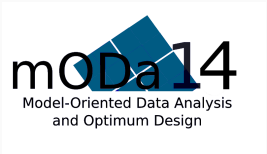
Optimal subsampling for prediction in presence of model misspecification



Carlos de la Calle-Arroyo[1], Laura Deldossi[2] and Chiara Tommasi[3]

16th June 2026

[1]University of Oviedo

[2]Università Cattolica del Sacro Cuore di Milano

[3]Università degli Studi di Milano

Background

The Subsampling Problem

Big Data with costly labels

In many modern applications predictor observations $\boldsymbol{x}_i \in \mathbb{R}^p$ are collected for $N \gg 1$ units, but obtaining responses y_i is **costly or time-consuming** (cloud queries, laboratory assays, human labelling...).

The natural strategy: select a **subsample** $\mathcal{S} \subset \{1, \ldots, N\}$ of size $n \ll N$, label those units, and fit a predictive model.

The key question

Which n units should be labelled to maximise prediction accuracy?

Existing methods (IBOSS, leverage-score sampling, Twinning...) rely on the **model being correctly specified.**

Model Misspecification

True data-generating process

In practice the linear model may be **misspecified**. The true model includes an unknown nonlinear term $h(\cdot)$:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{h} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

where $\boldsymbol{h} = [h(\boldsymbol{x}_1), \ldots, h(\boldsymbol{x}_N)]^\top$ with $h : \mathcal{X} \to \mathbb{R}$ **unknown**. At prediction points: $\boldsymbol{Y}_0 = \boldsymbol{X}_0\boldsymbol{\beta} + \boldsymbol{h}_0 + \boldsymbol{\varepsilon}_0$.

In our set up, we fit a **linear model**:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

Therefore, in the presence of model misspecification, since standard I-optimal subsampling ignores h, it can perform very poorly. As a precedent, Meng et al. (2021) proposed **LowCon** to address bias in estimation

Subsample and Regression Setup

OLS estimator on the subsample

The selection is encoded by a diagonal matrix $\boldsymbol{S}$ ($\delta_i = 1$ iff unit i is selected). The OLS estimator and predictions on the subsample are:

$$\hat{\beta}_S = \left(\boldsymbol{X}^\top \boldsymbol{S} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{S} \boldsymbol{y}, \qquad \hat{\boldsymbol{Y}}_0 = \boldsymbol{X}_0 \hat{\beta}_S$$

Performance is measured by the **Mean Squared Prediction Error**:

$$\mathrm{MSPE}(\boldsymbol{S}) = \mathbb{E}\left[\| \hat{\boldsymbol{Y}}_0 - \boldsymbol{Y}_0 \|^2 \mid \boldsymbol{X}, \boldsymbol{X}_0, \boldsymbol{S}\right]$$

MSPE under Misspecification

Theorem 1 — Exact MSPE

Let $\boldsymbol{M} = \boldsymbol{X}^\top \boldsymbol{S} \boldsymbol{X}$. The MSPE decomposes as:

$$\mathrm{MSPE}(\boldsymbol{S}) = \underbrace{\sigma^2 \left[m + \mathrm{tr}(\boldsymbol{X}_0 \boldsymbol{M}^{-1} \boldsymbol{X}_0^\top) \right]}_{\text{Variance } = \ \sigma^2(m + \Phi_I)} + \underbrace{\| \boldsymbol{X}_0 \boldsymbol{M}^{-1} \boldsymbol{X}^\top \boldsymbol{S} \boldsymbol{h} - \boldsymbol{h}_0 \|^2}_{\text{Bias}^2}$$

where $\Phi_I(\boldsymbol{S}) = \mathrm{tr}(\boldsymbol{X}_0 \boldsymbol{M}^{-1} \boldsymbol{X}_0^\top)$ is the **I-criterion**.

- **Variance**: minimised by I-optimal subsampling
- **Bias**: term due to $\boldsymbol{h}$, the unknown misspecification term

Bounding the Bias

First Bound: LowCon (bounded misspecification)

Assumption: bounded misspecification

Assume $|h(\boldsymbol{x})| \leq \alpha \|\boldsymbol{x}\|$ for all $\boldsymbol{x} \in \mathcal{X}$, so $\|\boldsymbol{h}\|^2 \leq \alpha^2 \operatorname{tr}(\boldsymbol{M})$.

Under this condition, a bias bound for the MSE of $\hat{\boldsymbol{\beta}}_S$ (Meng et al., 2021) involves the **condition number** $\kappa(\boldsymbol{M}) = \lambda_{\max}(\boldsymbol{M})/\lambda_{\min}(\boldsymbol{M})$. The analogous prediction bound is:

$$\mathrm{Bias}^2 \;\leq\; 2\alpha^2 \left[\lambda_{\max}(\boldsymbol{X}_0^\top \boldsymbol{X}_0) \cdot \frac{\operatorname{tr}(\boldsymbol{M})}{\lambda_{\min}(\boldsymbol{M})} + \operatorname{tr}(\boldsymbol{X}_0^\top \boldsymbol{X}_0) \right]$$

LowCon (Meng et al., 2021)

Minimise $\kappa(\boldsymbol{M})$ via orthogonal Latin hypercube-based sampling, which achieves low condition number

A New Bound — No Assumption on h

Theorem 2

For any h with $\|\boldsymbol{h}\|^2 < \infty$:

$$\mathrm{Bias}^2 \;\leq\; 2\|\boldsymbol{h}\|^2 \underbrace{\lambda_{\max}\left(\boldsymbol{X}_0 \boldsymbol{M}^{-1} \boldsymbol{X}_0^\top\right)}_{\Phi_B(\boldsymbol{S})} + 2\|\boldsymbol{h}_0\|^2.$$

Note that $\Phi_B(\boldsymbol{S})$ requires **no specific form** for h and it seems to define a **tighter** bound when compared to that derived applying the same assumptions adopted by Meng with LowCon approach

Multi-Objective Approach

Two Conflicting Objectives

The joint minimisation problem

Since σ^2 and $\|\boldsymbol{h}\|^2$ are unknown, scalarising the MSPE bound requires prior assumptions. Instead we solve:

$$\min_{\boldsymbol{S}} \left(\Phi_I(\boldsymbol{S}),\ \Phi_B(\boldsymbol{S})\right) \qquad \text{s.t. } |\mathcal{S}| = n$$

This yields the **Pareto front** of non-dominated subsamples.

Our approach

We use a **Memetic Algorithm** (GA + local search) to:

1. Compute **mono-objective** designs minimising Φ_I or Φ_B individually
2. Trace the full **Pareto front** in a multi-objective run
3. Extract a **compromise design** (Elbow) from the front

These designs are then compared against baseline competitors.

Memetic Algorithm

Selecting n units from N is combinatorial: $\binom{N}{n}$ candidates. We use a **Memetic Algorithm** (Genetic Algorithm with embedded local search), adapted for multi-objective problems (NSGA-II flavour).

Genetic operators:

- Individuals: $\boldsymbol{\delta} \in \{0,1\}^N$, $\|\boldsymbol{\delta}\|_1 = n$
- Tournament selection ($k = 2$)
- Uniform crossover (feasibility preserved by bit-swap)
- Random bit-swap mutation

Local search (greedy, per individual):

- Propose bit-swap neighbours
- Accept any (Pareto-)improving swap
- Run for $\ell_s = 20$ iterations

Mono-objective runs (one criterion at a time) yield the **I-opt** and **UpperBias-opt** designs; the multi-objective run traces the full **Pareto front**.

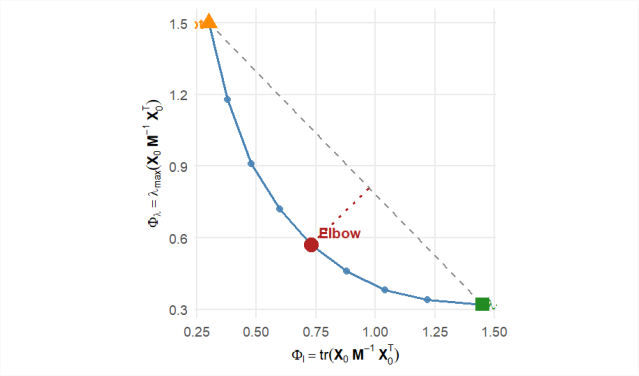
Pareto Front & Elbow Selection

Pareto front in criterion space

The bi-criteria space (Φ_I, Φ_B) shows a characteristic **decreasing convex trade-off**.

Elbow selection

1. Standardise the Pareto front to $[0, 1]^2$
2. Find the design with **maximum perpendicular distance** to the diagonal connecting the two extremes



Simulation Study

Simulation Setup

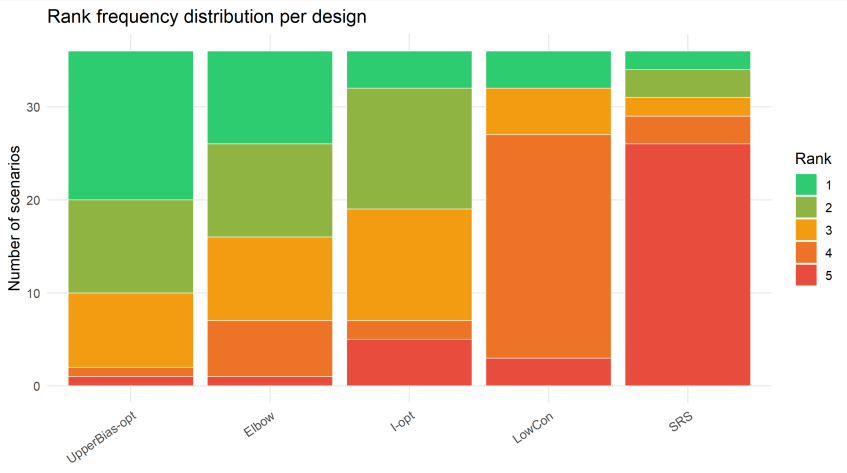
Component	Details
True model	$Y = \boldsymbol{x}^\top \boldsymbol{\beta} + h(\boldsymbol{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1), \quad \boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_2), \quad \boldsymbol{\beta} = (1,1,1)^\top$
Misspecification	$h_1(\boldsymbol{x}) = x_1 x_2$ (interaction); $\quad h_2(\boldsymbol{x}) = x_2^2 + x_1 x_2$ (quadratic + interaction); $h_3(\boldsymbol{x}) = x_1 \sin(x_2)$ (oscillatory)
Strength α	$\{0.1, 1, 10\}$; $\quad$ bounded h: $\lvert h(\boldsymbol{x})\rvert \leq \alpha \lVert \boldsymbol{x} \rVert$
Data sizes	$N \in \{10{,}000;\ 100{,}000\}$; $\quad n \in \{50, 100, 500\}$
Prediction	4 regions of $m = 5$ points each in $\mathbb{R}^2$
Evaluation	Exact analytical MSPE for each generated dataset (single run per scenario)

Competing Designs

Design	Description
I-opt	Mono-objective GA minimising $\Phi_I = \mathrm{tr}(\boldsymbol{X}_0 \boldsymbol{M}^{-1} \boldsymbol{X}_0^\top)$. Pure variance minimisation; ignores h.
UpperBias-opt	Mono-objective GA minimising $\Phi_B = \lambda_{\max}(\boldsymbol{X}_0 \boldsymbol{M}^{-1} \boldsymbol{X}_0^\top)$. Pure new bias bound minimisation.
Elbow	Extracted from the multi-objective Pareto front. Elbow point in the standardised (Φ_I, Φ_B) space: maximum perpendicular distance to the diagonal.
LowCon	Meng et al. (2021). Minimises $\kappa(\boldsymbol{M})$ via orthogonal Latin hypercube-based sampling. LowCon-inspired baseline (estimation-oriented).
SRS	Simple random sampling. Uninformative reference.

Simulation Results — Rank frequency

Overview — $N = 100{,}000$, $n = 100$



Simulation Results — MSPE rank by scenario

Mean rank over all 4 prediction regions — $N = 100{,}000$, $n = 100$, bounded h

h	α	I-opt	UpperBias-opt	LowCon	SRS	Elbow
h_1	0.1	**2.25**	**2.25**	3.25	5.00	**2.25**
h_1	1	3.50	**1.25**	3.00	4.50	2.75
h_1	10	2.75	**2.25**	3.50	4.00	2.50
h_2	0.1	3.25	**1.75**	3.75	4.25	2.00
h_2	1	3.50	**1.75**	3.25	4.00	2.50
h_2	10	3.75	**1.00**	3.25	4.00	3.00
h_3	0.1	**1.50**	3.00	3.75	5.00	1.75
h_3	1	**2.25**	2.50	3.50	4.00	2.75
h_3	10	2.75	**1.50**	3.75	4.00	3.00
Mean		2.81	**1.92**	3.44	4.31	2.53

Simulation Results — Robustness across (N, n)

Mean MSPE rank over all 36 scenarios (3 h-types $\times$ 3 α $\times$ 4 prediction regions), bounded h

Configuration	I-opt	UpperBias-opt	LowCon	SRS	Elbow
$N = 10{,}000,\ n = 50$	2.72	**2.00**	3.58	4.22	2.39
$N = 10{,}000,\ n = 100$	2.78	**2.25**	3.47	4.17	2.33
$N = 10{,}000,\ n = 500$	2.86	**2.31**	3.14	3.94	2.75
$N = 100{,}000,\ n = 50$	2.86	**2.03**	3.47	4.17	2.47
$N = 100{,}000,\ n = 100$	2.81	**1.92**	3.44	4.31	2.53
$N = 100{,}000,\ n = 500$	2.67	**2.28**	3.25	4.06	2.75

- UpperBias-opt wins in all 6 configurations
- There might be a sweet spot of observed proportion for the bias design

Real Data Application

Real Data: Soil Dataset

Dataset — Hengl et al. (2015)

- Soil survey in Africa: $N = 1{,}157$ observations
- $p = 5$ covariates: CTI, ELEV, RELI, TMAP, TMFI (topographic and climatic indices); response: Sand content (%)
- Genuine misspecification expected: complex soil-landscape relationships

MSPE approximated via held-out test set

In real data, h is genuinely unknown, hence we approximate the MSPE using a held-out test set within each neighbourhood.

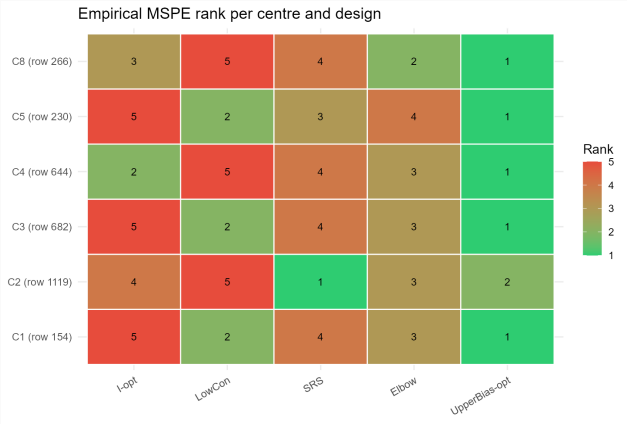
Evaluation protocol

1. Standardise all covariates
2. Define $k = 6$ neighbourhood centres via k-means
3. For each centre c:
   form a hypercube, and take $m = 30$ held-out test points as $\boldsymbol{X_0}$
4. Apply each design on the pool ($n = 50$)
5. Evaluate: $\widehat{\mathrm{MSPE}} = \frac{1}{m}\|\boldsymbol{X}_0\hat{\boldsymbol{\beta}}_S - \boldsymbol{Y}_0\|^2$

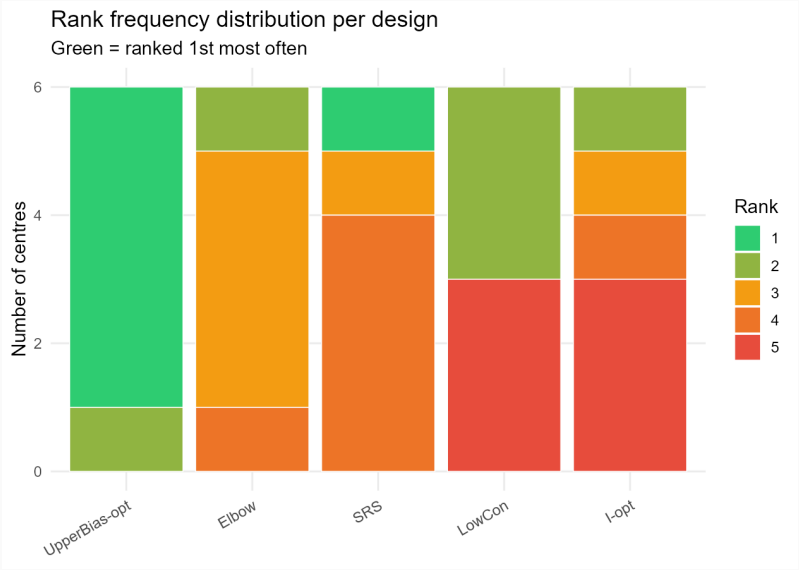
Real Data: Soil — Results

Mean emp. MSPE rank (1 = best)

Design	$\bar{r}$	#1st
UpperBias-opt	**1.17**	**5**
Elbow	3.00	0
SRS	3.33	1
LowCon	3.50	0
I-opt	4.00	0



Real Data: Soil Dataset — Rank Distribution



Conclusions

Conclusions

Key contributions

- **Prediction-oriented MSPE framework**: extends misspecification analysis from estimation (Meng et al., 2021) to the **MSPE** with arbitrary prediction sets $\boldsymbol{X}_0$.
- **Tighter bias bound** (Theorem 2): minimising $\lambda_{\max}(\boldsymbol{X}_0\boldsymbol{M}^{-1}\boldsymbol{X}_0^\top)$ is sharper than the LowCon-inspired bound and requires **no specific assumption on** h.
- **Multi-objective approach**: jointly minimising (Φ_I, Φ_B) via a memetic algorithm yields a Pareto front that covers all variance/bias preferences without tuning σ^2 or α.
- **Consistent empirical findings**: UpperBias-opt is the overall winner in simulations and on the Soil real dataset.

Thank you for your attention,
any questions?

References i

- Meng, C., Xie, R., Mandal, A., Zhang, X., Zhong, W., & Ma, P. (2021). LowCon: A Design-based Subsampling Approach in a Misspecified Linear Model. *Journal of Computational and Graphical Statistics*, 30(3), 694–708.
- Wang, H., Yang, M., & Stufken, J. (2019). Information-Based Optimal Subdata Selection for Big Data Linear Regression. *Journal of the American Statistical Association*, 114(525), 393–405.
- Hengl, T., Heuvelink, G. B. M., Kempen, B., et al. (2015). Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. *PLoS ONE*, 10(6).