

Target-Aware Linear Regression Under Distribution Shift

ZHEWEN HOU¹ AND TIAN ZHENG^{1,*}

¹*Department of Statistics, Columbia University, New York, NY, USA*

Abstract

Distribution shift between training and deployment is a pervasive challenge for modern AI systems. In many cases, the target marginals of covariates and response are known or specified through population-level observations, boundary conditions, properties of simulator configurations, or alignment-time distributional constraints. Such knowledge may provide valuable side information for regression estimation. We study this problem in the multivariate linear regression setting with a stable conditional mean $\mathbb{E}[Y | X]$ across source and target, and identify the hybrid-loss estimator, which jointly incorporates both target marginals, as a benchmark target-aware estimator. Its direct computation, however, requires solving a coupled nonlinear optimization that is expensive at scale. Our main contribution is to develop and evaluate two computationally tractable alternatives: a constrained moment-matching estimator and a two-stage estimator that augments ordinary least squares with a calibration step. For all three estimators, we derive and compare closed-form asymptotic mean squared errors, yielding conditions under which the tractable alternatives match or closely approximate the hybrid benchmark, and regimes in which they do not. Monte Carlo experiments across three controlled shift regimes validate the theoretical results, investigate the accuracy–runtime tradeoffs among the three estimators, and translate into guidance on estimator choice. In particular, the two-stage estimator nearly matches the hybrid benchmark in the high signal-to-noise regime at essentially no additional cost, providing theoretical grounding for empirical observations in nonlinear settings.

Keywords *calibration; distribution shift; knowledge informed learning*

1 Introduction

Modern AI models often face the challenge when the data they encounter during deployment differs from the data used during training. When a model is trained on data drawn from a source distribution but evaluated under a different target distribution, the resulting mismatch can substantially degrade predictive performance. In the literature, this problem has been addressed through importance weighting (Shimodaira, 2000), domain adaptation (Ben-David et al., 2010), and related techniques, typically assuming access to unlabeled samples from the target covariate distribution.

In many applications, however, the practitioner has access to richer side information about the target regime. Population level observations (e.g., income distributions) prescribe the marginal distribution of covariates in the deployment population. Boundary conditions and conservation laws in physical systems constrain the marginal behavior of the response variable (Finzi et al., 2023; Zhong and Meidani, 2023). Simulator configurations specify both input and out-

*Corresponding author. Email: zh2475@columbia.edu or tian.zheng@columbia.edu.

put marginals for digital twins and climate projections (Gloege et al., 2021; Kay et al., 2015). Alignment-time distributional constraints in generative models fix the target output distribution by design. In population studies, census data and administrative records provide precise marginal distributions of demographic and economic variables (e.g., age, income, education) for the target population, even when the training data come from a convenience sample or a different demographic. In observational medical and epidemiological research, disease registries, vital statistics, and large-scale surveys furnish the population-level marginal distributions of clinical outcomes (e.g., blood pressure, BMI, mortality rates) and risk factors, against which study cohorts may be calibrated. In survey methodology more broadly, known population totals from external sources are routinely used to calibrate survey weights so that weighted estimates match target marginals. In each of these settings, the *target marginals* of covariates X and response Y are known or can be specified, and they provide valuable side information for regression estimation that goes beyond what unlabeled covariate data alone can offer.

In this paper, we study how to incorporate such target marginal information into regression estimation. We work in the multivariate linear regression setting with Gaussian target marginals and a stable conditional mean $\mathbb{E}[Y | X]$ across source and target. These assumptions are chosen to enable a rigorous and exact theoretical investigation, with closed-form asymptotic mean squared errors for every estimator for precise comparisons and sharp characterization of accuracy-complexity tradeoffs. The insights derived in this paper on estimator ranking and the central role of signal-to-noise ratio confirm and provide theoretical grounding for empirical observations made in more complex, nonlinear settings. In particular, Hou et al. (2025) observe that post hoc calibration via the kernelized Stein discrepancy (KSD) yields the largest improvements when the base machine learning models are already reasonably accurate; our closed-form results explain precisely why this is so.

Within the proposed framework, we study three estimators. The *hybrid-loss estimator* jointly penalizes the empirical risk with a Wasserstein-type term that matches both target marginals simultaneously; it serves as a benchmark target-aware estimator, uniformly dominating the tractable alternatives studied here, in terms of accuracy. Its computation, however, requires solving a coupled nonlinear optimization that is expensive at scale. We therefore consider two computationally tractable alternatives. The *constrained moment-matching estimator* enforces the target marginal constraints exactly. The *two-stage estimator* first fits ordinary least squares and then calibrates the predictions to match the target marginals in a post-processing step.

Our main **contributions** are as follows.

1. We formulate the target-aware regression problem in a multivariate linear model and develop three estimators: hybrid, constrained moment-matching, and two-stage, all of which incorporate target marginal information in structurally different ways.
2. For all three estimators, we derive the asymptotic distribution, mean squared error, and prediction error, making the accuracy-complexity tradeoffs transparent.
3. We identify precise conditions under which the tractable alternatives match or closely approximate the hybrid benchmark, as well as regimes in which they do not. A key finding is that the two-stage estimator, which requires only a scalar rescaling of OLS, nearly matches the hybrid estimator in the high signal-to-noise regime, which is the practically relevant setting when the base model is already powerful.
4. Monte Carlo experiments across three controlled shift regimes validate the theoretical results, demonstrate accuracy-runtime tradeoffs, and provide guidance on estimator choice in practice.

2 Related Work

Covariate shift. When the covariate distribution changes from source to target but the conditional $Y | X$ remains stable, importance weighting (Shimodaira, 2000) reweights the training loss by the density ratio $p_t(x)/p_s(x)$. Estimating this ratio in high dimensions is itself challenging; approaches include low-dimensional projections (Stojanov et al., 2019), adaptive kernel methods (Zellinger et al., 2023), and nearest-neighbor sampling (Portier et al., 2024). These methods use knowledge of the target covariate distribution $\mathcal{P}_{t,x}$ but do not incorporate the target response marginal $\mathcal{P}_{t,y}$. However, importance-weighting estimators do not improve upon OLS and are less favorable in the regimes we analyze. This contrasts with the estimators studied in this paper, whose main advantage comes from incorporating the target response marginal $\mathcal{P}_{t,y}$ in addition to target covariate information. Thus, importance weighting and target-marginal estimation should be viewed as complementary: the former corrects covariate mismatch through reweighting, while the latter leverages response-marginal information to reduce estimation variance.

Distributional learning. Wasserstein Distributional Learning (WDL) (Tang et al., 2023) proposes a majorization-minimization algorithm for fitting semi-parametric conditional Gaussian mixture models for conditional density estimation, using the Wasserstein distance W_2 as a metric on density outcomes. It highlights the practical feasibility of Wasserstein-based distributional matching.

Post-hoc distribution calibration. In the nonlinear setting, Hou et al. (2025) propose a post hoc calibration method based on the kernelized Stein discrepancy (KSD) (Liu et al., 2016) that adjusts predictions from an arbitrary base model to match a known target distribution. A normalization step rescales model-generated predictions to match target moments before KSD refinement. This is a nonlinear analog of the two-stage estimator studied here. Hou et al. (2025) provided intuition on the MSE reduction of the KSD gradient step via Stein’s Lemma, directly connecting to the James-Stein mechanism.

Constrained and auxiliary-information estimation. The moment-matching estimator can be viewed as a constrained least-squares estimator (Judge et al., 1980). Many methods have been developed to use known population information to improve efficiency, including survey calibration (Devaud and Tillé, 2019; Deville and Särndal, 1992; Särndal, 2007), the generalized method of moments (GMM) with auxiliary moments (Chen et al., 2008; Hansen, 1982; Imbens, 1997), and empirical likelihood with auxiliary summary information (Gao and Chan, 2023; Han and Lawless, 2019; Qin and Lawless, 1994). Recent semi-supervised and calibrated-estimation work also exploits unlabeled covariates, calibration losses, or balancing constraints to reduce variance (Bruns-Smith et al., 2026; Chakraborty and Cai, 2018; Hirshberg and Wager, 2021; Tan, 2020). In our setting, the auxiliary information is the target response marginal. The moment-matching estimator enforces the target mean and variance exactly, while the hybrid estimator relaxes these constraints through a penalty. The main distinction is that the constraints are target-distributional rather than source-sample moment restrictions or known covariate totals.

Shrinkage estimation and external information. The two-stage estimator bears a structural resemblance to James–Stein shrinkage (Efron and Morris, 1973; Stein, 1956): both use external information to suppress unnecessary sampling variation in selected directions. Han et al. (2024) study James-Stein-type estimators that incorporate external population information into linear regression. In our setting, however, the target marginal constraints are valid population constraints satisfied by the true parameter, so the leading-order gain is an efficiency gain from auxiliary information rather than a leading-order bias effect.

3 Methods

We formalize the target-aware regression problem in a multivariate linear model with known target marginals. After establishing the setup and notation, we present the OLS baseline and three target-aware estimators. For every estimator we state the asymptotic distribution, mean squared error, and prediction error under the target distribution.

3.1 Setup and notation

We consider the multivariate linear regression model.

Assumption 3.1. *Let $X \in \mathbb{R}^d$ be a d -dimensional covariate vector. The response variable satisfies*

$$Y = \beta_0 + X^\top \beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

where $\beta \in \mathbb{R}^d$ and $X \perp \varepsilon$.

For simplicity, write the augmented regressor as $\tilde{X} := (1, X^\top)^\top \in \mathbb{R}^{d+1}$, $\theta := (\beta_0, \beta^\top)^\top \in \mathbb{R}^{d+1}$.

Assumption 3.2. *We observe i.i.d. data $\{(X_i, Y_i)\}_{i=1}^n$ generated (from the source distribution, subscript s) as*

$$Y_i = \beta_0 + X_i^\top \beta + \varepsilon_i, \quad i = 1, \dots, n.$$

Moreover, $X_i \sim \mathcal{P}_{s,x}$, $\mathbb{E}_{\mathcal{P}_{s,x}}[X] = \mu_{s,x} \in \mathbb{R}^d$, $\text{Cov}_{\mathcal{P}_{s,x}}(X) = \Sigma_{s,x} \in \mathbb{R}^{d \times d}$, where $\Sigma_{s,x}$ is positive definite. The distribution $\mathcal{P}_{s,x}$ may be unknown.

Assumption 3.3. *The training covariate satisfies $\mathbb{E}_{\mathcal{P}_{s,x}} \|X\|^4 < \infty$.*

Assumption 3.4. *The training covariate is sub-Gaussian under $\mathcal{P}_{s,x}$.*

Let $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$ and $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_n)^\top \in \mathbb{R}^{n \times (d+1)}$. Define the source moment matrix $Q_s := \mathbb{E}_{\mathcal{P}_{s,x}}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top)$.

A key structural assumption is that the conditional mean $\mathbb{E}[Y | X] = \beta_0 + X^\top \beta$ is *stable* across source and target: the regression coefficients (β_0, β) and the noise variance σ_ε^2 do not change, while the marginal distribution of X may differ.

In addition to the regression structure, we assume access to target marginals describing the marginal behavior of the model variables under the deployment distribution.

Assumption 3.5. *The target marginals of (X, Y) are known (subscript k) and specified as*

$$X \sim \mathcal{N}(\mu_{k,x}, \Sigma_{k,x}), \quad Y \sim \mathcal{N}(\mu_{k,y}, \sigma_{k,y}^2),$$

where $\mu_{k,x} \in \mathbb{R}^d$ and $\Sigma_{k,x} \in \mathbb{R}^{d \times d}$ is positive definite.

Write $\tilde{\mu}_{k,x} = (1, \mu_{k,x}^\top)^\top$ and $\tilde{\Sigma}_{k,x} = \text{Var}(\tilde{X}) = \text{diag}(0, \Sigma_{k,x})$. Under Assumptions 3.1 and 3.5, $\mathbb{E}[Y] = \beta_0 + \mu_{k,x}^\top \beta = \mu_{k,y}$ and $\text{Var}(Y) = \beta^\top \Sigma_{k,x} \beta + \sigma_\varepsilon^2 = \sigma_{k,y}^2$. The source and target marginals need not coincide.

Remark 3.1. By applying a linear transformation to X , we may assume without loss of generality that $\mu_{k,x} = 0$ and $\Sigma_{k,x} = I_d$. This changes θ , $\mu_{s,x}$, and $\Sigma_{s,x}$, but preserves the linear relationship, σ_ε , $\mu_{k,y}$, and $\sigma_{k,y}$.

3.2 OLS baseline

The ordinary least squares (OLS) estimator is $\hat{\theta}^{\text{OLS}} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^\top \mathbf{Y}$, with noise variance estimator $\hat{\sigma}_\varepsilon^2 = \|\mathbf{Y} - \tilde{\mathbf{X}} \hat{\theta}^{\text{OLS}}\|_2^2 / (n - d - 1)$. By standard arguments, $\sqrt{n}(\hat{\theta}^{\text{OLS}} - \theta) \Rightarrow \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{Q}_s^{-1})$, giving

$$\mathbb{E} \|\hat{\beta}^{\text{OLS}} - \beta\|^2 = (\sigma_\varepsilon^2 / n) \text{tr}(\Sigma_{s,x}^{-1}) + o(n^{-1}).$$

OLS ignores the target marginals, and its prediction error under the target distribution,

$$\mathbb{E} \left[\left(Y - \tilde{\mathbf{X}}^\top \hat{\theta}^{\text{OLS}} \right)^2 \right] = \sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n} \left[1 + \text{tr}(\Sigma_{k,x} \Sigma_{s,x}^{-1}) + (\mu_{k,x} - \mu_{s,x})^\top \Sigma_{s,x}^{-1} (\mu_{k,x} - \mu_{s,x}) \right] + o(n^{-1}), \quad (1)$$

grows with the source–target mismatch through both the covariance ratio and the squared Mahalanobis distance of the mean shift.

3.3 Hybrid-loss estimator

The hybrid estimator combines the empirical risk with a penalty that matches the model-implied output distribution to the target marginals:

$$L_H(\theta; \omega) = \frac{1}{n} \|\mathbf{Y} - \tilde{\mathbf{X}} \theta\|_2^2 + \omega (\tilde{\mu}_{k,x}^\top \theta - \mu_{k,y})^2 + \omega \left(\sqrt{\theta^\top \tilde{\Sigma}_{k,x} \theta + \frac{1}{n} \|\mathbf{Y} - \tilde{\mathbf{X}} \theta\|_2^2} - \sigma_{k,y} \right)^2, \quad (2)$$

where $\omega \geq 0$ controls the weight of the target constraints. The second term enforces mean matching, and the third enforces variance matching via the squared Wasserstein distance between two one-dimensional Gaussians. We define $\hat{\theta}^H(\omega) \in \arg \min_\theta L_H(\theta; \omega)$. Since the loss is nonlinear in θ , the minimizer is computed numerically by gradient descent initialized at $\hat{\theta}^{\text{OLS}}$; the explicit gradient is given in Section S.1.3 of the Supplement.

Let $\tilde{v}_{\sigma\beta} = \tilde{\Sigma}_{k,x} \theta$ and define

$$\mathcal{Q}(\omega) := \mathcal{Q}_s + \omega \left(\tilde{\mu}_{k,x} \tilde{\mu}_{k,x}^\top + \frac{1}{\sigma_{k,y}^2} \tilde{v}_{\sigma\beta} \tilde{v}_{\sigma\beta}^\top \right), \quad \Omega(\omega) := \mathcal{Q}_s + \frac{\omega^2 \sigma_\varepsilon^2}{2\sigma_{k,y}^4} \tilde{v}_{\sigma\beta} \tilde{v}_{\sigma\beta}^\top.$$

All proofs are deferred to Section S.2 of the Supplement; detailed derivations of the estimators, including the constraint geometry, the gradient and quartic characterization of the optimal weight, and the prediction error formulas, are given in Section S.1 of the Supplement.

Theorem 3.1. *Under Assumptions 3.1–3.5, for any $\omega \geq 0$, as $n \rightarrow \infty$,*

$$\sqrt{n}(\hat{\theta}^H(\omega) - \theta) \Rightarrow \mathcal{N}(0, \Sigma_H(\omega)), \quad \Sigma_H(\omega) = \sigma_\varepsilon^2 \mathcal{Q}(\omega)^{-1} \Omega(\omega) \mathcal{Q}(\omega)^{-1}.$$

Corollary 3.1. *Under the same assumptions, the asymptotic mean squared error is*

$$\mathbb{E} \left[\left\| \hat{\beta}^H(\omega) - \beta \right\|^2 \right] = \frac{\sigma_\varepsilon^2}{n} V_H(\omega) + o(n^{-1}), \quad (3)$$

where $V_H(\omega) := \text{tr}(P_\beta \mathcal{Q}(\omega)^{-1} \Omega(\omega) \mathcal{Q}(\omega)^{-1} P_\beta^\top)$ with $P_\beta := (0_{d \times 1}, I_d)$.

Under the normalization of Remark 3.1, define $M_H(\omega) := \Sigma_{s,x} + \frac{\omega}{1+\omega} \mu_{s,x} \mu_{s,x}^\top + \frac{\omega}{\sigma_{k,y}^2} \beta \beta^\top$. Then Eq. (3) becomes

$$\begin{aligned} \mathbb{E} \left[\left\| \hat{\beta}^H(\omega) - \beta \right\|^2 \right] = & \frac{\sigma_\varepsilon^2}{n} \left[\text{tr}(M_H(\omega)^{-1}) - \frac{\omega}{(1+\omega)^2} \mu_{s,x}^\top M_H(\omega)^{-2} \mu_{s,x} \right. \\ & \left. - \left(\frac{\omega}{\sigma_{k,y}^2} - \frac{\omega^2 \sigma_\varepsilon^2}{2\sigma_{k,y}^4} \right) \beta^\top M_H(\omega)^{-2} \beta \right] + o(n^{-1}). \end{aligned} \quad (4)$$

Remark 3.2. Here, the matrix $Q(\omega)$ augments the source moment Q_s with two rank-one terms from the mean and variance constraints. The matrix $\Omega(\omega)$ differs from $Q(\omega)$ because of the variance constraint. The three terms in Eq. (4) have clear roles: the first is the baseline variance under the augmented geometry. The second is the reduction from the mean constraint, weighted by the mean shift $\mu_{s,x}$; and the third is the “reduction” from the variance constraint, which is positive (beneficial) when $\omega\sigma_\varepsilon^2/(2\sigma_{k,y}^2) < 1$ and changes sign at large noise levels.

Optimal ω^* The asymptotically optimal tuning parameter is $\omega^* \in \arg \min_{\omega \geq 0} V_H(\omega)$. The function $V_H(\omega)$ is a rational function of ω (Proposition S.2 in the Supplement) whose derivative admits a quartic equation. In practice, one collects its nonnegative real roots and compares V_H at those roots and at $\omega = 0$.

3.4 Constrained moment-matching (MM) estimator

Rather than penalizing the Wasserstein distance, the moment-matching estimator enforces the target marginal constraints exactly; equivalently, it replaces the soft penalty in (2) by a hard constraint, so that the hybrid loss with $\omega \rightarrow \infty$ recovers the moment-matching objective. Define the loss

$$L_{\text{MM}}(\theta) = (\tilde{\mu}_{k,x}^\top \theta - \mu_{k,y})^2 + \left(\sqrt{\theta^\top \tilde{\Sigma}_{k,x} \theta + \frac{1}{n} \|Y - \tilde{X}\theta\|_2^2 - \sigma_{k,y}} \right)^2. \quad (5)$$

Let $\hat{\Theta}^{\text{MM}} = \arg \min_{\theta} L_{\text{MM}}(\theta)$. The mean constraint fixes $\beta_0 = \mu_{k,y} - \mu_{k,x}^\top \beta$, and the variance constraint reduces to a quadratic in β (Lemma S.1 in the Supplement). When $d > 1$ and the zero-loss set is non-singleton, we select the element with smallest residual sum of squares, which is characterized by a KKT system with a single Lagrange multiplier (Proposition S.1 in the Supplement):

$$\hat{\theta}^{\text{MM}} = \arg \min_{\theta \in \hat{\Theta}^{\text{MM}}} \sum_{i=1}^n (\check{Y}_i - \check{X}_i^\top \beta)^2, \quad \check{X}_i := X_i - \mu_{k,x}, \quad \check{Y}_i := Y_i - \mu_{k,y}.$$

Write $Q_{s|k} := \Sigma_{s,x} + (\mu_{s,x} - \mu_{k,x})(\mu_{s,x} - \mu_{k,x})^\top$, $v_{\sigma\beta} := \Sigma_{k,x}\beta$, $\kappa := v_{\sigma\beta}^\top Q_{s|k}^{-1} v_{\sigma\beta}$, $\chi := v_{\sigma\beta}^\top Q_{s|k}^{-1} \Sigma_{k,x} Q_{s|k}^{-1} v_{\sigma\beta}$, $J_{\mu_k} := \begin{pmatrix} \mu_{k,x}^\top \\ -I_d \end{pmatrix}$, and $\Omega := Q_{s|k} - \frac{v_{\sigma\beta} v_{\sigma\beta}^\top}{\kappa} + \frac{\sigma_\varepsilon^2}{2} \frac{v_{\sigma\beta} v_{\sigma\beta}^\top}{\kappa^2}$.

Theorem 3.2. Under Assumptions 3.1–3.5, if $\beta \neq 0$, as $n \rightarrow \infty$,

$$\sqrt{n} (\hat{\theta}^{\text{MM}} - \theta) \Rightarrow \mathcal{N}(0, \Sigma_{\text{MM}}), \quad \Sigma_{\text{MM}} = \sigma_\varepsilon^2 J_{\mu_k} Q_{s|k}^{-1} \Omega Q_{s|k}^{-1} J_{\mu_k}^\top.$$

Corollary 3.2. Under the same assumptions,

$$\mathbb{E} \left[\left\| \hat{\beta}^{\text{MM}} - \beta \right\|^2 \right] = \frac{\sigma_\varepsilon^2}{n} \text{tr}(Q_{s|k}^{-1} \Omega Q_{s|k}^{-1}) + o(n^{-1}).$$

Under the normalization of Remark 3.1, this becomes

$$\frac{\sigma_\varepsilon^2}{n} \left[\text{tr}(Q_{s|k}^{-1}) - \frac{\beta^\top Q_{s|k}^{-2} \beta}{\beta^\top Q_{s|k}^{-1} \beta} + \frac{\sigma_\varepsilon^2}{2} \frac{\beta^\top Q_{s|k}^{-2} \beta}{(\beta^\top Q_{s|k}^{-1} \beta)^2} \right] + o(n^{-1}). \quad (6)$$

The moment-matching estimator improves over OLS when $\sigma_\varepsilon^2 \leq 2\beta^\top Q_{s|k}^{-1} \beta$.

Remark 3.3. The quantity $\beta^\top Q_{s|k}^{-1} \beta$ measures the effective signal strength under the geometry induced by the training distribution, while σ_ε^2 is the noise level. When noise is not too large relative to this effective signal, the target marginal constraints shrink the feasible set of β and suppress unnecessary sampling variation, yielding a smaller asymptotic MSE than OLS.

3.5 Two-stage calibration estimator

The two-stage estimator first computes OLS and then calibrates the predictions to match the target marginals. We work under the normalization of Remark 3.1. Suppose an independent test sample $\{X_j^{\text{test}}\}_{j=1}^m$ is drawn from $\mathcal{N}(0, I_d)$ with empirical moments matching the target exactly.

Stage 1. Compute $\hat{\theta}^{\text{OLS}}$ and $\hat{\sigma}_\varepsilon^2$ from the training data.

Stage 2. Form raw predictions $\hat{Y}_j = \hat{\beta}_0^{\text{OLS}} + (X_j^{\text{test}})^\top \hat{\beta}^{\text{OLS}}$ and rescale them to match the target mean $\mu_{k,y}$ and target prediction variance $(\sigma_{k,y}^2 - \hat{\sigma}_\varepsilon^2)_+$. This yields calibrated predictions $\tilde{Y}_j = a + b \hat{Y}_j$ with

$$a = \mu_{k,y} - b \cdot \bar{\tilde{Y}}, \quad b = \sqrt{\frac{(\sigma_{k,y}^2 - \hat{\sigma}_\varepsilon^2)_+}{\|\hat{\beta}^{\text{OLS}}\|^2}}.$$

The implied coefficient estimator is $\hat{\beta}^{\text{cali}} = b \cdot \hat{\beta}^{\text{OLS}}$. The target mean constraint determines the intercept exactly ($\hat{\beta}_0^{\text{cali}} = \mu_{k,y} - \mu_{k,x}^\top \hat{\beta}^{\text{cali}}$), so the MSE below accounts only for the slope.

Theorem 3.3. Under Assumptions 3.1–3.5 and the normalization of Remark 3.1, with $\beta \neq 0$,

$$\mathbb{E} \left[\|\hat{\beta}^{\text{cali}} - \beta\|^2 \right] = \frac{\sigma_\varepsilon^2}{n} \left[\text{tr}(\Sigma_{s,x}^{-1}) - \frac{\beta^\top \Sigma_{s,x}^{-1} \beta}{\|\beta\|^2} + \frac{\sigma_\varepsilon^2}{2\|\beta\|^2} \right] + o(n^{-1}). \quad (7)$$

The prediction error satisfies

$$\mathbb{E} \left[\frac{1}{m} \sum_{j=1}^m (\tilde{Y}_j - Y_j^{\text{test}})^2 \right] = \sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n} \left[\text{tr}(\Sigma_{s,x}^{-1}) - \frac{\beta^\top \Sigma_{s,x}^{-1} \beta}{\|\beta\|^2} + \frac{\sigma_\varepsilon^2}{2\|\beta\|^2} \right] + o(n^{-1}). \quad (8)$$

Remark 3.4. The two-stage rescaling reduces the variance along the signal direction β at the cost of introducing a noise-dependent correction $\sigma_\varepsilon^2/(2\|\beta\|^2)$.

Remark 3.5 (The $\beta = 0$ boundary). Theorems 3.2 and 3.3 require $\beta \neq 0$. When $\beta = 0$, the response $Y = \beta_0 + \varepsilon$ is pure noise and $\sigma_{k,y}^2 = \sigma_\varepsilon^2$. The target variance constraint carries no directional information about β , and the rescaling factor b is ill-defined (division by $\|\hat{\beta}^{\text{OLS}}\|^2 \rightarrow 0$). In this degenerate regime, OLS is already efficient for estimating the slope, so the target marginals offer no additional gain for inference on β (although the mean constraint still suffices to identify β_0). When $\|\beta\|$ is small but nonzero, the correction terms $\sigma_\varepsilon^2/(2\|\beta\|^2)$ and $\sigma_\varepsilon^2/(2\kappa)$ grow, and the target-aware estimators may perform worse than OLS; the improvement condition $\sigma_\varepsilon^2 < 2\beta^\top Q_{s|k}^{-1} \beta$ correctly excludes this regime.

Computational complexity of estimators.

As shown in Table 1, OLS, two-stage, and moment-matching (MM) all have essentially the same computational cost as OLS (matrix inversion plus only a scalar rescaling or a simple root-finding step), whereas the hybrid estimator requires tolerance-dependent iterative optimization.

Table 1: Computational costs of different estimators.

Estimator	Computation	Relative cost
OLS	matrix inversion	$1 \times$
Two-stage (cali)	OLS + scalar rescaling	$\approx 1 \times$
Moment-matching (MM)	OLS + root-finding	$\approx 1 \times$
Hybrid	iterative optimization	tolerance-dependent

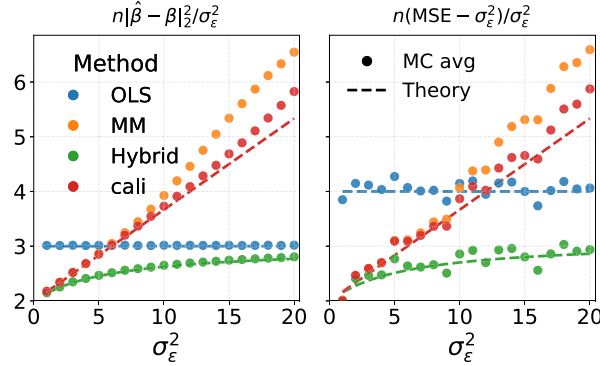


Figure 1: **Noise level.** Scaled coefficient error (left) and scaled excess prediction error (right) vs. σ_ϵ^2 . Points: Monte Carlo averages ($L = 10^6$, $n = 1000$); dashed lines: theoretical values. Monte Carlo standard errors are on the order of 10^{-3} and are visually negligible at this scale.

4 Numerical Experiments

Setup. We perform experiments in dimension $d = 3$ under the multivariate linear model with intercept $\beta_0 = 1$ and default coefficients $\beta = (0.970, -1.268, 0.671)^\top$. The target distribution is normalized: $\mu_{k,x} = 0$, $\Sigma_{k,x} = I_3$. Default source parameters are $\mu_{s,x} = 0$, $\Sigma_{s,x} = I_3$.

For each configuration, we generate $L = 10^6$ independent Monte Carlo repetitions with training sample size $n = 1000$ and test sample size $m = 100$ (with test covariates exactly matching target moments). We compare four estimators: OLS, moment-matching (MM), hybrid, and two-stage (cali). For the hybrid estimator, we compute a plug-in $\hat{\omega}$ by replacing the population quantities in the characterization of ω^* in Proposition S.2 in the Supplement with sample estimates. We then minimize $L_H(\theta; \hat{\omega})$ from the OLS initialization using L-BFGS-B with maximum iteration count 1000, `ftol` = 10^{-12} , and `gtol` = 10^{-8} . We report two metrics: the scaled coefficient error $n\|\hat{\beta} - \beta\|_2^2/\sigma_\epsilon^2$ and the scaled excess prediction error $n(\text{MSE}_{\text{test}} - \sigma_\epsilon^2)/\sigma_\epsilon^2$.

We consider three controlled regimes, each varying one aspect of the source–target relationship while holding the others fixed.

Noise level. We vary $\sigma_\epsilon^2 \in \{1, 2, \dots, 20\}$ with $\mu_{s,x} = 0$, $\Sigma_{s,x} = I_3$. Under the isotropic setting, the theoretical scaled MSEs are 3 for OLS, $2 + \sigma_\epsilon^2/(2\|\beta\|^2)$ for moment-matching and two-stage, and $2 + \sigma_\epsilon^2/(2\|\beta\|^2 + \sigma_\epsilon^2)$ for the hybrid. Moment-matching and two-stage coincide and beat OLS only when $\sigma_\epsilon^2 < 2\|\beta\|^2 \approx 6$. Figure 1 confirms the theoretical results: the OLS curve is flat at 3, the moment-matching and two-stage curves rise linearly, and the hybrid stays below them throughout. Target marginals provide the largest gains in the high signal-to-noise regime.

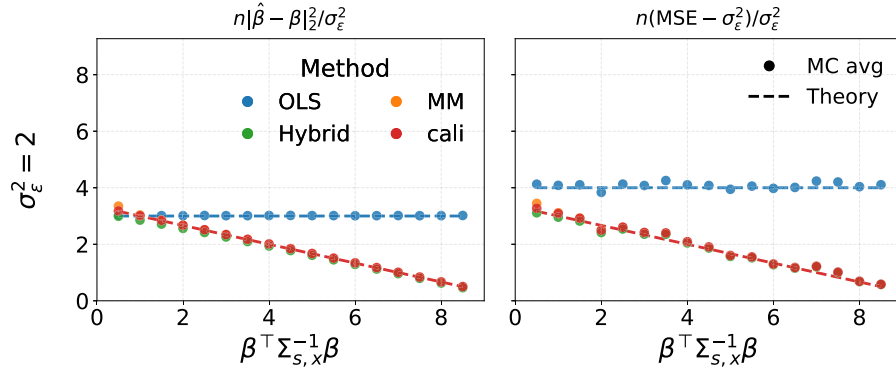


Figure 2: **Covariance geometry** ($\sigma_\epsilon^2 = 2$). Scaled coefficient error (left) and excess prediction error (right) vs. $\beta^\top \Sigma_{s,x}^{-1} \beta$. Monte Carlo standard errors are on the order of 10^{-3} and are visually negligible at this scale.

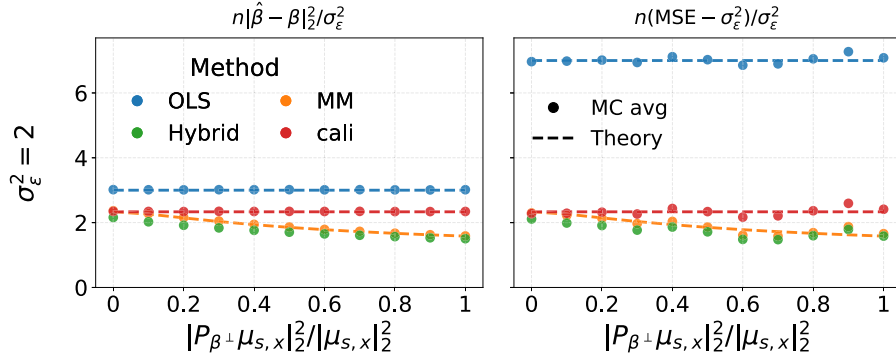


Figure 3: **Mean mismatch** ($\sigma_\epsilon^2 = 2$). Scaled errors vs. $\rho = \|P_{\beta^\perp} \mu_{s,x}\|^2 / \|\mu_{s,x}\|^2$. Monte Carlo standard errors are on the order of 10^{-3} and are visually negligible at this scale.

Covariance geometry. We vary $\Sigma_{s,x}$ subject to $\text{tr}(\Sigma_{s,x}^{-1}) = 3$ and $\beta^\top \Sigma_{s,x}^{-1} \beta \in \{0.5, 1.0, \dots, 8.5\}$, displaying results at $\sigma_\epsilon^2 = 2$ where the geometry effect is largest. The general MSE formulas (6) and (7) now depend on the alignment between $\Sigma_{s,x}$ and β , and the two estimators are no longer identical. Figure 2 shows $\sigma_\epsilon^2 = 2$, where the geometry effect is most pronounced. Notably, at low values of $\beta^\top \Sigma_{s,x}^{-1} \beta$, moment-matching is worse than OLS: since $\mu_{s,x} = 0$ here, the improvement condition $\sigma_\epsilon^2 < 2\beta^\top Q_{s|k}^{-1} \beta$ (Corollary 3.2) reduces to $\beta^\top \Sigma_{s,x}^{-1} \beta > \sigma_\epsilon^2/2 = 1$, and the figure confirms that moment-matching crosses OLS near this threshold. As $\beta^\top \Sigma_{s,x}^{-1} \beta$ increases, both moment-matching and two-stage improve, but moment-matching improves faster because it exploits the full geometry of $Q_{s|k}$, whereas two-stage projects onto β . The hybrid remains uniformly best. The full results across multiple noise levels (Figure S1 in the Supplement) reveal that at low noise the three target-aware estimators are nearly indistinguishable and they separate as the noise increases.

Mean mismatch. We fix $\|\mu_{s,x}\|^2 = 3$, $\Sigma_{s,x} = I_3$, and vary $\rho = \|P_{\beta^\perp} \mu_{s,x}\|^2 / \|\mu_{s,x}\|^2 \in \{0, 0.1, \dots, 1.0\}$, where $P_{\beta^\perp} = I_d - \beta\beta^\top / \|\beta\|^2$. Since $\Sigma_{s,x} = I_3$, the two-stage MSE (7) is flat in ρ , whereas the moment-matching MSE depends on the direction of the mean shift through $Q_{s|k}$.

Figure 3 shows $\sigma_\epsilon^2 = 2$: as more of the mean shift moves into the $\beta^\perp$ direction (larger ρ),

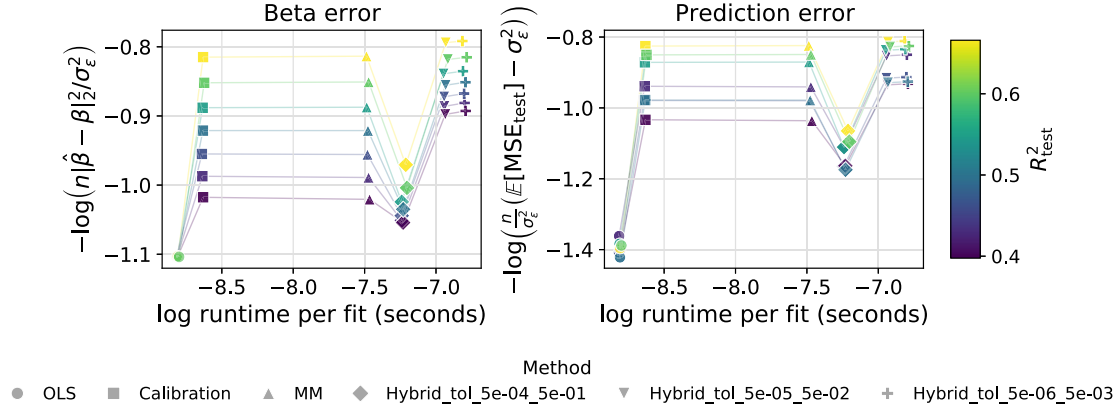


Figure 4: **Accuracy–runtime Pareto frontier.** Each point is one estimator at one noise level ($\sigma_\epsilon^2 \in \{1, \dots, 5\}$). Higher y-axis values indicate better accuracy; color indicates R_{test}^2 , which serves as a proxy for the signal-to-noise ratio. The three hybrid variants use the same sample plug-in ω^* and are optimized by L-BFGS-B from the OLS initialization with maximum iteration count 1000; their labels report the L-BFGS-B stopping tolerances (ftol, gtol): Hybrid_tol_5e-04_5e-01 uses $(5 \times 10^{-4}, 5 \times 10^{-1})$, Hybrid_tol_5e-05_5e-02 uses $(5 \times 10^{-5}, 5 \times 10^{-2})$, and Hybrid_tol_5e-06_5e-03 uses $(5 \times 10^{-6}, 5 \times 10^{-3})$. Monte Carlo standard errors are on the order of 10^{-3} and are visually negligible at this scale.

moment-matching improves relative to two-stage, while the hybrid stays uniformly below both. OLS is flat at $4 + \|\mu_{s,x}\|^2 = 7$; the target-aware methods are substantially lower. Across noise levels (Figure S2 in the Supplement), the two-stage MSE is flat in ρ at every σ_ϵ^2 since it depends on $\Sigma_{s,x}$ but not on $\mu_{s,x}$.

Accuracy-runtime tradeoffs. The theory predicts that two-stage and moment-matching achieve nearly the same accuracy as the hybrid at a fraction of the computational cost. To quantify this empirically, we run each estimator across noise levels $\sigma_\epsilon^2 \in \{1, \dots, 5\}$ and record wall-clock time per fit, varying the hybrid’s convergence tolerance across three settings.

Figure 4 shows that OLS and the two-stage estimator lie at the far left of the curve (fastest), with the two-stage method attaining substantially higher accuracy at essentially the same computational cost. The hybrid estimator traces a Pareto frontier across tolerance levels, with some looser settings already capturing most of the attainable gains. Overall, among the fast methods, the two-stage estimator delivers striking accuracy per unit of computation at high signal-to-noise ratios.

Computability. All experiments were conducted on an Intel Core i7-14700KF (3.40GHz, 32GB RAM) and an NVIDIA GeForce RTX 4080 GPU (16GB). Codes for reproducing results in this paper can be found at <https://github.com/TZstats-Columbia/target-aware-linear-regression>.

5 Discussion

We studied how target marginal information improves regression under distribution shift. Asymptotically, the hybrid estimator is uniformly best. Among the three target-aware estimators studied here, whether it achieves a semiparametric efficiency bound remains an open question. Moment-matching and two-stage coincide in the isotropic case but diverge under joint mean

and covariance shifts. Monte Carlo experiments corroborate these rankings across regimes.

Estimator choice depends critically on the signal-to-noise ratio. The improvement conditions differ by estimator: the moment-matching estimator improves over OLS when $\sigma_\varepsilon^2 < 2\beta^\top Q_{s|k}^{-1}\beta$ (Corollary 3.2), where $Q_{s|k}$ captures both the source covariance and the mean shift. The two-stage estimator improves when $\sigma_\varepsilon^2 < 2\beta^\top \Sigma_{s,x}^{-1}\beta$ (Theorem 3.3), which depends only on $\Sigma_{s,x}$. In the isotropic, zero-shift setting these conditions coincide at $\sigma_\varepsilon^2 < 2\|\beta\|^2$, but they diverge under covariance or mean shifts. In the high signal-to-noise regime, when a strong base model is available, the two-stage estimator nearly matches the hybrid at a fraction of the computational cost (Figure 4). This yields a simple rule: when the base predictor is accurate, post hoc rescaling captures most of the benefit from distributional information, and the additional hybrid optimization offers only small gains. As base models strengthen (through more data, richer features, or deeper architectures), the near-optimality region for two-stage expands, making cheap post hoc calibration increasingly appealing. When training-time optimization is affordable and signal-to-noise is moderate, the hybrid remains preferable.

The asymptotic formulas also reveal how the benefit of target marginals scales with the covariate dimension d . For estimating the d -dimensional slope β , the scaled coefficient MSE is d for OLS and $d - 1 + \sigma_\varepsilon^2/(2\|\beta\|^2)$ for moment-matching and two-stage in the isotropic, zero-shift setting. The absolute reduction is approximately one degree of freedom, independent of d , so the relative improvement is $O(1/d)$ and all target-aware estimators converge to OLS as $d \rightarrow \infty$. This is because $\sigma_{k,y}^2$ provides a single scalar constraint on the d -dimensional β . Across all simulation regimes, we also observe consistent reductions in the test mean squared error for predicting Y when incorporating target marginals, in addition to the gains attributable solely to improved estimation of β .

Although our derived asymptotic MSE results rely on linearity and Gaussian marginals, the general finding is expected to be applicable for nonlinear models, as we observed in machine learning applications (Hou et al., 2025): post hoc calibration is most effective in high signal-to-noise regimes and captures most of the value of target marginals. While the closed-form MSE expressions we derive rely on the Gaussian-linear setting, the key structural insights stem from the auxiliary-information mechanism rather than from any specific distributional assumptions. In particular, the dependence of performance gains on the signal-to-noise ratio, the near-optimality of inexpensive post hoc calibration when the base model is already strong, and the complementary effects of imposing constraints on the mean and variance are expected to extend to more complex and realistic modeling scenarios. Hou et al. (2025) observe exactly these patterns in nonlinear neural network models calibrated via the kernelized Stein discrepancy, without any Gaussian or linearity assumption. The present paper provides the theoretical explanation for why that pattern occurs.

Supplementary Material

The supplementary material for this article is available on the official website of the Journal of Data Science at <https://jds-online.org/>. It may include additional results, appendices, tables, figures, code, and data.

References

Ben-David S, Blitzer J, Crammer K, Kulesza A, Pereira F, Vaughan JW (2010). A theory of

- learning from different domains. *Machine Learning*, 79(1–2): 151–175. [MR3108150](#)
- Bruns-Smith D, Dukes O, Feller A, Ogburn EL (2026). Augmented balancing weights as linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(3), 699–723. <https://doi.org/10.1093/jrsssb/qkaf019>. [MR5102315](#)
- Chakraborty A, Cai T (2018). Efficient and adaptive linear regression in semi-supervised settings. *The Annals of Statistics*, 46(4): 1541–1572. <https://doi.org/10.1214/17-AOS1594>. [MR3819109](#)
- Chen X, Hong H, Tarozzi A (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics*, 36(2): 808–843. [MR2396816](#)
- Devaud D, Tillé Y (2019). Deville and Särndal’s calibration: revisiting a 25-years-old successful optimization problem. *Test*, 28(4): 1033–1065. [MR4036996](#)
- Deville JC, Särndal CE (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418): 376–382. [MR1173804](#)
- Efron B, Morris C (1973). Stein’s estimation rule and its competitors—an empirical Bayes approach. *Journal of the American Statistical Association*, 68(341): 117–130. [MR0388597](#)
- Finzi MA, Boral A, Wilson AG, Sha F, Zepeda-Núñez L (2023). In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J (eds.), User-defined event sampling and uncertainty quantification in diffusion models for physical dynamical systems. In: *International Conference on Machine Learning 2023*, 10136–10152. PMLR.
- Gao F, Chan K (2023). Noniterative adjustment to regression estimators with population-based auxiliary information for semiparametric models. *Biometrics*, 79(1): 140–150. [MR4572511](#)
- Gloege L, McKinley GA, Landschützer P, Fay AR, Frölicher TL, . . . , Takano Y (2021). Quantifying errors in observationally based estimates of ocean carbon sink variability. *Global Biogeochemical Cycles*, 35(4): e2020GB006788.
- Han P, Lawless JF (2019). Empirical likelihood estimation using auxiliary summary information with different covariate distributions. *Statistica Sinica*, 29(3): 1321–1342. [MR3932520](#)
- Han P, Li H, Park SK, Mukherjee B, Taylor JM (2024). Improving prediction of linear regression models by integrating external information from heterogeneous populations: James–Stein estimators. *Biometrics*, 80(3): ujae072. [MR4896878](#)
- Hansen LP (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4): 1029–1054. [MR0666123](#)
- Hirshberg DA, Wager S (2021). Augmented minimax linear estimation. *The Annals of Statistics*, 49(6): 3206–3227. [MR4352528](#)
- Hou Z, Sun J, Venkatasubramanian S, Jin P, Li S, Zheng T (2025). Calibrating geophysical predictions under constrained probabilistic distributions. arXiv preprint: [arXiv:2512.03081](#).
- Imbens GW (1997). One-step estimators for over-identified generalized method of moments models. *The Review of Economic Studies*, 64(3): 359–383. [MR1456135](#)
- Judge GG, Griffiths WE, Hill RC, Lee TC (1980). *The Theory and Practice of Econometrics*. Wiley. [MR0594435](#)
- Kay JE, Deser C, Phillips A, Mai A, Hannay C, . . . , Vertenstein M (2015). The Community Earth System Model (CESM) large ensemble project: A community resource for studying climate change in the presence of internal climate variability. *Bulletin of the American Meteorological Society*, 96(8): 1333–1349.
- Liu Q, Lee J, Jordan M (2016). A kernelized Stein discrepancy for goodness-of-fit tests. In: Balcan MF, Weinberger KQ (eds.), *International Conference on Machine Learning 2016*, 276–284. PMLR.

- Portier F, Truquet L, Yamane I (2024). Nearest neighbor sampling for covariate shift adaptation. *Journal of Machine Learning Research*, 25(410): 1–42. [MR4869955](#)
- Qin J, Lawless J (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1): 300–325. [MR1272085](#)
- Särndal CE (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33(2): 99–119.
- Shimodaira H (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2): 227–244. [MR1795598](#)
- Stein C (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Neyman J (ed.), *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, 197–206. University of California Press. [MR0084922](#)
- Stojanov P, Gong M, Carbonell J, Zhang K (2019). Low-dimensional density ratio estimation for covariate shift correction. In: Chaudhuri K, Sugiyama M (eds.), *International Conference on Artificial Intelligence and Statistics 2019*, 3449–3458. PMLR.
- Tan Z (2020). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1): 137–158. [MR4064145](#)
- Tang C, Lenssen N, Wei Y, Zheng T (2023). Wasserstein distributional learning via majorization-minimization. In: Ruiz F, Dy J, van de Meent J-W (eds.), *International Conference on Artificial Intelligence and Statistics 2023*, 10703–10731. PMLR.
- Zellinger W, Kindermann S, Pereverzyev SV (2023). Adaptive learning of density ratios in RKHS. *Journal of Machine Learning Research*, 24(395): 1–28. [MR4720851](#)
- Zhong W, Meidani H (2023). PI-VAE: Physics-informed variational auto-encoder for stochastic differential equations. *Computer Methods in Applied Mechanics and Engineering*, 403: 115664. [MR4500668](#)