A lava Attack on the Recovery of Sums of Dense and Sparse Signals

Victor Chernozhukov Christian Hansen Yuan Liao

INFORMS APS 2017

Introduction

- Sparse model :
  - many zeros and a few "large" components.
  - Lasso works well
- Dense model:
  - no large parameters and very many small non-zero parameters
  - Ridge works well

Motivation of this work: sparsity is restrictive in some cases:

- predictions
- nonparametric fitting
- Treatment effect inference with many controls.

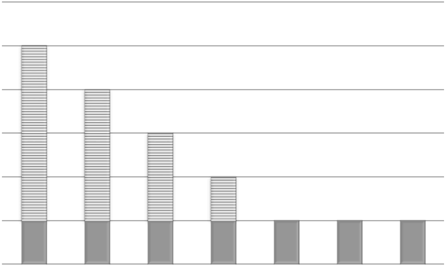
In these applications, variable selection is not a requirement.

A dense+sparse model

A basic assumption for non-sparse models:

$$\theta = \underbrace{\beta}_{\text{dense signal}} + \underbrace{\delta}_{\text{sparse signal}} .$$

Figure: dense+sparse decomposition



lava: a new technique for signal recovery

Let $\ell(\text{data}, \theta)$ be a loss function.

$$\widehat{\theta}_{\text{lava}} = \widehat{\beta} + \widehat{\delta},$$

where

$$(\widehat{\beta}, \widehat{\delta}) = \arg \min_{(\beta', \delta')' \in \mathbb{R}^{2p}} \Big\{ \ell(\text{data}, \beta + \delta) + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\delta\|_1 \Big\}.$$

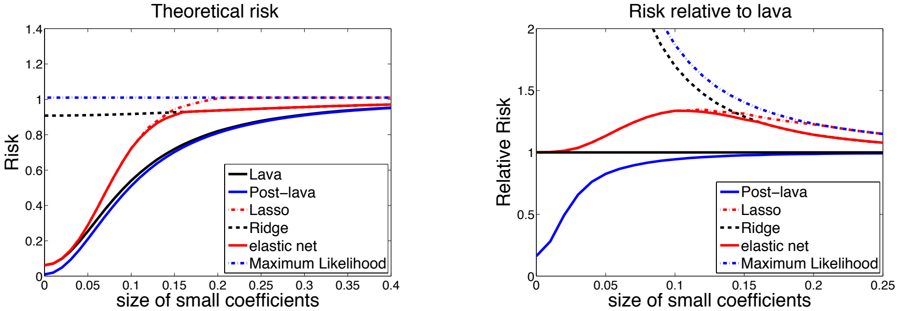
- ℓ_2-part captures dense signal; ℓ_1-part captures sparse signal.

Risk comparison in $Z \sim N(\theta, I)$

$$\theta = (3, q, ..., q)', \qquad q : \text{small coefficient}$$

$$\widehat{\theta} = \widehat{\beta} + \widehat{\delta}, \quad (\widehat{\beta}, \widehat{\delta}) = \arg\min \|Z - \beta - \delta\|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\delta\|_1$$

Figure: $\mathrm{E}\|\widehat{\theta}(Z) - \theta\|_2^2$, oracle tunings



one-dimensional case

Consider shrinkage estimation:

$$d(Z) = \arg\min_{\theta} (Z - \theta)^2 + P_\lambda(\theta)$$

We set

$$P_\lambda(\theta) = \lambda_2 |\beta|^2 + \lambda_1 |\delta|, \quad \theta = \beta + \delta$$

- To compare with related methods:
  - Lasso: $P_\lambda(\theta) = \lambda|\theta|$
  - elastic net: $P_\lambda(\theta) = \lambda_2|\theta|^2 + \lambda_1|\theta|$
  - Ridge: $P_\lambda(\theta) = \lambda|\theta|^2$

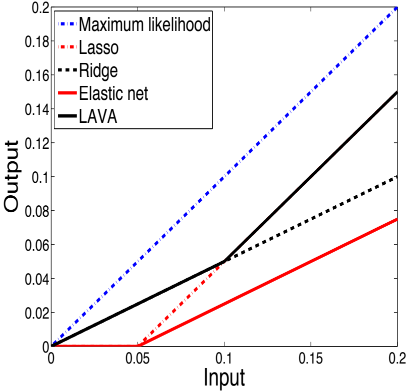
- Weighted average of the soft-thresholding and the data.

$$\begin{aligned} d_{\text{lava}}(Z) &= \widehat{\beta} + \widehat{\delta} \\ &= (1-k)Z + k(\text{soft th.}), \quad k = \frac{\lambda_2}{1+\lambda_2} \end{aligned}$$

By shrinking towards the data, robust to non-sparse signals.

- Does not produce sparse solutions.

Figure: Shrinkage functions



lava in the regression model

$$Y = X\theta_0 + U, \quad U \sim N(0, \sigma_u^2 I_n),$$

$$\begin{aligned} \widehat{\theta}_{\text{lava}} &= \widehat{\beta} + \widehat{\delta}, \\ (\widehat{\beta}, \widehat{\delta}) &= \arg\min_{\beta,\delta \in \mathbb{R}^p} \frac{1}{n}\|Y - X(\beta+\delta)\|_2^2 + \lambda_2\|\beta\|_2^2 + \lambda_1\|\delta\|_1. \end{aligned}$$

Computations

- If we knew δ, then ridge solution :

$$\widehat{\beta}(\delta) = (X'X + n\lambda_2 I_p)^{-1} X'(Y - X\delta).$$

- Substitute $\beta = \widehat{\beta}(\delta)$ into the objective function,

$$\widehat{\delta} = \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{n} \| Y - X(\widehat{\beta}(\delta) + \delta) \|_2^2 + \lambda_2 \| \widehat{\beta}(\delta) \|_2^2 + \lambda_1 \| \delta \|_1 \right\}.$$

- So lava is given by:

$$\widehat{\theta} = \widehat{\beta}(\widehat{\delta}) + \widehat{\delta}.$$

De-densify: another look at Lava

Theorem (A Key Characterization of the Profiled Lava Program)

Define ridge-projection matrices,

$$\mathsf{P}_{\lambda_2} = X(X'X + n\lambda_2 I_p)^{-1}X' \text{ and } \mathsf{K}_{\lambda_2} = I_n - \mathsf{P}_{\lambda_2},$$

and transformed data, $\widetilde{Y} = \mathsf{K}_{\lambda_2}^{1/2} Y$ *and* $\widetilde{X} = \mathsf{K}_{\lambda_2}^{1/2} X$. *Then*

$$\widehat{\delta} = \arg\min_{\delta \in \mathbb{R}^p} \left\{ \frac{1}{n} \| \widetilde{Y} - \widetilde{X}\delta \|_2^2 + \lambda_1 \|\delta\|_1 \right\}.$$

De-densify: another look at Lava

- In other words, "de-densify" first, then lasso

Step 1: Ridge-projection matrices,

$$\mathsf{P}_{\lambda_2} = X(X'X + n\lambda_2 I_p)^{-1}X' \text{ and } \mathsf{K}_{\lambda_2} = I_n - \mathsf{P}_{\lambda_2},$$

and transformed data, $\widetilde{Y} = \mathsf{K}_{\lambda_2}^{1/2} Y$ *and* $\widetilde{X} = \mathsf{K}_{\lambda_2}^{1/2} X$.

Step 2: Run lasso on $(\widetilde{Y}, \widetilde{X})$.

- Why are the signals for the "transformed data" sparse?

$$\tilde{Y} = \tilde{X}\delta + \tilde{U} + \underbrace{\mathsf{K}_{\lambda_2}^{1/2} X\beta_0}_{\text{projected off}}$$

- Taking the transformation $\mathsf{K}_{\lambda_2}^{1/2}$ removes the dense component.

Choices of tuning parameters

Data-driven choices

- min-SURE:
  Suppose $\widehat{R}(\widehat{\theta}_\lambda)$ is Stein's Unbiased Risk Estimator for method $\widehat{\theta}_\lambda$,

$$\arg\min_{\lambda} \widehat{R}(\widehat{\theta}_\lambda)$$

- K-fold cross validation.

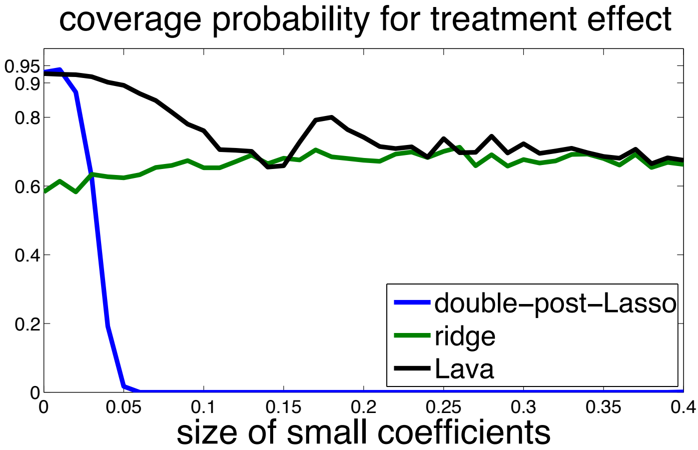
Inference with many controls

- Consider the model

$$\begin{aligned} y_i &= d_i\alpha + X_i'\theta + e_i \\ d_i &= X_i'\gamma + u_i \end{aligned}$$

  Belloni et al. (14) used double-post-selection.
- What if $\theta, \gamma =$dense $+$ sparse ?
- Obtain confidence intervals for α that is more robust to the signal
- Example: $\theta = \gamma = (3, q, ..., q)$; where q is the small coefficient.

Figure: Coverage probability: tuning chosen by 5-fold CV



Monte-Carlo

- $n = 100, p = 2n$.
- Gaussian regression,
$$\theta = (3, q, ..., q)',$$
- The tuning parameters are selected by numerically minimizing the SURE and 5-fold CV.
- Consider an independent design $X \sim N(0, I)$.
- Calculate averaged $\frac{1}{n}\|X\widehat{\theta} - X\theta_0\|_2^2$ from 100 replications.

Figure: Risk comparisons: tuning chosen by 5-fold CV

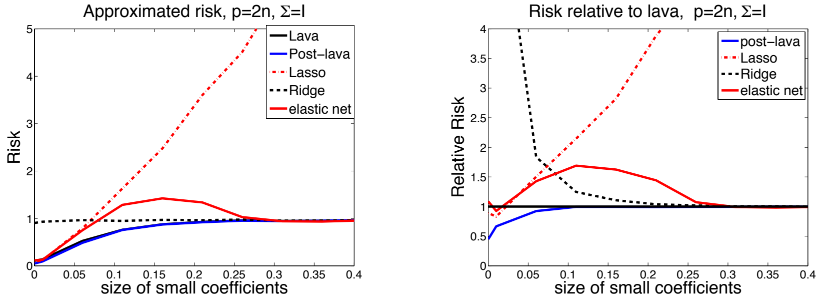
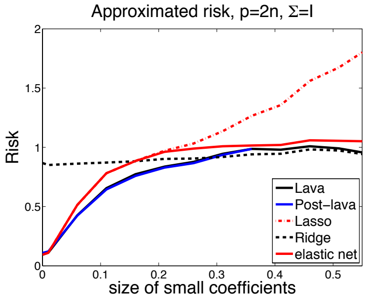


Figure: Risk comparisons: tuning chosen by min-SURE



Theorem (Deviation Bounds for Lava in Regression)

We have that with probability $1-\alpha-\epsilon$ *(note that* $\|\mathsf{K}_{\lambda_2}\| \leq 1$ *)*

$$\frac{1}{n}\|X\widehat{\theta}_{\text{lava}} - X\theta_0\|_2^2 \leq \frac{2}{n}\|\mathsf{K}_{\lambda_2}^{1/2}X(\widehat{\delta}-\delta_0)\|_2^2\|\mathsf{K}_{\lambda_2}\| + \frac{2}{n}\|\mathsf{D}_{\text{ridge}}(\lambda_2)\|_2^2$$

$$\leq \inf_{(\delta_0',\beta_0')'\in\mathbb{R}^{2p}:\delta_0+\beta_0=\theta_0}\left\{\Big(B_1(\delta_0)\vee B_2(\beta_0)\Big)\|\mathsf{K}_{\lambda_2}\| + \underbrace{B_3 + B_4(\beta_0)}_{\textit{bound of } \mathsf{D}_{\text{ridge}}(\lambda_2)}\right\},$$

$$B_1(\delta_0) = \frac{2^3\lambda_1^2}{\iota^2(c,\delta_0,\lambda_1,\lambda_2)} \leq \frac{2^5\sigma_u^2c^2\bar{V}_{\lambda_2}^2\log(2p/\alpha)}{n\iota^2(c,\delta_0,\lambda_1,\lambda_2)},$$

$$B_2(\beta_0) = \frac{2^5}{n}\|\mathsf{K}_{\lambda_2}^{1/2}X\beta_0\|_2^2 = 2^5\lambda_2\beta_0'S(S+\lambda_2 I)^{-1}\beta_0,$$

$$B_3 = \frac{2^2\sigma_u^2}{n}\left[\sqrt{\operatorname{tr}(\mathsf{P}_{\lambda_2}^2)} + \sqrt{2}\sqrt{\|\mathsf{P}_{\lambda_2}^2\|}\sqrt{\log(1/\epsilon)}\right]^2,$$

$$B_4(\beta_0) = \frac{2^2}{n}\|\mathsf{K}_{\lambda_2}X\beta_0\|_2^2 = 2^2\beta_0'V_{\lambda_2}\beta_0 \leq 2^3B_2(\beta_0)\|\mathsf{K}_{\lambda_2}\|.$$

Remarks

1. Does not require identification of (β_0, δ_0). "inf" finds the best split.
2. In dense models, lava works similarly to ridge.
3. In sparse models, lava works similarly to lasso.

Conclusions

- Lava is designed for "sparse+dense" models.
- Complements other approaches to structured sparsity: fused sparsity, matrix decomposition, etc.
- Extendable to more general M- and Z- estimations.