

Supplement to: Target-Aware Linear Regression Under Distribution Shift

Zhewen Hou and Tian Zheng

We use the notation from Section 3 throughout:

$$\begin{aligned} Q_s &= \mathbb{E}(\tilde{X}\tilde{X}^\top), \\ Q_{s|k} &= \Sigma_{s,x} + (\mu_{s,x} - \mu_{k,x})(\mu_{s,x} - \mu_{k,x})^\top, \\ v_{\sigma\beta} &= \Sigma_{k,x}\beta, \kappa = v_{\sigma\beta}^\top Q_{s|k}^{-1} v_{\sigma\beta}, \\ \chi &= v_{\sigma\beta}^\top Q_{s|k}^{-1} \Sigma_{k,x} Q_{s|k}^{-1} v_{\sigma\beta}, \\ \text{and } \tilde{v}_{\sigma\beta} &= \tilde{\Sigma}_{k,x}\theta. \end{aligned}$$

S.1 Detailed Estimator Derivations

S.1.1 OLS prediction error under the target distribution

Under Assumptions 3.1–3.2, $\sqrt{n}(\hat{\theta}^{\text{OLS}} - \theta) \Rightarrow \mathcal{N}(0, \sigma_\varepsilon^2 Q_s^{-1})$, so

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

If (X, Y) is independent of the training sample with $X \sim \mathcal{N}(\mu_{k,x}, \Sigma_{k,x})$ and $Y = \beta_0 + X^\top \beta + \varepsilon$, then

$$\begin{aligned} \mathbb{E}\left[\left(Y - \tilde{X}^\top \hat{\theta}^{\text{OLS}}\right)^2\right] &= \mathbb{E}[\varepsilon^2] + \mathbb{E}\left[\left(\tilde{X}^\top (\theta - \hat{\theta}^{\text{OLS}})\right)^2\right] \\ &= \sigma_\varepsilon^2 + \text{tr}\left(\mathbb{E}[\tilde{X}\tilde{X}^\top] \cdot \mathbb{E}[(\hat{\theta}^{\text{OLS}} - \theta)(\hat{\theta}^{\text{OLS}} - \theta)^\top]\right) \\ &= \sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n} \text{tr}\left((\tilde{\Sigma}_{k,x} + \tilde{\mu}_{k,x}\tilde{\mu}_{k,x}^\top)Q_s^{-1}\right) + o(n^{-1}). \end{aligned} \quad (\text{S.1})$$

Under the normalization $\mu_{k,x} = 0$, $\Sigma_{k,x} = I_d$, this gives

$$\sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n} \left[1 + \text{tr}(\Sigma_{s,x}^{-1}) + \mu_{s,x}^\top \Sigma_{s,x}^{-1} \mu_{s,x}\right] + o(n^{-1}),$$

which recovers (1) in the main text after writing

$$\Sigma_{s,x}^{-1} = \Sigma_{k,x} \Sigma_{s,x}^{-1}$$

and

$$\mu_{s,x}^\top \Sigma_{s,x}^{-1} \mu_{s,x} = (\mu_{k,x} - \mu_{s,x})^\top \Sigma_{s,x}^{-1} (\mu_{k,x} - \mu_{s,x}).$$

S.1.2 MM: constraint structure and minimizer selection

The MM loss $L_{\text{MM}}(\theta)$ defined in (5) is a sum of two nonnegative squares. It attains its minimum value zero if and only if

$$\tilde{\mu}_{k,x}^\top \theta = \mu_{k,y} \quad \text{and} \quad \theta^\top \tilde{\Sigma}_{k,x} \theta + \frac{1}{n} \|\mathbf{Y} - \tilde{\mathbf{X}}\theta\|_2^2 = \sigma_{k,y}^2.$$

The first equation (mean constraint) gives $\beta_0 = \mu_{k,y} - \mu_{k,x}^\top \beta$. Substituting into the second equation yields a quadratic in β .

Define the empirical moments

$$\hat{M}_{xx} := \frac{1}{n} \sum_{i=1}^n \check{X}_i \check{X}_i^\top, \quad \hat{m}_{xy} := \frac{1}{n} \sum_{i=1}^n \check{X}_i \check{Y}_i, \quad \hat{m}_{yy} := \frac{1}{n} \sum_{i=1}^n \check{Y}_i^2,$$

where $\check{X}_i := X_i - \mu_{k,x}$ and $\check{Y}_i := Y_i - \mu_{k,y}$. Let

$$A_n := \Sigma_{k,x} + \hat{M}_{xx}, \quad c_n := \hat{m}_{yy} - \sigma_{k,y}^2.$$

Lemma S.1 (Zero-loss characterization). *The loss $L_{\text{MM}}(\theta)$ attains its minimum value zero if and only if the quadratic*

$$q_n(\beta) := \beta^\top A_n \beta - 2\hat{m}_{xy}^\top \beta + c_n = 0 \tag{S.2}$$

has a real solution. In that case, $\hat{\Theta}^{\text{MM}} = \{\theta(\beta) : q_n(\beta) = 0\}$ where $\theta(\beta) = (\mu_{k,y} - \mu_{k,x}^\top \beta, \beta^\top)^\top$.

Since A_n is positive definite, we complete the square:

$$q_n(\beta) = (\beta - A_n^{-1} \hat{m}_{xy})^\top A_n (\beta - A_n^{-1} \hat{m}_{xy}) - \Delta_n, \tag{S.3}$$

where $\Delta_n := \hat{m}_{xy}^\top A_n^{-1} \hat{m}_{xy} - c_n$.

Case $\Delta_n = 0$. The quadratic (S.2) has the unique solution $\beta = A_n^{-1} \hat{m}_{xy}$.

Case $\Delta_n > 0$. When $d > 1$, equation (S.2) has infinitely many solutions forming an ellipsoid. The MM estimator selects the one that minimizes the residual sum of squares. On the constraint set $q_n(\beta) = 0$, we have $\frac{1}{n} \sum_{i=1}^n (\check{Y}_i - \check{X}_i^\top \beta)^2 = \sigma_{k,y}^2 - \beta^\top \Sigma_{k,x} \beta$. Therefore, (5) is equivalent to

$$\max_{\beta \in \mathbb{R}^d} \beta^\top \Sigma_{k,x} \beta \quad \text{s.t.} \quad q_n(\beta) = 0.$$

Proposition S.1 (KKT characterization of the MM minimizer). *Let $\rho_n := \lambda_{\max}(A_n^{-1/2} \Sigma_{k,x} A_n^{-1/2})$. If $\Delta_n > 0$, then there exists a unique $\hat{\lambda} \geq \rho_n$ such that*

$$(\hat{\lambda} A_n - \Sigma_{k,x}) \hat{\beta}^{\text{MM}} = \hat{\lambda} \hat{m}_{xy}, \quad q_n(\hat{\beta}^{\text{MM}}) = 0. \tag{S.4}$$

Moreover, if $\hat{\lambda} > \rho_n$, then $\hat{\beta}^{\text{MM}} = \hat{\lambda} (\hat{\lambda} A_n - \Sigma_{k,x})^{-1} \hat{m}_{xy}$.

The Lagrange multiplier $\hat{\lambda}$ is found by substituting $\hat{\beta}^{\text{MM}} = \hat{\lambda} (\hat{\lambda} A_n - \Sigma_{k,x})^{-1} \hat{m}_{xy}$ into the constraint $q_n(\hat{\beta}^{\text{MM}}) = 0$ and solving the resulting scalar equation numerically.

Case $\Delta_n < 0$. The quadratic has no real solution, so $\min L_{\text{MM}} > 0$. Define

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n \check{X}_i, \quad \bar{Y} := \frac{1}{n} \sum_{i=1}^n \check{Y}_i.$$

Lemma S.2. Suppose $\Delta_n < 0$. For $t \in \mathbb{R}$, define

$$\beta(t) := A_n^{-1} \left(\hat{m}_{xy} - t \bar{X} \right), \quad \beta_0(t) := \mu_{k,y} - \mu_{k,x}^\top \beta(t) + t,$$

and

$$u_n(t) := \hat{m}_{yy} - \hat{m}_{xy}^\top A_n^{-1} \hat{m}_{xy} + 2 \left(\hat{m}_{xy}^\top A_n^{-1} \bar{X} - \bar{Y} \right) t + \left(1 - \bar{X}^\top A_n^{-1} \bar{X} \right) t^2.$$

Then every element of $\hat{\Theta}^{\text{MM}}$ has the form $(\beta_0(t), \beta(t))$ for some real t satisfying

$$(2t + u'_n(t))^2 u_n(t) - \sigma_{k,y}^2 (u'_n(t))^2 = 0. \quad (\text{S.5})$$

With probability tending to one, $\Delta_n > 0$ when $\beta \neq 0$, so this case is asymptotically negligible.

S.1.3 Hybrid: gradient and optimal weight

The hybrid loss $L_{\text{H}}(\theta; \omega)$ in (2) does not admit a closed-form minimizer. Let $s(\theta) := n^{-1} \|\mathbf{Y} - \tilde{\mathbf{X}}\theta\|_2^2$ and $t(\theta) := \sqrt{\theta^\top \tilde{\Sigma}_{k,x} \theta + s(\theta)}$. A direct differentiation gives the gradient

$$\begin{aligned} \nabla_{\theta} L_{\text{H}}(\theta; \omega) = & -\frac{2}{n} \tilde{\mathbf{X}}^\top (\mathbf{Y} - \tilde{\mathbf{X}}\theta) + 2\omega \left(\tilde{\mu}_{k,x}^\top \theta - \mu_{k,y} \right) \tilde{\mu}_{k,x} \\ & + 2\omega \frac{t(\theta) - \sigma_{k,y}}{t(\theta)} \left(\tilde{\Sigma}_{k,x} \theta - \frac{1}{n} \tilde{\mathbf{X}}^\top (\mathbf{Y} - \tilde{\mathbf{X}}\theta) \right). \end{aligned} \quad (\text{S.6})$$

In practice, we solve $\nabla_{\theta} L_{\text{H}} = 0$ by gradient descent initialized at $\hat{\theta}^{\text{OLS}}$.

Optimal ω^* via quartic equation. By Corollary 3.1, the asymptotic MSE is $(\sigma_{\varepsilon}^2/n) V_{\text{H}}(\omega) + o(n^{-1})$. To find the optimal weight, define

$$m_1 := \tilde{\mu}_{k,x}^\top Q_s^{-1} \tilde{\mu}_{k,x}, \quad m_2 := \frac{1}{\sigma_{k,y}} \tilde{\mu}_{k,x}^\top Q_s^{-1} \tilde{v}_{\sigma\beta}, \quad m_3 := \frac{1}{\sigma_{k,y}^2} \tilde{v}_{\sigma\beta}^\top Q_s^{-1} \tilde{v}_{\sigma\beta},$$

$$n_1 := \tilde{\mu}_{k,x}^\top Q_s^{-1} P_\beta^\top P_\beta Q_s^{-1} \tilde{\mu}_{k,x}, \quad n_2 := \frac{1}{\sigma_{k,y}} \tilde{\mu}_{k,x}^\top Q_s^{-1} P_\beta^\top P_\beta Q_s^{-1} \tilde{v}_{\sigma\beta},$$

$$n_3 := \frac{1}{\sigma_{k,y}^2} \tilde{v}_{\sigma\beta}^\top Q_s^{-1} P_\beta^\top P_\beta Q_s^{-1} \tilde{v}_{\sigma\beta},$$

and the denominator polynomial

$$\Delta_{\text{H}}(\omega) := 1 + (m_1 + m_3)\omega + (m_1 m_3 - m_2^2)\omega^2.$$

Proposition S.2 (Rational form of V_H). *Under Assumptions 3.1–3.5,*

$$V_H(\omega) = \text{tr}\left(P_\beta Q_s^{-1} P_\beta^\top\right) + \frac{A_1\omega + A_2\omega^2 + A_3\omega^3 + A_4\omega^4}{\Delta_H(\omega)^2}, \quad (\text{S.7})$$

where

$$A_1 = -2(n_1 + n_3),$$

$$A_2 = -(n_1 + n_3)(m_1 + m_3) - 3(m_1n_3 + m_3n_1 - 2m_2n_2) + \frac{\sigma_\varepsilon^2}{2\sigma_{k,y}^2} n_3,$$

$$A_3 = -2(m_1 + m_3)(m_1n_3 + m_3n_1 - 2m_2n_2) + \frac{\sigma_\varepsilon^2}{\sigma_{k,y}^2} (m_1n_3 - m_2n_2),$$

$$A_4 = -(m_1m_3 - m_2^2)(m_1n_3 + m_3n_1 - 2m_2n_2) + \frac{\sigma_\varepsilon^2}{2\sigma_{k,y}^2} (m_1^2n_3 - 2m_1m_2n_2 + m_2^2n_1).$$

Any minimizer $\omega^\star > 0$ satisfies the quartic equation

$$B_0 + B_1\omega^\star + B_2(\omega^\star)^2 + B_3(\omega^\star)^3 + B_4(\omega^\star)^4 = 0, \quad (\text{S.8})$$

where

$$B_0 = A_1, \quad B_1 = 2A_2 - (m_1 + m_3)A_1,$$

$$B_2 = 3A_3 - 3(m_1m_3 - m_2^2)A_1,$$

$$B_3 = 4A_4 + (m_1 + m_3)A_3 - 2(m_1m_3 - m_2^2)A_2,$$

$$B_4 = 2(m_1 + m_3)A_4 - (m_1m_3 - m_2^2)A_3.$$

In practice, one collects the nonnegative real roots of (S.8) and compares V_H at those roots and at $\omega = 0$.

S.1.4 MM prediction error

If (X, Y) is independent of the training sample with $X \sim \mathcal{N}(\mu_{k,x}, \Sigma_{k,x})$ and $Y = \beta_0 + X^\top \beta + \varepsilon$, then by the same argument as (S.1),

$$\mathbb{E}\left[\left(Y - \tilde{X}^\top \hat{\theta}^{\text{MM}}\right)^2\right] = \sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n} \text{tr}\left(\left(\tilde{\Sigma}_{k,x} + \tilde{\mu}_{k,x}\tilde{\mu}_{k,x}^\top\right) J_{\mu_k}^\top Q_{s|k}^{-1} \Omega Q_{s|k}^{-1} J_{\mu_k}\right) + o(n^{-1}). \quad (\text{S.9})$$

S.1.5 Hybrid prediction error

Similarly, the prediction error of the hybrid estimator under the target distribution is

$$\mathbb{E}\left[\left(Y - \tilde{X}^\top \hat{\theta}^H(\omega)\right)^2\right] = \sigma_\varepsilon^2 + \frac{\sigma_\varepsilon^2}{n} \text{tr}\left(\left(\tilde{\Sigma}_{k,x} + \tilde{\mu}_{k,x}\tilde{\mu}_{k,x}^\top\right) Q(\omega)^{-1} \Omega(\omega) Q(\omega)^{-1}\right) + o(n^{-1}). \quad (\text{S.10})$$

S.2 Proofs

S.2.1 Proof of Theorem 3.1

The proof proceeds via a mean value expansion of the gradient of L_H .

Consistency. The population criterion $L_{H,\infty}(\theta; \omega) := \mathbb{E}[(Y - \tilde{X}^\top \theta)^2] + \omega(\tilde{\mu}_{k,x}^\top \theta - \mu_{k,y})^2 + \omega(\sqrt{\theta^\top \tilde{\Sigma}_{k,x} \theta + \mathbb{E}[(Y - \tilde{X}^\top \theta)^2] - \sigma_{k,y}^2})^2$ has θ as its unique minimizer because the first term is strictly convex and equals σ_ε^2 only at the true parameter. Since $L_H(\cdot; \omega)$ converges uniformly on compacts to $L_{H,\infty}(\cdot; \omega)$ by the law of large numbers, standard M-estimation arguments give $\hat{\theta}^H(\omega) \xrightarrow{p} \theta$.

Linearization. Let $\psi_n(\theta) := \frac{1}{2} \nabla_\theta L_H(\theta; \omega)$. Define

$$U_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{X}_i \varepsilon_i, \quad W_n := \frac{1}{\sqrt{n}} \sum_{i=1}^n (\varepsilon_i^2 - \sigma_\varepsilon^2).$$

Since $\tilde{\mu}_{k,x}^\top \theta = \mu_{k,y}$ and $t(\theta)^2 = \theta^\top \tilde{\Sigma}_{k,x} \theta + \frac{1}{n} \sum \varepsilon_i^2 = \sigma_{k,y}^2 + \frac{1}{\sqrt{n}} W_n$, a Taylor expansion gives

$$\sqrt{n} \psi_n(\theta) = -U_n + \frac{\omega}{2\sigma_{k,y}^2} W_n \tilde{v}_{\sigma\beta} + o_p(1).$$

By the multivariate central limit theorem, $(U_n, W_n) \Rightarrow \mathcal{N}(0, \text{diag}(\sigma_\varepsilon^2 Q_s, 2\sigma_\varepsilon^4))$, since $\mathbb{E}[\tilde{X}\varepsilon] = 0$, $\mathbb{E}[\tilde{X}\varepsilon(\varepsilon^2 - \sigma_\varepsilon^2)] = 0$, and $\text{Var}(\varepsilon^2 - \sigma_\varepsilon^2) = 2\sigma_\varepsilon^4$. Hence $-\sqrt{n} \psi_n(\theta) \Rightarrow Z$ with $\text{Var}(Z) = \sigma_\varepsilon^2 \Omega(\omega)$.

Hessian limit. Differentiating ψ_n and evaluating at a consistent intermediate point $\tilde{\theta}_n$ on the line segment between θ and $\hat{\theta}^H(\omega)$ yields $\nabla_\theta \psi_n(\tilde{\theta}_n) \xrightarrow{p} Q(\omega)$, which is positive definite.

Conclusion. The first-order condition $\psi_n(\hat{\theta}^H(\omega)) = 0$ and the mean value expansion give

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

S.2.2 Proof of Theorem 3.2

Write $\tilde{X}_i := X_i - \mu_{k,x}$, $\tilde{Y}_i := Y_i - \mu_{k,y}$, and define $\hat{M}_{xx} := \frac{1}{n} \sum_i \tilde{X}_i \tilde{X}_i^\top$, $\hat{m}_{xy} := \frac{1}{n} \sum_i \tilde{X}_i \tilde{Y}_i$, $\hat{m}_{yy} := \frac{1}{n} \sum_i \tilde{Y}_i^2$, $A_n := \Sigma_{k,x} + \hat{M}_{xx}$, and $c_n := \hat{m}_{yy} - \sigma_{k,y}^2 + \beta^\top \Sigma_{k,x} \beta$, so that the MM variance constraint becomes $q_n(\beta) := \beta^\top A_n \beta - 2\hat{m}_{xy}^\top \beta + c_n = 0$.

Consistency. Under the true parameter, the MM loss is exactly zero. The constraint set is nonempty with probability $1 - O(n^{-1})$ (shown by verifying the discriminant $\Delta_n = v_{\sigma\beta}^\top A_n^{-1} v_{\sigma\beta} \cdot (\beta^\top A_n \beta - c_n) > 0$ with high probability). The RSS-minimizing element converges to β by the continuous mapping theorem.

Linearization. Define $G_n(b, \lambda) := (\lambda(A_n b - \hat{m}_{xy}) - \Sigma_{k,x} b, q_n(b))^\top$. The MM first-order conditions are $G_n(\hat{\beta}^{\text{MM}}, \hat{\lambda}) = 0$ where $\hat{\lambda}$ is a Lagrange multiplier converging to 1. A mean value expansion gives

$$\sqrt{n} \begin{pmatrix} \hat{\beta}^{\text{MM}} - \beta \\ \hat{\lambda} - 1 \end{pmatrix} = -D^{-1} \sqrt{n} G_n(\beta, 1) + o_p(1), \quad D := \begin{pmatrix} Q_{s|k} & v_{\sigma\beta} \\ 2v_{\sigma\beta}^\top & 0 \end{pmatrix}.$$

Score evaluation. Substituting the true parameter,

$$\sqrt{n} G_n(\beta, 1) = \begin{pmatrix} -U_n \\ W_n \end{pmatrix}, \quad U_n := \frac{1}{\sqrt{n}} \sum_i \check{X}_i \varepsilon_i, \quad W_n := \frac{1}{\sqrt{n}} \sum_i (\varepsilon_i^2 - \sigma_\varepsilon^2),$$

which is jointly Gaussian with $\text{Var}(U_n) = \sigma_\varepsilon^2 Q_{s|k}$ and $\text{Var}(W_n) = 2\sigma_\varepsilon^4$ (independent).

Solving. Block inversion of D yields

$$\sqrt{n}(\hat{\beta}^{\text{MM}} - \beta) \Rightarrow Q_{s|k}^{-1} U - \frac{v_{\sigma\beta}^\top Q_{s|k}^{-1} U}{\kappa} Q_{s|k}^{-1} v_{\sigma\beta} - \frac{W}{2\kappa} Q_{s|k}^{-1} v_{\sigma\beta},$$

whose covariance is $\sigma_\varepsilon^2 Q_{s|k}^{-1} \Omega Q_{s|k}^{-1}$ with $\Omega = Q_{s|k} - \kappa^{-1} v_{\sigma\beta} v_{\sigma\beta}^\top + (\sigma_\varepsilon^2/2) \kappa^{-2} v_{\sigma\beta} v_{\sigma\beta}^\top$. Since $\hat{\beta}_0^{\text{MM}} - \beta_0 = -\mu_{k,x}^\top (\hat{\beta}^{\text{MM}} - \beta)$, the joint result follows.

S.2.3 Proof of Theorem 3.3

We work under the normalization $\mu_{k,x} = 0$ and $\Sigma_{k,x} = I_d$. Let $\delta_n := \hat{\beta}^{\text{OLS}} - \beta$, $\tau_n := \hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2$, and $r^2 := \|\beta\|^2$. Since $\beta \neq 0$ and $\sigma_{k,y}^2 = r^2 + \sigma_\varepsilon^2$,

$$b = \left(\frac{(r^2 - \tau_n)_+}{\|\beta + \delta_n\|^2} \right)^{1/2}.$$

By standard OLS theory, $\delta_n = O_p(n^{-1/2})$ and $\tau_n = O_p(n^{-1/2})$. Hence the positive part is inactive with probability tending to one, and Taylor expansion gives

$$b = 1 - \frac{\beta^\top \delta_n}{r^2} - \frac{\tau_n}{2r^2} + o_p(n^{-1/2}).$$

Therefore

$$\sqrt{n} (\hat{\beta}^{\text{cali}} - \beta) = \sqrt{n} \delta_n - \frac{\beta^\top \sqrt{n} \delta_n}{r^2} \beta - \frac{\sqrt{n} \tau_n}{2r^2} \beta + o_p(1).$$

Conditional on the design, the OLS coefficients and the residual variance estimator are independent, and $(n - d - 1) \hat{\sigma}_\varepsilon^2 / \sigma_\varepsilon^2 \sim \chi_{n-d-1}^2$. Thus

$$\sqrt{n} \delta_n \Rightarrow Z, \quad \sqrt{n} \tau_n \Rightarrow W,$$

where $Z \sim \mathcal{N}(0, \sigma_\varepsilon^2 \Sigma_{s,x}^{-1})$, $W \sim \mathcal{N}(0, 2\sigma_\varepsilon^4)$, and $Z \perp W$. It follows that

$$\sqrt{n} (\hat{\beta}^{\text{cali}} - \beta) \Rightarrow Z - \frac{\beta^\top Z}{r^2} \beta - \frac{W}{2r^2} \beta.$$

The limiting second moment is

$$\begin{aligned} \mathbb{E} \left\| Z - \frac{\beta^\top Z}{r^2} \beta - \frac{W}{2r^2} \beta \right\|^2 &= \mathbb{E} \|Z\|^2 - \frac{\mathbb{E}(\beta^\top Z)^2}{r^2} + \frac{\mathbb{E} W^2}{4r^2} \\ &= \sigma_\varepsilon^2 \text{tr}(\Sigma_{s,x}^{-1}) - \sigma_\varepsilon^2 \frac{\beta^\top \Sigma_{s,x}^{-1} \beta}{\|\beta\|^2} + \frac{\sigma_\varepsilon^4}{2\|\beta\|^2}. \end{aligned}$$

The required uniform integrability follows from the sub-Gaussian assumption on X and the Gaus-

sianity of ε . Hence

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

For prediction, since the test covariates have empirical mean zero and empirical covariance I_d , $\tilde{Y}_j = \mu_{k,y} + (X_j^{\text{test}})^\top \hat{\beta}^{\text{cali}}$. Also $\mu_{k,y} = \beta_0$ under the normalization. Therefore

$$\frac{1}{m} \sum_{j=1}^m \mathbb{E} \left[(\tilde{Y}_j - Y_j^{\text{test}})^2 \right] = \sigma_\varepsilon^2 + \mathbb{E} \left[\left\| \hat{\beta}^{\text{cali}} - \beta \right\|^2 \right],$$

which gives (8).

S.3 Additional Experimental Results

S.3.1 Covariance geometry across noise levels

Figure S1 displays the full covariance geometry experiment across multiple noise levels σ_ε^2 . At low noise (high signal-to-noise ratio), the three target-aware estimators are nearly indistinguishable and all substantially improve over OLS. As the noise increases, the estimators separate: moment-matching (MM) continues to exploit the full geometry of $Q_{s|k}$ and improves steadily as $\beta^\top \Sigma_{s,x}^{-1} \beta$ grows, whereas the two-stage estimator, which projects only onto β through $\Sigma_{s,x}$, improves more slowly. The hybrid remains uniformly best across all noise levels and geometry configurations.

S.3.2 Mean mismatch across noise levels

Figure S2 displays the full mean mismatch experiment across multiple noise levels. The two-stage estimator is flat in ρ at every noise level because its MSE (7) depends on $\Sigma_{s,x}$ but not on $\mu_{s,x}$. Moment-matching, by contrast, benefits from mean shifts in the $\beta^\perp$ direction through $Q_{s|k}$, and this advantage grows with noise. At low noise all three target-aware estimators cluster together; at higher noise the ordering hybrid < moment-matching < two-stage becomes pronounced, with moment-matching gaining the most from favorable mean-shift geometry.

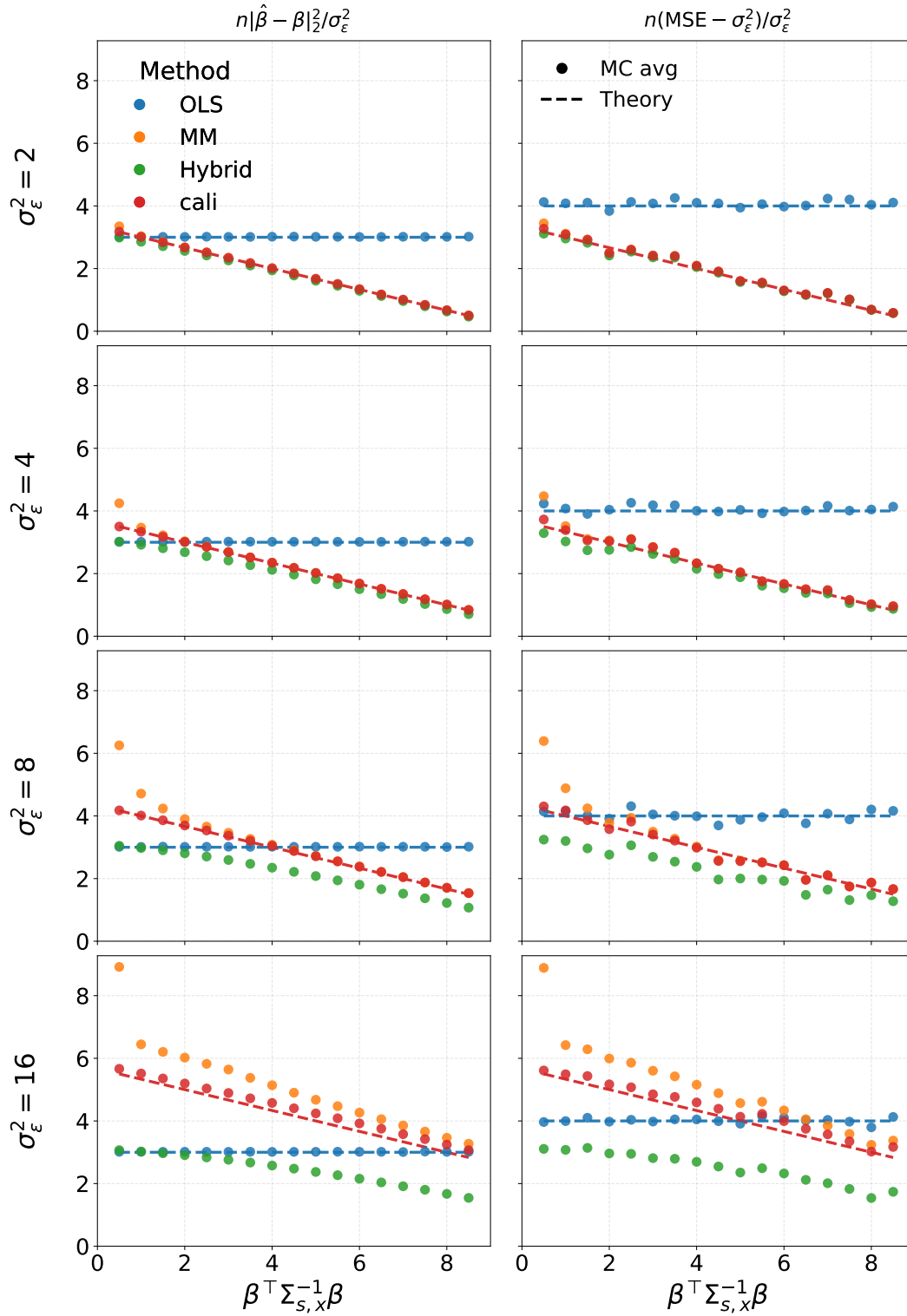


Figure S1: **Covariance geometry** across noise levels. Scaled coefficient error (left) and scaled excess prediction error (right) vs. $\beta^\top \Sigma_{s,x}^{-1} \beta$, for multiple values of σ_ϵ^2 . Points: Monte Carlo averages ($L = 10^6$, $n = 1000$); dashed lines: theoretical values.

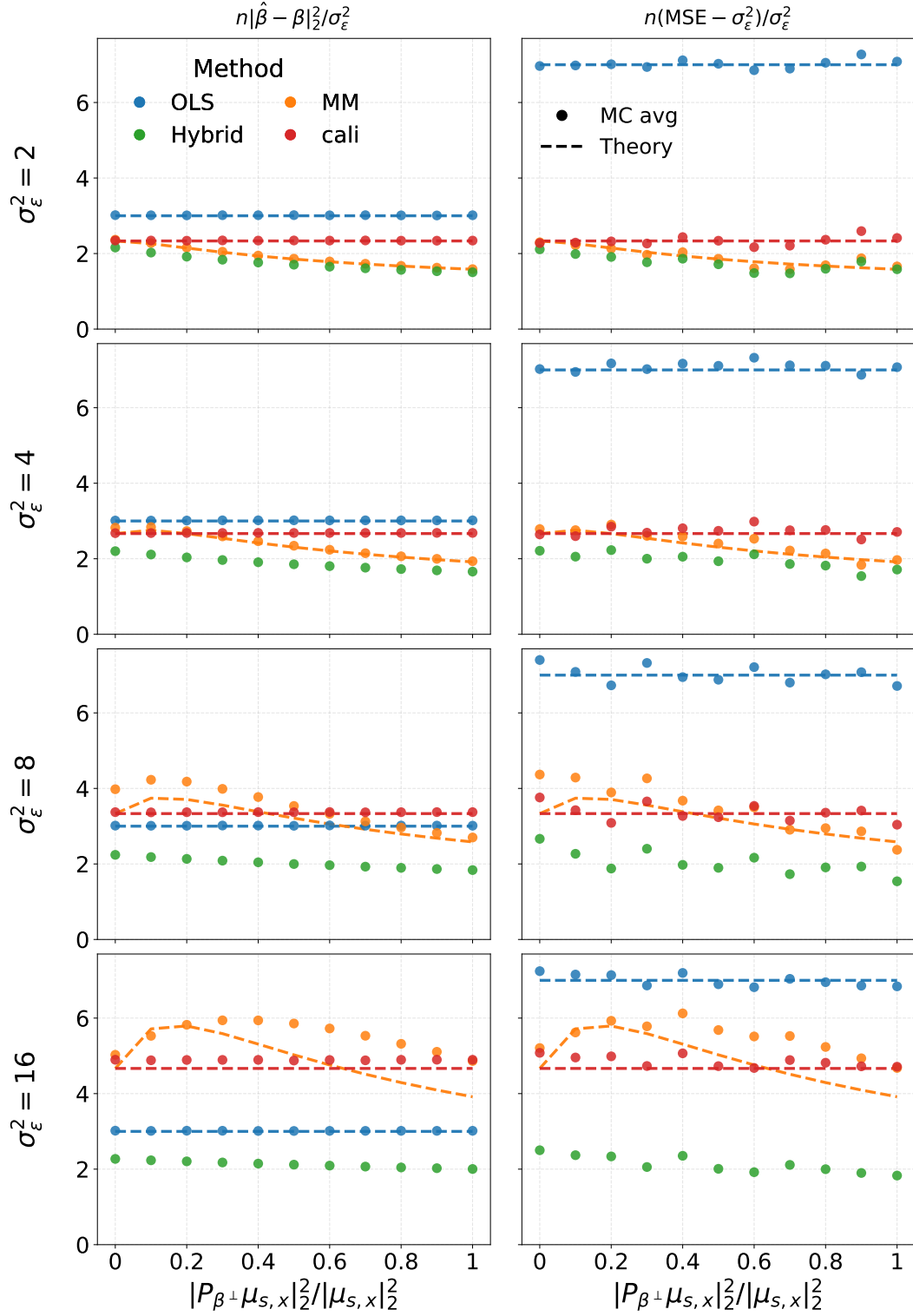


Figure S2: **Mean mismatch** across noise levels. Scaled errors vs. $\rho = \|P_{\beta^\perp} \mu_{s,x}\|^2 / \|\mu_{s,x}\|^2$, for multiple values of σ_ϵ^2 . Points: Monte Carlo averages ($L = 10^6$, $n = 1000$); dashed lines: theoretical values.