Pareto-optimal parameters in linear regression problems

Anja Mühlemann **Johanna F. Ziegel**

Institute of Mathematical Statistics and Actuarial Science, University of Bern, Switzerland

anja.muehlemann@stat.unibe.ch johanna.ziegel@stat.unibe.ch

Motivation

Let $Y \in \mathbb{R}$ and $X \in \mathbb{R}^p$ be two random elements. We are interested in a parametric model for

$$g(x) = \mathbb{E}(Y|X = x).$$

Let

$$m : \mathbb{R}^p \times \Theta \to \mathbb{R}, \quad (x, \theta) \to m(x; \theta)$$

be such a model.

Definition 1. *A parametric model, $m(x; \theta)$, is said to be correctly specified for the conditional expectation if:*

$$\mathbb{E}(Y|X = x) = m(x; \theta^*) \text{ a.s. for some } \theta^* \in \mathbb{R}^p.$$

Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be independent copies of (X, Y). Assuming some moment conditions and correct specification one can show that

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta \in \Theta} \frac{1}{n} \sum_{k=1}^{n} L_\phi(m(X_k; \theta), Y_k)$$

yields a consistent estimator of θ^* for any choice of Bregman loss function L_ϕ, that is

$$L_\phi(x, y) = \phi(y) - \phi(x) - \phi'(x)(y - x),$$

where ϕ is a (strictly) convex function with subgradient ϕ'. We are interested in estimation methods for model parameters that avoid the choice of a specific loss function but take all loss functions into account simultaneously.

Justification

The estimate $\hat{\theta}_n$ is a consistent estimator for θ^* because the class of Bregman loss functions is consistent for the mean, that is

$$\mathbb{E}(Z) = \operatorname*{argmin}_{x \in \mathbb{R}} \mathbb{E}L_\phi(x, Z)$$

for Z such that $L_\phi(x, Z)$ is integrable. The more general definition given by Gneiting (2011) reads as follows: Let $\mathcal{P}$ be a class of probability measures on $O \subset \mathbb{R}^d$ and let

$$T : \mathcal{P} \to A, \quad F \mapsto T(F),$$

be a functional where $A \subset \mathbb{R}^k$.

Definition 2. *A loss function $L : A \times O \to \mathbb{R}$ is consistent for the functional T relative to the class $\mathcal{P}$ if*

$$\mathbb{E}_F L(t, Y) \leq \mathbb{E}_F L(x, Y) \tag{1}$$

for all probability distributions $F \in \mathcal{P}$, all $t \in T(F)$ and all $x \in A$. It is strictly consistent if it is consistent and equality in (1) *implies that $x \in T(F)$.*

The result that the class of Bregman loss functions is consistent for the mean was shown by Savage (1971).

Pareto Optimality

Ehm et al. (2016) suggested a notion of forecast dominance relative to some class $\mathcal{S}$ where a forecast dominates another forecast if it minimizes the expected loss for all loss functions in $\mathcal{S}$. Adapting this definition, we obtain the following concept of model parameter dominance.

Definition 3. *A parameter θ_1 is dominated by a parameter θ_2 if*

$$\mathbb{E}L(Y, m(X; \theta_2)) \leq \mathbb{E}L(Y, m(X; \theta_1))$$

for all consistent loss functions. It is strictly dominated if the inequality is strict for some consistent loss function. A parameter θ is Pareto optimal if it is not strictly dominated by any other parameter.

Immediate observations:

- Any parameter that is the unique minimizer of some consistent loss function is Pareto optimal.
- If the model is correctly specified, then all correct parameters are Pareto optimal.
- If the model is point-identified, then there is only one Pareto optimal parameter.
- If a parameter minimizes all loss functions simultaneously, then it is Pareto optimal.

Ehm et al. (2016) showed that any Bregman loss functions can be written as a weighted average of a continuum of so-called elementary loss functions.

Theorem 4. *Any Bregman loss function can be written as*

$$L_\phi(y, \hat{y}) = \int S_\eta(y, \hat{y}) \, \mathrm{d}H_\phi(\eta)$$

where

$$S_\eta(y, \hat{y}) = (\mathbb{1}\{\eta \leq \hat{y}\} - \mathbb{1}\{\eta \leq y\})(\eta - y),$$

and H_ϕ is a uniquely determined non-negative measure on $\mathbb{R}$ that depends on ϕ.

This theorem allows us to define Pareto optimality in terms of the elementary loss functions.

Theoretical properties

The following proposition gives an idea on how strong the concept of dominance is. Namely, if some parameter dominates all other parameters then the model $m(X; \theta^*)$ is an autocalibrated prediction for $\mathbb{E}(Y)$.

Theorem 5. *Suppose that $\theta^* \in \Theta$ dominates all other parameters. In particular, it minimizes any consistent loss function. Under some assumptions on the model and some moment conditions, we obtain that*

$$m(X; \theta^*) = \mathbb{E}(Y|\sigma((m(X; \theta^*))).$$

In general it is not the case that one parameter dominates all the others. We therefore are interested in the Pareto optimal set, that is set of all non-dominated parameters. If we consider a simple linear model then under some conditions on the function $g(x) = \mathbb{E}(Y|X = x)$, we can explicitly calculate the Pareto optimal parameters.

Proposition 6. *Suppose that $X \in \mathbb{R}$ has a non-vanishing density $p(x)$ and that $g(x) = \mathbb{E}(Y|X = x)$ is strictly increasing and differentiable. Consider a linear model*

$$m : \mathbb{R} \times \Theta \to \mathbb{R}, \quad (x, \theta) \mapsto m(x; \theta) = \theta_0 + \theta_1 x,$$

where $\Theta = \mathbb{R} \times (0, \infty)$. The set of Pareto optimal parameters consists of all parameters of the form

$$\theta_0 = \eta - g^{-1}(\eta) g'(g^{-1}(\eta)), \quad \theta_1 = g'(g^{-1}(\eta)), \tag{2}$$

$\eta \in \mathbb{R}$, and all parameters θ such that the function

$$\eta \mapsto \eta - g((\eta - \theta_0)/\theta_1)$$

has at least two zeros.

Example 7. *Consider $g(x) = 1 - e^{-x} + x$. Clearly, this function is strictly increasing, differentiable and concave. The Pareto optimal parameters according to* (2) *are given by*

$$\theta_0 = \eta - \left(1 + W(e^{1-\eta})\right)\left(\eta + W(e^{1-\eta}) - 1\right)$$

and

$$\theta_1 = 1 + W(e^{1-\eta})$$

for $\eta \in \mathbb{R}$, where W is the Lambert-W function. Moreover, the function $\eta \mapsto \eta - g((\eta - \theta_0)/\theta_1)$ has two zeros for $\theta \in \Theta$ with $\theta_1 > 1$ and

$$\theta_0 < 2 + \theta_1 - \log(\theta_1 - 1)(1 - \theta_1).$$

The following pictures show the theoretical Pareto optimal set as well as the approximation of the Pareto optimal set using the algorithm NSGA-II both for $n = 50$ and $n = 500$. The data $X_1, \ldots X_n$ is a random sample from $\mathcal{U}(-2, 2)$ and $Y_i = f(X_i) + \varepsilon_i$ for all i with $\varepsilon_i \sim \mathcal{N}(0, 1)$, where $f(x) = g(x)$ for the misspecified case, and $f(x) = a + bx$ for the correctly specified case, where a and b are the OLS estimates for the misspecified case.

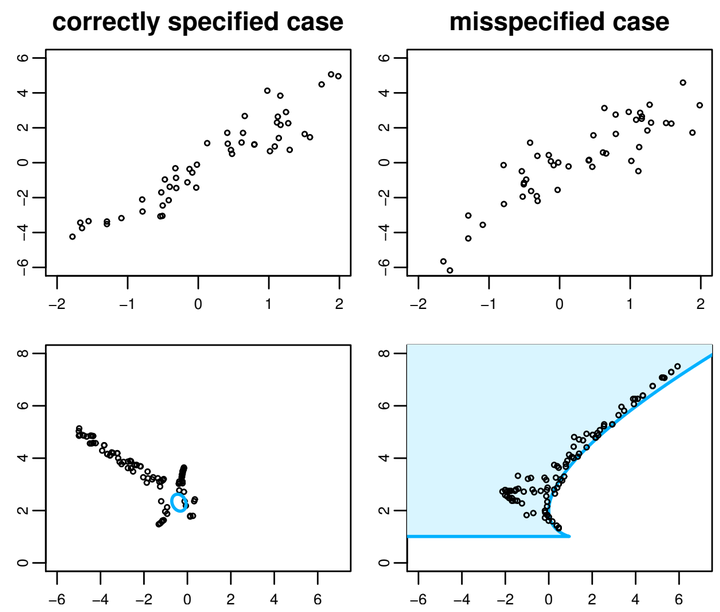

Figure 1: *Pareto optimal parameters for $n = 50$.*

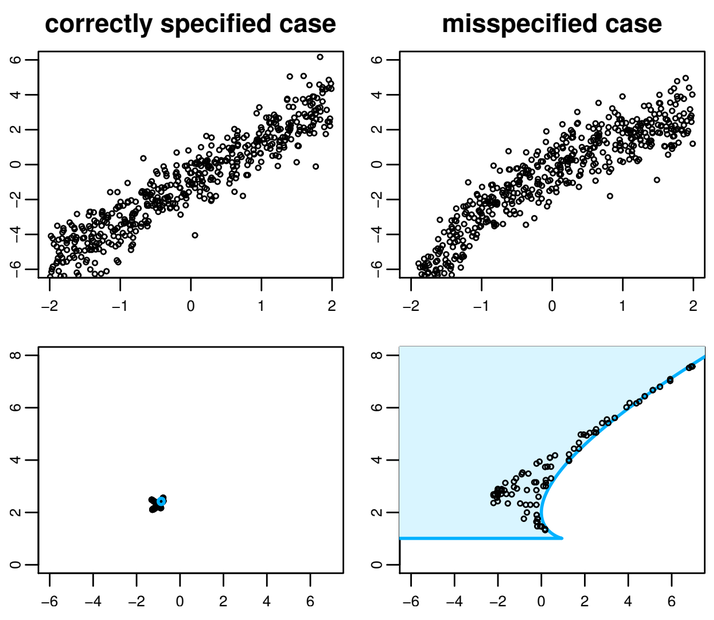

Figure 2: *Pareto optimal parameters for $n = 500$.*

Open Questions

- What is a good interpretation of the set of Pareto optimal parameters? Can we interpret the set of Pareto optimal parameters as a degree of misspecification? How?
- Let $\hat{F}_n$ be the empirical distribution of $(X_1, Y_1), \ldots (X_n, Y_n)$ and F the distribution of (X, Y). Does the set of Pareto optimal parameters of the $\hat{F}_n$ converge to the set of Pareto optimal Parameters of F? In which sense? How fast?
- How can we compute Pareto optimal parameters efficiently?
- Can we improve the model knowing the set of Pareto optimal parameters?

References

Werner Ehm, Tilmann Gneiting, Alexander Jordan, and Fabian Krüger. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(3):505–562, 2016.

Tilmann Gneiting. Making and evaluating point forecasts. *J. Amer. Statist. Assoc.*, 106(494):746–762, 2011.

Leonard J. Savage. Elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*, 66(336):783–801, 1971.