What a weighted subset of the data can fit

MANTIS DISCOVERY

MANTIS · MIT CSAIL

OCTOBER 2026

01

The question

A coreset is a small weighted subset of a dataset that should give the same model as the full data. For least squares, which fits can a fixed subset reach as its weights change? How well conditioned are they, and when does one subset keep working along the whole ridge path?

02

How we are approaching it

- Describe the reachable fits with a hyperplane arrangement.
- Set up convex reweighting problems that improve conditioning on a fixed support.
- Match moments on a Krylov subspace to keep the ridge path.

03

What we have so far

On a fixed support, the reachable weighted least-squares fits and their limits are the bounded faces of a hyperplane arrangement. With $d + 1$ rows they fill the convex hull of the d-row interpolants.

A subset keeps the entire ridge path exactly when it matches the response cross moment and the action of the covariance on a Krylov subspace.

04

Weighted fits with three rows

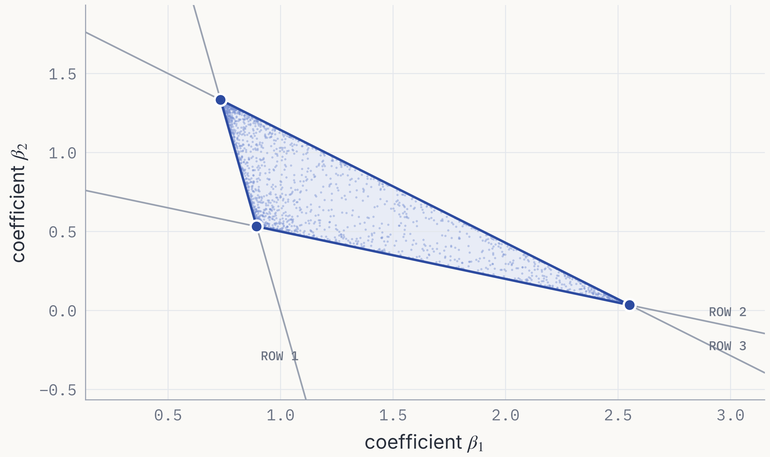


FIGURE 1. With two features and three rows, each row is a line of exact fits, and the triangle's corners are the fits through two rows. We solved 3,000 random positive weightings, and every weighted fit landed inside the triangle.

05

Ridge paths

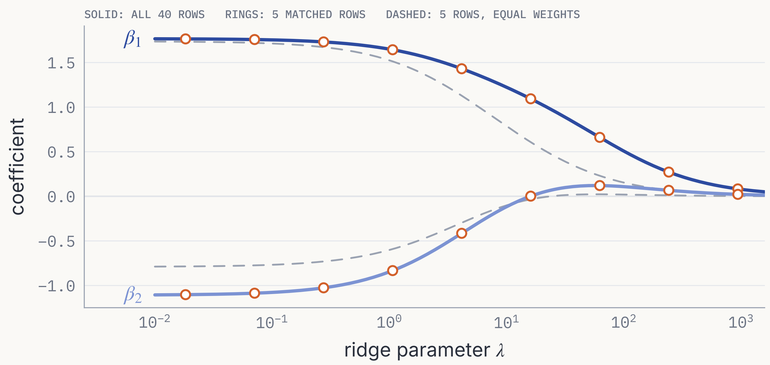
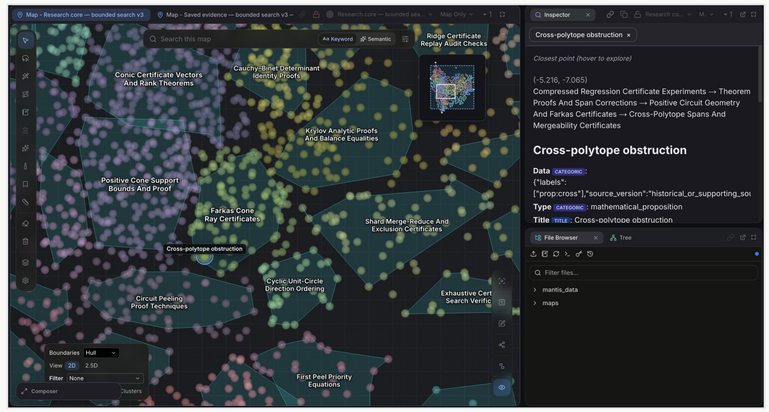


FIGURE 2. We computed ridge paths for 40 rows and two features. Five rows weighted to match the full data's covariance and cross moment (rings) follow the full path exactly, and five rows with equal weights (dashed) do not. In two dimensions the Krylov condition is this full match.

06

In Mantis



3,966 RECORDS · THE OPEN RECORD

Cross-polytope obstruction.

The vectors $\pm e_1, \ldots, \pm e_d$ positively span $\mathbb{R}^d$, yet every full-rank positive balance uses all $2d$ of them.

ALSO ON RECORD

Whitening needs a qualification, because it changes the minimum-norm selected-fit ratio from 3 to $25/6$.

New upper bounds for the scalar minimax risk give, for example, $\frac{8}{5} \le R_5(7) \le \frac{5}{3}$.

07

Open questions

- Reweighting improved every starting condition number in 16 configurations, by a median factor of 6.46, but usually made fits more sensitive to new labels. Can both be controlled at once?
- Where between $\frac{8}{5}$ and $\frac{5}{3}$ is $R_5(7)$?