\mathcal{C}}x_i\right)+ \\
& \frac{1}{n}\left(\Delta_\star^2\sum_{i\notin\mathcal{C}}x^2 - \Delta_c^2\sum_{i\in\mathcal{C}}x_i^2\right)+ \\
& \frac{2}{n}\left(\Delta_\star\sum_{i\notin\mathcal{C}}\gamma_i x_i - \Delta_c\sum_{i\in\mathcal{C}}\gamma_i x_i\right)+ \\
& \lambda_n^2\left(-\frac{n}{n(1-b)-2}+\frac{n}{nb-2}\right) \\
& \lambda_n\left(\beta_\star - \beta_c + \Delta_\star - \Delta_c\right).
\end{aligned}
$$

Setting the expectation of D to 0 yields

$$
\begin{aligned}
0 = \mathbb{E}[D] = &\rho_\infty(2b-1)+ \\
&(1-b)\sigma_\star^2 - b\sigma_c^2+ \\
&\lambda_n^2\left(\frac{n^2(1-b)}{(n(1-b)-2)^2} - \frac{n^2 b}{(nb-2)^2}\right)+ \\
&\Delta_\star^2(1-b) - \Delta_c^2\sigma_c^2 b+ \\
&\lambda_n^2\left(-\frac{n}{(n(1-b)-2)} + \frac{n}{(nb-2)}\right) \\
&\lambda_n\left(\beta_\star - \beta_c + \Delta_\star - \Delta_c\right) \\
=&\lambda_n^2\left[\frac{n^2(1-b)}{(n(1-b)-2)^2} - \frac{n}{(n(1-b)-2)} - \frac{n^2 b}{(nb-2)^2} + \frac{n}{(nb-2)}\right]+ \\
&\lambda_n\left(\beta_\star - \beta_c + \Delta_\star - \Delta_c\right)+ \\
&\rho_\infty(2b-1) + (1-b)\sigma_\star^2 - b\sigma_c^2.
\end{aligned}
$$

Setting $A_n = \left[\frac{n^2(1-b)}{(n(1-b)-2)^2} - \frac{n}{(n(1-b)-2)} - \frac{n^2 b}{(nb-2)^2} + \frac{n}{(nb-2)}\right]$, $B = \beta_\star - \beta_c + \Delta_\star - \Delta_c$ and $C = \rho(\infty)(b-(1-b)) + (1-b)\sigma_\star^2 - b\sigma_c^2$, we can see that $A_n < 0$ with $\lim_{n\to\infty} A_n = 0$ and $B > 0$. Further, we can choose σ_c and $\sigma_\star$ such that $C < 0$. Now if $\beta_\star - \beta_c = o((\sigma_\star^2 - \sigma_c^2)/n)$ then $B^2 - 4AC > 0$ and hence there exists a $\lambda_n > 0$ such that $\mathbb{E}[D] = 0$. Specifically,

$$
\lim_{n\to\infty} \lambda_n = \frac{b\sigma_c^2 - (1-b)\sigma_\star^2 + \rho_\infty(1-2b)}{\beta_\star - \beta_c + \Delta_\star - \Delta_c}.
$$

Moreover, $\mathrm{Var}[D] = \frac{\sqrt{2}b}{n}\left(\sigma_c^2 + C_{cn}^2\right) + \frac{\sqrt{2}(1-b)}{n}\left(\sigma_\star^2 + C_{\star n}^2\right)$ with $C_{cn} = \frac{\lambda_n n}{(nb-2)} + \Delta_c$ and $C_{\star n} = \frac{\lambda_n n}{(n(1-b)-2)} + \Delta_\star$ which goes to 0 as n goes to infinity. Therefore, $\lim_{n\to\infty} \mathbb{P}\left\{|D| > \epsilon\right\} = 0$. □

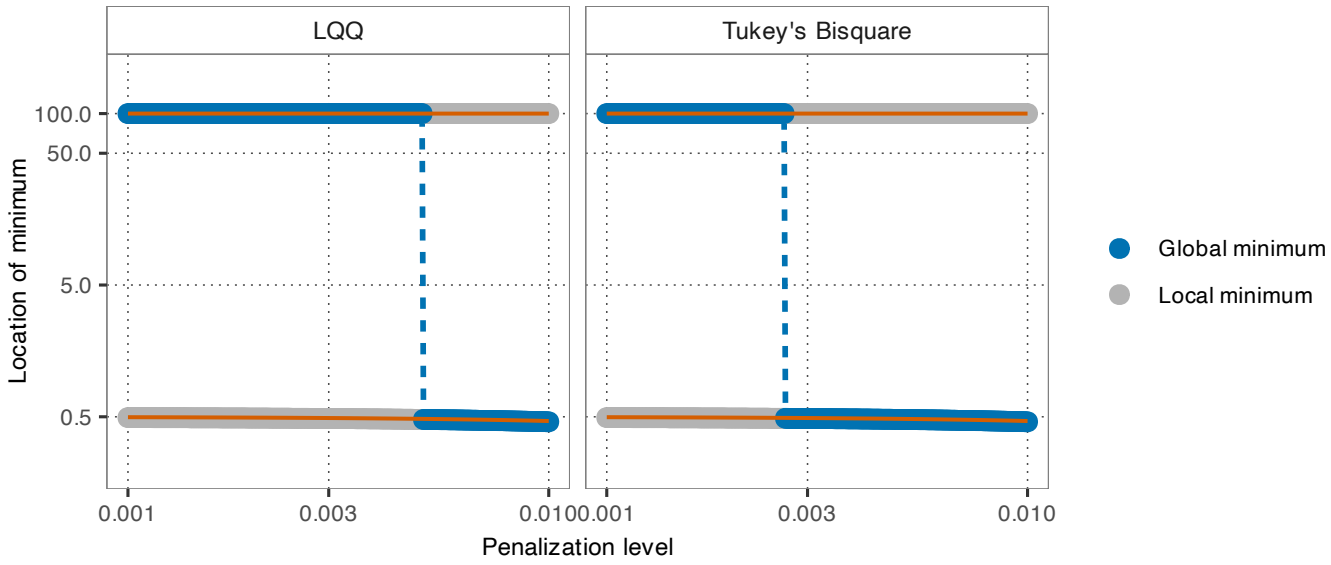

Figure S1: Demonstration of a non-smooth regularization path for a penalized M-estimator of regression in a simulation following the example scenario from Section S1.1.1 (with $\sigma_c = 0.01$, $\sigma_\star = 0.1$, $\beta_c = 0.5$, $\beta_\star = 100$, $b = 0.3$ and $n = 100$). We show the location of local minima when using the LQQ ρ function (left panel) and Tukey's bisquare ρ function (right panel). Gray dots represent local minima and blue dots indicate the global minimum. The orange lines depict the expected value of the minima at $\hat{\beta}_c = \beta_c - \frac{\lambda n}{nb-2}$ and $\hat{\beta}_\star = \beta_\star - \frac{\lambda n}{n(1-b)-2}$.

S2 Additional Empirical Studies

S2.1 Gene Pathway Recovery Analysis

This application is a gene pathway recovery analysis using data from Pfister et al. (2021) and the PENSE estimator (Kepplinger 2023). The data set contains preprocessed protein expression levels from 340 genes from seven different pathways for 315 subjects. Mimicking the original analysis (Pfister et al. 2021) we define as response the average expression of proteins on the *Cholesterol Biosynthesis* pathway. We further add Laplace-distributed noise to achieve a signal-to-noise ratio (SNR) of 1. We split the data set into a training data set comprising 165 randomly selected subjects and a test data set with the remaining 150 subjects. We then contaminate the training data set by replacing the response for 25 subjects (15%) with the average expression of proteins on the *Ribosome* pathway, again adding Laplace noise with a SNR of 1. The PENSE estimator is tuned to a breakdown

point of 25% and uses an EN penalty with $\alpha = 0.5$. For RIS-CV, we retain up to 50 local solutions. For both N-CV and RIS-CV, we utilize only the solutions on the complete training data as starting points for the CV folds, demonstrating that this computational shortcut can still yield reliable results.

In contrast to the CAV application, here we can evaluate the out-of-sample prediction performance on an independent test set. In Figure S2 we see both the CV estimated prediction errors and the out-of-sample MAPE evaluated on the independent test set. In this application, N-CV is not as badly affected by local optima as in the previous example, but there is nevertheless a noticeable change in slope around λ-index$= 13$ for N-CV. RIS-CV, on the other hand, yields a somewhat smoother CV curve and the 1-SE rule yields a minimally better prediction accuracy (0.196 vs. 0.194).

The regularization path in Figure S3 also shows the advantages of selecting a solution other than the "global" minimum in RIS-CV. N-CV exhibits a substantial discontinuity at λ-index$= 48$.

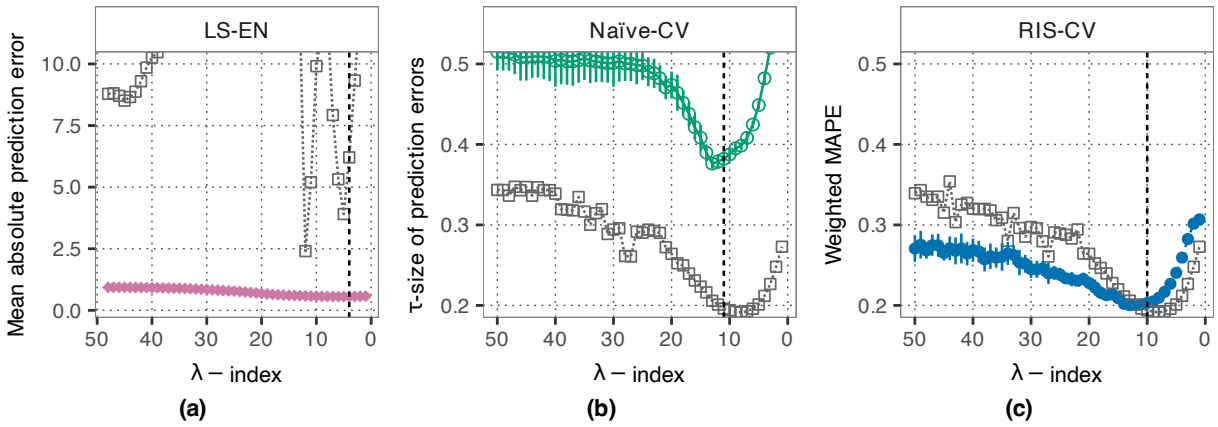

Figure S2: CV curves from ten replications of 10-fold CV in the gene pathway recovery analysis for the non-robust LS-EN estimator (left, magenta), the PENSE estimator with N-CV (middle, green) and the PENSE estimator with RIS-CV (right, blue). The gray curves show the out-of-sample (OOS) mean absolute error for the selected solutions.

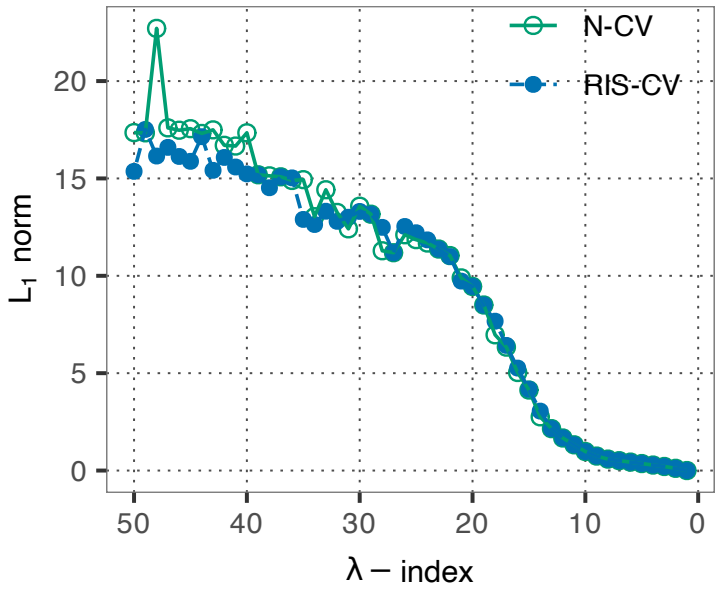

Figure S3: Regularization paths (L_1 norm) of the solutions selected by N-CV (green) and RIS-CV (blue).

S2.2 Determinants of Plasma Beta-Carotene Levels

In this application, we want to uncover determinants of plasma beta-carotene levels from data publicly available at http://lib.stat.cmu.edu/datasets/Plasma_Retinol (Nierenberg et al. 1989), using the PENSE estimator. We dummy-code the data and build a model with all available covariates and their interaction with the subject's sex (binary male/female). This leads to a total of 22 predictors for 315 subjects (42 male, 273 female). We randomly select a training data set of size $N = 100$, stratified among male/female such that 30 subjects in the training data are male and 70 are female. We deliberately oversample male subjects to the training data to ensure sufficient variation in all sex-dependent interaction terms.

Figure S4 indicates that both 10-fold RIS-CV and N-CV lead to smooth CV curves. However, the penalization level/solution selected by the 1-SE rule with the RIS-CV curve leads to a smaller out-of-sample (OOS) error than the solution selected with N-CV. Moreover, it is obvious that the non-robust least-squares EN estimator achieves substantially worse prediction accuracy than the robust estimator. In this application, RIS-CV always

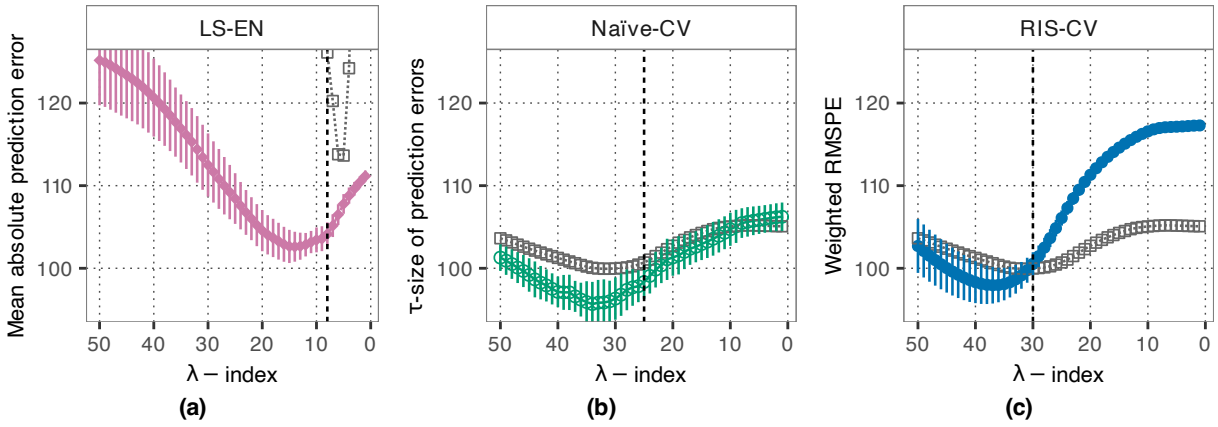

Figure S4: CV curves from ten replications of 10-fold CV for the analysis of plasma beta-carotene levels for the non-robust LS-EN estimator (left, magenta), the PENSE estimator with N-CV (middle, green) and the PENSE estimator with RIS-CV (right, blue). The gray curves show the out-of-sample (OOS) mean absolute error for the selected solutions.

selects the first solution and hence the regularization path is identical between N-CV and RIS-CV.

S3 Additional Details About the Simulation Study

S3.1 Details About the Simulation Settings

The contamination data generating process is defined as follows. For each contamination signal we first randomly select $\lfloor \log_2(p) \rfloor$ covariates (excluding the first s covariates), denoted by $\mathcal{J}^* \subset \{s+1, \ldots, p\}$. Then, for three different values of $u_1 = -1.5, u_2 = -1, u_3 = -0.5$ and the respective contamination indices $\mathcal{C}_1 = \{1, \ldots, 0.1n\}$, $\mathcal{C}_2 = \{0.1n+1, \ldots, 0.2n\}$, $\mathcal{C}_3 = \{0.2n+1, \ldots, 0.3n\}$, in observations $i \in \mathcal{C}_k$ the covariates and responses are replaced according to the following DGP:

$$x_{ij}^* = \begin{cases} x_{ij} & j \notin \mathcal{J}^* \\ k_l x_{ij} & j \in \mathcal{J}^* \end{cases}, \qquad \beta_{0j}^* = \begin{cases} 0 & j \notin \mathcal{J}^* \\ k_v & j \in \mathcal{J}^* \end{cases}, \qquad y_i^* = \mathbf{x}_i^{*\intercal} \boldsymbol{\beta}_0^* + \varepsilon_i^*.$$

The constant k_l is chosen such that $\mathbf{x}_i^*$ is at least twice as far from the center (in terms of the Mahalanobis distance) than all the other non-contaminated observations. The error term ε_i^* is Gaussian with variance such that $\boldsymbol{\beta}_0^*$ achieves a SNR of 10.

S3.2 Additional Simulation Results

Here we present additional results from the simulation study.

Figure S5 shows the relative prediction error achieved by N-CV using the two-step approximation for the CV fits as proposed in Khan et al. (2010). While this strategy leads to some improvement over N-CV, the gains are most often very small.

Figures S6 and S7 show the CV curves and the corresponding true prediction errors for simulation runs from two different settings. The number in each box indicates the index of the chosen solution (1 being the global minimum). In the setting with Laplace errors, both RIS-CV and N-CV lead to CV curves with the typical "U" shape, but it is clearly advantageous to consider more than just the global minimum as evident from the true prediction errors. The solutions selected by RIS-CV (at the minimum and using the 1-SE rule) are not the global minima at the respective penalization level. For the Gaussian case, N-CV fails to capture the actual shape of the prediction error, and the global minimum has a clear discontinuity between $\lambda = 1$ to $\lambda = 3$. RIS-CV, on the other hand, again matches the "U" shape of the true prediction error and, except at a single penalization level, identify the local minima that lead to a smoother prediction error curve.

We further compute measures of smoothness of the CV curves and regularization paths across the 1,800 simulation runs, namely (1) the maximum absolute difference between two consecutive penalization levels and (2) the sum of the squared second-order differences

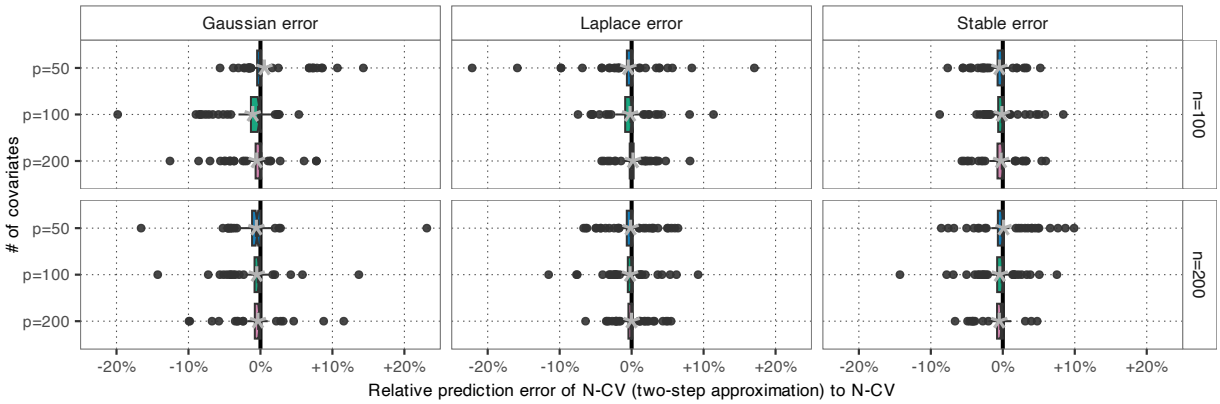

Figure S5: Prediction accuracy of the PENSE solution selected by N-CV with the two-step approximation relative to the prediction accuracy of the PENSE solution selected by N-CV re-computing and fully iterating the ENPY starting points. Prediction accuracy is measured by the RMSPE for normal errors and by the MAPE for other error distributions, computed on an independent test set of size 10,000. The star in each boxplot denotes the arithmetic mean.

(i.e., the "wiggliness" of the curve), i.e.,

$$S_1 = \max_k \left| \hat{E}(\lambda_k) - \hat{E}(\lambda_{k-1}) \right|, \tag{S5}$$

$$S_2 = \sum_k \left(\hat{E}(\lambda_{k-1}) - 2\hat{E}(\lambda_k) - \hat{E}(\lambda_{k+1}) \right)^2. \tag{S6}$$

In Figure S8, we plot these summaries across all settings from the simulation study in Section 5in the revised manuscript. In the top panel, we can see that the RIS-CV curve is overall smoother than the N-CV curve according to these measures. We want to note, however, that N-CV sometimes leads to fairly flat CV curves, which are smooth but not very informative as to a good selection of the penalization level. In the bottom panel, there does not seem to be a large difference in the smoothness of the L_1 norm of the coefficients along the regularization path. This suggests that in the simulations, the local minima selected by RIS-CV are similar to the global minima. Combining this insight with the prediction accuracies in Figure 4it seems that RIS-CV is able to better quantify the prediction error of these minima by matching them with a more appropriate surrogate CV solution.

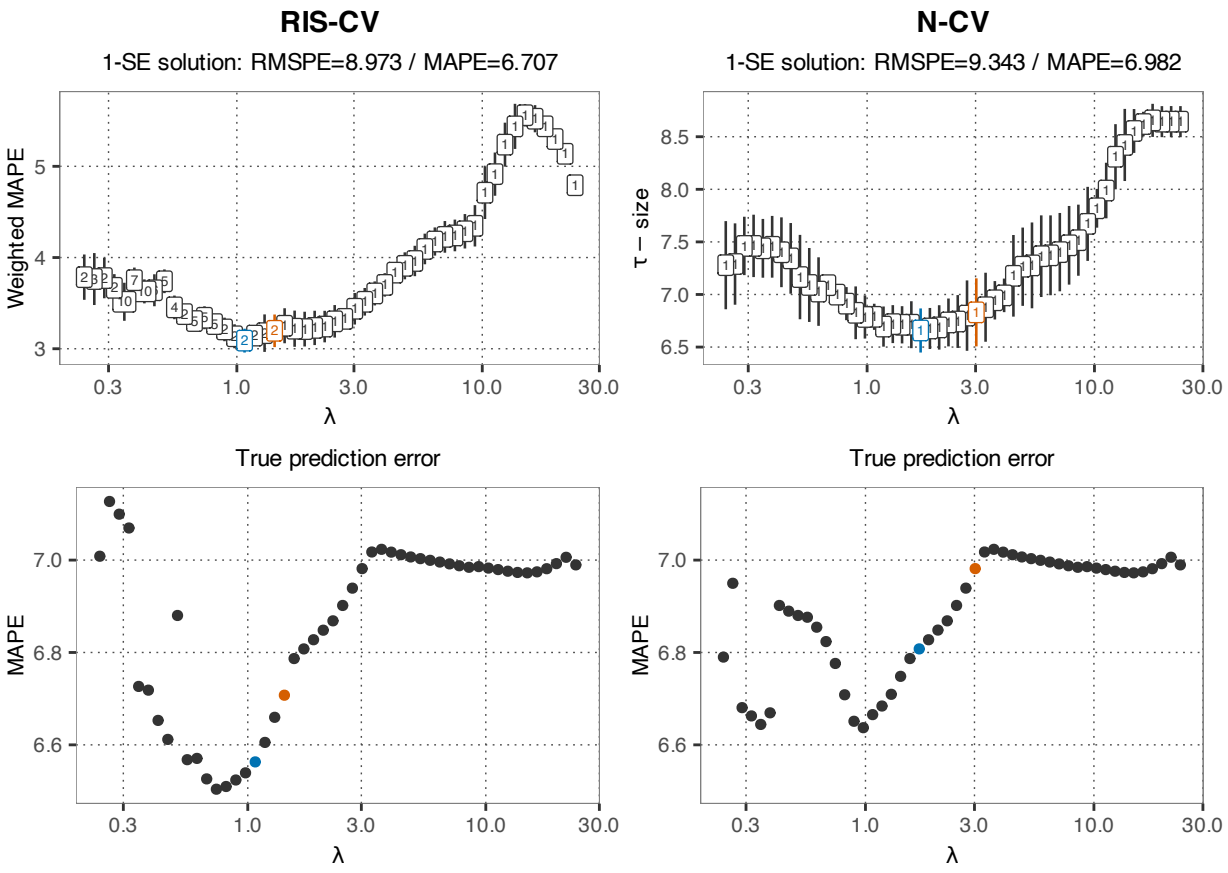

Figure S6: Examples of CV curves estimated by RIS-CV (left) and N-CV (right) for a setting with Laplace errors, $p = 50$ and $n = 100$. The numbers in the boxes denote the index of the selected solution (1 is the global minimum). The blue solution indicates the minimum of the CV curve, whereas the orange solution is chosen by the 1-SE rule (3).

S3.3 Using only shared starts

We therefore compare the selection and achieved prediction accuracy of RIS-CV, where new starting points are computed on each CV fold, and *RIS-CV (warm)* where we use only the solutions from the fit to the complete training data, in the same simulation study as presented in the main manuscript (Section 5).

Figure S9 summarizes the differences across all settings in the simulation study. As expected, the computations with *RIS-CV (cold)* take substantially longer than with *RIS-CV (warm)*, on average more than 20 times longer (Figure S9a). However, we also observe in Figure S9b that RIS-CV leads to slightly better model selection. This better prediction accuracy may be due to RIS-CV sometimes choosing larger penalization levels and hence

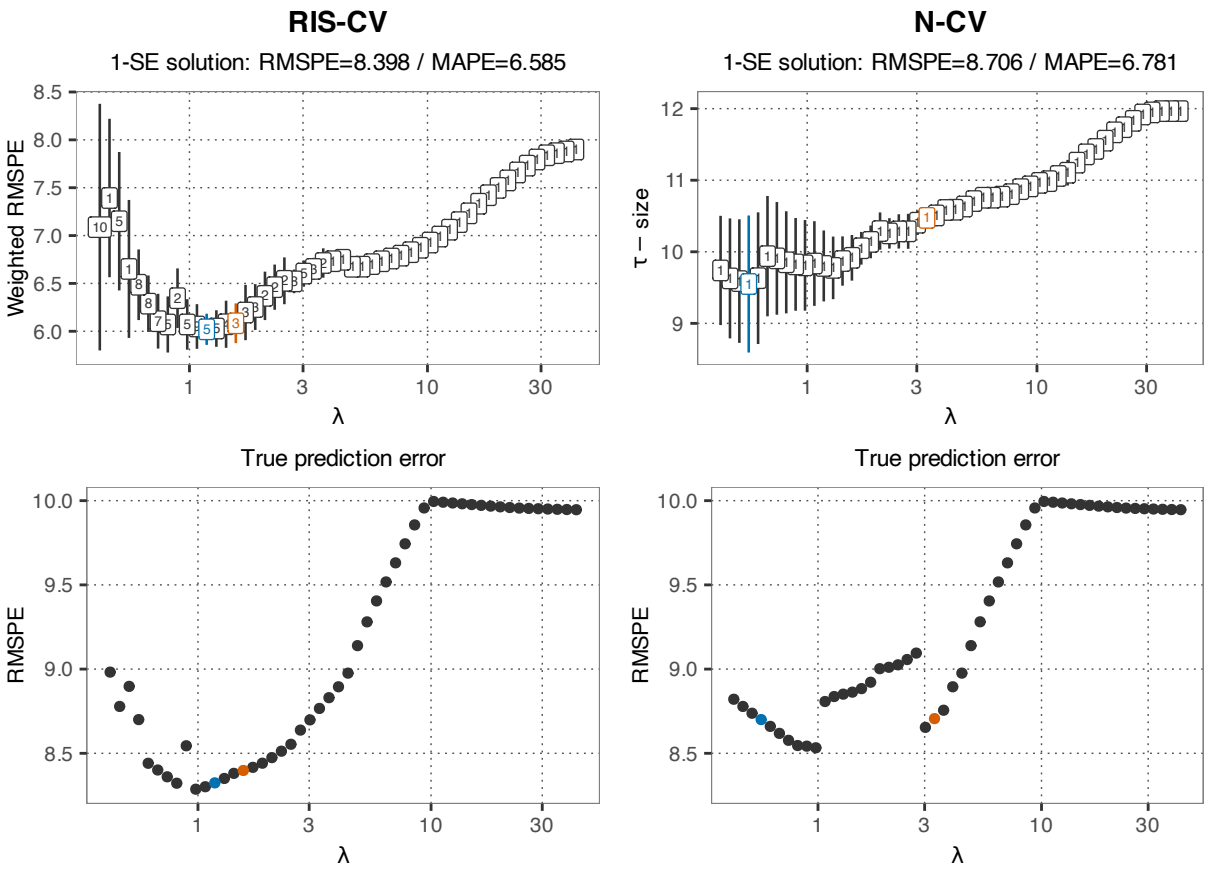

Figure S7: Examples of CV curves estimated by RIS-CV (left) and N-CV (right) for a setting with Gaussian errors, $p = 50$ and $n = 200$. The numbers in the boxes denote the index of the selected solution (1 is the global minimum). The blue solution indicates the minimum of the CV curve, whereas the orange solution is chosen by the 1-SE rule (3).

possibly reducing overfitting in these cases (Figure S9d). While the matched CV surrogates are overall more similar to the minima on the complete training data with *RIS-CV (warm)* (Figure S9c), the matched solutions are much more similar with RIS-CV at the "important" penalization levels (i.e., where the estimated prediction error attains its minimum and where it is within 1 standard error of that minimum). Therefore, the advantages of RIS-CV are possibly due to finding "better" surrogates by considering a larger number of starting points, but also because *RIS-CV (warm)* could induce some bias towards overfitting.

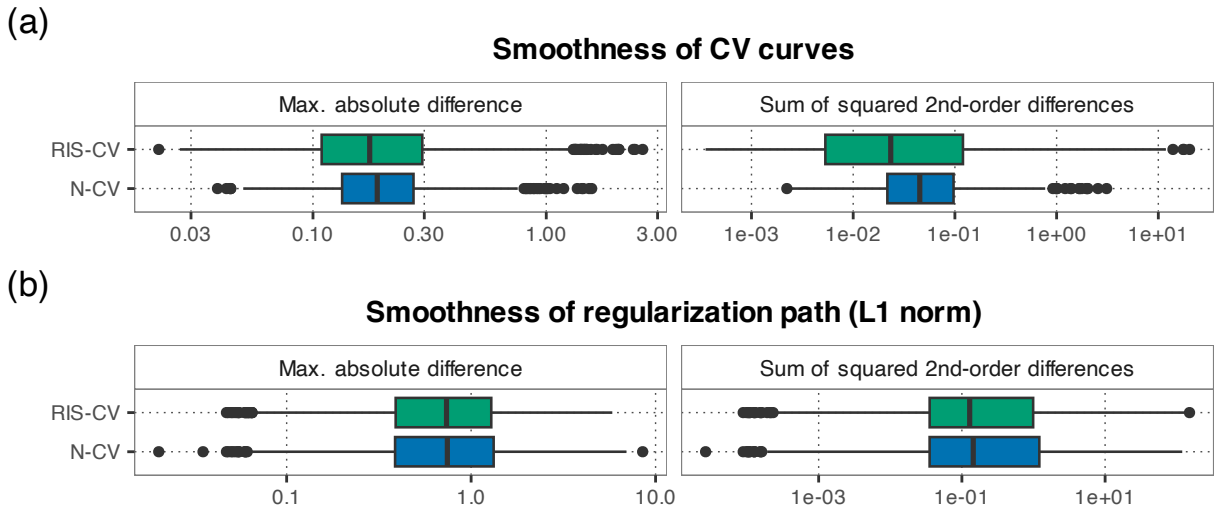

Figure S8: Measures of smoothness of the CV curves and the regularization path for RIS-CV (blue) and N-CV (green), respectively, across all 1,800 simulation runs from Section 5.The left panels show the maximum absolute difference (S5), whereas the right panel shows the sum of the squared second order differences (S6) (i.e., the overall "wigglyness").

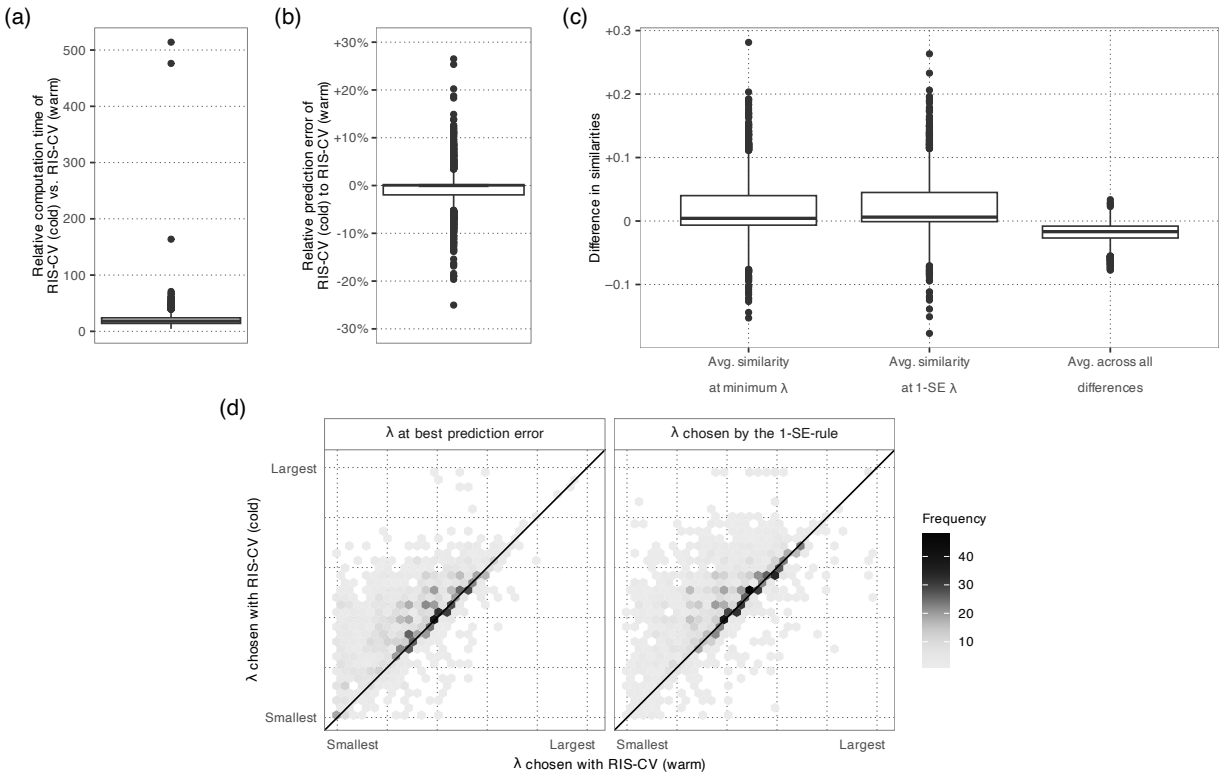

Figure S9: Comparisons of RIS-CV against RIS-CV (warm) across all settings in the simulation study. Boxplot (a) depicts the increase in computation time for RIS-CV relative to RIS-CV (warm), while (b) shows the overall improvement in prediction error attained by RIS-CV over RIS-CV (warm). The boxplots in (c) show the difference in the similarities of the matched minima: the difference in the average similarity between the chosen solution (at the minimum prediction error and using the 1-SE rule) and the average differences across all similarities for all minima on the complete training data. Sub-figure (d) shows the difference in the selected penalization level at the level with smallest estimated prediction error and the prediction error with 1 standard deviation of the minimum.

References

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Kepplinger, D. (2023). "Robust Variable Selection and Estimation via Adaptive Elastic Net S-Estimators for Linear Regression". *Computational Statistics & Data Analysis*, *183*, 107730. https://doi.org/10.1016/j.csda.2023.107730

Khan, J. A., Van Aelst, S., and Zamar, R. H. (2010). "Fast Robust Estimation of Prediction Error Based on Resampling". *Computational Statistics & Data Analysis*, *54*(12), 3121–3130. https://doi.org/10.1016/j.csda.2010.01.031

Koller, M., and Stahel, W. A. (2011). "Sharpening Wald-Type Inference in Robust Regression for Small Samples". *Computational Statistics & Data Analysis*, *55*(8), 2504–2515. https://doi.org/10.1016/j.csda.2011.02.014

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., Greenberg, E. R., and The Skin Cancer Prevention Study Group. (1989). "Determinants of Plasma Levels of Beta-Carotene and Retinol". *American Journal of Epidemiology*, *130*(3), 511–521. https://doi.org/10.1093/oxfordjournals.aje.a115365

Pfister, N., Williams, E. G., Peters, J., Aebersold, R., and Bühlmann, P. (2021). "Stabilizing Variable Selection and Regression". *The Annals of Applied Statistics*, *15*(3), 1220–1246. https://doi.org/10.1214/21-AOAS1487