PNAS
www.pnas.org

Supporting Information for

Cooperative Learning for Multiview Analysis

Daisy Yi Ding, Shuangning Li, Balasubramanian Narasimhan and Robert Tibshirani

Robert Tibshirani.
E-mail: tibs@stanford.edu

This PDF file includes:

Supporting Information Text

1. Adaptive cooperative learning

In this section, we outline an adaptive strategy for optimizing over λ_x and λ_z for different data views. We call this adaptive cooperative learning. The method incorporates the values of λ_x and λ_z that have been adaptively optimized by the one-at-a-time algorithm (Algorithm S1) as a penalty factor in the direct algorithm (Algorithm S2). In the two-dimensional grid of λ_x and λ_z, our proposed strategy works by iteratively searching along one axis of λ while fixing the other constant.

Algorithm S1 *One-at-a-time algorithm for adaptive cooperative learning (regression).*

Input: $X \in \mathcal{R}^{n\times p_x}$ and $Z \in \mathcal{R}^{n\times p_z}$, the response $\boldsymbol{y} \in \mathcal{R}^n$, and a fixed hyperparameter $\rho \in \mathcal{R}$.
Output: $\hat{\boldsymbol{\theta}_x}$ and $\hat{\boldsymbol{\theta}_z}$ from the last iteration, along with the hyperparameters λ_x^* and λ_z^* and the corresponding CV errors.

1. Initialize $\boldsymbol{\theta}_x^{(0)} \in \mathcal{R}^{p_x}$ and $\boldsymbol{\theta}_z^{(0)} \in \mathcal{R}^{p_z}$.
2. For $k = 0, 1, 2, \ldots$ until convergence:
   (a) Set $\boldsymbol{y_x^*} = \frac{\boldsymbol{y}}{1+\rho} - \frac{(1-\rho)Z\boldsymbol{\theta_z}}{(1+\rho)}$. Solve Lasso$(X, \boldsymbol{y}_x^*, \lambda)$ over a decreasing grid of λ values. Update $\boldsymbol{\theta}_x^{(k+1)}$ to be the solution and record the hyperparameter λ_x^* that minimizes the CV error.
   (b) Set $\boldsymbol{y_z^*} = \frac{\boldsymbol{y}}{1+\rho} - \frac{(1-\rho)X\boldsymbol{\theta_x}}{(1+\rho)}$. Solve Lasso$(Z, \boldsymbol{y}_z^*, \lambda)$ over a decreasing grid of λ values. Update $\boldsymbol{\theta}_z^{(k+1)}$ to be the solution and record the hyperparameter λ_z^* that minimizes the CV error.

Algorithm S2 *Direct algorithm for adaptive cooperative learning (regression).*

Input: $X \in \mathcal{R}^{n\times p_x}$ and $Z \in \mathcal{R}^{n\times p_z}$, the response $\boldsymbol{y} \in \mathcal{R}^n$, and a grid of hyperparameter values $(\rho_{\texttt{min}}, \ldots, \rho_{\texttt{max}})$.

for $\rho \leftarrow \rho_{\texttt{min}}, \ldots, \rho_{\texttt{max}}$ **do**
  Run `Algorithm S1` with both (X,Z) and (Z,X) with the same folds for CV. Select the one with the lower sum of the two CV errors. Get the corresponding λ_x^* and λ_z^*.

  Set

$$\tilde{X} = \begin{pmatrix} X & Z \\ -\sqrt{\rho}X & \sqrt{\rho}Z \end{pmatrix}, \tilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}.$$

  Solve Lasso$(\tilde{X}, \tilde{\boldsymbol{y}}, \lambda)$ over a decreasing grid of λ values, with a penalty factor of $(1, \ldots, 1, \frac{\lambda_z^*}{\lambda_x^*}, \ldots, \frac{\lambda_z^*}{\lambda_x^*})$. Note that we form folds from the rows of X and Z and then construct the corresponding $\tilde{X}$.
end
Select the optimal value of ρ based on the CV error and get the final fit.

2. More comprehensive simulation studies on cooperative regularized regression

Fig. S1-S3 show more simulation results of the high-dimensional settings ($p = 1000, n = 200$), and Fig. S4-S6 show the simulation results of the lower-dimensional settings ($p = 200, n = 500$).

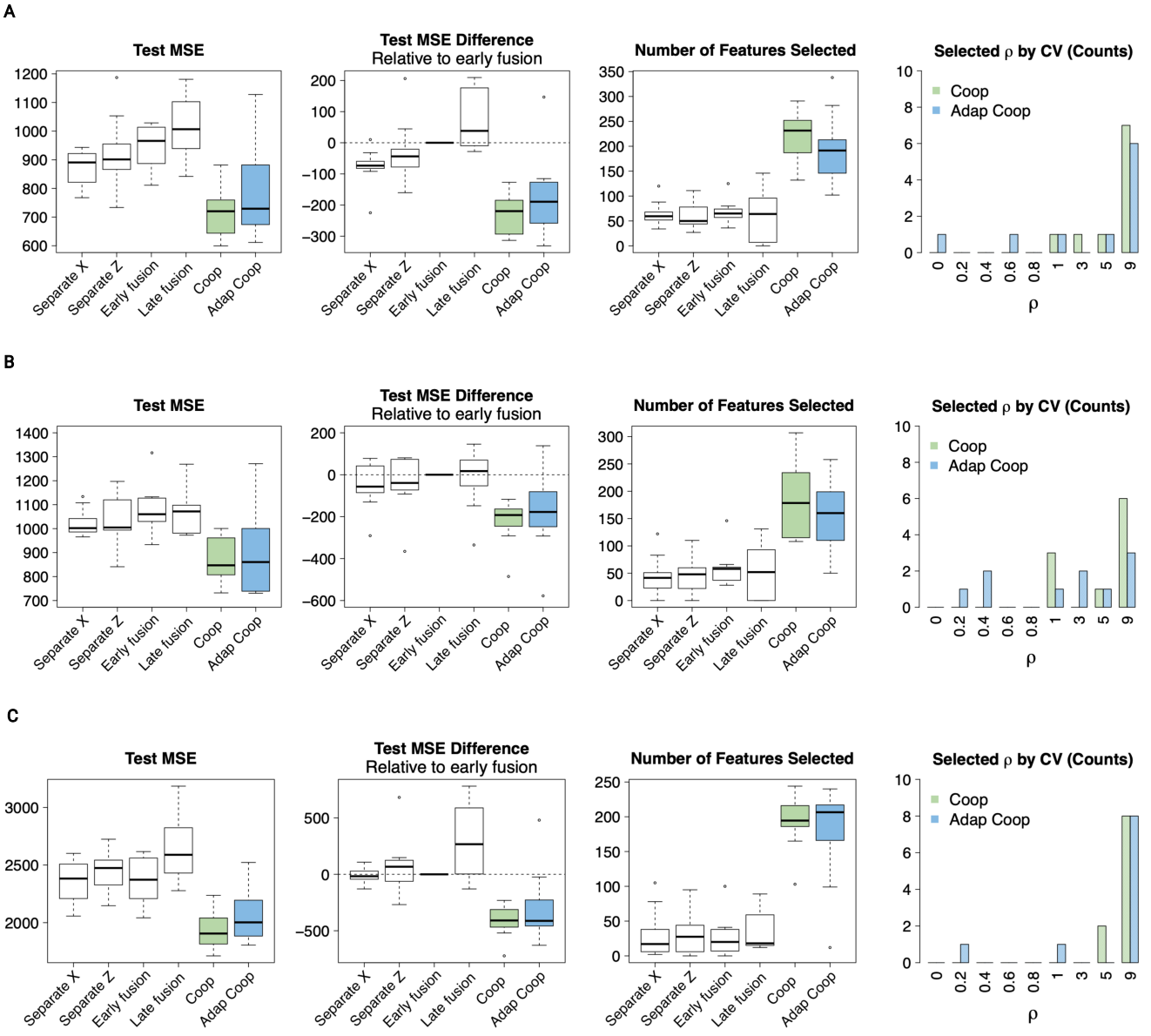

Fig. S1. Simulation studies on cooperative regularized linear regression when X and Z are high-dimensional and have a high level of correlation with each other. *(A)* Simulation results when X and Z have a high level of correlation and both contain signal ($t_x = t_z = 6$), $n = 200, p = 1000$, SNR $= 1.0$. The first panel shows MSE on a test test; the second panel shows the MSE difference on the test set relative to early fusion; the third panel shows the number of features selected; the fourth panel shows the ρ values selected by CV in cooperative learning. Here "Coop" refers to cooperative learning outlined in Algorithm 1 and "Adap Coop" refers to adaptive cooperative learning outlined in Algorithm S2. *(B)* Simulation results when X and Z have a high level of correlation and both contain signal ($t_x = t_z = 6$), $n = 200, p = 1000$, SNR $= 0.6$. *(C)* Simulation results when X and Z have a high level of correlation, X contains more signal than Z ($t_x = 4, t_z = 2$), $n = 200, p = 1000$, SNR $= 0.6$.

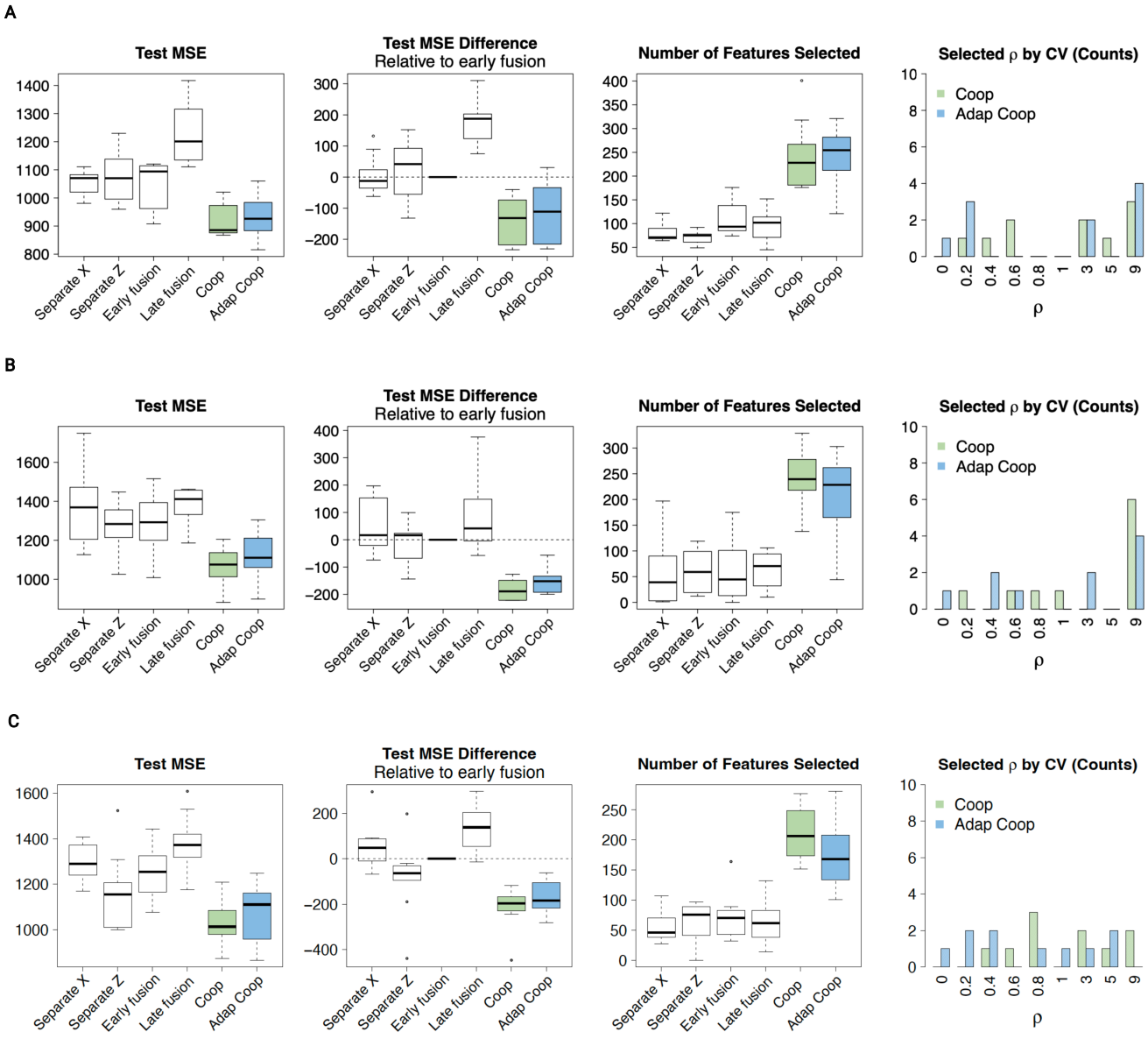

Fig. S2. Simulation studies on cooperative regularized linear regression when X and Z are high-dimensional and have a medium level of correlation with each other. *(A)* Simulation results when X and Z have a medium level of correlation and both contain signal ($t_x = t_z = 2$), $n = 200, p = 1000$, SNR $= 3.5$. The setup is the same as in Figure S1. *(B)* Simulation results when X and Z have a medium level of correlation and both contain signal ($t_x = t_z = 2$), $n = 200, p = 1000$, SNR $= 1.6$. *(C)* Simulation results when X and Z have a medium level of correlation, and Z contains more signal than X ($t_x = 2, t_z = 3$), $n = 200, p = 1000$, SNR $= 1.5$.

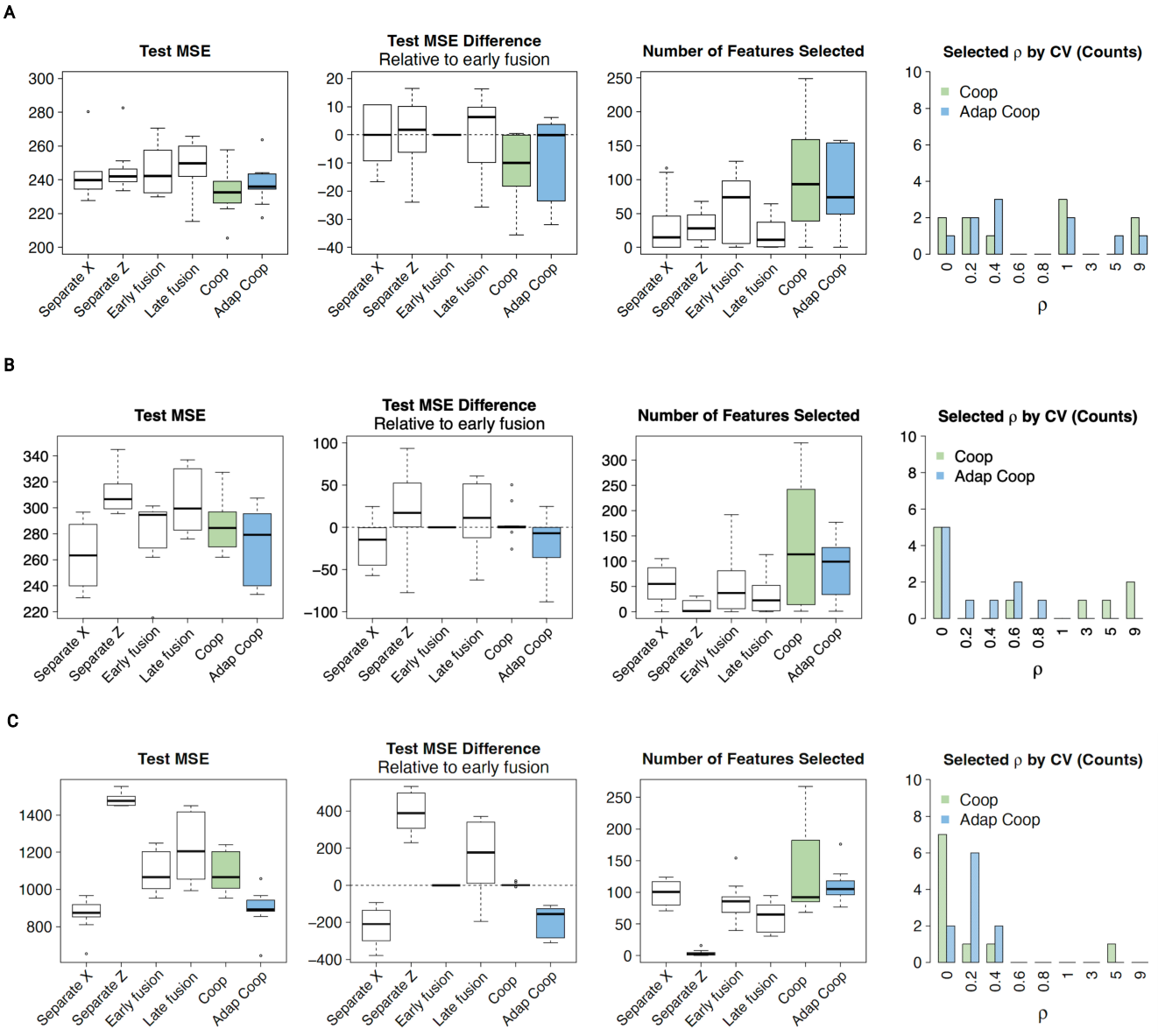

Fig. S3. Simulation studies on cooperative regularized linear regression when X and Z are high-dimensional and have no correlation. *(A)* Simulation results when X and Z have no correlation, and both X and Z contain signal (here we generated y as a linear combination of X and Z instead of the latent factors), $n = 200, p = 1000$, SNR $= 1.0$. The setup is the same as in Figure S1. *(B)* Simulation results when X and Z have no correlation; X contains more signal than Z (here we generated y as a linear combination of X and Z instead of the latent factors), $n = 200, p = 1000$, SNR $= 1.1$. *(C)* Simulation results when X and Z have no correlation; only X contains signal ($t_x = 2, t_z = 0$), $n = 200, p = 1000$, SNR $= 3.5$.

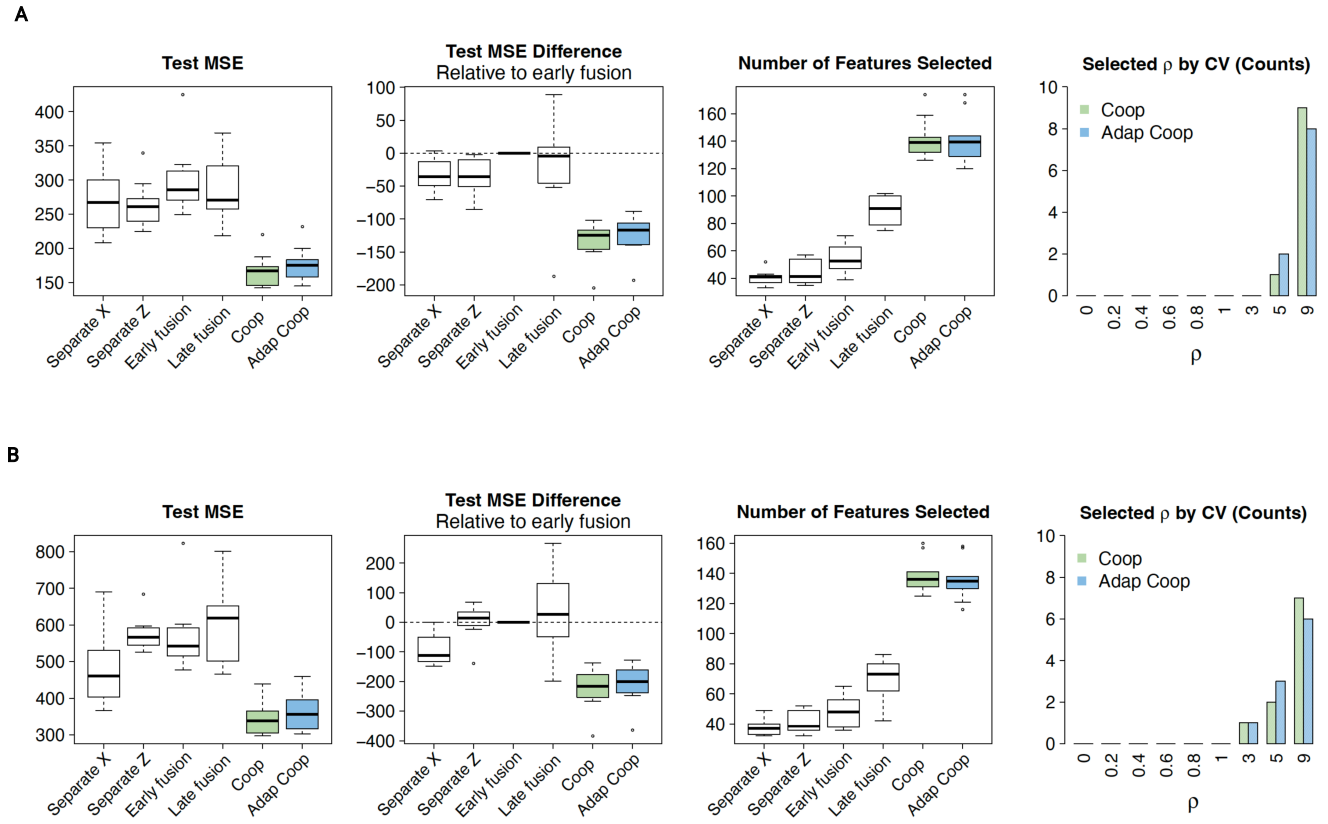

Fig. S4. Simulation studies on cooperative regularized linear regression when X and Z are of a lower dimension and have a high level of correlation with each other. *(A)* Simulation results when X and Z have a high level of correlation and both contain signal ($t_x = t_z = 6$), $n = 500, p = 200$, SNR $= 1.2$. The first panel shows MSE on a test test; the second panel shows the MSE difference on the test set relative to early fusion; the third panel shows the number of features selected; the fourth panel shows the ρ values selected by CV in cooperative learning. Here "Coop" refers to cooperative learning outlined in Algorithm 1 and "Adap Coop" refers to adaptive cooperative learning outlined in Algorithm S2. *(B)* Simulation results when X and Z have a high level of correlation and X contains more signal than Z ($t_x = 5, t_z = 3$), $n = 500, p = 200$, SNR $= 0.7$.

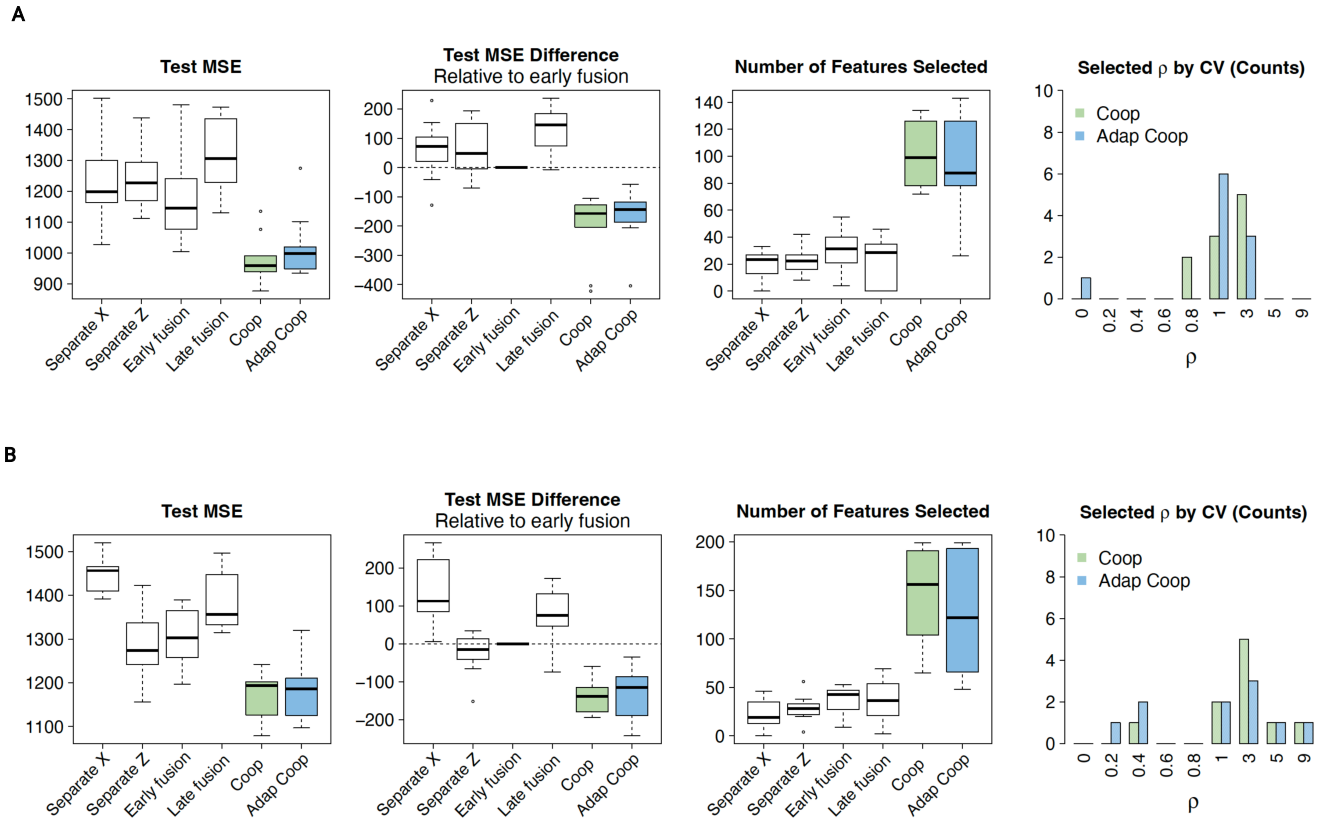

Fig. S5. Simulation studies on cooperative regularized linear regression when X and Z are of a lower dimension and have a medium level of correlation with each other. *(A)* Simulation results when X and Z have a medium level of correlation and both contain signal ($t_x = t_z = 1$), $n = 500, p = 200$, SNR $= 0.8$. The setup is the same as in Figure S4. *(B)* Simulation results when X and Z have a medium level of correlation, and Z contains more signal than X ($t_x = 0.6, t_z = 0.9$), $n = 500, p = 200$, SNR $= 0.5$.

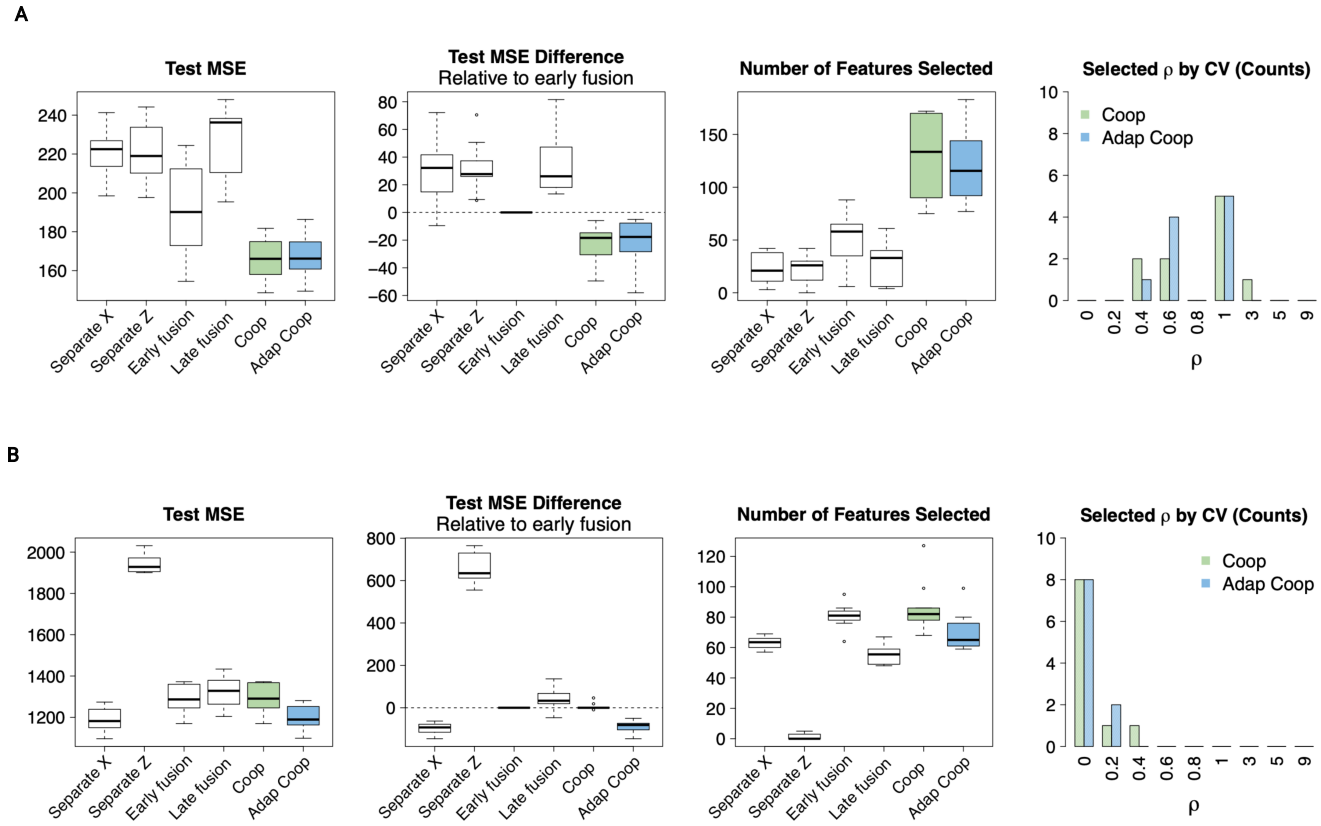

Fig. S6. Simulation studies on cooperative regularized linear regression when X and Z are of a lower dimension and have no correlation with each other. *(A)* Simulation results when X and Z have no correlation ($t_x = t_z = 0$), and both X and Z contain signal (here we generated y as a linear combination of X and Z instead of the latent factors), $n = 500, p = 200$, SNR $= 0.3$. The setup is the same as in Figure S4. *(B)* Simulation results when X and Z have no correlation and only X contains signal ($t_x = 2, t_z = 0$), $n = 500, p = 200$, SNR $= 3.0$.

3. Simulation studies on cooperative learning for more than two data views

Here we conduct simulation studies on cooperative learning for more than two data views. Specifically, we consider the setting of three data views, and this generalizes easily to more data views. We generated Gaussian data with $n = 200$ and $p = 300$ in each of the views X_1, X_2 and X_3, and created correlation between them using latent factors. The response $\boldsymbol{y}$ was generated as a linear combination of the latent factors, corrupted by Gaussian noise.

A. Simulation procedure. The simulation for 3 data views is set up as follows. Given values for parameters $n, p_{x_1}, p_{x_2}, p_{x_3}, p_u, s_u, t_{x_1}, t_{x_2}, t_{x_3}, \boldsymbol{\beta}_u, \sigma$, we generate data according to the following procedure:

1. $x_{1j} \in \mathcal{R}^n$ distributed i.i.d. MVN$(0, I_n)$ for $j = 1, 2, \ldots, p_{x_1}$.
2. $x_{2j} \in \mathcal{R}^n$ distributed i.i.d. MVN$(0, I_n)$ for $j = 1, 2, \ldots, p_{x_2}$.
3. $x_{3j} \in \mathcal{R}^n$ distributed i.i.d. MVN$(0, I_n)$ for $j = 1, 2, \ldots, p_{x_3}$.
4. For $i = 1, 2, \ldots, p_u$ (p_u corresponds to the number of latent factors):
   (a) $u_i \in \mathcal{R}^n$ distributed i.i.d. MVN$(0, s_u^2 I_n)$;
   (b) $x_{1i} = x_{1i} + t_{x_1} * u_i$;
   (c) $x_{2i} = x_{2i} + t_{x_2} * u_i$;
   (d) $x_{3i} = x_{3i} + t_{x_3} * u_i$.
5. $X_1 = [x_{11}, x_{12}, \ldots, x_{1p_{x_1}}]$, $X_2 = [x_{21}, x_{22}, \ldots, x_{2p_{x_2}}]$, $X_3 = [x_{31}, x_{32}, \ldots, x_{3p_{x_3}}]$.
6. $U = [u_1, u_2, \ldots, u_{p_u}]$, $\boldsymbol{y} = U\boldsymbol{\beta}_u + \epsilon$ where $\epsilon \in \mathcal{R}^n$ distributed i.i.d. MVN$(0, \sigma^2 I_n)$.

We compare the following methods: (1) separate X_1, separate X_2, and separate X_3: the standard lasso is applied on the separate data views of X_1, X_2 and X_3 with 10-fold CV; (2) early fusion: the standard lasso is applied on the concatenated data views of X_1, X_2 and X_3 with 10-fold CV (note that this is equivalent to cooperative learning with $\rho = 0$); (3) late fusion: separate lasso models are first fitted on X_1, X_2 and X_3 independently with 10-fold CV, and the three resulting predictors are then combined through linear least squares for the final prediction; (4) cooperative learning (regression) and adaptive cooperative learning.

We evaluated the performance based on the mean-squared error (MSE) on a test set and conducted each simulation experiment 10 times.

B. Simulation results. Fig. S7-S8 show the simulation results for 3 data views. Overall, the simulation results can be summarized as follows:

- Cooperative learning performs the best in terms of test MSE across the range of SNR and correlation settings. It is most helpful when the data views are correlated and contain signal. When the correlation between data views is higher, higher values of ρ are more likely to be selected.
- When only two data views are correlated and contain signal (as in Fig. S7*B* and Fig. S8*C*), cooperative learning also gives performance gains by leveraging the correlation through the agreement penalty, while early fusion can be outperformed by the separate models fit on the data views containing the signal.
- When only one view contains signal and the views are not correlated (as in Fig. S7*C*), cooperative learning is outperformed by the separate model fit on the view containing the signal, but adaptive cooperative learning is able to perform on par with the separate model, outperforming early and late fusion.

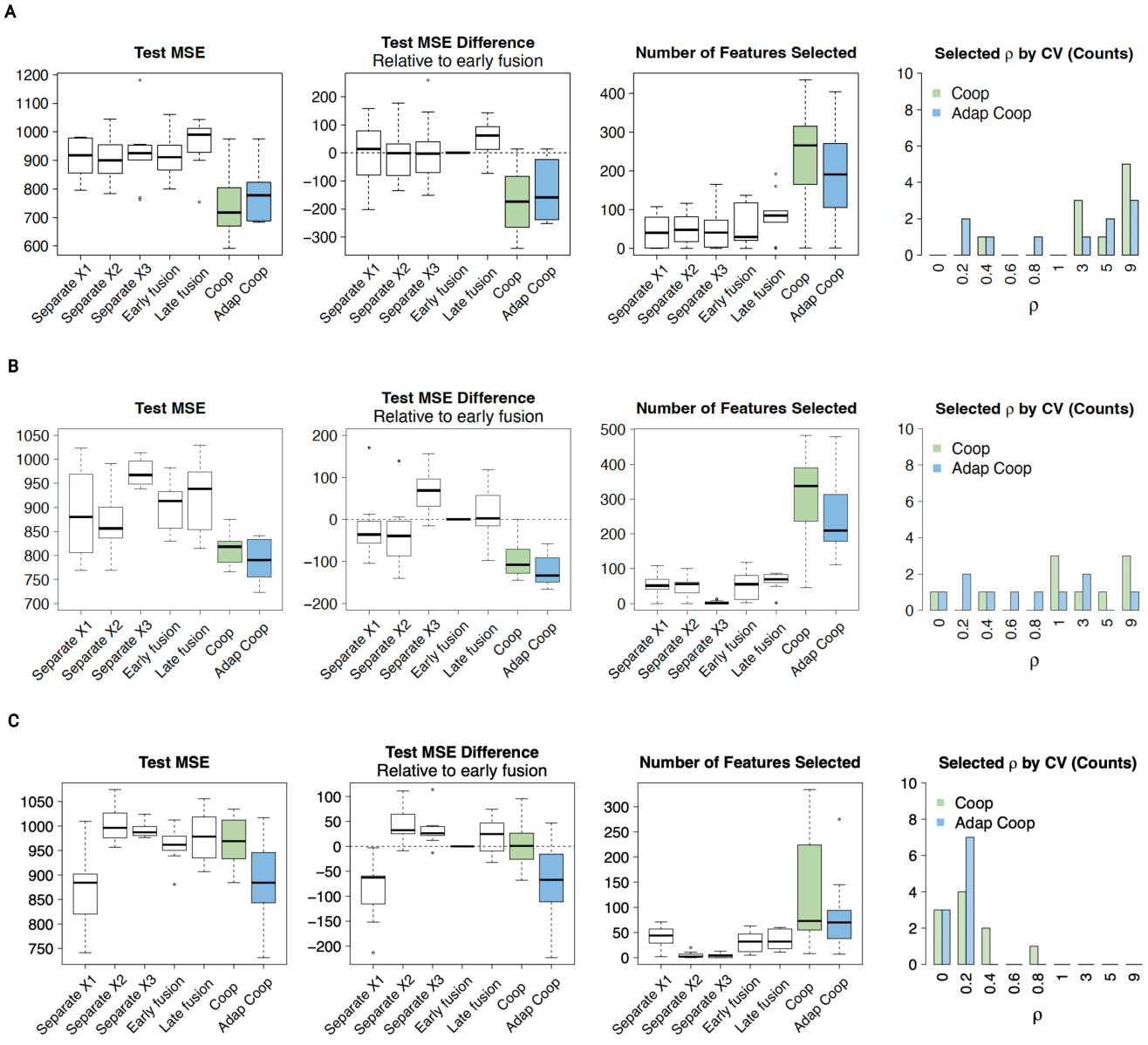

Fig. S7. Simulation studies on cooperative regularized linear regression for more than two data views. *(A)* Simulation results when X_1, X_2 and X_3 are correlated and all contain signal ($t_{x_1} = t_{x_2} = t_{x_3} = 2$), $n = 200, p = 900$, SNR $= 1.5$. The first panel shows MSE on a test test; the second panel shows the MSE difference on the test set relative to early fusion; the third panel shows the number of features selected; the fourth panel shows the ρ values selected by CV in cooperative learning. Here "Coop" refers to cooperative learning outlined in Algorithm 1 and "Adap Coop" refers to adaptive cooperative learning outlined in Algorithm S2. *(B)* Simulation results when only X_1 and X_2 are correlated and contain signal ($t_{x_1} = t_{x_2} = 2, t_{x_3} = 0$), $n = 200, p = 900$, SNR $= 1.5$. *(C)* Simulation results when X_1, X_2 and X_3 are uncorrelated, and only X_1 contains signal ($t_{x_1} = 2, t_{x_2} = t_{x_3} = 0$), $n = 200, p = 900$, SNR $= 1.5$.

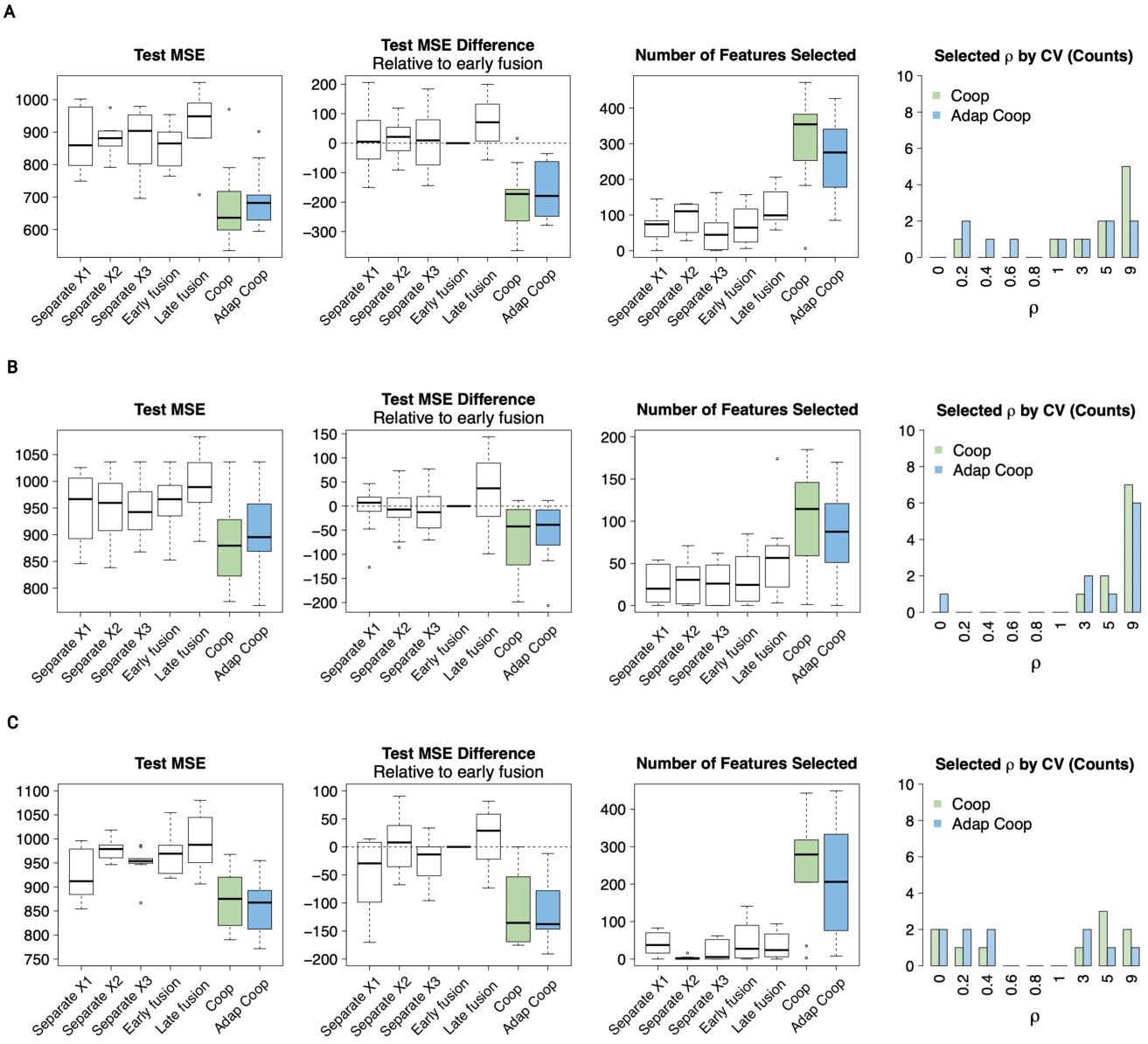

Fig. S8. Simulation studies on cooperative regularized linear regression for more than two data views. *(A)* Simulation results when X_1, X_2 and X_3 are correlated and all contain signal ($t_{x_1} = t_{x_2} = t_{x_3} = 2$), $n = 200, p = 900$, SNR $= 2.5$. The first panel shows MSE on a test test; the second panel shows the MSE difference on the test set relative to early fusion; the third panel shows the number of features selected; the fourth panel shows the ρ values selected by CV in cooperative learning. Here "Coop" refers to cooperative learning outlined in Algorithm 1 and "Adap Coop" refers to adaptive cooperative learning outlined in Algorithm S2. *(B)* Simulation results when X_1, X_2 and X_3 are correlated and all contain signal ($t_{x_1} = t_{x_2} = t_{x_3} = 2$), $n = 200, p = 900$, SNR $= 0.6$. *(C)* Simulation results when only X_1 and X_3 are correlated; X_1 contains more signal than X_3, X_2 does not contain signal ($t_{x_1} = 2, t_{x_2} = 0, t_{x_3} = 1.5$), $n = 200, p = 900$, SNR $= 1.0$.

4. Theoretical analysis under the factor model

To understand the role of the agreement penalty from a theoretical perspective, we consider the following latent factor model. Let $\boldsymbol{u} = (U_1, U_2, \ldots, U_n)$ be a vector of n i.i.d. random variables with $U_i \sim \mathcal{N}(0,1)$. Let $\boldsymbol{y} = (Y_1, \ldots, Y_n)$, $\boldsymbol{x} = (X_1, \ldots, X_n)$, and $\boldsymbol{z} = (Z_1, \ldots, Z_n)$, with

$$Y_i = \gamma_y U_i + \varepsilon_{yi}, \qquad X_i = \gamma_x U_i + \varepsilon_{xi}, \qquad \text{and} \qquad Z_i = \gamma_z U_i + \varepsilon_{zi}, \tag{1}$$

where $\varepsilon_{yi} \sim \mathcal{N}\left(0, \sigma_y^2\right)$, $\varepsilon_{xi} \sim \mathcal{N}\left(0, \sigma_x^2\right)$, $\varepsilon_{zi} \sim \mathcal{N}\left(0, \sigma_z^2\right)$ independently.

In this section, we study the mean squared error (MSE) of the cooperative learning algorithm. More precisely, let

$$\hat{\theta} = \text{argmin}_\theta \sum_{i=1}^{n} \left[\frac{1}{2} \left(Y_i - X_i\theta_x - Z_i\theta_z\right)^2 + \frac{\rho}{2} \left(X_i\theta_x - Z_i\theta_z\right)^2 \right]. \tag{2}$$

Let U_{new}, X_{new}, Y_{new}, Z_{new} be some new random variables generated from Eq. (1) independently of the previous data, i.e.,

$$Y_{\text{new}} = \gamma_y U_{\text{new}} + \varepsilon_{y\,\text{new}}, \qquad X_{\text{new}} = \gamma_x U_{\text{new}} + \varepsilon_{x\,\text{new}}, \qquad \text{and} \qquad Z_{\text{new}} = \gamma_z U_{\text{new}} + \varepsilon_{z\,\text{new}}, \tag{3}$$

where $U_{\text{new}} \sim \mathcal{N}(0,1)$, $\varepsilon_{y\,\text{new}} \sim \mathcal{N}\left(0, \sigma_y^2\right)$, $\varepsilon_{x\,\text{new}} \sim \mathcal{N}\left(0, \sigma_x^2\right)$, $\varepsilon_{z\,\text{new}} \sim \mathcal{N}\left(0, \sigma_z^2\right)$ independently. We focus on the MSE conditioning on $\boldsymbol{x}$ and $\boldsymbol{z}$:

$$\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho) = \text{E}\left[\left(Y_{\text{new}} - \left(X_{\text{new}}\hat{\theta}_x + Z_{\text{new}}\hat{\theta}_z \right) \right)^2 \mid \boldsymbol{x}, \boldsymbol{z} \right]. \tag{4}$$

The case of $\rho = 0$ corresponds to the linear regression with no agreement penalty. We will study the behavior of $\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)$ when ρ is around 0.

Proposition 1. *The derivative of* $\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)$ *satisfies*

$$\frac{d}{d\rho} \left[\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)\right] |_{\rho=0} = \sigma^{\star 2}(C_2 B_1 - 2C_1 B_2)/C_2^3, \tag{5}$$

where

$$\begin{aligned}
\sigma^{\star 2} &= \frac{\gamma_y^2}{1 + \gamma_x^2/\sigma_x^2 + \gamma_z^2/\sigma_z^2} + \sigma_y^2, \\
C_1 &= \left[(\gamma_x^2 + \sigma_x^2)(\boldsymbol{z}^\top\boldsymbol{z}) + (\gamma_z^2 + \sigma_z^2)(\boldsymbol{x}^\top\boldsymbol{x}) - 2\gamma_x\gamma_z(\boldsymbol{x}^\top\boldsymbol{z})\right] ((\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) - (\boldsymbol{x}^\top\boldsymbol{z})^2), \\
B_1 &= 2\left[(\gamma_x^2 + \sigma_x^2)(\boldsymbol{z}^\top\boldsymbol{z}) + (\gamma_z^2 + \sigma_z^2)(\boldsymbol{x}^\top\boldsymbol{x}) + 2\gamma_x\gamma_z(\boldsymbol{x}^\top\boldsymbol{z})\right] ((\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) - (\boldsymbol{x}^\top\boldsymbol{z})^2), \\
C_2 &= (\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) - (\boldsymbol{x}^\top\boldsymbol{z})^2, \\
B_2 &= 2\left((\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) + (\boldsymbol{x}^\top\boldsymbol{z})^2\right).
\end{aligned} \tag{6}$$

Proposition 2. *The derivative of* $\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)$ *at* $\rho = 0$ *satisfies*

$$\frac{d}{d\rho} \left[\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)\right] |_{\rho=0} = -\frac{4}{n} \left(1 + \frac{2\gamma_x^2\gamma_z^2}{\sigma_x^2\gamma_z^2 + \sigma_z^2\gamma_x^2 + \sigma_x^2\sigma_z^2}\right) \left(\sigma_y^2 + \frac{\gamma_y^2\sigma_x^2\sigma_z^2}{\sigma_x^2\gamma_z^2 + \sigma_z^2\gamma_x^2 + \sigma_x^2\sigma_z^2}\right) + \mathcal{O}_p\left(n^{-\frac{3}{2}}\right). \tag{7}$$

Here the notation $\mathcal{O}_p(\cdot)$ *is used with the following meaning:* $X_n = \mathcal{O}_p(a_n)$ *as* $n \to \infty$ *means that for any* $\varepsilon > 0$*, there exists a finite* $M > 0$ *and a finite* $N > 0$ *such that* $\text{P}\left[|X_n/a_n| > M\right] < \varepsilon, \forall n > N$.

The proposition establishes that the MSE is a decreasing function of ρ around 0 with high probability. In other words, the agreement penalty is helpful in reducing the mean squared error. To further interpret the above results, we study the ratio of the derivative to the MSE itself.

Proposition 3. *The ratio of the derivative of* $\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)$ *at* $\rho = 0$ *and* $\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; 0)$ *satisfies*

$$\frac{\frac{d}{d\rho} \left[\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)\right] |_{\rho=0}}{\text{MSE}(\boldsymbol{x}, \boldsymbol{z}; 0)} = -\frac{4}{n} \left(1 + \frac{2\gamma_x^2\gamma_z^2}{\sigma_x^2\gamma_z^2 + \sigma_z^2\gamma_x^2 + \sigma_x^2\sigma_z^2}\right) + \mathcal{O}_p\left(n^{-\frac{3}{2}}\right). \tag{8}$$

Here the notation $\mathcal{O}_p(\cdot)$ *is used with the same meaning as in Proposition 2.*

Proposition 3 presents a simple form of the ratio of the derivative to the MSE itself. The ratio quantifies by what percentage the "agreement" penalty decreases the MSE. It can been seen from this representation that this ratio depends on the structure of the factor model, and that the agreement penalty is more helpful when the sample size n is smaller. In the extreme case, when we have infinite data, i.e., when $n = \infty$, the derivative of the MSE is 0; in this case, we learn all the signals from the data even without the agreement penalty.

A. Proof of Proposition 1. Here we present a lemma that is used in the proof of Proposition 1.

Lemma 1. *Assume that $c_2 > 0$. Let $g(\rho) = (a_1\rho^2 + b_1\rho + c_1)/(a_2\rho^2 + b_2\rho + c_2)^2$, then*

$$g'(\rho)|_{\rho=0} = \frac{b_1c_2 - 2b_2c_1}{c_2^3}. \tag{9}$$

Proof. We compute the derivative of the function g:

$$g'(\rho) = \frac{(2a_1\rho + b_1)(a_2\rho^2 + b_2\rho + c_2) - 2(2a_2\rho + b_2)(a_1\rho^2 + b_1\rho + c_1)}{(a_2\rho^2 + b_2\rho + c_2)^3}. \tag{10}$$

Evaluating at $\rho = 0$, we get

$$g'(\rho)|_{\rho=0} = \frac{b_1c_2 - 2b_2c_1}{c_2^3}. \tag{11}$$

□

With this lemma, we are ready to prove the proposition. We start with writing down an explicit expression for the estimator $\hat{\theta}$. Let

$$\tilde{X} = \begin{pmatrix} \boldsymbol{x} & \boldsymbol{z} \\ -\sqrt{\rho}\boldsymbol{x} & \sqrt{\rho}\boldsymbol{z} \end{pmatrix}, \quad \tilde{\boldsymbol{y}} = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}. \tag{12}$$

Then Eq. (2) implies that $\hat{\theta} = (\hat{\theta}_x, \hat{\theta}_z)^\mathsf{T}$ takes the following form:

$$\hat{\theta} = \left(\tilde{X}^\mathsf{T}\tilde{X}\right)^{-1}\left(\tilde{X}^\mathsf{T}\tilde{\boldsymbol{y}}\right). \tag{13}$$

In particular,

$$\tilde{X}^\mathsf{T}\tilde{X} = \begin{pmatrix} (1+\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{x} & (1-\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{z} \\ (1-\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{z} & (1+\rho)\boldsymbol{z}^\mathsf{T}\boldsymbol{z} \end{pmatrix}, \tag{14}$$

and

$$\tilde{X}^\mathsf{T}\tilde{\boldsymbol{y}} = (\boldsymbol{x}^\mathsf{T}\boldsymbol{y}, \boldsymbol{z}^\mathsf{T}\boldsymbol{y})^\mathsf{T}. \tag{15}$$

Therefore,

$$\begin{aligned} \hat{\theta} = \begin{pmatrix} \hat{\theta}_x \\ \hat{\theta}_z \end{pmatrix} &= \begin{pmatrix} (1+\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{x} & (1-\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{z} \\ (1-\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{z} & (1+\rho)\boldsymbol{z}^\mathsf{T}\boldsymbol{z} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{x}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{z}^\mathsf{T}\boldsymbol{y} \end{pmatrix} = \frac{1}{\det} \begin{pmatrix} (1+\rho)\boldsymbol{z}^\mathsf{T}\boldsymbol{z} & -(1-\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{z} \\ -(1-\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{z} & (1+\rho)\boldsymbol{x}^\mathsf{T}\boldsymbol{x} \end{pmatrix} \begin{pmatrix} \boldsymbol{x}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{z}^\mathsf{T}\boldsymbol{y} \end{pmatrix} \\ &= \frac{1}{\det} \begin{pmatrix} (1+\rho)(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\boldsymbol{y}) - (1-\rho)(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{z}^\mathsf{T}\boldsymbol{y}) \\ (1+\rho)(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{y}) - (1-\rho)(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\boldsymbol{y}) \end{pmatrix}, \end{aligned} \tag{16}$$

where $\det = (1+\rho)^2(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (1-\rho)^2(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2$.

We then move on to analyze the conditional distribution of $\boldsymbol{u}$ and $\boldsymbol{y}$ on $\boldsymbol{x}$ and $\boldsymbol{z}$. By Eq. (1), we can write down a joint distribution of (U_i, X_i, Z_i):

$$\begin{pmatrix} U_i \\ X_i \\ Z_i \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \gamma_x & \gamma_z \\ \gamma_x & \gamma_x^2 + \sigma_x^2 & \gamma_x\gamma_z \\ \gamma_z & \gamma_x\gamma_z & \gamma_z^2 + \sigma_z^2 \end{pmatrix}\right). \tag{17}$$

Using formulas from conditional distribution of multivariate gaussian, we get that

$$U_i \mid X_i, Z_i \sim \mathcal{N}(\mathrm{E}\left[U_i \mid X_i, Z_i\right], \mathrm{Var}\left[U_i \mid X_i, Z_i\right]), \tag{18}$$

where

$$\begin{aligned} \mathrm{E}\left[U_i \mid X_i, Z_i\right] &= \begin{pmatrix} \gamma_x & \gamma_z \end{pmatrix} \begin{pmatrix} \gamma_x^2 + \sigma_x^2 & \gamma_x\gamma_z \\ \gamma_x\gamma_z & \gamma_z^2 + \sigma_z^2 \end{pmatrix}^{-1} \begin{pmatrix} X_i \\ Z_i \end{pmatrix} \\ &= \frac{1}{\sigma_x^2\sigma_z^2 + \gamma_x^2\sigma_z^2 + \gamma_z^2\sigma_x^2} \begin{pmatrix} \gamma_x & \gamma_z \end{pmatrix} \begin{pmatrix} \gamma_z^2 + \sigma_z^2 & -\gamma_x\gamma_z \\ -\gamma_x\gamma_z & \gamma_x^2 + \sigma_x^2 \end{pmatrix} \begin{pmatrix} X_i \\ Z_i \end{pmatrix} \\ &= \frac{\gamma_x X_i}{\sigma_x^2 + \gamma_x^2 + \gamma_z^2\sigma_x^2/\sigma_z^2} + \frac{\gamma_z Z_i}{\sigma_z^2 + \gamma_z^2 + \gamma_x^2\sigma_z^2/\sigma_x^2}, \end{aligned} \tag{19}$$

and

$$\begin{aligned} \mathrm{Var}\left[U_i \mid X_i, Z_i\right] &= 1 - \begin{pmatrix} \gamma_x & \gamma_z \end{pmatrix} \begin{pmatrix} \gamma_x^2 + \sigma_x^2 & \gamma_x\gamma_z \\ \gamma_x\gamma_z & \gamma_z^2 + \sigma_z^2 \end{pmatrix}^{-1} \begin{pmatrix} \gamma_x \\ \gamma_z \end{pmatrix} \\ &= 1 - \frac{1}{\sigma_x^2\sigma_z^2 + \gamma_x^2\sigma_z^2 + \gamma_z^2\sigma_x^2} \begin{pmatrix} \gamma_x & \gamma_z \end{pmatrix} \begin{pmatrix} \gamma_z^2 + \sigma_z^2 & -\gamma_x\gamma_z \\ -\gamma_x\gamma_z & \gamma_x^2 + \sigma_x^2 \end{pmatrix} \begin{pmatrix} \gamma_x \\ \gamma_z \end{pmatrix} \\ &= 1 - \frac{\gamma_x^2\sigma_z^2 + \gamma_z^2\sigma_x^2}{\sigma_x^2\sigma_z^2 + \gamma_x^2\sigma_z^2 + \gamma_z^2\sigma_x^2} = \frac{1}{1 + \gamma_x^2/\sigma_x^2 + \gamma_z^2/\sigma_z^2}. \end{aligned} \tag{20}$$

Since $Y_i = \gamma_y U_i + \varepsilon_{yi}$, the above analysis implies that

$$Y_i \mid X_i, Z_i \sim \mathcal{N}(\mathrm{E}\left[Y_i \mid X_i, Z_i\right], \mathrm{Var}\left[Y_i \mid X_i, Z_i\right]), \tag{21}$$

where

$$\mathrm{E}\left[Y_i \mid X_i, Z_i\right] = \gamma_y \mathrm{E}\left[U_i \mid X_i, Z_i\right] = \frac{\gamma_x \gamma_y X_i}{\sigma_x^2 + \gamma_x^2 + \gamma_z^2 \sigma_x^2/\sigma_z^2} + \frac{\gamma_z \gamma_y Z_i}{\sigma_z^2 + \gamma_z^2 + \gamma_x^2 \sigma_z^2/\sigma_x^2}, \tag{22}$$

and

$$\mathrm{Var}\left[Y_i \mid X_i, Z_i\right] = \gamma_y^2 \mathrm{Var}\left[U_i \mid X_i, Z_i\right] + \sigma_y^2 = \frac{\gamma_y^2}{1 + \gamma_x^2/\sigma_x^2 + \gamma_z^2/\sigma_z^2} + \sigma_y^2. \tag{23}$$

Let

$$\theta_x^\star = \frac{\gamma_x \gamma_y}{\sigma_x^2 + \gamma_x^2 + \gamma_z^2 \sigma_x^2/\sigma_z^2}, \qquad \theta_z^\star = \frac{\gamma_z \gamma_y}{\sigma_z^2 + \gamma_z^2 + \gamma_x^2 \sigma_z^2/\sigma_x^2}, \qquad \sigma^{\star 2} = \frac{\gamma_y^2}{1 + \gamma_x^2/\sigma_x^2 + \gamma_z^2/\sigma_z^2} + \sigma_y^2, \tag{24}$$

then the above shows that we can express Y_i as

$$Y_i = \theta_x^\star X_i + \theta_z^\star Z_i + \varepsilon_i^\star, \tag{25}$$

where $\varepsilon_i^\star \perp\!\!\!\perp (X_i, Z_i)$ and $\varepsilon_i^\star \sim \mathcal{N}(0, \sigma^{\star 2})$. In words, Y_i can be decomposed into two independent terms: a linear combination of X_i and Z_i, and an error term independent of (X_i, Z_i).

With the above tools, we are ready to study the MSE. Using Eq. (25), we can write

$$\begin{aligned}
\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho) &= \mathrm{E}\left[\left(Y_{\mathrm{new}} - \left(X_{\mathrm{new}}\hat{\theta}_x + Z_{\mathrm{new}}\hat{\theta}_z\right)\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] \\
&= \mathrm{E}\left[\left(\theta_x^\star X_{\mathrm{new}} + \theta_z^\star Z_{\mathrm{new}} + \varepsilon_{\mathrm{new}}^\star - \left(X_{\mathrm{new}}\hat{\theta}_x + Z_{\mathrm{new}}\hat{\theta}_z\right)\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] \\
&= \mathrm{E}\left[\left((\theta_x^\star - \hat{\theta}_x)X_{\mathrm{new}} + (\theta_z^\star - \hat{\theta}_z)Z_{\mathrm{new}} + \varepsilon_{\mathrm{new}}^\star\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] \\
&= \mathrm{E}\left[\left((\theta_x^\star - \hat{\theta}_x)X_{\mathrm{new}} + (\theta_z^\star - \hat{\theta}_z)Z_{\mathrm{new}}\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] + \mathrm{E}\left[{\varepsilon_{\mathrm{new}}^\star}^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] \\
&= \mathrm{E}\left[\left((\theta_x^\star - \hat{\theta}_x)X_{\mathrm{new}} + (\theta_z^\star - \hat{\theta}_z)Z_{\mathrm{new}}\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] + \sigma^{\star 2}.
\end{aligned} \tag{26}$$

Here the cross terms vanish because $\varepsilon_{\mathrm{new}}^\star \perp\!\!\!\perp (X_{\mathrm{new}}, Z_{\mathrm{new}})$. Since the new dataset is independent of the training dataset, we can further simply the above:

$$\begin{aligned}
\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho) &= \mathrm{E}\left[\left((\theta_x^\star - \hat{\theta}_x)X_{\mathrm{new}}\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] + \mathrm{E}\left[\left((\theta_z^\star - \hat{\theta}_z)Z_{\mathrm{new}}\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] \\
&\quad + 2\mathrm{E}\left[\left(\theta_z^\star - \hat{\theta}_z\right)\left(\hat{\theta}_x - \theta_x^\star\right) Z_{\mathrm{new}} X_{\mathrm{new}} \mid \boldsymbol{x}, \boldsymbol{z}\right] + \sigma^{\star 2} \\
&= \mathrm{E}\left[\left(\hat{\theta}_x - \theta_x^\star\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] \mathrm{E}\left[X_{\mathrm{new}}^2\right] + \mathrm{E}\left[\left(\hat{\theta}_z - \theta_z^\star\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] \mathrm{E}\left[Z_{\mathrm{new}}^2\right] \\
&\quad + 2\mathrm{E}\left[\left(\hat{\theta}_z - \theta_z^\star\right)\left(\hat{\theta}_x - \theta_x^\star\right) \mid \boldsymbol{x}, \boldsymbol{z}\right] \mathrm{E}\left[Z_{\mathrm{new}} X_{\mathrm{new}}\right] + \sigma^{\star 2} \\
&= \mathrm{E}\left[\left(\hat{\theta}_x - \theta_x^\star\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] (\gamma_x^2 + \sigma_x^2) + \mathrm{E}\left[\left(\hat{\theta}_z - \theta_z^\star\right)^2 \mid \boldsymbol{x}, \boldsymbol{z}\right] \left(\gamma_z^2 + \sigma_z^2\right) \\
&\quad + 2\mathrm{E}\left[\left(\hat{\theta}_z - \theta_z^\star\right)\left(\hat{\theta}_x - \theta_x^\star\right) \mid \boldsymbol{x}, \boldsymbol{z}\right] (\gamma_x \gamma_z) + \sigma^{\star 2}.
\end{aligned} \tag{27}$$

We can then further decompose the terms into squared bias plus variance.

$$\begin{aligned}
\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho) &= \mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]^2 (\gamma_x^2 + \sigma_x^2) + \mathrm{Var}\left[\hat{\theta}_x \mid \boldsymbol{x}, \boldsymbol{z}\right] (\gamma_x^2 + \sigma_x^2) \\
&\quad + \mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]^2 \left(\gamma_z^2 + \sigma_z^2\right) + \mathrm{Var}\left[\hat{\theta}_z \mid \boldsymbol{x}, \boldsymbol{z}\right] \left(\gamma_z^2 + \sigma_z^2\right) \\
&\quad + 2\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right] \mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right] (\gamma_x \gamma_z) + 2\,\mathrm{Cov}\left[\hat{\theta}_z, \hat{\theta}_x \mid \boldsymbol{x}, \boldsymbol{z}\right] (\gamma_x \gamma_z) + \sigma^{\star 2} \\
&= B^2(\boldsymbol{x}, \boldsymbol{z}; \rho) + V(\boldsymbol{x}, \boldsymbol{z}; \rho) + \sigma^{\star 2},
\end{aligned} \tag{28}$$

where $B^2(\boldsymbol{x}, \boldsymbol{z}; \rho) = \mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]^2 (\gamma_x^2 + \sigma_x^2) + \mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]^2 \left(\gamma_z^2 + \sigma_z^2\right) + 2\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right] \mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right] (\gamma_x \gamma_z)$ is the sum of bias related terms, and $V(\boldsymbol{x}, \boldsymbol{z}; \rho) = \mathrm{Var}\left[\hat{\theta}_x \mid \boldsymbol{x}, \boldsymbol{z}\right] (\gamma_x^2 + \sigma_x^2) + \mathrm{Var}\left[\hat{\theta}_z \mid \boldsymbol{x}, \boldsymbol{z}\right] \left(\gamma_z^2 + \sigma_z^2\right) + 2\,\mathrm{Cov}\left[\hat{\theta}_z, \hat{\theta}_x \mid \boldsymbol{x}, \boldsymbol{z}\right] (\gamma_x \gamma_z)$ is the sum of variance related terms.

We can then use Eq. (21) - Eq. (23) to study the bias and variance of the estimators $\hat{\theta}_x$ and $\hat{\theta}_z$. We start with the bias. By Eq. (25), we have $\mathrm{E}\left[\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}\right] = \theta_x^\star \boldsymbol{x} + \theta_z^\star \boldsymbol{z}$. Therefore,

$$\begin{aligned}
\mathrm{E}\left[\hat{\theta}_x \mid \boldsymbol{x}, \boldsymbol{z}\right] &= \mathrm{E}\left[\frac{1}{\det}\left((1+\rho)(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\boldsymbol{y}) - (1-\rho)(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{z}^\mathsf{T}\boldsymbol{y})\right) \mid \boldsymbol{x}, \boldsymbol{z}\right] \\
&= \frac{1}{\det}\left((1+\rho)(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\mathrm{E}\left[\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}\right]) - (1-\rho)(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{z}^\mathsf{T}\mathrm{E}\left[\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{z}\right])\right) \\
&= \frac{1}{\det}\left((1+\rho)(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\left(\theta_x^\star \boldsymbol{x} + \theta_z^\star \boldsymbol{z}\right)) - (1-\rho)(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{z}^\mathsf{T}\left(\theta_x^\star \boldsymbol{x} + \theta_z^\star \boldsymbol{z}\right))\right) \\
&= \frac{1}{\det}\left((1+\rho)\left[\theta_x^\star(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) + \theta_z^\star(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})\right] - (1-\rho)\left[\theta_x^\star(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2 + \theta_z^\star(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})\right]\right) \\
&= \frac{1}{\det}\left(\theta_x^\star\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + \rho\left(\theta_x^\star\left[(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\boldsymbol{x}) + (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + 2\theta_z^\star(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})\right)\right).
\end{aligned} \tag{29}$$

Note that

$$\begin{aligned}
\det &= (1+\rho)^2(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (1-\rho)^2(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2 \\
&= (\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2 + 2\rho\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) + (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + \rho^2\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right]
\end{aligned} \tag{30}$$

Therefore,

$$\begin{aligned}
&\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right] \\
&= \frac{\theta_x^\star\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + \rho\left(\theta_x^\star\left[(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\boldsymbol{x}) + (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + 2\theta_z^\star(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})\right)}{(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2 + 2\rho\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) + (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + \rho^2\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right]} - \theta_x^\star \\
&= \frac{\rho\left(\theta_x^\star\left[-(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})(\boldsymbol{x}^\mathsf{T}\boldsymbol{x}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + 2\theta_z^\star(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})\right) - \rho^2\theta_x^\star\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right]}{(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2 + 2\rho\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) + (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + \rho^2\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right]} \\
&= \frac{b_1\rho + a_1\rho^2}{c_2 + b_2\rho + a_2\rho^2},
\end{aligned} \tag{31}$$

where b_1, a_1, c_2, b_2, a_2 are expressions depending on $\boldsymbol{x}$ and $\boldsymbol{z}$ but not on ρ. We can then clearly see that when $\rho = 0$, $\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right] = 0$. By symmetry, we have the same property for $\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]$. Therefore,

$$\begin{aligned}
\frac{d}{d\rho}B^2(\boldsymbol{x}, \boldsymbol{z}; \rho) &= (\gamma_x^2 + \sigma_x^2)\frac{d}{d\rho}\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]^2|_{\rho=0} + \left(\gamma_z^2 + \sigma_z^2\right)\frac{d}{d\rho}\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]^2|_{\rho=0} \\
&\quad + 2\left(\gamma_x\gamma_z\right)\frac{d}{d\rho}\left(\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]\right)|_{\rho=0} \\
&= 2(\gamma_x^2 + \sigma_x^2)\left(\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0}\right)\frac{d}{d\rho}\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0} \\
&\quad + 2\left(\gamma_z^2 + \sigma_z^2\right)\left(\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0}\right)\frac{d}{d\rho}\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0} \\
&\quad + 2\left(\gamma_x\gamma_z\right)\left(\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0}\right)\frac{d}{d\rho}\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0} \\
&\quad + 2\left(\gamma_x\gamma_z\right)\left(\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0}\right)\frac{d}{d\rho}\mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0}.
\end{aligned} \tag{32}$$

Since $\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0} = \mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right]|_{\rho=0} = 0$, we have

$$\frac{d}{d\rho}B^2(\boldsymbol{x}, \boldsymbol{z}; \rho) = 0. \tag{33}$$

It remains to study $\frac{d}{d\rho}V(\boldsymbol{x}, \boldsymbol{z}; \rho)$. From the form of $\hat{\theta}_x$, we can get that

$$\begin{aligned}
\mathrm{Var}\left[\hat{\theta}_x \mid \boldsymbol{x}, \boldsymbol{z}\right] &= \frac{\sigma^{\star 2}}{\det^2}\left\|(1+\rho)(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})\boldsymbol{x} - (1-\rho)(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})\boldsymbol{z}\right\|_2^2 \\
&= \frac{\sigma^{\star 2}}{\det^2}\left[(1+\rho)^2(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})^2(\boldsymbol{x}^\mathsf{T}\boldsymbol{x}) + (1-\rho)^2(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - 2(1+\rho)(1-\rho)(\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})\right] \\
&= \frac{\sigma^{\star 2}(\boldsymbol{z}^\mathsf{T}\boldsymbol{z})}{\det^2}\left(\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + 2\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right]\rho + a_{v1}\rho^2\right),
\end{aligned} \tag{34}$$

for some a_{v1} depending on $\boldsymbol{x}$ and $\boldsymbol{z}$ but not on ρ. Similarly, we get that

$$\mathrm{Var}\left[\hat{\theta}_z \mid \boldsymbol{x}, \boldsymbol{z}\right] = \frac{\sigma^{\star 2}(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})}{\det^2}\left(\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right] + 2\left[(\boldsymbol{x}^\mathsf{T}\boldsymbol{x})(\boldsymbol{z}^\mathsf{T}\boldsymbol{z}) - (\boldsymbol{x}^\mathsf{T}\boldsymbol{z})^2\right]\rho + a_{v2}\rho^2\right), \tag{35}$$

for some a_{v2} depending on $\boldsymbol{x}$ and $\boldsymbol{z}$ but not on ρ. For the covariance term,

$$\begin{aligned}\operatorname{Cov}\left[\hat{\theta}_x, \hat{\theta}_z \mid \boldsymbol{x}, \boldsymbol{z}\right] &= \frac{\sigma^{\star 2}}{\det^2}\left[(1+\rho)(\boldsymbol{z}^\top\boldsymbol{z})\boldsymbol{x} - (1-\rho)(\boldsymbol{x}^\top\boldsymbol{z})\boldsymbol{z}\right]^\top\left[(1+\rho)(\boldsymbol{x}^\top\boldsymbol{x})\boldsymbol{z} - (1-\rho)(\boldsymbol{z}^\top\boldsymbol{x})\boldsymbol{x}\right] \\ &= \frac{\sigma^{\star 2}}{\det^2}\left((-1+3\rho)(1+\rho)(\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z})(\boldsymbol{x}^\top\boldsymbol{z}) + (1-\rho)^2(\boldsymbol{x}^\top\boldsymbol{z})^3\right) \\ &= \frac{\sigma^{\star 2}(\boldsymbol{x}^\top\boldsymbol{z})}{\det^2}\left(-\left[(\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) - (\boldsymbol{x}^\top\boldsymbol{z})^2\right] + 2\left[(\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) - (\boldsymbol{x}^\top\boldsymbol{z})^2\right]\rho + a_{v3}\rho^2\right),\end{aligned} \tag{36}$$

for some a_{v3} depending on $\boldsymbol{x}$ and $\boldsymbol{z}$ but not on ρ. Combining the three terms, we get

$$\begin{aligned}V(\boldsymbol{x}, \boldsymbol{z}; \rho) &= \operatorname{Var}\left[\hat{\theta}_x \mid \boldsymbol{x}, \boldsymbol{z}\right](\gamma_x^2 + \sigma_x^2) + \operatorname{Var}\left[\hat{\theta}_z \mid \boldsymbol{x}, \boldsymbol{z}\right](\gamma_z^2 + \sigma_z^2) + \operatorname{Cov}\left[\hat{\theta}_z, \hat{\theta}_x \mid \boldsymbol{x}, \boldsymbol{z}\right](\gamma_x\gamma_z) \\ &= \sigma^{\star 2}\frac{C_1 + B_1\rho + A_1\rho^2}{\det^2} = \sigma^{\star 2}\frac{C_1 + B_1\rho + A_1\rho^2}{(C_2 + B_2\rho + A_2\rho^2)^2},\end{aligned} \tag{37}$$

where

$$\begin{aligned}C_1 &= \left[(\gamma_x^2+\sigma_x^2)(\boldsymbol{z}^\top\boldsymbol{z}) + (\gamma_z^2+\sigma_z^2)(\boldsymbol{x}^\top\boldsymbol{x}) - 2\gamma_x\gamma_z(\boldsymbol{x}^\top\boldsymbol{z})\right]((\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) - (\boldsymbol{x}^\top\boldsymbol{z})^2), \\ B_1 &= 2\left[(\gamma_x^2+\sigma_x^2)(\boldsymbol{z}^\top\boldsymbol{z}) + (\gamma_z^2+\sigma_z^2)(\boldsymbol{x}^\top\boldsymbol{x}) + 2\gamma_x\gamma_z(\boldsymbol{x}^\top\boldsymbol{z})\right]((\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) - (\boldsymbol{x}^\top\boldsymbol{z})^2), \\ C_2 &= (\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) - (\boldsymbol{x}^\top\boldsymbol{z})^2, \\ B_2 &= 2\left((\boldsymbol{x}^\top\boldsymbol{x})(\boldsymbol{z}^\top\boldsymbol{z}) + (\boldsymbol{x}^\top\boldsymbol{z})^2\right).\end{aligned} \tag{38}$$

By Lemma 1, $\frac{d}{d\rho}V(\boldsymbol{x}, \boldsymbol{z}; \rho)|_{\rho=0} = \sigma^{\star 2}(C_2B_1 - 2C_1B_2)/C_2^3$.

Finally by Eq. (28)

$$\begin{aligned}\frac{d}{d\rho}\operatorname{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)|_{\rho=0} &= \frac{d}{d\rho}B^2(\boldsymbol{x}, \boldsymbol{z}; \rho)|_{\rho=0} + \frac{d}{d\rho}V(\boldsymbol{x}, \boldsymbol{z}; \rho)|_{\rho=0} \\ &= \frac{d}{d\rho}V(\boldsymbol{x}, \boldsymbol{z}; \rho)|_{\rho=0} = \sigma^{\star 2}(C_2B_1 - 2C_1B_2)/C_2^3.\end{aligned} \tag{39}$$

B. Proof of Proposition 2. By the central limit theorem, we have that

$$\boldsymbol{x}^\top\boldsymbol{x} = n\left(\gamma_x^2 + \sigma_x^2\right) + \mathcal{O}_p\left(\sqrt{n}\right), \qquad \boldsymbol{z}^\top\boldsymbol{z} = n\left(\gamma_z^2 + \sigma_z^2\right) + \mathcal{O}_p\left(\sqrt{n}\right), \qquad \boldsymbol{x}^\top\boldsymbol{z} = n\left(\gamma_x\gamma_z\right) + \mathcal{O}_p\left(\sqrt{n}\right). \tag{40}$$

Plugging into Eq. (6) gives

$$\begin{aligned}C_1 &= 2n^3\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right]^2 + \mathcal{O}_p\left(n^{5/2}\right), \\ B_1 &= 4n^3\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) + \gamma_x^2\gamma_z^2\right]\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right] + \mathcal{O}_p\left(n^{5/2}\right), \\ C_2 &= n^2\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right] + \mathcal{O}_p\left(n^{3/2}\right), \\ B_2 &= 2n^2\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) + \gamma_x^2\gamma_z^2\right] + \mathcal{O}_p\left(n^{3/2}\right).\end{aligned} \tag{41}$$

Thus we have that

$$\begin{aligned}C_1B_2 &= 4n^5\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) + \gamma_x^2\gamma_z^2\right]\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right]^2 + \mathcal{O}_p\left(n^{9/2}\right), \\ C_2B_1 &= 4n^5\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) + \gamma_x^2\gamma_z^2\right]\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right]^2 + \mathcal{O}_p\left(n^{9/2}\right).\end{aligned} \tag{42}$$

Therefore,

$$C_2B_1 - 2C_1B_2 = -4n^5\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) + \gamma_x^2\gamma_z^2\right]\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right]^2 + \mathcal{O}_p\left(n^{9/2}\right). \tag{43}$$

Now we also know that

$$C_2^3 = n^6\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right]^3 + \mathcal{O}_p\left(n^{11/2}\right). \tag{44}$$

Hence

$$\begin{aligned}&\frac{d}{d\rho}\left[\operatorname{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)\right]|_{\rho=0} \\ &= \sigma^{\star 2}(C_2B_1 - 2C_1B_2)/C_2^3 \\ &= \frac{-4n^5\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) + \gamma_x^2\gamma_z^2\right]\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right]^2}{n^6\left[(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2\right]^3}\sigma^{\star 2} + \mathcal{O}_p\left(n^{-3/2}\right) \\ &= -\frac{4}{n}\frac{(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) + \gamma_x^2\gamma_z^2}{(\gamma_x^2+\sigma_x^2)(\gamma_z^2+\sigma_z^2) - \gamma_x^2\gamma_z^2}\sigma^{\star 2} + \mathcal{O}_p\left(n^{-3/2}\right) \\ &= -\frac{4}{n}\left(1 + \frac{2\gamma_x^2\gamma_z^2}{\sigma_x^2\gamma_z^2 + \sigma_z^2\gamma_x^2 + \sigma_x^2\sigma_z^2}\right)\sigma^{\star 2} + \mathcal{O}_p\left(n^{-3/2}\right) \\ &= -\frac{4}{n}\left(1 + \frac{2\gamma_x^2\gamma_z^2}{\sigma_x^2\gamma_z^2 + \sigma_z^2\gamma_x^2 + \sigma_x^2\sigma_z^2}\right)\left(\sigma_y^2 + \frac{\gamma_y^2\sigma_x^2\sigma_z^2}{\sigma_x^2\gamma_z^2 + \sigma_z^2\gamma_x^2 + \sigma_x^2\sigma_z^2}\right) + \mathcal{O}_p\left(n^{-3/2}\right).\end{aligned} \tag{45}$$

C. Proof of Proposition 3. For $\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{z}; 0)$, we have that by Eq. (28),

$$\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{z}; 0) = B^2(\boldsymbol{x}, \boldsymbol{z}; 0) + V(\boldsymbol{x}, \boldsymbol{z}; 0) + \sigma^{\star 2} = V(\boldsymbol{x}, \boldsymbol{z}; 0) + \sigma^{\star 2}. \quad [46]$$

Here we make use of the fact that when $\rho = 0$, $\mathrm{E}\left[\hat{\theta}_x - \theta_x^\star \mid \boldsymbol{x}, \boldsymbol{z}\right] = \mathrm{E}\left[\hat{\theta}_z - \theta_z^\star \mid \boldsymbol{x}, \boldsymbol{z}\right] = 0$ and that $B^2(\boldsymbol{x}, \boldsymbol{z}; 0) = 0$. For $V(\boldsymbol{x}, \boldsymbol{z}; 0)$, we have that by Eq. (37) and Eq. (40),

$$V(\boldsymbol{x}, \boldsymbol{z}; 0) = \sigma^{\star 2} \frac{C_1}{C_2^2} = \frac{4\sigma^{\star 2}}{n} + \mathcal{O}_p\left(n^{-3/2}\right) = \mathcal{O}_p\left(\frac{1}{n}\right). \quad [47]$$

Therefore,

$$\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{z}; 0) = V(\boldsymbol{x}, \boldsymbol{z}; 0) + \sigma^{\star 2} = \sigma^{\star 2} + \mathcal{O}_p\left(\frac{1}{n}\right). \quad [48]$$

Thus, together with the result in Proposition 2, we have

$$\frac{\frac{d}{d\rho}\left[\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{z}; \rho)\right]|_{\rho=0}}{\mathrm{MSE}(\boldsymbol{x}, \boldsymbol{z}; 0)} = -\frac{4}{n}\left(1 + \frac{2\gamma_x^2\gamma_z^2}{\sigma_x^2\gamma_z^2 + \sigma_z^2\gamma_x^2 + \sigma_x^2\sigma_z^2}\right) + \mathcal{O}_p\left(n^{-\frac{3}{2}}\right). \quad [49]$$

5. Distribution of predicted versus true time to delivery for the labor onset prediction example

We show in Figure S9 the distribution of predicted and true time to delivery for each patient, which gives a better sense of the quality of the predictions for the regression task. The left plot shows the distribution of time to delivery for all patients at their first time points in the longitudinal study; the right plot shows the predicted versus true time to delivery for the training and test samples. This is based on one random split of the training and test sets of 40 and 13 patients, respectively.

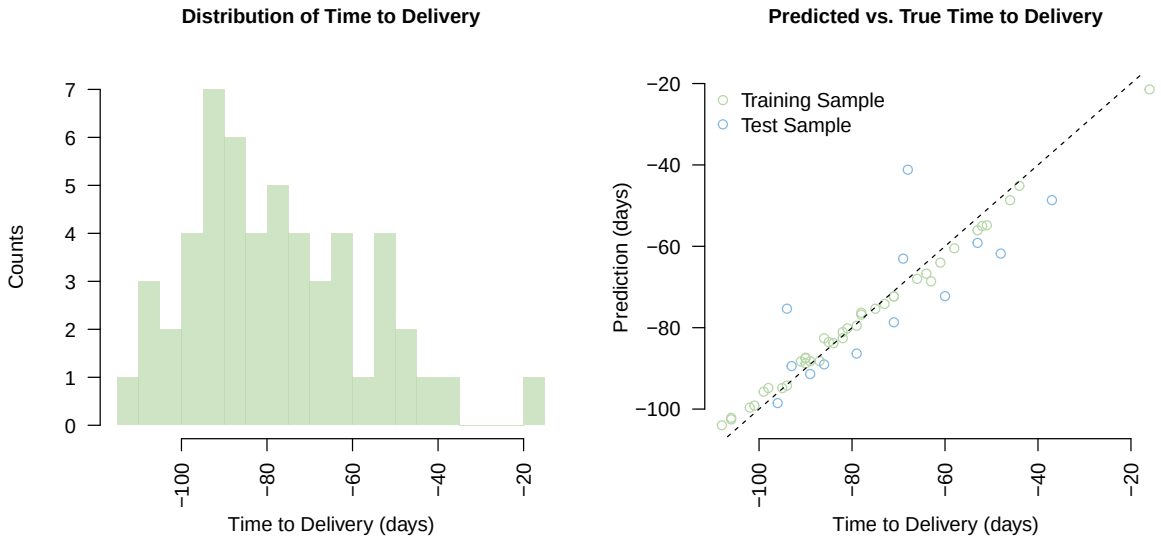

Fig. S9. Distribution of time to delivery and predicted versus true time to delivery for training and test samples. The predictions were derived from cooperative learning.

6. Procedure for generating the imaging and "omics" data

Here we outline the detailed procedure for data generation in the simulation study with imaging and "omics" data in Algorithm S3. The "omics" data (X), imaging data (Z), and the response $\boldsymbol{y}$ are generated such that there are correlations between X, Z, and $\boldsymbol{y}$.

Algorithm S3 *Simulation procedure for generating the imaging and "omics" data.*

Input: Parameters $n, p_x, p_u, s_u, t, \sigma, \boldsymbol{\beta}_u, I_{\max}, \text{ndim}, \text{threshold}$.
Output: $X \in \mathcal{R}^{n \times p_x}$ (omics), $Z \in \mathcal{R}^{n \times \text{ndim} \times \text{ndim} \times 1}$ (images assuming one color channel), $\boldsymbol{y} \in \mathcal{R}^n$.

1. $x_j \in \mathcal{R}^n$ distributed i.i.d. MVN$(0, I_n)$ for $j = 1, 2, \ldots, p_x$
2. For $i = 1, 2, \ldots, p_u$ ($p_u < p_x$, where p_u corresponds to the number of factors):
   (a) $u_i \in \mathcal{R}^n$ distributed i.i.d. MVN$(0, s_u^2 I_n)$
   (b) $x_i = x_i + t * u_i$
3. $U = [u_1, u_2, \ldots, u_{p_u}]$, $X = [x_1, x_2, \ldots, x_{p_x}]$
4. $\boldsymbol{y}_u = U\boldsymbol{\beta}_u + \epsilon$ where $\epsilon \in \mathcal{R}^n$ distributed i.i.d. MVN$(0, \sigma^2 I_n)$
5. For $i = 1, 2, \ldots, n$:
   (a) $P_i = \frac{1}{1+\exp(\boldsymbol{y}_{u_i})}$, $\boldsymbol{y}_i \sim$ Bernoulli(P_i)
   (b) Generate a 2D pixel matrix of image $Z_i \in \mathcal{R}^{\text{dim} \times \text{dim} \times 1}$
   (c) Generate a polygon PG$_i$ inside Z_i, defined by 4 vertices $[v_1, v_2, v_3, v_4]$ on the axes, i.e. $v_1 = [0, a_1], v_2 = [0, a_2], v_3 = [a_3, 0], v_4 = [a_4, 0]$, where $a_1 \sim \text{Unif}(\frac{\text{ndim}}{2}, \text{ndim}), a_2 \sim \text{Unif}(-\text{ndim}, -\frac{\text{ndim}}{2}), a_3 \sim \text{Unif}(\frac{\text{ndim}}{2}, \text{ndim}), a_4 \sim \text{Unif}(-\text{ndim}, -\frac{\text{ndim}}{2})$
   (d) Randomly sample points from Z_i: if the point $[x', y']$ falls inside the polygon PG$_i$, i.e. $[x', y'] \in \text{PG}_i$, then $Z_i[x', y'] \sim \text{Unif}(0, 1)$
   (e) If $\boldsymbol{y}_i = 1$, $I_{\text{disease}} = I_{\max} \times \boldsymbol{y}_{u_i}$, where $I_{\max}$ is the maximum intensity of pixel values for images,
   - For $x' = 1, 2, \ldots, \text{ndim}$:
     - For $y' = 1, 2, \ldots, \text{ndim}$:
       * $P(x', y') \sim \text{Unif}(0, 1)$
       * If $[x', y'] \in \text{PG}_i$ and $P(x', y') < \text{threshold}$, $Z_i[x', y'] = I_{\text{disease}}$