ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ
ΟΙΚΟΝΟΜΙΚΩΝ
ΕΠΙΣΤΗΜΩΝ
SCHOOL OF
ECONOMIC
SCIENCES

ΤΜΗΜΑ
ΟΙΚΟΝΟΜΙΚΗΣ
ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF
ECONOMICS

Department of Economics

Athens University of Economics and Business

WORKING PAPER no. 19-2025

Robust Multiobjective Model Averaging in Predictive Regressions

Stelios Arvanitis, Mehmet Pinar, Thanasis Stengos,

and Nikolas Topaloglou

Otober 2025

Robust Multiobjective Model Averaging in Predictive Regressions

Stelios Arvanitis[*], Mehmet Pinar[†], Thanasis Stengos[‡], Nikolas Topaloglou [§]

October 13, 2025

Abstract

This paper develops a framework for robust model averaging under multiobjective dominance relations defined over a set of predictive and information-theoretic scoring criteria. We introduce and analyze a set of model averaging estimators that are constructed as Pareto-optimal solutions to scalarized formulations of a vector-valued criterion function. First-order estimators—including fixed and optimized linear blends, minimax estimators, and Nash bargaining solutions—are shown to be first-order approximations to an ideal but infeasible robust estimator defined via influence minimization under contamination. Second-order estimators, such as the optimally weighted ℓ_1 and Dirichlet-Nash procedures, are formally characterized as improved approximations with asymptotically adaptive behavior and tighter influence bounds. We provide sufficient conditions under which these second-order estimators dominate first-order procedures in terms of approximating the optimal predictive mean squared error. The proposed estimators are also evaluated in Monte Carlo simulations under contamination and in an empirical application to cross-country growth regressions.

Keywords– Model averaging; Multiobjective optimization; Dominance relations; Scalarization; Nash estimator; Robust estimation; Influence function; Local contamination; Cross-validation; Approximate bound estimator; Growth regressions.

1 Introduction

Model averaging has emerged as a flexible and principled framework to address model uncertainty in statistical and econometric applications. Broadly speaking, two methodological strands exist: Bayesian Model Averaging (BMA) and Frequentist Model Averaging (FMA).

[*] Athens University of Economics and Business, Greece. Email: stelios@aueb.gr

[†] Universidad de Sevilla, Spain. Email:mpinar@us.es

[‡] University of Guelph, Canada. Email: tstengos@uoguelph.ca

[§] Athens University of Economics and Business, Greece. Email: nikolas@aueb.gr

While the former leverages prior beliefs and posterior weights-see Steel (2020) for a relevant comprehensive survey, FMA operates directly on prediction error criteria to combine models in a data-driven way. This paper falls within the FMA tradition and advances it by developing a robust, multiobjective approach to model averaging.

The motivation for model averaging stems from the recognition that in many applied settings—notably macroeconomics, forecasting, and growth regressions—numerous competing models coexist, reflecting distinct theoretical perspectives or empirical strategies. Each model captures only part of the underlying data-generating process (DGP), and no single model is guaranteed to be correctly specified. This leads to estimation and inference risks if one commits to a single model. Averaging across models can reduce these risks by pooling information, hedging against misspecification, and enhancing out-of-sample performance.

Within the FMA literature, foundational estimators include Mallows Model Averaging (MMA), introduced by Hansen (2007), which minimizes an asymptotic risk bound using a Mallows-type correction; Jackknife Model Averaging (JMA), proposed by Hansen and Racine (2012), which minimizes a cross-validated loss and is robust to heteroskedasticity and collinearity; and information criterion–based methods, which use potentialy smoothes versions of AIC or BIC to generate penalty-adjusted model weights-see for example Claeskens and Hjort (2008).[1] While successful, these approaches rely on single scoring rules, which raises two interrelated challenges. First, criterion dependence implies that different criteria (e.g., AIC vs. JMA) often yield markedly different weight vectors, introducing ambiguity in the absence of a universally optimal rule. Second, single-criterion estimators may be sensitive to data contamination or model misspecification, particularly when the underlying assumptions of the selected criterion are violated.

This paper tries to address these limitations by proposing a multiobjective model averaging (MOMA) framework-see Hwang and Masud (2012)-that generalizes existing FMA methods. The central idea is to treat model evaluation as a vector-valued optimization problem, where multiple criteria jointly define a dominance relation over the simplex of model weights. We aim to identify estimators that are approximately Pareto-optimal with respect to this vector of criteria, thus hedging against the drawbacks of relying on a single score.

Our contributions are several. We introduce a formal dominance relation over model weights, based on simultaneous comparison across multiple scoring criteria. This motivates the search for Pareto-efficient model averaging estimators, which cannot be jointly dominated by any alternative vector of weights. We then construct a number of such estimators based on scalarization techniques or equilibrium formulations. These include convex combinations of criteria (ℓ_1), worst-case criterion minimization (ℓ_∞), geometric regret minimization (Nash), maximum regret minimization (AB, inspired by Condorcet comparisons), and second-order estimators such as validation-optimized ℓ_1^{opt} and the randomized Dirichlet-Nash, which perform a second-stage optimization over criterion weights using cross-validation.

We develop an approximation theory for robust model averaging under distributional contamination, grounded in local expansions of the risk functional. The core result characterizes the excess risk of any estimator defined by a smooth, strongly convex criterion,

[1] Several studies develop asymptotic theories for the averaging estimators at hand, for example Zhang and Liu (2019) derive the asymptotic properties of least-squares-based averaging estimators in linear regression models under local asymptotic frameworks. Usually the limiting weights are stochastic, usually due to identification failures, rendering non-standard asymptotic theories for the averaging estimators.

evaluated under a locally contaminated distribution, relative to the robust optimal estimator minimizing the (locally contaminated) risk. The approximation bounds typically hold for large enough sample size and depend on three key components: (i) the first-order gradient of the risk at the oracle, (ii) the difference between the influence function derivatives between the criteria, and (iii) the local curvature of the criteria, via their Hessians.

These results provide a first step towards the justification for viewing robust model averaging as a problem of the uncontaminated risk gradient and the influence-function approximation in function space. Second-order estimators—such as ℓ_1^{opt} and Dirichlet–Nash—are thus interpreted as selecting criterion weights that approximately minimize those discrepancies, leading to improved approximation bounds even under contamination. Their second-order nature stems from their reliance on a nested validation step used to adaptively choose a composite criterion. The theoretical framework is formulated under a high-order regularity regime, formalized via Gateaux differentiability strong convexity of criteria.

The theoretical insights are complemented by a comprehensive Monte Carlo study across varying contamination levels and sample sizes. We compare classical estimators (JMA, MMA, AIC, BIC, naive) with multiobjective ones. The results demonstrate that second-order estimators consistently outperform others in terms of predictive MSE, mean absolute error (MAE), and robustness.

Furthermore, we apply our methods to cross-country growth regressions, a setting where model uncertainty and data contamination are both present. There, model uncertainty plays a crucial role encompassing multiple competing perspectives. The standard production function-based Solow growth model, the institutional perspective proposed by Acemoglu et al. (2001), and the geographic determinants emphasized by Sachs (2003) all represent valid but distinct explanatory approaches. The importance of model uncertainty in assessing the relative influence of these growth theories has been studied in Kourtellos et al. (2010). Susceptibility to measurement error, omitted variables, and sample heterogeneity is emphasized in the relevant literature-see for example Durlauf et al. (2005), rendering the robustness issue relevant. As argued by Brock and Durlauf (2001)... "one of the main roles of cross-country growth data would be to compute predictive distributions for the consequences of policy outcomes, distributions that can then be combined with a policymaker's welfare function to assess alternative policy scenarios..." The standard cross-country growth regression framework looks at model uncertainty as a means of combining different theoretical approaches into a weighted average so to arrive at more reliable inferences in sample. However, it is the predictive nature of the entire (averaged) model distribution that can be useful for policy purposes and not so much individual variables. Making predictions about economic growth in general in the context of data that at times are undergoing repeated revisions, Barro and Lee (2013) and also rely on statistical agencies that may not be similar in their collection capabilities, introduces uncertainty about the use of the available data sets that are routinely used in the empirical literature, beyond the model uncertainty that was mentioned earlier. In this context, it is also data contamination that appears to be an issue that needs to be taken into account within the FMA framework. The proposed estimators yield more stable and accurate forecasts of GDP growth, confirming the practical value of the multiobjective framework.

Our work extends the FMA literature in several directions. It builds on Hansen (2007), Hansen (2007), Claeskens and Hjort (2003), but moves beyond fixed-criterion averaging toward a dominance-based and multiobjective perspective-see for example Arvanitis et al. (2021) for some similar concepts in stochastic dominance and portfolio analysis. We formally distinguish between first-order estimators (e.g., fixed scalarizations such as ℓ_1, Nash, ℓ_∞, and AB) and second-order estimators (e.g., ℓ_1^{opt}, Dirichlet-Nash) that optimize over criterion weight space. We also discuss the possibility of third-order estimators that combine second-order scores via additional validation or learning-based aggregation steps.

Importantly, despite its generality and conceptual richness, the multiobjective framework proposed in this paper is computationally efficient. In the linear model setting, all candidate model predictions can be precomputed, and the optimization problems involved—whether convex combinations, max regret, or geometric averages—are convex or differentiable. Second-order procedures involve low-dimensional optimization over the criterion weight simplex, often handled efficiently by grid search or stochastic sampling. As a result, the practical implementation cost of the proposed estimators is mild even in large-scale applications.

The remainder of the paper is structured as follows. Section 2 defines the dominance relation, Pareto efficiency, and the resulting class of multiobjective estimators. Section 3 presents the theoretical foundations based on optimization approximation bounds towards a benchmark optimal estimator under infinitesimal contamination. Section 4 reports Monte Carlo results. Section 5 applies the estimators to cross-country growth data. Section 6 concludes and discusses future directions.

2 Framework: Linear Models and Notation

Let $y \in \mathbb{R}^n$ denote the response vector and $X \in \mathbb{R}^{n\times p}$ the matrix of predictors. We consider a collection of linear models indexed by $m = 1, \dots, M$, each associated with a selection matrix $R_m \in \mathbb{R}^{p_m \times p}$, mapping full regressors X to submodels: $X_m := XR_m^\top \in \mathbb{R}^{n\times p_m}$.

Each model m specifies a linear regression:

$$y = X_m\beta_m + \varepsilon_m,$$

with $\mathbb{E}[\varepsilon_m] = 0$, and $\mathrm{Var}(\varepsilon_m)$ is well defined and pd. Let $\hat{\beta}_m$ denote the OLS estimator and $\hat{y}_m := X_m\hat{\beta}_m$ the fitted values.

We define the model averaging prediction as:

$$\hat{y}(w) = \sum_{m=1}^{M} w_m \hat{y}_m, \quad w \in \Delta^M := \left\{ w \in \mathbb{R}^M : w_m \geq 0, \ \sum_{m=1}^{M} w_m = 1 \right\}.$$

We are primarily interested in predictive performance, as evaluated on test data or through resampling-based out-of-sample error estimates. Our goal is to choose $w \in \Delta^M$ minimizing a vector of criteria.

3 Dominance Relations and Multiobjective Motivation

Let $\{J_k(w)\}_{k=1}^K$ denote a set of real-valued scoring functions defined on Δ^M, such as MSE, MAE, KL-divergence, or information criteria. We define the criterion vector:

$$\mathbf{J}(w) := (J_1(w), \dots, J_K(w)) \in \mathbb{R}^K.$$

Define a dominance relation over Δ^M by:

$$w \prec w' \quad \text{if} \quad J_k(w) \leq J_k(w') \ \forall k, \ \text{with strict inequality for some } k.$$

We are interested in identifying (approximately) Pareto-optimal weight vectors $w^* \in \Delta^M$, i.e., weights not dominated by any other vector. This motivates our introduction of multiobjective estimators as solutions to specific scalarizations or dominance-based approximations of this vector criterion problem.

In this context, the approximate bound (AB) estimator, Nash-type estimators, and weighted ℓ_1 solutions with cross-validated criterion weights aim to balance trade-offs across criteria.

3.1 Basis Scoring Criteria Used

We now formally specify the set of scoring criteria $\{J_k(w)\}_{k=1}^K$ used throughout the paper, both in the theoretical development of multiobjective estimators and in the Monte Carlo and empirical analyses. Each criterion $J_k(w)$ assigns a score to a weight vector $w \in \Delta^M$, corresponding to a particular model averaging estimator based on a distinct information-theoretic or prediction-based principle. Each of these scoring rules reflects a distinct estimation philosophy — uniform averaging, resampling-based cross-validation, penalized prediction based on model complexity, and exponentially smoothed versions of classical information criteria — and all serve as the atomic basis for multiobjective combinations. The selected basis criteria actually used in this paper are:

1. Naive Averaging Criterion $J_{\textbf{naive}}$: To formalize the equal-weight benchmark as a minimization problem, we define

$$J_{\text{naive}}(w) := \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \sum_{m=1}^{M} w_m \hat{y}_t^{(m)} \right)^2 + \gamma \sum_{m=1}^{M} \left(w_m - \frac{1}{M} \right)^2, \quad \gamma > 0.$$

The regularization term penalizes deviations from uniform weighting, ensuring a unique minimizer at $w_m = 1/M$ for all m for large γ.

2. Jackknife Model Averaging Criterion (JMA): Originally proposed by Hansen and Racine (2012), the JMA criterion minimizes a leave-one-out cross-validation estimate of prediction error:

$$J_{\text{JMA}}(w) := \frac{1}{n} \sum_{t=1}^{n} \left(y_t - \sum_{m=1}^{M} w_m \hat{y}_t^{(-t,m)} \right)^2,$$

where $\hat{y}_t^{(-t,m)}$ denotes the prediction from model m estimated without observation t. This criterion is minimized under the constraint $w \in \Delta^M$, often using quadratic programming. JMA is known to be consistent under general heteroskedasticity and collinearity, provided sufficient sample size and smoothness conditions on the models.

3. Mallows Model Averaging Criterion (MMA): The MMA criterion–see for example Zhang and Liu (2019)–penalizes model complexity using a Mallows-type correction and is defined as:

$$J_{\mathrm{MMA}}(w) := \frac{1}{n}\sum_{t=1}^{n}\left(y_t - \sum_{m=1}^{M} w_m \hat{y}_t^{(m)}\right)^2 + \sum_{m=1}^{M} w_m \cdot \mathrm{pen}_m,$$

where $\mathrm{pen}_m := 2\hat{\sigma}^2 \cdot d_m/n$, with d_m denoting the number of regressors in model m, and $\hat{\sigma}^2$ is a common estimate of the noise variance; in our framework typically the OLSE variance estimator from the full model incorporating all candidate regressors. This criterion accounts for overfitting by discouraging large weights on complex models.

4. KL-Regularized Smoothed AIC Criterion: To incorporate the Akaike Information Criterion (AIC) into a convex optimization framework while ensuring differentiability and strong convexity, we adopt an exponential smoothing formulation with Kullback–Leibler (KL) regularization:

$$J_{\mathrm{AIC}}^{\mathrm{KL}}(w) := \sum_{m=1}^{M} w_m \cdot \exp\left(\frac{\mathrm{AIC}_m}{c}\right) + \rho \cdot \sum_{m=1}^{M} w_m \log\left(\frac{w_m}{u_m}\right),$$

where $\mathrm{AIC}_m = n\log(\hat{\sigma}_m^2) + 2d_m$ is the AIC score of model m, $c > 0$ is a temperature parameter controlling smoothness, and $\rho > 0$ is a regularization coefficient. The reference distribution $u = (u_1, \ldots, u_M)$ is typically uniform over Δ^M. The second term introduces a KL divergence penalty that ensures ρ-strong convexity of $J_{\mathrm{AIC}}^{\mathrm{KL}}$ over the simplex, even though the exponential criterion is linear in w.

5. KL-Regularized Smoothed BIC Criterion: Analogously, the Bayesian Information Criterion (BIC) can be incorporated through a smoothed and regularized version:

$$J_{\mathrm{BIC}}^{\mathrm{KL}}(w) := \sum_{m=1}^{M} w_m \cdot \exp\left(\frac{\mathrm{BIC}_m}{c}\right) + \rho \cdot \sum_{m=1}^{M} w_m \log\left(\frac{w_m}{u_m}\right),$$

where $\mathrm{BIC}_m = n\log(\hat{\sigma}_m^2) + \log(n) \cdot d_m$, and the parameters c, ρ, and u are as defined above. BIC imposes a stricter complexity penalty than AIC, favoring more parsimonious models. The KL regularization again guarantees strong convexity.

All five criteria are differentiable, convex in w, and consistently defined over the simplex Δ^M. They provide a comprehensive base for constructing robust and flexible model averaging estimators via multiobjective optimization. Their convexity and regularity ensure the theoretical guarantees derived in our influence-function-based analysis apply across all examined estimators.

These criteria jointly define the vector-valued performance mapping

$$\mathbf{J}(w) := (J_1(w), \ldots, J_K(w)) \in \mathbb{R}^K,$$

which serves as the foundation for all scalarization-based multiobjective estimators constructed in later sections. They also enable the definition of dominance relations and Pareto-optimality structures, which motivate the approximate equilibrium formulations used throughout.

3.2 Multiobjective Estimators: Definitions and Characterizations

Given the vector of scoring functions $\mathbf{J}$, we define several classes of multiobjective model averaging estimators. Each estimator corresponds to a particular scalarization or equilibrium principle.

Linear Aggregation Estimator (ℓ_1): Given criterion weights $\lambda \in \Delta^K$, the estimator is defined by

$$w^{\ell_1}(\lambda) := \arg\min_{w \in \Delta^M} \sum_{k=1}^{K} \lambda_k J_k(w).$$

This estimator represents a convex scalarization of the multiobjective problem. The solution lies on the Pareto frontier for every strictly positive λ, and yields different trade-offs depending on the weighting. The adaptive choice of the hyperparameter λ is addressed in the following version of the criterion. In what follows, the non-adaptive standard choice is the uniform one. Furthermore, each of the basis criteria is recoverable as an ℓ_1 case for the appropriate choice of λ.

Optimal Criterion Weighting ($\ell_1^{\mathbf{opt}}$): Building on the previous, now, the criterion weights themselves are chosen optimally via validation:

$$\lambda^{\text{opt}} := \arg\min_{\lambda \in \Delta^K} \text{MSE}_{\text{val}}(w^{\ell_1}(\lambda)),$$

$$w^{\ell_1^{\text{opt}}} := w^{\ell_1}(\lambda^{\text{opt}}).$$

This estimator is considered as second-order, in the sense that it involves a nested optimization: the outer level selects λ, while the inner computes the optimal aggregation given λ.

Max-Min Estimator (ℓ_∞):

$$w^{\ell_\infty} := \arg\min_{w \in \Delta^M} \max_k J_k(w).$$

This estimator minimizes the worst-case loss among criteria. It reflects a conservative scalarization approach and corresponds to a uniform scaling Chebyshev approximation to the Pareto front–see Zhang and Golovin (2020).

Nash Estimator: the Nash bargaining solution (see Aumann and Hart (1992)) is now considered for the determination of Pareto points. In this framework, the basis criteria act as cooperative players negotiating over model averaging choices. Each player's utility is defined as the relative improvement over their fallback (least satisfactory) option, producing the overall averaging optimization problem,

$$w^{\text{Nash}} := \arg\min_{w\in\Delta^M} \left(\prod_{k=1}^{K} \left[J_k(w) - \min_{w'} J_k(w')\right]\right)^{1/K} .$$

This estimator represents a geometric compromise across all criteria, balancing their relative deviations from optimality. It obviously results in an estimator that is invariant to re-scalings of the basis criteria.

Dirichlet-Nash Estimator: Building on the above, criterion exponent vectors are sampled, $\lambda^{(s)} \sim \text{Dir}(\alpha)$ and then the resulting Nash estimator is computed:

$$w^{(s)} := \arg\min_{w\in\Delta^M} \prod_{k=1}^{K} \left[J_k(w) - \min_{w'} J_k(w')\right]^{\lambda_k^{(s)}} ,$$

Then, the estimator is finalized by selecting the best-performing $w^{(s)}$ via the validation step:

$$w := w^{(s^*)}, \quad s^* := \arg\min_{s} \text{MSE}_{\text{val}}(w^{(s)}).$$

This estimator is also second-order, due to its data-driven sampling and validation step. It generalizes the Nash formulation via random convex scalarizations.

Approximate Bound Estimator (AB): A particularly relevant concept in stochastic dominance optimization is the Approximate Bound (AB), introduced by Arvanitis et al. (2021). There the weight vector is as close as possible to being maximal, meaning that it minimizes the worst-case deviation from dominance. It is defined as the solution to:

$$w^{\text{AB}} := \arg\min_{w\in\Delta^M} \max_{k} \left[J_k(w) - \min_{w'} J_k(w')\right] .$$

This estimator can be viewed as a Condorcet winner–see for example Brandl et al. (2016)–in a social choice framework: each criterion votes against a weight vector if it performs worse relative to its best score. The AB estimator selects the weight vector that minimizes maximum opposition across all criteria.

3.3 Classification: First-Order vs Second-Order Estimators

The estimators above can be categorized into:

(i) First-order estimators: ℓ_1, ℓ_∞, AB, Nash. These are direct minimizers of fixed scalarizations and require no additional training/validation splits. The aforementioned connection of ℓ_1 with each of the basis estimators classifies also the latter as first order estimators.

(ii) Second-order estimators: ℓ_1^{opt}, Dirichlet-Nash. These involve nested optimization over the space of criterion weights, using out-of-sample validation to optimize criterion combinations.

Second-order estimators could by construction exhibit improved predictive MSE robustness; the theory section below shows that this is plausible when their influence function approximation better aligns with that of an infeasible robust MSE estimator.

3.4 Indicative Extensions

Future work may explore further scalarization methods for defining multiobjective estimators:

(i) Chebyshev Scalarizations: These generalize ℓ_∞ by replacing the max with smooth approximations:

$$w^{\text{smooth-Cheb}} := \arg\min_{w \in \Delta^M} \left(\sum_{k=1}^{K} \left(\frac{J_k(w) - \min J_k}{\rho_k} \right)^p \right)^{1/p}, \quad p \gg 1.$$

These can for example be tuned to interpolate between ℓ_1 and ℓ_∞ behavior. The criterion can be also extended to incorporate basis criteria weights that can be optimally chosen via validation methods, rendering second-order estimators.

(ii) Deep Learning: Hypernetwork learning of the Pareto set under predictive MSE restrictions may be feasible using machine learning technologies as in Navon et al. (2020). Regularization and validation-based weight selection connected to predictive MSE or MAE could help avoiding overfitting.

These extensions retain the multiobjective interpretation by searching for data-driven approximations to Pareto-efficient combinations.

4 Statistical Theory

This section develops a theoretical framework for model averaging under local distributional contamination. Our approach is grounded in a functional approximation perspective: we analyze how well different estimators approximate an ideal robust benchmark by comparing the influence functions of their defining loss criteria. The key tools are expansions of empirical risk under infinitecimal contamination and risk gradient as well as influence function characterizations of estimator behavior. These results are developed under a high-order assumption framework, whose practical validity is explored in the discussion section.

We begin by defining an ideal estimator that minimizes the predictive mean squared error (MSE) under local contamination of the empirical distribution. This benchmark induces a target influence function whose behavior under perturbations serves as a reference for evaluating other estimators.

This leads to a general theory of approximation under contamination: any estimator defined as the minimizer of an admissible convex criterion over the simplex Δ^M admits a contamination expansion whose first-order term involves the influence function of the criterion. The predictive MSE of such estimators deviates from that of the ideal benchmark by a tractable approximation error. The magnitude of this error depends explicitly–among others–on the functional proximity between the influence function of the chosen criterion and that of the robust population risk. In this sense, model averaging estimators can be characterized as approximate influence-minimizers of the robust target.

We apply this framework to the class of model averaging estimators. These include fixed-criterion procedures (e.g., JMA, MMA, smoothed AIC and BIC), multiobjective methods based on convex or log-convex combinations of criteria (such as ℓ_1-weighted or Dirichlet–Nash estimators), and the data-driven second-order procedures that estimate criterion weights through validation-based optimization. Our results show that under mild contamination conditions, second-order estimators that select criterion weights via predictive loss minimization can better approximate the influence-function behavior of the validation-optimal surrogate, and thus achieve improved approximation to the robust benchmark.

Hopefully, these results offer a first attempt towards unified and extensible framework for model averaging under contamination, grounded in high-order functional analysis and validated through both theoretical approximation bounds and empirical performance. In the final paragraph of this section we provide a discussion on the low order validation of the high order assumption framework upon which the results are based, as well as on potential extensions.

4.1 Definition of the Optimal Estimator

Let $\widehat{P}_n$ denote the empirical distribution of the observed data $z_1, \dots, z_n \in \mathcal{Z} \subseteq \mathbb{R}^{p+1}$, where each $z_i = (x_i, y_i)$. Motivated by concerns over localized contamination or model misspecification, we define a family of perturbed distributions of the form:

$$\widehat{P}_n^{\delta} := (1-\delta)\widehat{P}_n + \delta\delta_z, \quad \text{for } z \in \mathcal{Z},\ \delta \in (0,1).$$

Here, δ_z denotes the Dirac measure at the point $z \in \mathcal{Z}$. This defines a local contamination neighborhood of $\widehat{P}_n$ along individual data points. The exact contaminated points are unknown.

We consider the distributionally robust model averaging estimator $w^\star$ as the solution to the following problem:

$$w^\star \in \arg\min_{w \in \Delta^M} \mathcal{R}(w; \widehat{P}_n^{\delta}), \tag{1}$$

for the empirical risk $\mathcal{R}(w; \widehat{P}_n^{\delta}) := \mathbb{E}_{\widehat{P}_n^{\delta}}[\ell(z'; w)]$ and the loss is defined as

$$\ell(z'; w) := \left(y' - \sum_{m=1}^{M} w_m \hat{y}_m(x') \right)^2,$$

and each $\hat{y}_m(x)$ is the m^{th} model predictor. The set Δ^M denotes the standard $M-1$-dimensional simplex over model weights.

This estimator serves as the benchmark for our robustification analysis. In particular, we approximate its behavior using Gâteaux (functional) expansions and influence function representations of the risk.

Remark 4.1 (Adversarial Contamination and Envelope Theorems). The definition (1) considers contamination at an arbitrary fixed point $z \in \mathcal{Z}$. This local formulation allows us to derive clean Gâteaux expansions and influence–function representations of the resulting estimators.

In practice, one may be interested in the more conservative adversarial formulation

$$\sup_{z \in \mathcal{Z}} \mathbb{E}_{\widehat{P}_n^\delta(z)}[\ell(z'; w)],$$

which seeks protection against the worst–case contamination direction. Under mild regularity assumptions on $\ell(z; w)$ (continuity in z, differentiability in w), Danskin's theorem and related envelope results (cf. (Rockafellar and Wets, 1998, Ch. X)) can be used to characterize the (sub)gradient of the supremum. In particular, if the maximizer $z^\star(w)$ is unique and varies continuously with w, then

$$\nabla_w \sup_z \mathbb{E}_{Q_z^\delta}[\ell(z'; w)] = \nabla_w \mathbb{E}_{Q^\delta_{z^\star(w)}}[\ell(z'; w)],$$

whereas in the non-unique case the subdifferential is the convex hull of the gradients at all active maximizers. Thus the influence–function arguments developed below for fixed z extend to adversarial contamination by replacing ordinary gradients with these (sub)gradients or by working with a smoothed sup operator.

Remark 4.2 (Wasserstein Neighborhoods and Conservative Robustification). An alternative more conservative distributional robustness formulation, potentially more tractable computationally, replaces the pointwise Dirac contamination with a Wasserstein ball:

$$\mathcal{B}_\delta(\widehat{P}_n) := \left\{ Q \in \mathcal{P}_1(\mathcal{Z}) : W_1(Q, \widehat{P}_n) \leq \delta \right\},$$

where W_1 is the 1-Wasserstein distance on $\mathcal{Z}$ and $\mathcal{P}_1(\mathcal{Z})$ is the space of Borel probability measures with finite first moments. Under mild assumptions (e.g., Lipschitz losses), this leads to a dual representation of the robust risk via penalization (cf. Gao et al. (2017)). Moreover, the optimization over $\mathcal{B}_\delta(\widehat{P}_n)$ is amenable to discretization via linear programming, as discussed in (Galichon, 2016, Ch. 3).

Although Wasserstein-based DRO may be too conservative for certain contamination scenarios, it can provide outer approximations to the localized model in (1), and serve as a tractable bound–see Blanchet et al. (2024).

4.2 Approximate Optimality

We now formalize the sense in which any of the estimators considered above can be interpreted as an approximate minimizer of the optimal (infeasible) criterion introduced in the previous section, albeit with different optimization errors.

Let $J(w; \widehat{P}_n^\delta)$ denote a general criterion defined over weights $w \in \Delta^M$, evaluated on the contaminated $\widehat{P}_n^\delta$. Our goal is to identify sufficient conditions under which minimizing $J(w; \widehat{P}_n^\delta)$ leads to predictive performance close to that of the optimal estimator defined in (1). To do this, we employ among others influence function expansions, along with properties of strong convexity.

The following theorem presents the sufficient conditions along with lower and upper bounds for the empirical risk discrepancy between the optimal averaging estimator and the approximation–in what follows $\rightsquigarrow$ denotes weak convergence:

Theorem 4.3 (General Approximation Theorem under Local Contamination)**.** *Let*

$$w_J^\delta \in \arg\min_{w\in\Delta^M} J(w;\widehat{P}_n^\delta), \qquad w^\star \in \arg\min_{w\in\Delta^M} \mathcal{R}(w;\widehat{P}_n^\delta)$$

denote, respectively, the minimizers of a general criterion J and of the robust risk $\mathcal{R}$. Suppose the following hold:

(A1) Uniform Strong Convexity and Differentiability. $J(\cdot;P)$ is Gâteaux differentiable and α–strongly convex on Δ^M, for each P in a contamination neighbourhood of $\widehat{P}_n$. $\mathcal{R}(\cdot;P)$ is Gâteaux differentiable, $\alpha^\star$–strongly convex and β-smooth on Δ^M, in this neighborhood. The strong convexity and smoothness indices α, $\alpha^\star$ and β are independent of n.

(A2) Influence Function Regularity. For each $w\in\Delta^M$, the directional influence functions

$$\mathrm{IF}_J(z;w) := \lim_{\delta\to 0}\frac{1}{\delta}\big(J(w;\widehat{P}_n^\delta) - J(w;\widehat{P}_n)\big), \quad \mathrm{IF}_{\mathcal{R}}(z;w) := \lim_{\delta\to 0}\frac{1}{\delta}\big(\mathcal{R}(w;\widehat{P}_n^\delta) - \mathcal{R}(w;\widehat{P}_n)\big)$$

exist, are bounded uniformly over z, and their weak derivatives $\nabla_w\mathrm{IF}_J, \nabla_w\mathrm{IF}_{\mathcal{R}}$ belong to a Sobolev space $W^{1,2}$ with uniformly bounded norms.

(A3) Vanishing Local Contamination $\delta = \delta_n \to 0$ as $n\to\infty$.

(A4) Interior Minimizers and Hessians. The minimizers w_J^δ and $w^\star$ lie in the interior of Δ^M. Let $H_{\mathcal{R}} = \nabla_w^2\mathcal{R}(w_0;P_0)$ be the limiting Hessian of $\mathcal{R}$ at P_0, the uncontaminated weak limit of $\widehat{P}_n$ as $n\to\infty$, and at $w_0 = \arg\min_{w\in\Delta^M} J(w,P_0)$, where w_0 is also interior. Also, $H_J = \nabla_w^2\mathcal{J}(w_0;P_0)$ is defined analogously w.r.t. J. Assume

$$\nabla_w^2\mathcal{R}(w_0;\widehat{P}_n^\delta) \rightsquigarrow H_{\mathcal{R}}, \quad \nabla_w^2 J(w_0;\widehat{P}_n^\delta) \rightsquigarrow H_J,$$

in operator norm as $n\to\infty$.

Then, as $\delta\to 0$:

i Upper bound.

$$\mathcal{R}(w_J^\delta;\widehat{P}_n^\delta) - \mathcal{R}(w^\star;\widehat{P}_n^\delta) \le \frac{1}{2\alpha^\star}\big\|\partial_w\mathcal{R}(w_0;\widehat{P}_n)\big\| + \frac{\delta}{2\alpha^\star}\Big\|\partial_w\mathrm{IF}_{\mathcal{R}}(z;w_0,\widehat{P}_n) - \alpha^\star H_J^{-1}\partial_w\mathrm{IF}_J(z;w_0,\widehat{P}_n)\Big\| + o_p(\delta). \tag{2}$$

ii Lower bound.

$$\mathcal{R}(w_J^\delta;\widehat{P}_n^\delta) - \mathcal{R}(w^\star;\widehat{P}_n^\delta) \ \ge \frac{1}{2\alpha^\star}\left\|\delta\cdot H_J^{-1}\partial_w\mathrm{IF}_J(z;w_0,\widehat{P}_n) - H_{\mathcal{R}}^{-1}\Big(\partial_w\mathcal{R}(w_0,\widehat{P}_n) + \partial_w\mathrm{IF}_{\mathcal{R}}(z;w_0,\widehat{P}_n)\Big)\right\|^2 + o_p(\|\partial_w\mathcal{R}(w_0,\widehat{P}_n) + \partial_w\mathrm{IF}_{\mathcal{R}}(z;w_0,\widehat{P}_n)\|^2) + o_p(\delta^2). \tag{3}$$

Proof. We expand the risk difference using Taylor's theorem and strong convexity. First, using the differentiability of $\mathcal{R}$:

$$\mathcal{R}(w_J^\delta; \widehat{P}_n^\delta) = \mathcal{R}(w_0; \widehat{P}_n^\delta) + \partial_w \mathcal{R}(w_0; \widehat{P}_n^\delta)^\top v + O_p(\|v\|^2),$$

where $v := w_J^\delta - w_0$ and the asymptotic order of the remainder is justified by the β-smoothness assumption in (A1). We also obtain the influence function representation:

$$\partial_w \mathcal{R}(w_0; \widehat{P}_n^\delta) = \partial_w \mathcal{R}(w_0; \widehat{P}_n) + \delta \cdot \partial_w \mathrm{IF}_{\mathcal{R}}(z, w_0, \widehat{P}_n) + o_p(\delta),$$

where the asymptotic order of the remainder is justified by (A3). Similarly, under first-order optimality of w_J^δ for J, we have:

$$\mathbf{0} = \partial_w J(w_J^\delta; \widehat{P}_n^\delta) = \partial_w J(w_0; \widehat{P}_n) + \delta \cdot \partial_w \mathrm{IF}_J(z, w_0, \widehat{P}_n) + o_p(\delta).$$

Solving this for v, we get:

$$v = -\delta \cdot H_J^{-1} \partial_w \mathrm{IF}_J(z, w_0, \widehat{P}_n) + o_p(\delta).$$

Substituting into the expansion of $\mathcal{R}(w_J^\delta; \widehat{P}_n^\delta)$ we get:

$$\mathcal{R}(w_J^\delta; \widehat{P}_n^\delta) - \mathcal{R}(w^\star; \widehat{P}_n^\delta) = \mathcal{R}(w_0; \widehat{P}_n^\delta) - \mathcal{R}(w^\star; \widehat{P}_n^\delta) - \delta \cdot \partial_w \mathcal{R}(w_0; \widehat{P}_n^\delta)^\top H_J^{-1} \partial_w \mathrm{IF}_J(z, w_0, \widehat{P}_n) + o(\delta).$$

The upper is obtained by employing the Polyak-Lojasiewicz inequality–see for example Zhou (2018)–on the discrepancy $\mathcal{R}(w_0; \widehat{P}_n^\delta) - \mathcal{R}(w^\star; \widehat{P}_n^\delta)$, then using the influence function expansion and collecting terms of order $O_p(\delta)$; the lower bound utilizes that $w^\star$ is interior, and employs the definition of strong convexity to obtain a bound of the form $\mathcal{R}(w_J^\delta; \widehat{P}_n^\delta) - \mathcal{R}(w^\star; \widehat{P}_n^\delta) \geq \frac{1}{2\alpha^\star}\|w_J^\delta - w_\star\|^2$. It then utilizes the asymptotic expression for v above, along with an analogous expression obtained from the focs $\partial_w \mathcal{R}(w_J^\delta; \widehat{P}_n^\delta) = \mathbf{0}$ by expanding it around w_0, and then insering the influence function expansion. □

The results assumes that the (local) contamination level is infinitesimal (e.g., of order δ)–see for example Ch. 11 in Huber and Ronchetti (1981); contamination becomes negligible as the sample size augments. In this strong convexity framework, the optimality bounds reflect the proximity of the uncontaminated J-optimizer w_0 to be the uncontaminated optimum for the risk criterion $\mathcal{R}(w, \widehat{P}_n)$, as well first order influence functions gaps between the involved criteria (and estimators), present due to contamination. The level of strong convexity for the risk criterion, namely $\alpha^\star$, mitigates the influence function gaps' weight on approximate optimality. Notice that w_0 need not have a deterministic limit in distribution as $n \to \infty$, a typical case in several model averaging estimators; see for example Zhang and Liu (2019).

4.3 General Approximation in Second-Order Procedures

Second-order estimators are constructed by minimizing a surrogate validation loss function defined over model averaging weights $w \in \Delta^M$, where the loss itself arises as a convex or log-convex combination of base risk criteria. These estimators do not directly minimize predictive mean squared error (MSE), but rather approximate the behavior of a robust first-order estimator via optimization over suitably constructed composite objectives. This section establishes an asymptotic theory for the criterion weights used to construct these estimators as the validation sample size increases.

Let $Z_{n_\tau} := \{(x_t, y_t)\}_{t=1}^{n_\tau}$ denote the training sample, and $Z_{n_v} := \{(x_t, y_t)\}_{t=1}^{n_v}$ denote the validation sample; $\widehat{P}_{n_\tau}$ and $\widehat{P}_{n_v}$ denote the respective empirical distributions. We consider the model averaging basis criterion family $J_k(w; \widehat{P}_{n_\tau})$, for $k = 1, \dots, K$, evaluated at training. Recall that second-order procedures may operate not directly on these raw criteria but on transformed versions, such as

$$\widetilde{J}_k(w; \widehat{P}_{n_\tau}) = \log\big(J_k(w; \widehat{P}_{n_\tau}) - \min_w J_k(w; \widehat{P}_{n_\tau})\big),$$

to induce criterion selection curvature; recall the case of the Dirichlet-Nash estimator.

Given these transformed criteria $\{\widetilde{J}_k(w; \widehat{P}_{n_\tau})\}_{k=1}^K$, and a criteria weight vector λ, the training sample multiobjective averaging criterion can be defined as $J_{n_\tau}(w; \lambda, \widehat{P}_{n_\tau}) := \sum_k \lambda_k \widetilde{J}_k(w; \widehat{P}_{n_\tau})$. For the validation data Z_{n_τ}, let $\mathcal{L}(w, \widehat{P}_{n_\tau}, \widehat{P}_{n_v}) := \frac{1}{n_v}\sum_{t=1}^{n_v}(y_t - \sum_{m=1}^M w_m \beta_m'(Z_{n_\tau}) x_{m,t})^2$ denote the predictive validation error associated with a given weight vector w. This depends on the training sample through the training part OLS estimators used for the formation of the validation fold prediction. Then the optimal criterion weights $\lambda_n^{\text{opt}} \in \Delta^K$ solve:

$$\lambda_{n_v,n_\tau}^{\text{opt}} \in \arg\min_{\lambda\in\Delta^K} \mathcal{L}(w_{n_\tau}(\lambda), \widehat{P}_{n_\tau}, \widehat{P}_{n_v}),\ w_{n_\tau}(\lambda) \in \arg\min_{w\in\Delta^M} J_{n_\tau}(w; \lambda, \widehat{P}_{n_\tau}),$$

while the second order estimator solves:

$$w_n^{(2)} \in \arg\min_{w\in\Delta^M} J(w; \lambda_{n_v,n_\tau}^{\text{opt}}, \widehat{P}_n).$$

In the analysis, every empirical distribution present is allowed to be contaminated similarly to $\widehat{P}_n^\delta$; the respective contamination levels are denoted by δ_{n_τ} and δ_{n_v}. The local contamination framework allows for any of them to be zero. The key object of interest is the behavior of $\lambda_{n_v,,n_\tau}^{\text{opt}}$ for large enough $n^\star(n) := \min(n_v, n_\tau)$. In particular, we are interested in the impact on the influence function gaps that appear in the bounds of Theorem 4.3 of the training and validation folds contamination through the estimation of the optimal criterion weights:

Theorem 4.4 (Influence Function Expansions under Fold-Wise Contamination)**.** *Assume:*

(B1) The training and validation folds are contaminated at levels δ_{n_τ} and δ_{n_v} satisfying $\delta_{n_\tau}, \delta_{n_v} = o(\delta)$, where δ is the contamination level appearing in Theorem 4.3.

(B2) For large enough $n^\star(n)$, the mapping $(\widehat{P}_{n_\tau}, \widehat{P}_{n_v}) \mapsto \lambda_{n_v,n_\tau}^{\text{opt}}$ is Gâteaux differentiable and locally Lipschitz (in the product weak topology).

Then for large enough $n^\star(n)$,

$$\partial_w \mathrm{IF}_{J(\lambda_{n_v,n_\tau}^{\text{opt}})}(z; w_0, \widehat{P}_n) = \partial_w \mathrm{IF}_{J_n(\bar{\lambda})}(z; w_0, \widehat{P}_n) + O_p(\max(\delta_{n_\tau}, \delta_{n_v})), \tag{4}$$

$$H_{J_n(\lambda_{n_v,n_\tau}^{\text{opt}})}(w_0) = H_{J_n(\bar{\lambda})}(w_0) + O_p(\max(\delta_{n_\tau}, \delta_{n_v})), \tag{5}$$

and the bounds of Theorem 4.3 remain valid with these $O_p(\delta)$ modifications.

Proof. (B1) and the Lipschitz property in (B2) imply,

$$\|\lambda^{\mathrm{opt}}_{n_v,n_\tau} - \bar{\lambda}\| = O_p(\max(\delta_{n_\tau}, \delta_{n_v})) = o_p(\delta).$$

By linearity of differentiation in λ, and the Gateaux differentiability in (B2), along with (B1),

$$\partial_w \mathrm{IF}_{J_n(\lambda^{\mathrm{opt}}_{n_v,n_\tau})} = \sum_k \lambda^{\mathrm{opt}}_{n_v,n_\tau,k}\, \partial_w \mathrm{IF}_{\widetilde{J}_k} + O_p(\max(\delta_{n_\tau}, \delta_{n_v})),$$

which-along with the previous, yields (4); the same argument applies to Hessians, giving (5). □

Substituting these into the bounds of Theorem 4.3 shows that fold-wise contamination affects the influence-function terms only through $O_p(\max(\delta_{n_\tau}, \delta_{n_v}))$ corrections, which vanish faster than the $O_p(\delta)$ contamination level of interest:

Proposition 4.5 (Finite-Sample Oracle Interpretation and Improved Approximation Guarantees)**.** *Consider the premises of Theorems 4.3 and 4.4. Let* $w_0(\lambda) \in \arg\min_{w\in\Delta^M} J(w; \lambda, \widehat{P}_n)$ *be the training-sample minimizer of the composite criterion* $J(w; \lambda, \widehat{P}_n) = \sum_k \lambda_k \widetilde{J}_k(w; \widehat{P}_{n_\tau})$*, and let* $\bar{\lambda} = \lambda^{\mathrm{opt}}_{n_v,n_\tau}|_{\delta_{n_\tau}=\delta_{n_v}=0}$ *be the uncontaminated validation-optimal criterion weight vector. Then, for large enough* $n^\star(n) = \min(n_\tau, n_v)$*:*

(i) Empirical oracle characterization. Conditional on the observed folds, $\bar{\lambda}$ *minimizes the empirical validation loss* $\mathcal{L}(w_{n_\tau}(\lambda), \widehat{P}_{n_\tau}, \widehat{P}_{n_v})$ *up to* $O_p(\max(\delta_{n_\tau}, \delta_{n_v}))$*, and therefore its associated minimizer* $w_0(\bar{\lambda})$ *satisfies the local risk optimality inequality*

$$\left\|\partial_w \mathcal{R}(w_0(\bar{\lambda}); \widehat{P}_n)\right\| \le \left\|\partial_w \mathcal{R}(w_0(\lambda); \widehat{P}_n)\right\| + O_p(\max(\delta_{n_\tau}, \delta_{n_v})), \qquad \forall \lambda \in \Delta^K.$$

(i) Improved approximation bounds. Substituting $J(w; \bar{\lambda})$ *into the upper and lower bounds of Theorem 4.3 yields smaller influence functions involving terms in both inequalities.*

Proof. By definition of $\lambda^{\mathrm{opt}}_{n_v,n_\tau}$ and the assumption $\delta_{n_\tau}, \delta_{n_v} = o(\delta)$, we have for every $\lambda \in \Delta^K$:

$$\mathcal{L}(w_{n_\tau}(\bar{\lambda}), \widehat{P}_{n_\tau}, \widehat{P}_{n_v}) \le \mathcal{L}(w_{n_\tau}(\lambda), \widehat{P}_{n_\tau}, \widehat{P}_{n_v}) + O_p(\max(\delta_{n_\tau}, \delta_{n_v})).$$

Since the validation loss $\mathcal{L}$ is strongly convex in w and consistent (up to $O_p(\delta)$) with $\mathcal{R}$, its gradient at the corresponding minimizer satisfies

$$\left\|\partial_w \mathcal{R}(w_0(\bar{\lambda}); \widehat{P}_n)\right\| \le \left\|\partial_w \mathcal{R}(w_0(\lambda); \widehat{P}_n)\right\| + O_p(\max(\delta_{n_\tau}, \delta_{n_v})),$$

establishing part (i).

For (ii), Theorem 4.4along with the derivation of the bounds in the proof of Theorem 4.3, and taking into account the strong convexity of $\mathcal{R}(w, \widehat{P}_n)$ and the result in the first part of the theorem yield the claimed inequalities. □

The result says that $w_0(\bar{\lambda})$ is the empirical minimizer whose first-order optimality gap for $\mathcal{R}$ is smallest among all convex combinations of the base criteria, up to contamination terms of smaller order. Also, the composite criterion $J(w;\bar{\lambda}) = \sum_k \bar{\lambda}_k \widetilde{J}_k(w)$ provides, in finite samples, the best local surrogate of the empirical robust risk $\mathcal{R}(w;\widehat{P}_n)$ among all convex combinations of base criteria, leading to the tightest contamination-robust approximation bounds. Hence, under the assumption framework above, second-order estimators that aggregate base risk criteria via validation-optimal weights (e.g., λ^{opt} estimated from cross-fold minimization of out-of-sample predictive loss) inherits the asymptotic approximation behavior of the oracle weight vector $\bar{\lambda}$. Thereby, the second-order estimators (such as ℓ_1^{opt} and Dirichlet–Nash) attain for large enough n tighter upper and lower risk bounds than any fixed-λ or basis estimator, given that any first order estimator can be approximated as a minimizer of $J(w;\lambda,\widehat{P}_n)$ for some fixed choice of the criterion weights vector.

4.4 Discussion

4.4.1 Assumption frameowrk

The theoretical results established above rely on a number of high-order structural assumptions that merit further elaboration.

The core results—Theorems 4.3 and 4.4—require that the empirical robust risk $\mathcal{R}(w;\widehat{P}_n)$ and the composite criterion $J(w;\lambda,\widehat{P}_{n_\tau}) = \sum_k \lambda_k \widetilde{J}_k(w;\widehat{P}_{n_\tau})$ are both strongly convex and twice differentiable with respect to the weight vector $w \in \Delta^M$, for all λ in the interior of the simplex Δ^K. The validity of these assumptions hinges on both the form of the loss functions and the regularity of the model space. In particular, if the cross-model prediction vectors are linearly independent across models—something ensured when the models differ in variables or interaction terms—then the Hessian of $\mathcal{R}$ is strictly positive definite on Δ^M, even when the models are misspecified. Thus, the assumption of strong convexity for $\mathcal{R}$ can be justified for large enough n under mild regularity of the design distribution. [2] Then the form of the loss functions (or their logarithmic transforms) actually used in any of the estimators examined above, implies strong convexity, at least for large enough n. (Gateux) Differentiablity is valid for each criterion used, except for the case of the AB estimator:

Remark 4.6. While convex in w, the AB criterion involves a non-differentiable maximum over convex functions and thus may fail to satisfy the smoothness assumptions underpinning the second-order approximation theory developed in Section 4.3. In particular, the required Hessian behaviors may not be available, as the pointwise maximum introduces kinks at points where multiple J_k coincide. Nevertheless, two perspectives are relevant:

First, one may introduce a smooth relaxation of the AB criterion using the log-sum-exp function:

$$J^{\text{AB,soft}}(w;\tau) := \frac{1}{\tau}\log\left(\sum_{k=1}^{K}\exp\left(\tau\left[J_k(w) - \min_{w'} J_k(w')\right]\right)\right), \quad \tau > 0.$$

[2] Notice that such a requirement typically precludes asymptotic analysis in nested linear models; many model averaging estimators have then stochastic limits–see for example Zhang and Liu (2019), which implies either stochastic limiting criteria and/or asymptotic non identification–the latter should preclude strong convexity under deterministic criterion limits.

This function is differentiable and strongly convex under mild assumptions on the J_k, and converges pointwise to the AB criterion as $\tau \to \infty$. Theorems 4.3-4.4 can be applied to this smooth approximation.

Second, even without theoretical inclusion, the AB estimator can be evaluated empirically. Our Monte Carlo simulations indicate that second-order procedures such as ℓ_1^{opt} and Dirichlet–Nash consistently outperform AB in predictive loss and stability under contamination.

In any case, a formal theoretical comparison between the second order estimators approximate optimality and AB, or the estimators briefly described as extensions in Paragraph 3.4 is left for future research. Notice though, that the Monte Carlo simulations below, provide some evidence on that the second-order framework yields practical dominance over AB in terms of approximate optimality for $\mathcal{R}(w; \widehat{P}_n)$.

The statements of Theorem 4.3 were given under the interiority assumption for clarity. In practice many model averaging estimators (in particular sparsity-inducing procedures or those that concentrate mass on a few base models) place the optimizer on the boundary of the simplex. Below we summarise how the previous expansions and bounds are modified in the boundary (subdifferential) case and what additional regularity is needed to recover first-order approximations.

1) Optimality and subgradients. If w_0 lies on the boundary of Δ^M then the first-order optimality conditions are expressed in terms of subgradients: for the criterion J we require

$$0 \in \partial_w J(w_J^\delta; \widehat{P}_n^\delta) + N_{\Delta^M}(w_J^\delta),$$

where $\partial_w J$ denotes the (Clarke or convex) subdifferential and $N_{\Delta^M}(w)$ is the normal cone to the simplex at w–see for example Rockafellar and Wets (1998). Analogous conditions hold for $\mathcal{R}$ and its minimizer $w^\star$. All gradient-type objects in the interior proof should be replaced by appropriate selections from the subdifferentials.

2) Tangent-space projection and pseudo-inverse. Let T_{w_0} denote the tangent space of Δ^M at w_0 (the affine subspace tangent to the active constraints) and Π_T the orthogonal projector onto T_{w_0}. Under the usual constraint qualification (e.g. strict complementarity / active-set stability—see below), one can restrict all derivatives to T_{w_0} and work with the reduced (projected) Hessian. Define

$$H_T := \Pi_T \, \nabla_w^2 J(w_0; \widehat{P}_n) \, \Pi_T, \qquad H_{\mathcal{R},T} := \Pi_T \, \nabla_w^2 \mathcal{R}(w_0; \widehat{P}_n) \, \Pi_T,$$

which are positive definite on T_{w_0} under a projected strong-convexity assumption. Then the inverse H_T^{-1} above is to be understood as the inverse on the tangent subspace (equivalently a Moore–Penrose pseudo-inverse of the full Hessian restricted to T_{w_0}).

3) Projected first-order expansion. If the active set is stable for small contamination (i.e. the set of indices with zero weight at w_0 remains identical for w_J^δ and $w^\star$ for sufficiently small δ), and if one can choose measurable selections $g_J \in \partial_w J$ and $g_{\mathcal{R}} \in \partial_w \mathcal{R}$ that are Gâteaux differentiable in the directions considered, then the expansion in projected coordinates reads

$$\Pi_T(w_J^\delta - w_0) = -\delta \, H_T^{-1} \, \Pi_T \, \mathbb{E}_Q\big[\nabla_w \mathrm{IF}_J(z; w_0, \widehat{P}_n)\big] + o(\delta).$$

An entirely analogous expression holds for $\Pi_T(w^\star - w_0)$ with $H_{\mathcal{R},T}^{-1}$ and $\mathrm{IF}_{\mathcal{R}}$. All risk-difference bounds in the interior case continue to hold upon replacing Hessians and gradient/influence objects by their projected counterparts; constants such as $1/\alpha^\star$ are computed with respect to the smallest eigenvalue of the projected Hessian $H_{\mathcal{R},T}$.

The assumptions about the limiting Hessians can be justified in frameworks of stationarity/ergodicity or appropriate exhangeability for the undelying stochastic processes.

Finally, the results in Theorem 4.4 hinge on the Lipschitz continuity and Gâteaux differentiability of the mapping $(\widehat{P}_{n_\tau}, \widehat{P}_{n_v}) \mapsto \lambda^{\mathrm{opt}}_{n_v,n_\tau}$. These properties hold whenever the validation loss function $\mathcal{L}$ is smooth and strictly convex in w, and the inner minimization over $w_{n_\tau}(\lambda)$ is well-posed with a unique minimizer for each λ. In the present setup, $\mathcal{L}$ is a quadratic function of w evaluated over an independent validation sample, and the minimizer $w_{n_\tau}(\lambda)$ exists and is unique due to the strong convexity of $J(w;\lambda)$ in w. The composite functional $\lambda \mapsto \mathcal{L}(w_{n_\tau}(\lambda), \cdot)$ is thus continuously differentiable on the interior of Δ^K, ensuring differentiability of the minimizer λ^{opt} with respect to the underlying data distribution.

4.4.2 Higher-Order Expansions

The influence–function expansions established in Theorem 4.3 are first–order in the local contamination level δ. Extending the theory to higher orders requires uniform control of third derivatives of both $\mathcal{R}$ and J with respect to w. Specifically, if $\mathcal{R}(w; \widehat{P}_n^\delta)$ admits a third Gâteaux derivative that is continuous in (w, δ) and uniformly bounded on a neighborhood of w_0, then the remainder in the Taylor expansion of $\mathcal{R}$ satisfies

$$\mathcal{R}(w_J^\delta; \widehat{P}_n^\delta) = \mathcal{R}(w_0; \widehat{P}_n) + \delta\, \partial_w \mathrm{IF}_{\mathcal{R}}(z; w_0, \widehat{P}_n)^\top v + \tfrac{1}{2} v^\top H_{\mathcal{R}} v + O_p(\delta^2),$$

with $v = -\delta\, H_J^{-1} \partial_w \mathrm{IF}_J(z; w_0, \widehat{P}_n) + o_p(\delta)$. This refinement yields $O_p(\delta^2)$ accuracy of the risk difference and would permit sharper bounds distinguishing between estimators whose first–order influence functions coincide but whose higher–order curvature corrections differ. In practice, such results are relevant for studying bias corrections and second–order robustness of validation-based estimators.

4.4.3 Third-Order Estimators and Functional Extensions

The second-order estimators analyzed above minimize a validation-weighted composite criterion and can thus be viewed as functionally adaptive convex aggregators of first-order procedures. A natural next step is to consider third-order estimators, constructed as convex combinations of second-order criteria themselves. Formally, letting $J^{(2)}_{\ell_1}$ and $J^{(2)}_{\mathrm{DirNash}}$ denote the convex and log-convex aggregators, respectively, one may define

$$J^{(3)}(w;\theta) = \theta J^{(2)}_{\ell_1}(w) + (1-\theta) J^{(2)}_{\mathrm{DirNash}}(w), \qquad \theta \in [0,1],$$

and consider $w^{(3)} = \arg\min_{w \in \Delta^M} J^{(3)}(w; \theta^{\mathrm{opt}})$, where θ^{opt} is selected by minimizing an out-of-sample predictive loss or a functional discrepancy with the robust risk $\mathcal{R}(w)$. Such a construction increases the expressive capacity of the estimator space while preserving convexity, and can be interpreted as optimizing within a higher-dimensional convex hull of influence functions.

At a conceptual level, the hierarchy of first-, second-, and third-order estimators corresponds to successive approximations of the ideal robust estimator $w^\star$ through functional projections of terms in the expansion of its uncontaminated risk and of its influence function regarding local contamination. Each level in the hierarchy expands the functional span of available approximations, from single criteria, to convex mixtures, and finally to mixtures of mixtures, approaching the ideal of a universal approximation of $\mathrm{IF}_{\mathcal{R}}(z; w)$ within the class of the available smooth convex risk functionals. These ideas suggest that higher-order estimators, possibly regularized or constrained by entropy or Wasserstein penalties, could attain improved robustness properties.

5 Monte Carlo Experiments: Design and Results

We now present a Monte Carlo study designed to evaluate the performance of various model averaging estimators under covariate contamination. The abovementioned model averaging estimators are considered, namely the basis estimators (such as JMA, MMA, smooth AIC/BIC, and naive averaging) as well as their multiobjective counterparts (ℓ_1, ℓ_1^{opt}, ℓ_∞, Nash, AB, and Dirichlet-Nash). We emphasize that our focus is on predictive performance, not parameter estimation per se.

5.1 Design Setup

Each data-generating process (DGP) is constructed using a linear model with $p = 10$ regressors and an additive noise term. Covariates are drawn from a mixture of clean and contaminated distributions to simulate persistent measurement error. Specifically, let $X_t = (1-\delta)Z_t + \delta C_t$,[3] where Z_t is clean and C_t is contaminated, and $\delta \in 0.0, 0.5, 1.5$ denotes the contamination strength.

We consider three sample sizes: $T \in 100, 500, 1000$. For each T, the data is split into 70% training and 30% test sets. The training set is used to compute model averaging weights, while the test set is used to evaluate predictive MSE, mean absolute error (MAE), squared bias, and entropy of the final weight vector (as a proxy for estimator sparsity).

Each Monte Carlo simulation is repeated for 500 replications. All estimators are evaluated under the same DGP realizations to enable paired comparisons. The figures in Figure 1 report the average performance metrics across all replications.[4]

[3] We are thus considering more complicated scenarios of contamination compared to the theory section.

[4] Additional simulation scenarios involving alternative forms of contamination (e.g., in the response variable or model misspecification) were also performed, yielding qualitatively analogous results. These are available from the authors upon request.

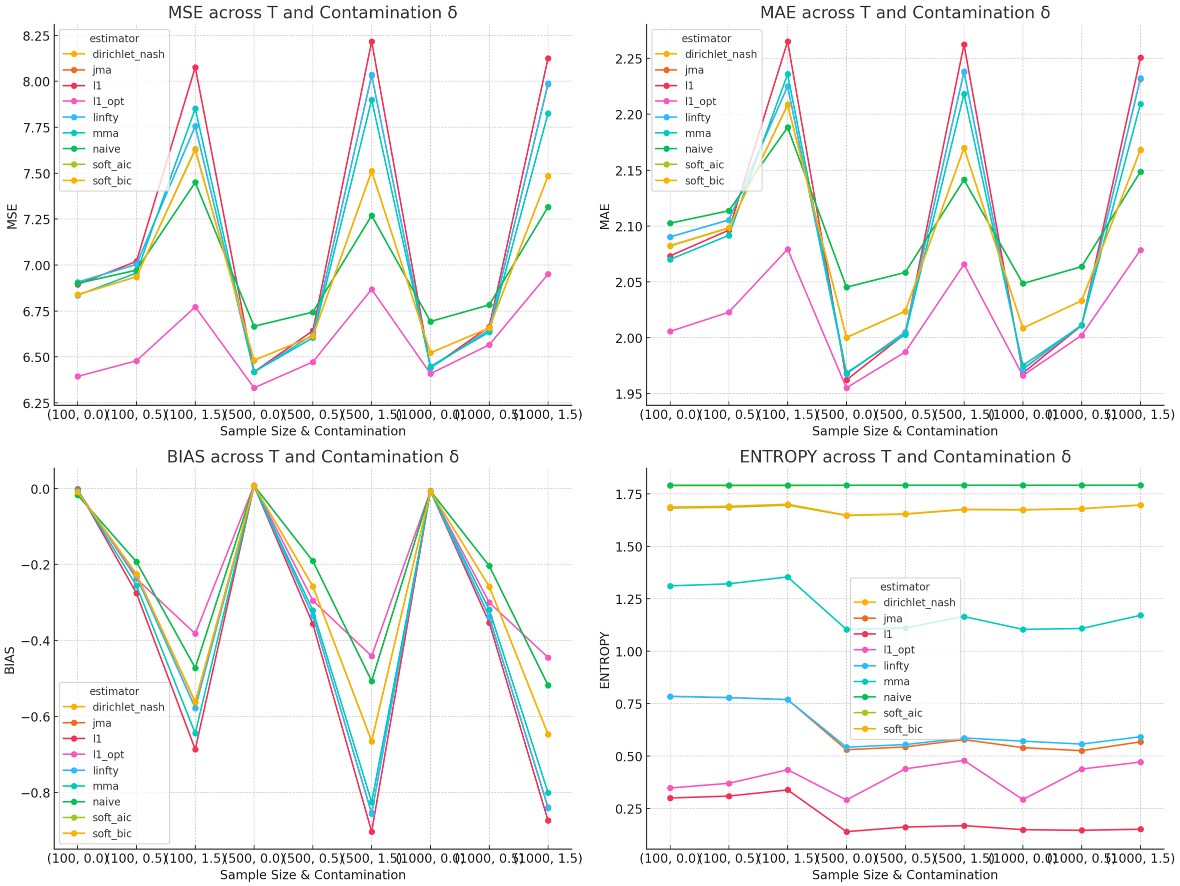

Figure 1: Predictive performance of all estimators across sample sizes T and contamination strengths δ. Top-left: MSE; Top-right: MAE; Bottom-left: Bias; Bottom-right: Entropy of weight vector.

5.2 Main Findings and Discussion

The most salient finding is the clear dominance of the ℓ_1^{opt} estimator (optimal criterion-weighted linear combination) in terms of predictive MSE and MAE. Across all (T, δ) configurations, ℓ_1^{opt} achieves the lowest or near-lowest average MSE and MAE. This suggests that criterion-based second-order selection, tuned on out-of-sample performance, provides substantial robustness against contamination and overfitting.

The Dirichlet-Nash estimator consistently performs as a near second-best method, particularly when contamination is high ($\delta = 1.5$). This is potentially related to the effectiveness of stochastic exponent search over multiobjective criteria. Importantly, Dirichlet-Nash maintains high performance while exhibiting greater entropy than ℓ_1^{opt}, indicating better weight dispersion and reduced sensitivity to extreme weights.

Standard estimators like JMA and MMA show more significant degradation under increasing contamination. Their limited robustness arises from being optimized relative to a single model fit rather than cross-validated predictions. Similarly, the naive average, while simple and stable, is uniformly suboptimal in terms of MSE.

Interestingly, the Nash estimator (unweighted geometric mean of criteria) achieves moderate robustness but is outperformed by its Dirichlet-weighted version in all configurations.

As expected, entropy metrics support the interpretation of ℓ_1, along with its adaptive version ℓ_1^{opt} as favoring sparsity. It consistently yields the lowest entropy, suggesting selection of a minimal effective subset of models. This aligns with the second-order tuning mechanism.

In summary, the simulation results suggest that predictive robustness can be enhanced by incorporating multiobjective model averaging mechanisms, especially those leveraging validation-driven selection over the simplex of criterion weights.

6 Empirical Application

This section presents an empirical application of the proposed robust multiobjective model averaging estimators in the context of cross-country growth regressions. This is a standard setting in the model averaging literature (see, e.g., Fernandez et al. (2001), Magnus et al. (2010), Liu (2015), Gunby et al. (2017)) due to the presence of high model uncertainty and limited sample sizes. The dataset—originally compiled by Magnus et al. (2010) and later reused by Liu (2015)—comprises 74 countries with a rich set of economic, institutional, and geographic indicators commonly used in explaining long-run economic growth.

We estimate the following canonical model:

$$\mathrm{GROWTH}_i = \mathbf{X}_{1i}\beta_1 + \mathbf{X}_{2i}\beta_2 + \varepsilon_i, \tag{6}$$

where GROWTH_i is the average growth rate of GDP per capita between 1960 and 1996. The core regressors $\mathbf{X}_1$ follow classical growth theory and include: (i) initial income level (GDP60), (ii) investment in equipment (INV), (iii) primary school enrollment (SCHOOL60), and (iv) life expectancy (LIFE60). The auxiliary regressors $\mathbf{X}_2$ capture structural, institutional, and geographic heterogeneity: (v) population growth (POP), (vi) rule of law (RULE), (vii) tropical location (TROPICS), (viii) ethnolinguistic fragmentation (ETHNO), and (ix) Confucian population share (CONFUC).

Although the primary goal of cross-sectional growth regressions is often structural interpretation, predictive performance is increasingly emphasized–see for example Brock and Durlauf (2001); Fernandez et al. (2001). Predictive success under model uncertainty serves two purposes: (1) it validates the explanatory content of the covariates under uncertainty and finite samples, and (2) it provides guidance for policy applications where forecasts of country growth outcomes are required. In this context, robust model averaging estimators are valuable, as they balance goodness-of-fit, overfitting control, and structural heterogeneity without requiring arbitrary model selection.

Model Hierarchy and Estimation Procedure

We define six nested linear models of increasing complexity, reflecting the incremental inclusion of auxiliary regressors beyond the core economic variables. Each model is estimated via OLS. The hierarchy is summarized below.

This model hierarchy reflects a natural ordering from purely economic fundamentals to increasingly institutional and cultural variables. However, this choice is obviously not unique. Given the aforementioned dichotomy of the regressors into core–which capture

Model	Included Variables
Model 1	GDP60, INV, SCHOOL60, LIFE60
Model 2	Model 1 + POP
Model 3	Model 2 + RULE
Model 4	Model 3 + TROPICS
Model 5	Model 4 + ETHNO
Model 6	Model 5 + CONFUC

Table 1: Hierarchy of Nested Models Used in the Averaging Procedure

basic economic determinants of growth, and auxiliary, the chosen nested structure is thus indicative. In principle, one could explore all $2^6 = 64$ possible combinations of the core model with subsets of the auxiliary regressors. Our choice reflects a tractable and interpretable progression of economic complexity and is consistent with the experimental setup used in several comparative studies in model averaging literature.

MA Estimation and Leave-One-Out Predictive Analysis

This section evaluates the robustness and predictive relevance of our model averaging estimators using a Leave-One-Out (LOO) cross-validation framework, applied to cross-country growth regressions. The LOO predictive analysis is motivated by the core theoretical contribution of this paper: the construction of multi-objective estimators that approximate the optimal solution to a robust prediction problem under contamination and model uncertainty.

In cross-sectional settings—such as the global growth regression dataset used here—parameter estimates are especially vulnerable to sample heterogeneity and local misspecification. To address this, we test whether model averaging estimators that combine information from different criteria (JMA, MMA, soft AIC/BIC, ℓ_1, AB, Dirichlet-Nash, etc.) improve predictive robustness.

Let $y_i \in \mathbb{R}$ denote the observed outcome (e.g., GDP growth) for country $i = 1, \dots, n$, with $n = 74$, and let $x_i \in \mathbb{R}^p$ be the corresponding covariate vector for country i.

For each estimator and each country i, the Leave-One-Out (LOO) prediction procedure proceeds as follows:

1. Remove country i from the dataset. This gives a training sample of size $n - 1 = 73$, denoted by $\{(x_j, y_j)\}_{j \neq i}$.
2. Re-estimate the model averaging weights and coefficients using this training sample. This involves:
    - Estimating each of the six base models on the reduced sample;
    - Computing the model-averaging weights specific to the estimator (e.g., JMA, ℓ_1, Nash) using only the 73 countries;
    - Generating the averaged coefficient vector $\hat{\beta}^{(-i)} \in \mathbb{R}^p$.

3. Predict the outcome for country i by computing:

$$\hat{y}_i^{(-i)} = x_i^\top \hat{\beta}^{(-i)}$$

where x_i is the covariate vector for the left-out country.

This procedure is repeated for each $i = 1, \ldots, n$, yielding a full set of leave-one-out predictions $\{\hat{y}_i^{(-i)}\}_{i=1}^n$ and corresponding LOO coefficient vectors $\{\hat{\beta}^{(-i)}\}_{i=1}^n$ for each estimator. The predictive performance of each estimator is then evaluated by aggregating the errors:

$$\text{LOO-MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{(-i)}\right)^2, \qquad \text{LOO-MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i^{(-i)}\right|$$

In addition to prediction metrics, we compute the standard deviation of the estimated coefficients across LOO folds, denoted $\text{SD}(\hat{\beta}_j)$, to assess coefficient stability and approximate standard errors. This allows for an approximate significance evaluation by comparing the average coefficient $\bar{\beta}_j$ to its LOO standard deviation:

$$t_j = \frac{\bar{\beta}_j}{\text{SD}(\hat{\beta}_j)}$$

Significance levels are then annotated using the conventional thresholds (e.g., * for $p < 0.10$, ** for $p < 0.05$, *** for $p < 0.01$), assuming approximate normality. This provides a way to assess the relative importance and robustness of the variables selected by each averaging estimator.

Overall, this setup ensures that each prediction and coefficient estimate is made using a model trained without access to the test observation, providing a stringent and unbiased estimate of out-of-sample generalization performance. The combination of LOO prediction errors and coefficient variability enables a deeper comparison of estimator robustness, sparsity, and inferential reliability under realistic data-limited scenarios.

Results. The results in Table 2 indicate that Dirichlet-Nash achieves the best predictive performance, both in MSE and MAE. Traditional estimators such as Naive, MMA, and soft AIC/BIC perform similarly, but exhibit mild inefficiencies in absolute prediction error. Estimators like ℓ_1 and AB, while competitive, appear to trade off predictive accuracy for increased sparsity or interpretability. Nash and generalized Nash retain competitive performance, supporting their theoretical appeal.

Table 2: Leave-One-Out Predictive Performance (MSE and MAE)

Estimator	**LOO-MSE**	**LOO-MAE**
naive	0.00014748	0.0088284
jma	0.00014408	0.0086185
mma	0.00014748	0.0088284
soft_aic	0.00014748	0.0088284
soft_bic	0.00014925	0.0089105
ℓ_1	0.00017375	0.0099765
ℓ_1^{opt}	0.00017391	0.0099816
ℓ_∞	0.00017375	0.0099765
AB	0.00016117	0.0093804
gen_nash	0.00014748	0.0088284
dirichlet_nash	**0.00012499**	**0.0076883**

Model Averaging Weights. Table 3 reports the model averaging weights for the six base models (m_1 to m_6) used in each estimator. Uniform weights (e.g., Naive, MMA) contrast with adaptive schemes (e.g., Dirichlet-Nash, AB), which exhibit selective weighting patterns.

Table 3: LOO-Averaged Model Weights per Estimator (rounded to 4 decimals)

Estimator	m_1	m_2	m_3	m_4	m_5	m_6
naive	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
jma	0.1766	0.0000	0.1150	0.0951	0.0006	0.6126
mma	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
soft_aic	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
soft_bic	0.1921	0.1811	0.1708	0.1610	0.1518	0.1432
ℓ_1	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ℓ_1^{opt}	0.8959	0.0938	0.0090	0.0012	0.0000	0.0000
ℓ_∞	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
AB	0.3136	0.2548	0.1960	0.1373	0.0785	0.0198
gen_nash	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
dirichlet_nash	0.3606	0.0698	0.0860	0.0964	0.0818	0.3054

Estimated Coefficients. To aid interpretation, Table 4 presents the top 5 coefficients (by absolute magnitude) for each estimator, revealing distinct patterns of variable selection and shrinkage.

Table 4: Top 5 Estimated Coefficients by Absolute Value (Per Estimator)

Estimator	**INV**	**POP**	**GDP60**	**RULE**	**CONFUC**
naive	0.1791^{***}	0.2227^{***}	-0.0173^{***}	0.0135^{***}	0.0094^{***}
jma	0.1591^{***}	0.2721^{***}	-0.0162^{***}	0.0150^{***}	0.0347^{***}
mma	0.1791^{***}	0.2227^{***}	-0.0173^{***}	0.0135^{***}	0.0094^{***}
soft_aic	0.1791^{***}	0.2227^{***}	-0.0173^{***}	0.0135^{***}	0.0094^{***}
soft_bic	0.1832^{***}	0.2061^{***}	-0.0172^{***}	0.0128^{***}	0.0081^{***}
ℓ_1	0.2405^{***}	0.0000	-0.0159^{***}	0.0000	0.0000
ℓ_1^{opt}	0.2397^{***}	-0.0012	-0.0160^{***}	0.0002	0.0000
ℓ_∞	0.2405^{***}	0.0000	-0.0159^{***}	0.0000	0.0000
AB	0.2033^{***}	0.1240^{***}	-0.0170^{***}	0.0090^{***}	0.0011^{***}
gen_nash	0.1791^{***}	0.2227^{***}	-0.0173^{***}	0.0135^{***}	0.0094^{***}
dirichlet_nash	0.1859^{***}	0.1921^{***}	-0.0165^{***}	0.0109^{***}	0.0173^{***}

The Dirichlet-Nash estimator delivers the best out-of-sample predictive performance among all model averaging schemes, as evidenced by its lowest LOO-MSE. Its coefficient estimates highlight the consistent importance of investment (INV), initial GDP (GDP60), population growth (POP), and institutional/cultural variables such as rule of law (RULE) and Confucian legacy (CONFUC). Notably, all five are statistically significant under the leave-one-out distribution, underscoring their robust predictive contribution. The predictive strength of institutional and cultural factors, especially CONFUC, suggests that long-term growth is possibly shaped not only by capital accumulation and convergence mechanisms but also by persistent historical and institutional structures. This is inline with analogous results in empirical studies emplying BMA–see for example Fernandez et al. (2001). The Dirichlet-Nash's success reflects its ability to flexibly balance multiple model criteria while regularizing the weight distribution. This provides a case for its use in cross-country growth forecasting exercises where model uncertainty and small-sample challenges are pervasive.

7 Further Research Directions

The theoretical and empirical results of this paper motivate a range of further developments in the modeling of robustness, aggregation, and predictive inference under model uncertainty. One natural extension concerns the treatment of contamination. The present analysis operates within the classical gross-error model, in which small fractions of covariate observations are corrupted by localized noise or adversarial distortion. However, many realistic settings involve more structured forms of contamination, such as latent omitted variables, misclassification in clusters, or regime-driven changes in the data-generating process. These types of uncertainty may be more appropriately modeled through distributional uncertainty sets, particularly those defined via Wasserstein distances or f-divergences. Extending the influence-function framework developed here to accommodate such robust formulations would yield a principled connection between our estimator classes and the literature on distributionally robust optimization. In particular, influence function expansions could serve as local approximations to the worst-case losses over these uncertainty sets, thereby offering computationally tractable characterizations of distributional robustness in model averaging.

In parallel, the current framework for multiobjective model averaging—grounded in the dominance relation induced by a vector of performance criteria—suggests deeper theoretical exploration of aggregation rules that respect or approximate Pareto efficiency in this induced partial order. While our analysis focuses on specific estimators such as the Nash and AB, both of which operationalize distinct compromise principles across the objective space, a general characterization of such estimators remains open. For instance, the Nash estimator minimizes a geometric mean of normalized excess losses, which evokes connections with bargaining solutions and variational inequalities in game theory. Formalizing this analogy may yield new classes of equilibrium-based aggregation procedures grounded in economic theory. Similarly, the AB estimator acts as a conservative envelope over model performances and can be interpreted as a Condorcet-consistent procedure in a social choice framework, where models are viewed as alternatives and criteria as evaluators. Axiomatizing such procedures could clarify under what conditions they ensure desirable aggregation properties such as monotonicity, neutrality, or resistance to domination reversal.

Further development is also warranted in the adaptive estimation of criterion weights. The optimal ℓ_1 and Dirichlet-Nash estimators rely on weight selection based on out-of-sample prediction errors, typically using cross-validated mean squared error. While effective in practice, this approach may be sensitive to sample partitioning and does not exploit the potential structure of validation residuals. Alternative weight inference methods based on empirical likelihood or moment conditions over residuals may offer greater robustness, interpretability, and statistical efficiency. These approaches could also support hierarchical weighting structures—for instance, assigning greater trust to criteria derived from large-sample or more stable estimates, or allowing weights to vary across model classes or groups of similar specifications. A comprehensive theoretical analysis of such adaptive schemes, particularly with respect to their robustness to sample splitting choices, would significantly enhance the statistical foundation of the methodology.

Empirically, the results obtained in the cross-sectional growth regression application provide several insights and raise additional questions. The good performance of the Nash, ℓ_1, and AB estimators in the presence of data contamination suggests they may be particularly suitable in real-world macroeconomic settings where data reliability varies across sources and variables. Moreover, while the application here is static and cross-sectional, many macroeconomic and financial environments are subject to time-varying structure. Regime shifts, structural breaks, or evolving data quality can modify the relative performance of candidate models and criteria. In such cases, dominance relations across objectives may shift over time, requiring estimators that can adapt accordingly. Extending the multiobjective framework to dynamic settings would allow for real-time model monitoring and estimator re-weighting based on rolling validation performance or structural diagnostics. Additionally, influence-based sensitivity analysis, already used here to assess robustness, could be repurposed to detect model drift or to diagnose changes in the local geometry of the model space that necessitate updates in the aggregation rule.

Taken together, these directions highlight the potential of the multiobjective model averaging framework not merely as a collection of robust estimators, but as a flexible and extensible paradigm for statistical decision-making under model uncertainty. Its integration with modern tools from robust statistics, dynamic forecasting, and economic aggregation theory could offer fertile ground for both theoretical development and practical application in high-stakes econometric modeling.

References

Acemoglu, D., S. Johnson, and J. A. Robinson (2001): "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review*, 91, 1369–1401.

Arvanitis, S., T. Post, and N. Topaloglou (2021): "Stochastic Bounds for Reference Sets in Portfolio Analysis," *Management Science*, 67, 7737–7754.

Aumann, R. J. and S. Hart, eds. (1992): *Handbook of Game Theory with Economic Applications*, vol. 2, Elsevier.

Barro, R. J. and J. W. Lee (2013): "A new data set of educational attainment in the world, 1950–2010," *Journal of Development Economics*, 104, 184–198.

Blanchet, J., J. Li, S. Lin, and X. Zhang (2024): "Distributionally robust optimization and robust statistics," *arXiv preprint arXiv:2401.14655*.

Brandl, F., F. Brandt, and H. G. Seedig (2016): "Consistent probabilistic social choice," *Econometrica*, 84, 1839–1880.

Brock, W. A. and S. N. Durlauf (2001): "What have we learned from a decade of empirical research on growth? Growth empirics and reality," *the world bank economic review*, 15, 229–272.

Claeskens, G. and N. L. Hjort (2003): "Frequentist model average estimators," *Journal of the American Statistical Association*, 98, 879–899.

——— (2008): *Model selection and model averaging*, Cambridge University Press.

Durlauf, S. N., P. A. Johnson, and J. R. Temple (2005): "Chapter 8 Growth Econometrics," Elsevier, vol. 1 of *Handbook of Economic Growth*, 555–677.

Fernandez, C., E. Ley, and M. F. J. Steel (2001): "Model uncertainty in cross-country growth regressions," *Journal of Applied Econometrics*, 16, 563–576.

Galichon, A. (2016): *Optimal transport methods in economics*, Princeton University Press.

Gao, R., X. Chen, and A. Kleywegt (2017): "Distributional robustness and regularization in statistical learning," *arXiv preprint*.

Gunby, P., Y. Jin, and W. R. Reed (2017): "Did FDI Really Cause Chinese Economic Growth? A Meta-Analysis," *World Development*, 90, 242–255.

Hansen, B. E. (2007): "Least squares model averaging," *Econometrica*, 75, 1175–1189.

Hansen, B. E. and J. Racine (2012): “Jackknife model averaging,” *Journal of Econometrics*, 167, 38–46.

Huber, P. J. and E. M. Ronchetti (1981): “Robust statistics, ser,” *Wiley Ser Probab Math Stat New York, NY, USA Wiley-IEEE*, 52, 54.

Hwang, C.-L. and A. S. M. Masud (2012): *Multiple Objective Decision Making: Methods and Applications*, vol. 164 of *Lecture Notes in Economics and Mathematical Systems*, Springer.

Kourtellos, A., T. Stengos, and C. M. Tan (2010): “Do institutions rule? The role of heterogeneity in the institutions vs. geography debate,” *Economics Bulletin*, 30, 1710–1719.

Liu, C. A. (2015): “Distribution theory of the least squares averaging estimator,” *Journal of Econometrics*, 186, 142–159.

Magnus, J. R., O. Powell, and P. Prüfer (2010): “A comparison of two model averaging techniques with an application to growth empirics,” *Journal of Econometrics*, 154, 139–153.

Navon, A., A. Shamsian, G. Chechik, and E. Fetaya (2020): “Learning the pareto front with hypernetworks,” *arXiv preprint arXiv:2010.04104*.

Rockafellar, R. T. and R. J. Wets (1998): *Variational analysis*, Springer.

Sachs, J. D. (2003): “Institutions Don’t Rule: Direct Effects of Geography on Per Capita Income,” Working Paper 9490, National Bureau of Economic Research.

Steel, M. F. J. (2020): “Model Averaging and Its Use in Economics,” *Journal of Economic Literature*, 58, 644–719.

Zhang, R. and D. Golovin (2020): “Random hypervolume scalarizations for provable multi-objective black box optimization,” in *International conference on machine learning*, PMLR, 11096–11105.

Zhang, X. and C. A. Liu (2019): “Inference after model averaging in linear regression models,” *Econometric Theory*, 35, 816–841.

Zhou, X. (2018): “On the fenchel duality between strong convexity and lipschitz continuous gradient,” *arXiv preprint arXiv:1803.06573*.

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ
ΟΙΚΟΝΟΜΙΚΩΝ
ΕΠΙΣΤΗΜΩΝ
SCHOOL OF
ECONOMIC
SCIENCES

ΤΜΗΜΑ
ΟΙΚΟΝΟΜΙΚΗΣ
ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF
ECONOMICS

Department of Economics
Athens University of Economics and Business

List of Recent Working Papers

2023

01-23 Real interest rate and monetary policy in the post Bretton Woods United States, George C. Bitros and Mara Vidali

02-23 Debt targets and fiscal consolidation in a two-country HANK model: the case of Euro Area, Xiaoshan Chen, Spyridon Lazarakis and Petros Varthalitis

03-23 Central bank digital currencies: Foundational issues and prospects looking forward, George C. Bitros and Anastasios G. Malliaris

04-23 The State and the Economy of Modern Greece. Key Drivers from 1821 to the Present, George Alogoskoufis

05-23 Sparse spanning portfolios and under-diversification with second-order stochastic dominance, Stelios Arvanitis, Olivier Scaillet, Nikolas Topaloglou

06-23 What makes for survival? Key characteristics of Greek incubated early-stage startup(per)s during the Crisis: a multivariate and machine learning approach,
Ioannis Besis, Ioanna Sapfo Pepelasis and Spiros Paraskevas

07-23 The Twin Deficits, Monetary Instability and Debt Crises in the History of Modern Greece, George Alogoskoufis

08-23 Dealing with endogenous regressors using copulas; on the problem of near multicollinearity, Dimitris Christopoulos, Dimitris Smyrnakis and Elias Tzavalis

09-23 A machine learning approach to construct quarterly data on intangible investment for Eurozone, Angelos Alexopoulos and Petros Varthalitis

10-23 Asymmetries in Post-War Monetary Arrangements in Europe: From Bretton Woods to the Euro Area, George Alogoskoufis, Konstantinos Gravas and Laurent Jacque

11-23 Unanticipated Inflation, Unemployment Persistence and the New Keynesian Phillips Curve, George Alogoskoufis and Stelios Giannoulakis

12-23 Threshold Endogeneity in Threshold VARs: An Application to Monetary State Dependence, Dimitris Christopoulos, Peter McAdam and Elias Tzavalis

13-23 A DSGE Model for the European Unemployment Persistence, Konstantinos Giakas

14-23 Binary public decisions with a status quo: undominated mechanisms without coercion, Efthymios Athanasiou and Giacomo Valletta

15-23 Does Agents' learning explain deviations in the Euro Area between the Core and the Periphery? George Economides, Konstantinos Mavrigiannakis and Vanghelis Vassilatos

16-23 Mild Explocivity, Persistent Homology and Cryptocurrencies' Bubbles: An Empirical Exercise, Stelios Arvanitis and Michalis Detsis

17-23 A network and machine learning approach to detect Value Added Tax fraud, Angelos Alexopoulos, Petros Dellaportas, Stanley Gyoshev, Christos Kotsogiannis, Sofia C. Olhede, Trifon Pavkov

18-23 Time Varying Three Pass Regression Filter, Yiannis Dendramis, George Kapetanios, Massimiliano Marcellino

19-23 From debt arithmetic to fiscal sustainability and fiscal rules: Taking stock, George Economides, Natasha Miouli and Apostolis Philippopoulos

20-23 Stochastic Arbitrage Opportunities: Set Estimation and Statistical Testing, Stelios Arvanitis and Thierry Post

21-23 Behavioral Personae, Stochastic Dominance, and the Cryptocurrency Market, Stelios Arvanitis, Nikolas Topaloglou, and Georgios Tsomidis

22-23 Block Empirical Likelihood Inference for Stochastic Bounding: Large Deviations Asymptotics Under *m*-Dependence, Stelios Arvanitis and Nikolas Topaloglou

23-23 A Consolidation of the Neoclassical Macroeconomic Competitive General Equilibrium Theory via Keynesianism (Part 1 and Part 2), Angelos Angelopoulos

24-23 Limit Theory for Martingale Transforms with Heavy-Tailed Noise, Stelios Arvanitis and Alexandros Louka

2024

01-24 Market Timing & Predictive Complexity, Stelios Arvanitis, Foteini Kyriazi, Dimitrios Thomakos

02-24 Multi-Objective Frequentistic Model Averaging with an Application to Economic Growth, Stelios Arvanitis, Mehmet Pinar, Thanasis Stengos, Nikolas Topaloglou

03-24 State dependent fiscal multipliers in a Small Open Economy, Xiaoshan Chen, Jilei Huang, Petros Varthalitis

04-24 Public debt consolidation: Aggregate and distributional implications in a small open economy of the Euro Area, Eleftherios-Theodoros Roumpanis

05-24 Intangible investment during the Global Financial Crisis in the EU, Vassiliki Dimakopoulou, Stelios Sakkas and Petros Varthalitis

06-24 Time will tell! Towards the construction of instantaneous indicators of different agent-types, Iordanis Kalaitzoglou, Stelios Arvanitis

07-24 Norm Constrained Empirical Portfolio Optimization with Stochastic Dominance: Robust Optimization Non-Asymptotics, Stelios Arvanitis

08-24 Asymptotics of a QLR-type test for optimal predictive ability, Stelios Arvanitis

09-24 Strongly Equitable General Equilibrium Allocations, Angelos Angelopoulos

10-24 The Greek macroeconomy: A note on the current situation and future outlook, Apostolis Philippopoulos

11-24 Evolution of Greek Tax System, A Survey of Legislated Tax Changes from 1974 to 2018, Panagiotis Asimakopoulos

12-24 Macroeconomic Impact of Tax Changes, The case of Greece from 1974 to 2018, Panagiotis Asimakopoulos

13-24 `Pareto, Edgeworth, Walras, Shapley' Equivalence in a Small Economy, Angelos Angelopoulos

14-24 Stimulating long-term growth and welfare in the U.S, James Malley and Apostolis Philippopoulos

15-24 Distributionally Conservative Stochastic Dominance via Subsampling, Stelios Arvanitis

16-24 Before and After the Political Transition of 1974. Institutions, Politics, and the Economy of Post-War Greece, George Alogoskoufis

17-24 Gaussian Stochastic Volatility, Misspecified Volatility Filters and Indirect Inference Estimation, Stelios Arvanitis, Antonis Demos

18-24 Endogenous Realtor Intermediation and Housing Market Liquidity, Miroslav Gabrovski, Ioannis Kospentaris, Victor Ortego-Marti

19-24 Universal Choice Spaces and Expected Utility: A Banach-type Functorial Fixed Point, Stelios Arvanitis, Pantelis Argyropoulos, Spyros Vassilakis

2025

01-25 Democracy, redistribution, and economic growth: Some evidence from post-1974 Greece, George C. Bitros

02-25 Assessing Downside Public Debt Risks in an Environment of Negative Interest Rates Growth Differentials, Yiannis Dendramis, Georgios Dimitrakopoulos and Elias Tzavalis

03-25 Wicksellian General Equilibrium, Angelos Angelopoulos

04-25 Institutions and Ethics, Angelos Angelopoulos

05-25 General Equilibrium in a Capitalist Economy, Angelos Angelopoulos

06-25 Public investment multipliers revisited: The role of production complementarities, Vasiliki Dimakopoulou, George Economides, and Apostolis Philippopoulos

07-25 The heterogeneous causal effects of the EU's Cohesion Fund, Angelos Alexopoulos, Ilias Kostarakos, Christos Mylonakis, and Petros Varthalitis

08-25 Gaussian Invariant Markov Chain Monte Carlo, Michalis K. Titsias, Angelos Alexopoulos , Siran Liu, and Petros Dellaportas

09-25 Financial Intermediation, Investment, and Monetary Policy, Ioannis Kospentaris and Florian Madison

10-25 Temperature as a premium factor. Insurance incorporating climate change. A comparative case study between Europe and North America, Nomikou-Lazarou Irene

11-25 How High will our Buildings Become? Angelos Angelopoulos

12-25 Emergency Room General Equilibrium, Angelos Angelopoulos

13-25 Updating Density Estimates using Conditional Information Projection: Stock Index Returns and Stochastic Dominance Relations, Stelios Arvanitis, Richard McGee, and Thierry Post

14-25 Measuring Textual Cohesion via Graph Curvature, Weighted Dominance, and Embedding-Based Coherence, Stelios Arvanitis

15-25 Universal Preference Spaces for Expected Utility via a Banach-type Functorial Fixed Point Stelios Arvanitis, Pantelis Argyropoulos, and Spyros Vassilakis

16-25 Planet Earth's General Equilibrium, Angelos Angelopoulos

17-25 Electable and Stable Insiders' Coalition Governments, Tryphon Kollintzas and Lambros Pechlivanos

18-25 Long Memory from Cheeger Bottlenecks and Long Cycles in Network Dynamics, Stelios Arvanitis

19-25 Robust Multiobjective Model Averaging in Predictive Regressions, Stelios Arvanitis, Mehmet Pinar, Thanasis Stengos, and Nikolas Topaloglou

ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ΣΧΟΛΗ
ΟΙΚΟΝΟΜΙΚΩΝ
ΕΠΙΣΤΗΜΩΝ
SCHOOL OF
ECONOMIC
SCIENCES

ΤΜΗΜΑ
ΟΙΚΟΝΟΜΙΚΗΣ
ΕΠΙΣΤΗΜΗΣ
DEPARTMENT OF
ECONOMICS

Department of Economics
Athens University of Economics and Business

The Department is the oldest Department of Economics in Greece with a pioneering role in organising postgraduate studies in Economics since 1978. Its priority has always been to bring together highly qualified academics and top quality students. Faculty members specialize in a wide range of topics in economics, with teaching and research experience in world-class universities and publications in top academic journals.

The Department constantly strives to maintain its high level of research and teaching standards. It covers a wide range of economic studies in micro-and macroeconomic analysis, banking and finance, public and monetary economics, international and rural economics, labour economics, industrial organization and strategy, economics of the environment and natural resources, economic history and relevant quantitative tools of mathematics, statistics and econometrics.

Its undergraduate program attracts high quality students who, after successful completion of their studies, have excellent prospects for employment in the private and public sector, including areas such as business, banking, finance and advisory services. Also, graduates of the program have solid foundations in economics and related tools and are regularly admitted to top graduate programs internationally. Three specializations are offered:1. Economic Theory and Policy, 2. Business Economics and Finance and 3. International and European Economics. The postgraduate programs of the Department (M.Sc and Ph.D) are highly regarded and attract a large number of quality candidates every year.

For more information:

https://www.dept.aueb.gr/en/econ/