SUPPLEMENTAL DOCUMENT FOR COMBINING ESTIMATES OF CONDITIONAL TREATMENT EFFECTS

This supplemental document has two parts: a theoretical justification for TEEM and additional simulation studies comparing TEEM to other model selection and combination methods. Section S1 develops a risk bound for an alternative version of the TEEM algorithm. TEEM_A, a version of TEEM intended for theoretical development (see Section 3.2 of the main article for a discussion of the differences in the algorithms), is presented, and a risk bound for its estimator's performance under squared error loss is given. Section S2 presents three sets of simulation results. In Section S2.1, all candidate models to be combined are misspecified; in Section S2.1.3, the error distributions are misspecified as well. Finally, Section S2.2 demonstrates how applying ideas of sufficient dimension reduction prior to the pairing step (see Section 3.3 of the main article for details) may improve the performance of TEEM. As in Sections 4 and 5 of the main article, the version of TEEM used in Section S2.1 is the nearest-neighbor version described in Section 3.1 of the main article.

S1. TEEM FOR THEORETICAL DEVELOPMENT

In this section, we present a version of the TEEM algorithm for theoretical development, as opposed to the version of TEEM presented in Section 3.1 of the main article that is recommended for practical use. Section 3.2 of the main article discusses the major differences between the two algorithms and the motivations for both.

S1.1 THE TEEM_A ALGORITHM

Here we describe in detail the version of the TEEM algorithm with independent pairs for which we derive the risk bound in Section S1.2. For our theoretical development, the support of the covariates $\mathcal{U}$ is assumed to be a compact subset of $\mathbb{R}^p$ and the covariate distribution $P_\mathbf{U}$ is assumed to have a density bounded below by a constant $\underline{c} > 0$ on $\mathcal{U}$ almost surely.

Without further loss of generality, we set $\mathcal{U} = [0,1]^p$. Note that these restrictions on $\mathcal{U}$ and $P_{\mathbf{U}}$ are not required or assumed for the version of the algorithm described in Section 3.1 of the main article.

Step 0. Select a fraction $\rho \in (0,1)$ of the n observations that will be used to fit the models. Denote $\lfloor \rho n + 0.5 \rfloor$ by n_1; n_1 is the number of observations used to fit the models. Similarly denote the size of the evaluation set, $n - n_1$, by n_2. Note that asymptotically, n_1 and n_2 are both of order n.

Step 1. Randomly permute the order of the n observations; call this permutation π. Split the resulting ordered data into two parts: the training part $\mathbf{Z}^{(1)} = (Y_i, T_i, \mathbf{U}_i)_{i=1}^{n_1}$ and the evaluation part $\mathbf{Z}^{(2)} = (Y_i, T_i, \mathbf{U}_i)_{i=n_1+1}^{n}$.

Step 2. Within the evaluation data $\mathbf{Z}^{(2)}$, let n_{t_2} denote the number of observations for which $T_i = t$ and n_{c_2} the number for which $T_i = c$. Let $n_2^* = \min(n_{t_2}, n_{c_2})$. Partition $\mathcal{U} = [0,1]^p$ into hypercubes each with side length h such that

$$\frac{1}{h} = \left\lfloor \left(\frac{\underline{c} n_2^*}{2 \log n_2^*} \right)^{1/p} \right\rfloor . \tag{1}$$

Let $\widetilde{n}_2$ denote the number of these hypercubes containing at least one realized covariate value from each treatment group in $\mathbf{Z}^{(2)}$. Within each of these $\widetilde{n}_2$ cells, randomly select a pair of observations (i, i^*) such that $T_i = t$ and $T_{i^*} = c$. Use the indices i from these pairs to create the ordering $m = 1, \ldots, \widetilde{n}_2$, where each m represents the treatment-control pair (i, i^*) with the mth-smallest value of i among the pairs created in this step. Using this index, hereafter denote the treatment and control observations (i, i^*) in pair m by (m_t, m_c).

Step 3. For each resulting matched pair (m_t, m_c), create approximate treatment effects $\widetilde{\delta}_m = Y_{m_t} - Y_{m_c}$. These approximate local treatment effects will be used to evaluate the candidate procedures and assign them weights.

Step 4. Fit the J candidate procedures $\psi_1, \ldots, \psi_J$ to the data $\mathbf{Z}^{(1)}$ to obtain J estimates of the treatment effect function (denoted by $\widehat{\Delta}_{n_1,1}, \ldots, \widehat{\Delta}_{n_1,J}$). Let $\hat{\sigma}_{\widetilde{\delta}_m, n_1, j} = \sqrt{\hat{\sigma}^2_{t,n_1,j}(\mathbf{U}_{m_t}) + \hat{\sigma}^2_{c,n_1,j}(\mathbf{U}_{m_c})}$ denote the estimated standard deviation of $\widetilde{\delta}_m$ from proce-

dure j applied to $\mathbf{Z}^{(1)}$.

Step 5. For each procedure indexed by $j = 1, 2, \ldots, J$, assign initial weights (or prior probabilities) $W_{1,j} = \omega_j$, where the ω_j's are positive numbers that sum to 1. Then for $2 \leq m \leq \widetilde{n}_2$, let

$$W_{m,j} = \frac{\omega_j \prod_{l=1}^{m-1} \phi\left\{\left[\widetilde{\delta}_l - \widehat{\Delta}_{n_1,j}(\mathbf{U}_{l_t})\right] / \hat{\sigma}_{\widetilde{\delta}_l,n_1,j}\right\} / \hat{\sigma}_{\widetilde{\delta}_l,n_1,j}}{\sum_{k=1}^{J} \omega_k \prod_{l=1}^{m-1} \phi\left\{\left[\widetilde{\delta}_l - \widehat{\Delta}_{n_1,k}(\mathbf{U}_{l_t})\right] / \hat{\sigma}_{\widetilde{\delta}_l,n_1,k}\right\} / \hat{\sigma}_{\widetilde{\delta}_l,n_1,k}}, \tag{2}$$

where ϕ is the pdf of the error distribution. Note that $\sum_{j \geq 1} W_{m,j} = 1$ for each $m = 1, \ldots, \widetilde{n}_2$.

Step 6. For $m = 1, \ldots, \widetilde{n}_2$, let

$$\widetilde{\Delta}_m(\mathbf{u}) = \sum_{j=1}^{J} W_{m,j} \widehat{\Delta}_{n_1,j}(\mathbf{u}). \tag{3}$$

Step 7. For every cell m containing at least one treatment-control pair, let $\mathcal{U}_m$ denote the region of the covariate space representing the cell. Then let

$$\widetilde{\widetilde{\Delta}}_\pi(\mathbf{u}) = \begin{cases} \widetilde{\Delta}_m(\mathbf{U}_{m_t}) & \text{if } \mathbf{u} \in \mathcal{U}_m \\ 0 & \text{if the cell containing } \mathbf{u} \text{ has no treatment-control pair in } \mathbf{Z}^{(2)}. \end{cases}$$

The subscript π indicates the estimator's dependence on the permutation π applied in Step 1.

Step 8. Repeat Steps 1-7 a total of P times for some $P \geq 1$, and average the resulting $\widetilde{\widetilde{\Delta}}_\pi$ to obtain the TEEM$_{\text{A}}$ estimator

$$\overline{\widetilde{\Delta}}(\mathbf{u}) = \frac{1}{P} \sum_{p=1}^{P} \widetilde{\widetilde{\Delta}}_{\pi_p}(\mathbf{u}), \tag{4}$$

where for each iteration $1 \leq p \leq P$, π_p denotes the permutation applied in Step 1 of the iteration.

The partition size given in Step 2 is not a bandwidth in the traditional sense. It takes the form given in (1) so that, asymptotically, the partition becomes finer (allowing for more precise estimation of local treatment effects) while each cell continues to contain at least

one treatment-control pair with high probability. The technical aspect of this partition in establishing the risk bound for the TEEM algorithm can be understood from the proof.

Equation (2) gives the formula for the combining weight of each procedure j in cell m. Intuitively, the numerator in (2) represents the product of the predictive densities (likelihoods) of the first $m-1$ treatment-control pairs in the evaluation set given procedure j. The greater the likelihoods, the more trustworthy the procedure, and the higher the weight in (2). The denominator in (2) scales the weights so that they sum to one. The weight of each procedure $W_{m,j}$ in (2) depends on the arbitrary ordering of the treatment-control pairs done in Step 2. However, the effect of the ordering is averaged out by repeating Steps 1-7 over P independent random permutations.

Whether or not the weight of the best procedure goes to one (and thus the method reduces to model selection) asymptotically depends on the relationship between n_1 and n_2 and the relative performance of the candidate procedures; see Rolling and Yang (2014) for a related discussion regarding the ability to identify the best treatment effect estimation procedure among the candidates. Unlike in Rolling and Yang (2014), the current work does not require one of the candidate procedures to be asymptotically better than the others in order for the main theoretical result (presented in the next section) to hold.

S1.2 Risk Bound for the TEEM$_{\mathbf{A}}$ Estimator

In this section we bound the risk of the estimator produced by the TEEM$_{\mathrm{A}}$ algorithm described in Section S1.1. Our proof uses the following assumptions on the data-generating process:

Regularity Conditions

1. Covariates of treatment and control groups: Let $P_{\mathbf{U}_t}$ and $P_{\mathbf{U}_c}$ denote the covariate distributions, conditional on treatment status, for the treatment and control groups, respectively. Note that we allow treatment to be associated with covariates, as is

the case in many observational studies, so $P_{\mathbf{U}_t}$ and $P_{\mathbf{U}_c}$ may differ from each other. We assume that the realizations $\mathbf{U}_i|T_i = t$ are i.i.d. from $P_{\mathbf{U}_t}$ and, similarly, that $\mathbf{U}_i|T_i = c$ are i.i.d. from $P_{\mathbf{U}_c}$. We assume that $P_{\mathbf{U}_t}$ and $P_{\mathbf{U}_c}$ are continuous and bounded above and below by $\overline{c} < \infty$ and $\underline{c} > 0$, respectively, on $\mathcal{U}$.

2. Size of treatment and control groups: For n large enough, there exist constants (a, b) not depending on n such that $0 < a < n_t/n < b < 1$, where n_t is the number of the n observations for which $T_i = t$.

3. Error distribution: The error density ϕ has the property that for each pair $0 < s_0 < 1$ and $T > 0$, there exists a constant B_0 (depending on s_0 and T) such that

$$\int \phi(x) \log \frac{\phi(x)}{(1/s)\phi[(x-t)/s]} dx \leq B_0[(1-s)^2 + t^2]$$

for all $s_0 \leq s \leq 1/s_0$ and $-T < t < T$ (see Assumption A2 in Yang, 2001). Many distributions satisfy this condition, including the Gaussian, t (with degrees of freedom greater than 2), and double-exponential (Laplace) distributions.

4. Boundedness: The regression functions f_t and f_c are uniformly bounded in absolute value by $A < \infty$, and the standard deviation functions σ_t and σ_c each are bounded above and below by $\overline{\sigma} < \infty$ and $\underline{\sigma} > 0$, respectively. We assume correspondingly that the estimators $\widehat{\Delta}_{l,j}$, $\hat{\sigma}_{t,l,j}$, and $\hat{\sigma}_{c,l,j}$ satisfy $\|\widehat{\Delta}_{l,j}\|_\infty \leq 2A$, $\hat{\sigma}_{t,l,j} \in [\underline{\sigma}, \overline{\sigma}]$, and $\hat{\sigma}_{c,l,j} \in [\underline{\sigma}, \overline{\sigma}]$, for each $l \geq 1$ and $j \geq 1$.

5. Smoothness: The regression functions for the treatment and control groups, f_t and f_c, and the estimators $\widehat{\Delta}_{l,j}$ for $l \geq 1$ and $j \geq 1$ have all p first-order partial derivatives, and each of these first-order partial derivatives is upper bounded in absolute value by a constant L on $\mathcal{U}$. The same smoothness is assumed for the standard deviation functions σ_t and σ_c and their corresponding estimators.

The theorem below bounds the risk of the TEEM_A estimator in terms of the minimum

risks of the individual candidate procedures, the dimension of the covariate vector, and the size of the evaluation set.

Theorem 1. *Under regularity conditions 1-5, the risk of* $\overline{\widehat{\Delta}}$ *from the TEEM*$_A$ *algorithm described in Section S1.1 has the following bound:*

$$\mathbb{E}\|\Delta - \overline{\widehat{\Delta}}\|_2^2 \leq C\Bigg\{ \left(\frac{\log n_2}{n_2}\right)^{1/p} + \inf_j \left[\left(\frac{\log n_2}{n_2}\right) \log \frac{1}{\omega_j} + \mathbb{E}\|\sigma_t - \hat{\sigma}_{t,n_1,j}\|_2^2 + \mathbb{E}\|\sigma_c - \hat{\sigma}_{c,n_1,j}\|_2^2 + \mathbb{E}\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 \right] \Bigg\},$$

where the constant C *depends on* a, b, $\underline{c}$, $\overline{c}$, $\underline{\sigma}$, $\overline{\sigma}$, A, p, *and* L *(but not on* n*).*

Proof. A detailed proof is provided in the Appendix of this supplemental document.

Remarks

1. By choosing a fixed fraction of n to fit the estimators and using the remainder to construct the combining weights, n_1 and n_2 both are of order n. Therefore, if one of the candidate models (say j^*) is a correctly specified parametric representation of the data-generating process, then $\mathbb{E}\|\Delta - \widehat{\Delta}_{n_1,j^*}\|_2^2$, $\mathbb{E}\|\sigma_t - \hat{\sigma}_{t,n_1,j^*}\|_2^2$, and $\mathbb{E}\|\sigma_c - \hat{\sigma}_{c,n_1,j^*}\|_2^2$ each will converge to zero at a rate of n^{-1}. In this case, if $p = 1$, the risk of the combined estimator will converge to zero at rate $(\log n)n^{-1}$, almost as fast as an oracle that knows the true model in advance.

2. The finite-sample performance of the method may be sensitive to ρ, the proportion of observations used for training data. In our experience, a 50/50 splitting of the data into estimation and evaluation provides a good balance for achieving the goals of estimating Δ and evaluating competing procedures for estimating Δ. More discussion regarding the allocation between training and testing data can be found in Yang (2007) and Zhang and Yang (2015). The number of permutations P should be as large as computationally feasible to average out the variability due to data splitting, but as is

typical for resampling methods, accuracy gains are subject to diminishing returns for large P. We use $P = 100$ in our numerical work.

3. Increasing the dimension of $\mathbf{U}$ slows the convergence of the combined estimator due to the "curse of dimensionality" in constructing the treatment-control pairs. This suggests that more efficient estimation can be achieved by reducing the dimension of the covariate vector before constructing the pairs if the dimension reduction does not result in any loss of information about Δ. See Sections 3.3 and S2.2 for more details about applying dimension reduction techniques to improve the performance of TEEM in high-dimensional settings.

4. The risk bound of TEEM can be simplified if one is willing to assume the error variances for the treatment and control groups are the same and are homoscedastic with respect to $\mathbf{U}$, as in the following corollary.

Corollary 1. *Assuming homoscedastic errors (that is, σ_t and σ_c are equal constants) and regularity conditions 1-5, the risk bound of $\overline{\widehat{\Delta}}$ can be expressed as*

$$\mathbb{E}\|\Delta - \overline{\widehat{\Delta}}\|_2^2 \leq C\Bigg\{ \left(\frac{\log n_2}{n_2}\right)^{1/p} + \inf_j \left[\left(\frac{\log n_2}{n_2}\right) \log \frac{1}{\omega_j} + \mathbb{E}(\sigma - \hat{\sigma}_{n_1,j})^2 + \mathbb{E}\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 \right] \Bigg\},$$

where $\sigma = \sigma_t = \sigma_c$ and the constant C depends on a, b, $\underline{c}$, $\overline{c}$, $\underline{\sigma}$, $\overline{\sigma}$, A, p, and L (but not on n).

Proof. The proof follows a similar path as that of Theorem 1.

5. In the setting of Corollary 1, with homoscedastic errors and smoothness conditions on f_t and f_c, the standard deviation can be estimated at rate n_1^{-1} independently of the candidate procedures (see, e.g., Rice, 1984). Thus, the term $\mathbb{E}(\sigma - \hat{\sigma}_{n_1,j})^2$ could be removed from the risk bound by incorporating this independent estimation of σ

into the algorithm. However, in practice, separate model-based estimators of σ often are helpful in assigning proper weights to each of the candidate procedures.

6. Under the regularity conditions of Theorem 1 (and Corollary 1), the minimax rate of convergence for Δ is $n^{-2/(2+p)}$. This rate can be achieved by a procedure j^{**} that sets $\widehat{\Delta}_{j^{**}} = \widehat{f}_t - \widehat{f}_c$, where $\widehat{f}_t$ and $\widehat{f}_c$ are appropriate nonparametric estimators that converge to f_t and f_c, respectively, at the minimax rate for each regression function. By including such a procedure j^{**} among the candidates and utilizing independent estimation of σ (see Remark 5), TEEM can automatically achieve the minimax rate for Δ when $p = 1$. However, because of the term $((\log n_2)/n_2)^{1/p}$ in the risk bound, when $p > 1$ inclusion of such a j^{**} is not sufficient to guarantee that the TEEM estimator will converge to Δ at the minimax rate.

S2. Additional Simulation Studies

S2.1 Misspecified Models

Nonlinearity in conditional treatment effects may exist but be difficult to detect through exploratory data analysis. For example, the effectiveness of a retailer's marketing treatment may depend on a customer's level of engagement with the retailer, and the treatment may be most effective for those moderately engaged; very occasional customers may ignore communications from the retailer, while very loyal customers may purchase regardless of the marketing treatment. Such a relationship cannot be captured by linear interaction terms, but this can be difficult to observe graphically. In such situations, it is possible that all candidate models may be misspecified, and that models best for predicting the response may not be best for estimating the treatment effect.

S2.1.1 Setup

We generate data from the following process:

$$Y_i = 0.5U_{i,1}^2 + 0.5U_{i_2} + I(T_i = t) * (0.5U_{i,1} + 0.5U_{i,2}^2) + \varepsilon_i, \tag{A.1}$$

where $(U_{i,1}, U_{i,2}, \varepsilon_i)$ are i.i.d. $N(\mathbf{0}, \mathbf{I}_3)$ and the T_i are i.i.d. with $P(T_i = t) = 0.5$. The nine candidate models contain different subsets of the two covariates and their interactions with treatment; they are enumerated in Table 1. The candidate models are hierarchical in the sense that if a treatment-covariate interaction is included in the model, the main effect of that covariate also is included.

Table 1: Candidate models for the simulation study in Section S2.1.

Model Number	Model Terms
1	$T, U_1, U_2, T * U_1, T * U_2$
2	$T, U_1, U_2, T * U_1$
3	$T, U_1, U_2, T * U_2$
4	T, U_1, U_2
5	$T, U_1, T * U_1$
6	T, U_1
7	$T, U_2, T * U_2$
8	T, U_2
9	T

We create 100 realizations of (A.1) at each of two sample sizes: $n = 100$ and $n = 300$. For each realization, we use the model selection and combination methods described in Section 4.3 to choose a model/combination and use the chosen model/combination to estimate Δ. The squared L_2 risks (for Δ) for each candidate model and for each selection/combination method at each sample size are estimated by averaging the risks over the 100 realizations,

where each realization-risk is estimated from the sample mean of $(\Delta(\mathbf{U}_i) - \widehat{\Delta}(\mathbf{U}_i))^2$ based on an independent evaluation data set of 1 million independent draws from the distribution of $(U_{i,1}, U_{i,2})$.

S2.1.2 Results

In this setting, there is a conflict between the goals of estimating the full regression function and estimating the treatment effect. For example, Model 5 is a relatively effective model for treatment effect estimation, because it contains the linear interaction term between T and U_1, but it is not effective for estimating the full regression function because it omits the main effect for U_2. Because of this conflict, we should expect that the selection and combination methods targeted toward Δ will perform better than the methods targeted toward the full regression function.

Table 2 shows the risks of the model selection and combination methods, as well as the risks of the individual models, at $n = 100$ and $n = 300$. Among the model selection methods, TECV performs the best at both sample size levels. Its performance is much better than that of traditional CV at both sample sizes. Overall, the TEEM algorithm features the lowest risk among all nine methods of selection and combination methods. At $n = 100$, TEEM results in much better performance than any of the model selection methods due to the high model selection instability at this sample size. At $n = 300$, the methods targeted to treatment effect estimation that do not assume the true model is among the candidates (TECV and TEEM) feature the lowest estimated risks for Δ.

S2.1.3 Misspecified Error Distribution

Because it is a likelihood-based method, our algorithm TEEM requires the error distribution to be known. Of course, knowing the true error distribution is unlikely in practice. To evaluate the robustness of our method to incorrect specification of the error distribution, the simulation described in this section was repeated with the random errors being

Table 2: Section S2.1 results with Gaussian errors.

		Estimated Risk of $\widehat{\Delta}$ (SE)	
	Model/Method	$n = 100$	$n = 300$
Candidate Models	Model 1	0.835 (0.025)	0.631 (0.010)
	Model 2	0.712 (0.017)	0.587 (0.009)
	Model 3	0.961 (0.024)	0.825 (0.007)
	Model 4	0.824 (0.009)	0.779 (0.003)
	Model 5	0.732 (0.019)	0.588 (0.009)
	Model 6	0.827 (0.010)	0.780 (0.004)
	Model 7	0.970 (0.023)	0.829 (0.007)
	Model 8	0.831 (0.009)	0.781 (0.004)
	Model 9	0.833 (0.010)	0.781 (0.004)
Model Selection Methods	AIC	0.847 (0.025)	0.629 (0.012)
	BIC	0.856 (0.021)	0.656 (0.013)
	CV	0.858 (0.021)	0.682 (0.014)
	wFIC	0.860 (0.023)	0.633 (0.012)
	TECV	0.834 (0.017)	0.622 (0.012)
Model Combination Methods	cAIC	0.805 (0.021)	0.627 (0.011)
	BMA	0.790 (0.018)	0.644 (0.011)
	ARM	0.733 (0.016)	0.635 (0.009)
	TEEM	0.714 (0.015)	0.607 (0.009)

generated from a double-exponential (Laplace) distribution with variance one, instead of a normal distribution. The model selection and combination methods all incorrectly assume that the errors follow a Gaussian distribution in this study.

The results of the analysis with a misspecified error distribution are shown in Table 3. In this setting, the performance of TEEM is fairly robust to the misspecification of the error distribution. Among the model selection and combination methods considered, TEEM again features the lowest mean squared error, which is close to that of the best candidate procedure, at both sample sizes. This example shows that TEEM can estimate the conditional treatment effect with similar accuracy as the best candidate, even when all candidate model forms and the assumed error distribution are incorrectly specified.

S2.2 TEEM with Dimension Reduction

According to Theorem 1, the risk bound of TEEM may grow with the number of covariates p. This property, common to nonparametric methods, is due to the increased difficulty of finding nearby neighbors in high-dimensional space. As discussed in Section 3.3 of the main article, one solution to this problem when p is moderate or large is to estimate a dimension reduction subspace for Δ and create the pairings necessary for TEEM using the projection of $\mathbf{U}$ onto this lower-dimensional subspace. This section demonstrates the use of dimension reduction techniques with TEEM.

S2.2.1 Setup

In this simulation study, we set $n = 500$ and $p = 10$; these are fairly moderate values typical of what one might encounter in an observational study. While a setting with $p = 10$ may not typically be considered high-dimensional, a sample size much higher than $n = 500$ often is needed to reliably find nearby neighbors in 10-dimensional space. Applying a dimension reduction technique prior to identifying neighbors may therefore be beneficial.

The $p = 10$ covariates are mean-zero normal with covariance matrix $\Sigma_{ij} = 0.7^{|i-j|}$. We

Table 3: Section S2.1 results with Laplace[a] errors.

		Estimated Risk of $\widehat{\Delta}$ (SE)	
	Model/Method	$n = 100$	$n = 300$
Candidate Models	Model 1	0.848 (0.034)	0.610 (0.010)
	Model 2	0.733 (0.028)	0.573 (0.008)
	Model 3	0.974 (0.027)	0.812 (0.006)
	Model 4	0.834 (0.015)	0.774 (0.003)
	Model 5	0.759 (0.030)	0.576 (0.009)
	Model 6	0.844 (0.016)	0.777 (0.003)
	Model 7	0.979 (0.027)	0.814 (0.006)
	Model 8	0.837 (0.014)	0.775 (0.003)
	Model 9	0.847 (0.016)	0.776 (0.003)
Model Selection Methods	AIC	0.854 (0.033)	0.603 (0.010)
	BIC	0.881 (0.030)	0.632 (0.013)
	CV	0.893 (0.029)	0.644 (0.013)
	wFIC	0.860 (0.033)	0.603 (0.010)
	TECV	0.845 (0.025)	0.592 (0.010)
Model Combination Methods	cAIC	0.816 (0.030)	0.602 (0.009)
	BMA	0.824 (0.028)	0.621 (0.009)
	ARM	0.748 (0.021)	0.617 (0.009)
	TEEM	0.717 (0.019)	0.586 (0.008)

[a] The model selection and combination methods, including TEEM, that require the specification of an error distribution incorrectly assume Gaussian errors in this study.

allow the probability that $T = t$ to depend on the covariates in this example; specifically,

$$P(T_i = t) = \frac{\exp\left(\frac{1}{10}\sum_{j=1}^{10} U_{ij}\right)}{1 + \exp\left(\frac{1}{10}\sum_{j=1}^{10} U_{ij}\right)}. \tag{A.2}$$

The outcome Y_i is generated according to

$$Y_i = 2\sum_{j=1}^{6} U_{ij} + I(T_i = t)\sum_{j=1}^{3} U_{ij} + \varepsilon_i, \tag{A.3}$$

where the ε_i are i.i.d. normal with $\sigma = 10$. The 20 candidate models considered are linear regressions with different subsets, including 10 main-effects only models with progressively larger numbers of covariates,

$$\begin{aligned} &T, U_1 \\ &T, U_1, U_2 \\ &\vdots \\ &T, U_1, U_2, U_3, \ldots, U_{10}, \end{aligned}$$

and the same models with interaction terms between the treatment variable and each covariate in the model,

$$\begin{aligned} &T, U_1, T * U_1 \\ &T, U_1, T * U_1, U_2, T * U_2 \\ &\vdots \\ &T, U_1, T * U_1, U_2, T * U_2, U_3, T * U_3, \ldots, U_{10}, T * U_{10}. \end{aligned}$$

As in the previous section, we compare the estimated risks of model selection and combination methods for estimating Δ by selecting one of, or combining, these models. All methods that require the specification of an error distribution assume normal errors. The sample size available for selecting or combining models is $n = 500$, and 100 independent realizations of such samples are generated. For each method-realization, the sample mean

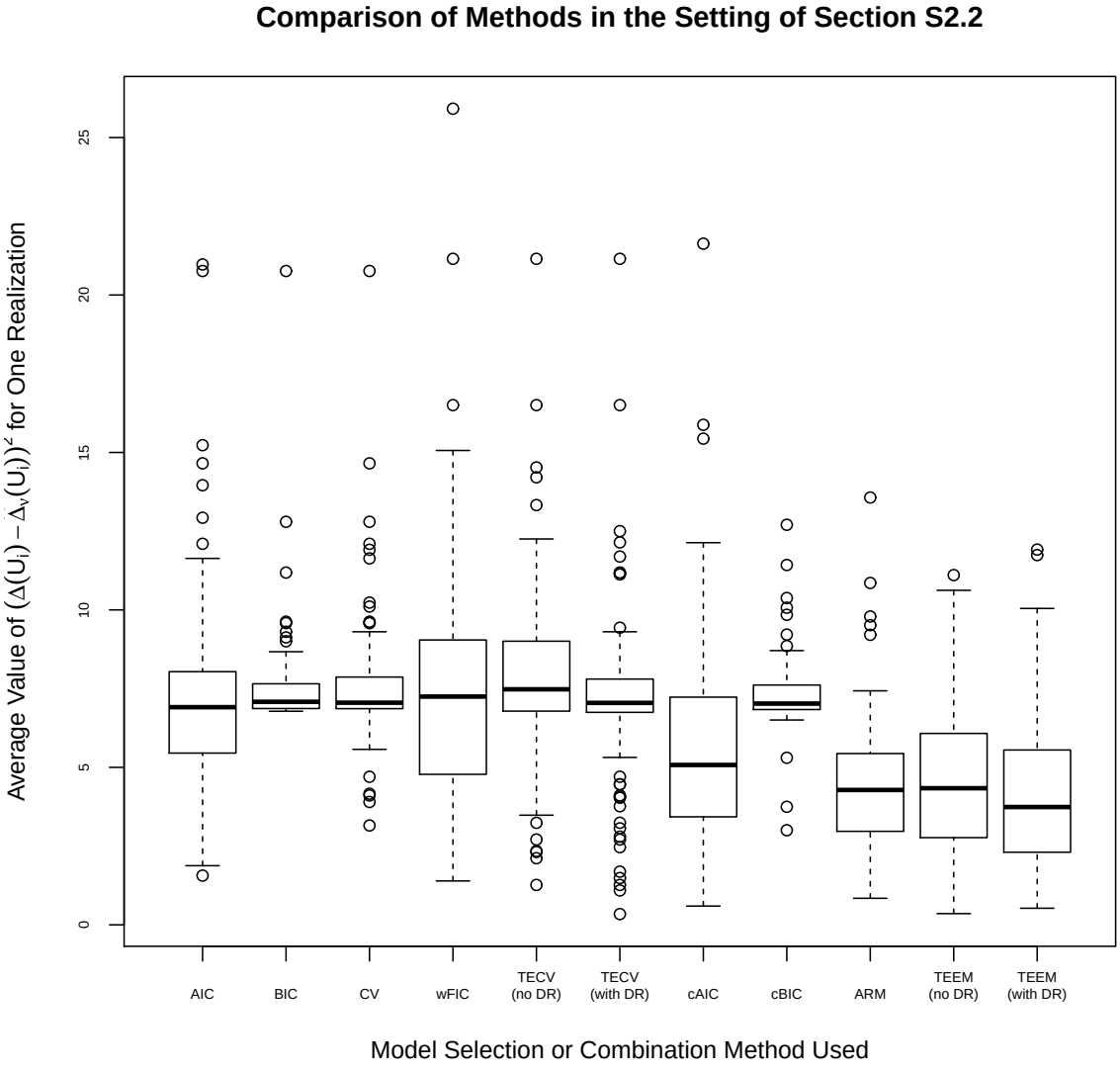
of $(\Delta(\mathbf{U}_i) - \widehat{\Delta}(\mathbf{U}_i))^2$ is calculated for an independent evaluation data set of 1 million draws from the distribution of $\mathbf{U}$. These estimated mean squared errors are averaged over the 100 realizations to estimate the risks of each method. The TEEM algorithm with a dimension reduction step (described below) is compared with the basic TEEM algorithm and the same competitors studied elsewhere in this paper.

For TEEM with dimension reduction, the observations in the treatment and control groups are used to create $\hat{f}_t = \hat{\beta}_t^T\mathbf{U}$ and $\hat{f}_c = \hat{\beta}_c^T\mathbf{U}$, respectively, via OLS regression using all 10 covariates. Assuming the regression functions under treatment and control are linear (which they are in this case), the two-column matrix $(\beta_t^T\mathbf{U}, \beta_c^T\mathbf{U})$ contains all of the information about Δ contained in the original 10-column $\mathbf{U}$. Let $\hat{\beta}_{tc}$ denote the 10×2 matrix with columns $\hat{\beta}_t$ and $\hat{\beta}_c$. If $\hat{\beta}_t$ and $\hat{\beta}_c$ are accurate estimates of their targets, then $\hat{\beta}_{tc}\mathbf{U}$ will be an approximate dimension reduction subspace for Δ. Thus in the pairing step of TEEM with dimension reduction, the distance between two observations $\mathbf{U}_i$ and $\mathbf{U}_{i'}$ is measured by $d(\hat{\beta}_{tc}\mathbf{U}_i, \hat{\beta}_{tc}\mathbf{U}_{i'})$, and for each observation the nearest neighbor in the other treatment group with respect to this distance is used as its pair. The same dimension reduction prior to pairing is used for the TECV algorithm, in addition to the usual TECV with no dimension reduction.

S2.2.2 Results

The boxplot of Figure 1 shows that model combination is generally more effective than model selection in this setting because of the small signal-to-noise ratio. Among the model combination methods, ARM and TEEM achieve the highest accuracy in this setting. For TEEM and TECV, the dimension reduction step prior to pairing is effective in producing pairs that are more similar with respect to $\Delta(\mathbf{u})$, thereby enabling more effective model combination and selection for the purpose of accurately estimating Δ.

Figure 1: Results of the dimension-reduction simulation setting of Section S2.2.



APPENDIX

Proof of Theorem 1

First let $P = 1$, where P is the number of permutations from Step 8 of the algorithm. For each pair m created in Step 2 of the algorithm, let $\widetilde{\delta}_m = Y_{m_t} - Y_{m_c}$ and $\sigma_{\widetilde{\delta}_m} = \sqrt{\sigma_t^2(\mathbf{u}_{m_t}) + \sigma_c^2(\mathbf{u}_{m_c})}$. Conditional on $(\mathbf{U}_{m_t}, \mathbf{U}_{m_c}) = (\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$, the density of $\widetilde{\delta}_m$ under Δ, f_c, σ_t^2 and σ_c^2 can be expressed as

$$p_{\Delta, f_c, \sigma_t^2, \sigma_c^2}(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c}) = \frac{1}{\sigma_{\widetilde{\delta}_m}} \phi \left\{ \frac{\widetilde{\delta}_m - \Delta(\mathbf{u}_{m_t}) - [f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})]}{\sigma_{\widetilde{\delta}_m}} \right\}.$$

The estimated density of $\widetilde{\delta}_m$ under $\widehat{\Delta}$, $\hat{\sigma}_t^2$, $\hat{\sigma}_c^2$, and supposing $f_c(\mathbf{u}_{m_t}) = f_c(\mathbf{u}_{m_c})$ is

$$p_{\widehat{\Delta}, \hat{\sigma}_t^2, \hat{\sigma}_c^2}(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c}) = \frac{1}{\hat{\sigma}_{\widetilde{\delta}_m}} \phi \left\{ \frac{\widetilde{\delta}_m - \widehat{\Delta}(\mathbf{u}_{m_t})}{\hat{\sigma}_{\widetilde{\delta}_m}} \right\},$$

where $\hat{\sigma}_{\widetilde{\delta}_m} = \sqrt{\hat{\sigma}_t^2(\mathbf{u}_{m_t}) + \hat{\sigma}_c^2(\mathbf{u}_{m_c})}$.

Define

$$q_1(\widetilde{\delta}_1 | \mathbf{u}_{1_t}, \mathbf{u}_{1_c}) = \sum_{j=1}^{J} \omega_j p_{\widehat{\Delta}_{n_1,j}, \hat{\sigma}^2_{t,n_1,j}, \hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_1 | \mathbf{u}_{1_t}, \mathbf{u}_{1_c}),$$

and for $2 \leq m \leq \widetilde{n}_2$, define

$$q_m(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c}) = \frac{\sum_{j=1}^{J} \omega_j \left[\prod_{l=1}^{m-1} p_{\widehat{\Delta}_{n_1,j}, \hat{\sigma}^2_{t,n_1,j}, \hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_l | \mathbf{u}_{l_t}, \mathbf{u}_{l_c}) \right] p_{\widehat{\Delta}_{n_1,j}, \hat{\sigma}^2_{t,n_1,j}, \hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c})}{\sum_{j=1}^{J} \omega_j \prod_{l=1}^{m-1} p_{\widehat{\Delta}_{n_1,j}, \hat{\sigma}^2_{t,n_1,j}, \hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_l | \mathbf{u}_{l_t}, \mathbf{u}_{l_c})}.$$

The error density ϕ has mean 0; therefore, given π, $\mathbf{Z}^{(1)}$, $(\mathbf{u}_{l_t}, \mathbf{u}_{l_c}, y_{l_t}, y_{l_c})_{l=1}^{m-1}$, and $(\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$, $q_m(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c})$ has mean $\sum_j W_{m,j} \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t}) = \widetilde{\Delta}_m(\mathbf{u}_{m_t})$, where $W_{m,j}$ represent the weights defined in Step 5 of the TEEM$_{\text{A}}$ algorithm.

Let

$$g_j \left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2} \right] = \prod_{m=1}^{\widetilde{n}_2} p_{\widehat{\Delta}_{n_1,j}, \hat{\sigma}^2_{t,n_1,j}, \hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m | \mathbf{u}_{m_t}, \mathbf{u}_{m_c}),$$

and let

$$\widetilde{g} \left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2} \right] = \sum_{j=1}^{J} \omega_j g_j \left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2} \right].$$

Note that $\prod_{m=1}^{\widetilde{n}_2} q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) = \widetilde{g}\left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right]$. One can view $q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})$ as an estimator of the conditional density of $\widetilde{\delta}_m$ given $(\mathbf{u}_{m_t},\mathbf{u}_{m_c})$. The cumulative risk, under the Kullback-Leibler divergence, of $q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})$ at the design points $(\mathbf{u}_{m_t},\mathbf{u}_{m_c})_{m=1}^{\widetilde{n}_2}$ can be bounded in terms of the risks of the individual procedures using an idea from Barron (1987). Letting $\mathbb{E}_\pi$ denote the expectation conditional on the permutation π and $D(f||g)$ the Kullback-Leibler divergence of g from f, we have

$$
\begin{aligned}
&\sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi D[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})] \\
&= \sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi \int p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \log \frac{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} d\widetilde{\delta}_m \\
&= \sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \log \frac{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} d\widetilde{\delta}_m \\
&= \mathbb{E}_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \left\{ \sum_{m=1}^{\widetilde{n}_2} \log \frac{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} \right\} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2} \\
&= \mathbb{E}_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{\prod_{m=1}^{\widetilde{n}_2} q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2} \\
&= \mathbb{E}_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{\widetilde{g}\left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right]} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2}.
\end{aligned}
$$

Since ϕ is a positive-valued function and $\log(x)$ is an increasing function, we have that for any $j \geq 1$,

$$
\begin{aligned}
&\mathbb{E}_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{\widetilde{g}\left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right]} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2} \\
&\leq \mathbb{E}_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{\omega_j g_j\left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right]} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2} \\
&= \log \frac{1}{\omega_j} \\
&\quad + \mathbb{E}_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{g_j\left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right]} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2}.
\end{aligned}
$$

The last term in the preceding equation is the cumulative risk, under the Kullback-Leibler divergence, of $p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}^2_{t,n_1,j},\hat{\sigma}^2_{c,n_1,j}}$ at the design points $(\mathbf{u}_{m_t},\mathbf{u}_{m_c})_{m=1}^{\widetilde{n}_2}$, given the permutation

π. This is because

$$\begin{aligned}
&\mathbb{E}_\pi \int \left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \log \frac{\prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{g_j\left[(\widetilde{\delta}_m)_{m=1}^{\widetilde{n}_2}\right]} d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2} \\
=&\mathbb{E}_\pi \int \left[\left\{ \prod_{m=1}^{\widetilde{n}_2} p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \right\} \right. \\
&\qquad \left. \times \left\{ \sum_{m=1}^{\widetilde{n}_2} \log \frac{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}^2_{t,n_1,j},\hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} \right\} \right] d\widetilde{\delta}_1 \cdots d\widetilde{\delta}_{\widetilde{n}_2} \\
=&\sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi \int p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \log \frac{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}{p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}^2_{t,n_1,j},\hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} d\widetilde{\delta}_m \\
=&\sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi D[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}^2_{t,n_1,j},\hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})].
\end{aligned}$$

By definition,

$$\begin{aligned}
&D[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}^2_{t,n_1,j},\hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})] \\
=&\int \left(\frac{1}{\sigma_{\widetilde{\delta}_m}} \phi \left\{ \frac{\widetilde{\delta}_m - \Delta(\mathbf{u}_{m_t}) - [f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})]}{\sigma_{\widetilde{\delta}_m}} \right\} \right. \\
&\qquad \left. \times \log \frac{(1/\sigma_{\widetilde{\delta}_m})\phi\left(\left\{ \widetilde{\delta}_m - \Delta(\mathbf{u}_{m_t}) - [f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})] \right\} / \sigma_{\widetilde{\delta}_m} \right)}{(1/\hat{\sigma}_{\widetilde{\delta}_m,n_1,j})\phi\left\{ \left[\widetilde{\delta}_m - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t}) \right] / \hat{\sigma}_{\widetilde{\delta}_m,n_1,j} \right\}} \right) d\widetilde{\delta}_m.
\end{aligned}$$

Letting

$$z = \frac{\widetilde{\delta}_m - \Delta(\mathbf{u}_{m_t}) - [f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})]}{\sigma_{\widetilde{\delta}_m}},$$

we perform an integral transformation to obtain

$$\begin{aligned}
&D[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}^2_{t,n_1,j},\hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})] \\
=&\int \phi(z) \log \left[\frac{\hat{\sigma}_{\widetilde{\delta}_m,n_1,j}}{\sigma_{\widetilde{\delta}_m}} \right. \\
&\qquad \left. \times \frac{\phi(z)}{\phi\left\{ \left(\sigma_{\widetilde{\delta}_m} z + \Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t}) + [f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})] \right) / \hat{\sigma}_{\widetilde{\delta}_m,n_1,j} \right\}} \right] dz.
\end{aligned}$$

Using the condition provided for the error distribution ϕ and taking

$$s_0 = \underline{\sigma}/\overline{\sigma},\ s = \hat{\sigma}_{\widetilde{\delta}_m,n_1,j}/\sigma,\ T = 4A/(\sqrt{2}\underline{\sigma}),\ \text{and}$$

$$t = -\left\{ \frac{\left[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right] + \left[f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right]}{\sigma_{\widetilde{\delta}_m}} \right\},$$

it follows that

$$\begin{aligned}
&D[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}^2_{t,n_1,j},\hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})] \\
&\leq B_0 \left(\left\{ 1 - \frac{\hat{\sigma}_{\widetilde{\delta}_m,n_1,j}}{\sigma_{\widetilde{\delta}_m}} \right\}^2 + \left\{ \frac{\left[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right] + \left[f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right]}{\sigma_{\widetilde{\delta}_m}} \right\}^2 \right),
\end{aligned}$$

for a constant B_0 depending on A, $\underline{\sigma}$, and $\overline{\sigma}$. Using $(\sigma_{\widetilde{\delta}_m})^2 \geq 2\underline{\sigma}^2$ and the parallelogram law, we obtain that for any $j \geq 1$,

$$\begin{aligned}
&D[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||p_{\widehat{\Delta}_{n_1,j},\hat{\sigma}^2_{t,n_1,j},\hat{\sigma}^2_{c,n_1,j}}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})] \\
&\leq \frac{B_0}{\underline{\sigma}^2} \left\{ \frac{1}{2}\left[\sigma_{\widetilde{\delta}_m} - \hat{\sigma}_{\widetilde{\delta}_m,n_1,j}\right]^2 + \left[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right]^2 + \left[f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right]^2 \right\}.
\end{aligned}$$

By the reverse triangle inequality,

$$\begin{aligned}
\left[\sigma_{\widetilde{\delta}_m} - \hat{\sigma}_{\widetilde{\delta}_m,n_1,j}\right]^2 &= \left[\sqrt{\sigma_t^2(\mathbf{u}_{m_t}) + \sigma_c^2(\mathbf{u}_{m_c})} - \sqrt{\hat{\sigma}^2_{t,n_1,j}(\mathbf{u}_{m_t}) + \hat{\sigma}^2_{c,n_1,j}(\mathbf{u}_{m_c})}\right]^2 \\
&\leq \left[\sigma_t(\mathbf{u}_{m_t}) - \hat{\sigma}_{t,n_1,j}(\mathbf{u}_{m_t})\right]^2 + \left[\sigma_c(\mathbf{u}_{m_c}) - \hat{\sigma}_{c,n_1,j}(\mathbf{u}_{m_c})\right]^2.
\end{aligned}$$

Thus we have shown

$$\begin{aligned}
&\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi D[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})||q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})] \\
&\quad\leq \frac{B_0}{\underline{\sigma}^2\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi\left[f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right]^2 + \inf_j \left(\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} \right. \\
&\qquad + \frac{B_0}{\underline{\sigma}^2}\left\{ \frac{1}{2\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi\left(\left[\sigma_t(\mathbf{u}_{m_t}) - \hat{\sigma}_{t,n_1,j}(\mathbf{u}_{m_t})\right]^2 + \left[\sigma_c(\mathbf{u}_{m_c}) - \hat{\sigma}_{c,n_1,j}(\mathbf{u}_{m_c})\right]^2 \right) \right. \\
&\qquad\quad \left. \left. + \frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi\left[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right]^2 \right\} \right). \qquad \text{(A.4)}
\end{aligned}$$

Let $d_H^2(f,g) = \int(\sqrt{f} - \sqrt{g})^2 d\nu$ denote the squared Hellinger distance between the densities f and g with respect to the measure ν. The squared Hellinger distance is upper bounded by the K-L divergence, so

$$\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2} \mathbb{E}_\pi d_H^2[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})]$$

is bounded above by (A.4).

As mentioned earlier, for each m, given π, $\mathbf{Z}^{(1)}$, $(\mathbf{u}_{l_t}, \mathbf{u}_{l_c}, y_{l_t}, y_{l_c})_{l=1}^{m-1}$, and $(\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$, $q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t}, \mathbf{u}_{m_c})$ has mean $\widetilde{\Delta}_m(\mathbf{u}_{m_t})$ with respect to $\widetilde{\delta}_m$. For this estimator, we have

$$
\begin{aligned}
&\left[\int \widetilde{\delta}_m p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m - \int \widetilde{\delta}_m q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right]^2 \\
=&\left\{\int \widetilde{\delta}_m \left[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) - q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right] d\widetilde{\delta}_m\right\}^2 \\
=&\Bigg\{\int \widetilde{\delta}_m \left[\sqrt{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} + \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right] \\
&\qquad \times \left[\sqrt{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} - \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right] d\widetilde{\delta}_m\Bigg\}^2 \\
\le& \int \widetilde{\delta}_m^2 \left[\sqrt{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} + \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right]^2 d\widetilde{\delta}_m \\
&\qquad \times \int \left[\sqrt{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} - \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right]^2 d\widetilde{\delta}_m \\
\le& 2\left[\int \widetilde{\delta}_m^2 p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) + \int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right] \\
&\qquad \times \int \left[\sqrt{p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})} - \sqrt{q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})}\right]^2 d\widetilde{\delta}_m \\
=& 2\left[\mathbb{E}(\widetilde{\delta}_m^2|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) + \int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right] \\
&\qquad \times d_H^2\left[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right] \\
=& 2\left\{\left[\mathbb{E}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right]^2 + \sigma_t^2(\mathbf{u}_{m_t}) + \sigma_c^2(\mathbf{u}_{m_c}) + \int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right\} \\
&\qquad \times d_H^2\left[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right] \\
=& 2\left\{\left[\Delta(\mathbf{u}_{m_t}) + f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\right]^2 + \sigma_t^2(\mathbf{u}_{m_t}) + \sigma_c^2(\mathbf{u}_{m_c}) + \int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right\} \\
&\qquad \times d_H^2\left[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right],
\end{aligned}
$$

where the first and second inequalities follow from the Cauchy-Schwarz inequality and the parallelogram law, respectively.

By the fourth regularity condition, $[\Delta(\mathbf{u}_{m_t}) + f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})]^2 \le (4A)^2$. Now $\int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m = \mathbb{E}_{q_m}(\widetilde{\delta}_m^2|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) \le [\mathbb{E}_{q_m}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})]^2 + (2\overline{\sigma})^2$, and $q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})$ is a convex combination of J densities in the location-scale family $\phi[(x -$

$b)/a]/a$, each with mean $\widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})$ with respect to $\widetilde{\delta}_m$. Therefore, $\int \widetilde{\delta}_m^2 q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m$ is bounded above by $(2A)^2+(2\overline{\sigma})^2$. It follows that

$$\left[\int \widetilde{\delta}_m p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m - \int \widetilde{\delta}_m q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m\right]^2$$
$$\leq (40A^2+16\overline{\sigma}^2)d_H^2\left[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right].$$

Together with

$$\int \widetilde{\delta}_m p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m = \mathbb{E}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}) = \Delta(\mathbf{u}_{m_t})+f_c(\mathbf{u}_{m_t})-f_c(\mathbf{u}_{m_c})$$

and

$$\int \widetilde{\delta}_m q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})d\widetilde{\delta}_m = \widetilde{\Delta}_m(\mathbf{u}_{m_t}),$$

we have, for each $1\leq m\leq \widetilde{n}_2$,

$$\left[\Delta(\mathbf{u}_{m_t})+f_c(\mathbf{u}_{m_t})-f_c(\mathbf{u}_{m_c})-\widetilde{\Delta}_m(\mathbf{u}_{m_t})\right]^2$$
$$\leq (40A^2+16\overline{\sigma}^2)d_H^2\left[p_{\Delta,f_c,\sigma_t^2,\sigma_c^2}(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c}), q_m(\widetilde{\delta}_m|\mathbf{u}_{m_t},\mathbf{u}_{m_c})\right]. \quad (A.5)$$

The expression (A.5) also is an upper bound for $\left\{\Delta(\mathbf{u}_{m_t})-[f_c(\mathbf{u}_{m_t})-f_c(\mathbf{u}_{m_c})]-\widetilde{\Delta}_m(\mathbf{u}_{m_t})\right\}^2$. So by the parallelogram law, (A.5) is an upper bound for $\left[\Delta(\mathbf{u}_{m_t})-\widetilde{\Delta}_m(\mathbf{u}_{m_t})\right]^2$. Then by using the earlier risk bound on the average squared Hellinger distance and combining constants, we obtain

$$\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\left[\Delta(\mathbf{u}_{m_t})-\widetilde{\Delta}_m(\mathbf{u}_{m_t})\right]^2$$
$$\leq B_2\left(\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\left[f_c(\mathbf{u}_{m_t})-f_c(\mathbf{u}_{m_c})\right]^2+\inf_j\left\{\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j}\right.\right.$$
$$+\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\left(\left[\sigma_t(\mathbf{u}_{m_t})-\hat{\sigma}_{t,n_1,j}(\mathbf{u}_{m_t})\right]^2+\left[\sigma_c(\mathbf{u}_{m_c})-\hat{\sigma}_{c,n_1,j}(\mathbf{u}_{m_c})\right]^2\right)$$
$$\left.\left.+\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\left[\Delta(\mathbf{u}_{m_t})-\widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\right]^2\right\}\right), \quad (A.6)$$

where B_2 depends on $\underline{\sigma}$, $\overline{\sigma}$, and A.

Now we connect the global risk of the estimator $\widetilde{\widetilde{\Delta}}_\pi$ to the average risk of the individual estimators $\widetilde{\Delta}_m$ at the design points. Let D_π denote the event that $\widetilde{n}_2 = (1/h)^p$; that is, the event that every cell in the partition of $\mathcal{U}$ contains at least one treatment-control pair from $\mathbf{Z}^{(2)}$ after the permutation π. Let $\mathcal{U}_m$ denote the cell in the partition containing the mth treatment-control pair. Conditional on D_π,

$$\begin{aligned}
&\mathbb{E}_\pi \|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \\
&= \mathbb{E}_\pi \int_{\mathcal{U}} \Big[\Delta(\mathbf{u}) - \widetilde{\widetilde{\Delta}}_\pi(\mathbf{u})\Big]^2 dP_{\mathbf{U}} \\
&= \mathbb{E}_\pi \sum_{m=1}^{\widetilde{n}_2} \int_{\mathcal{U}_m} \Big[\Delta(\mathbf{u}) - \widetilde{\widetilde{\Delta}}_\pi(\mathbf{u})\Big]^2 dP_{\mathbf{U}}.
\end{aligned}$$

By the definition of $\widetilde{\widetilde{\Delta}}_\pi$, for any $\mathbf{u} \in \mathcal{U}_m$, $\widetilde{\widetilde{\Delta}}_\pi(\mathbf{u}) = \widetilde{\Delta}_m(\mathbf{u}_{m_t})$. Therefore, for $\mathbf{u} \in \mathcal{U}_m$,

$$\begin{aligned}
&\Big[\Delta(\mathbf{u}) - \widetilde{\widetilde{\Delta}}_\pi(\mathbf{u})\Big]^2 \\
&= \Big\{\Big[\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big] + \Big[\Delta(\mathbf{u}_{m_t}) - \widetilde{\Delta}_m(\mathbf{u}_{m_t})\Big]\Big\}^2 \\
&\quad \le 2\Big[\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big]^2 + 2\Big[\Delta(\mathbf{u}_{m_t}) - \widetilde{\Delta}_m(\mathbf{u}_{m_t})\Big]^2.
\end{aligned}$$

Combining the previous two displays and using the fact that for any m, $\int_{\mathcal{U}_m} dP_{\mathbf{U}} \le \overline{c}/\widetilde{n}_2$, we have

$$\begin{aligned}
&\mathbb{E}_\pi \|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \\
&\le 2\mathbb{E}_\pi \left\{ \sum_{m=1}^{\widetilde{n}_2} \int_{\mathcal{U}_m} \Big[\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big]^2 dP_{\mathbf{U}} + \frac{\overline{c}}{\widetilde{n}_2} \sum_{m=1}^{\widetilde{n}_2} \Big[\Delta(\mathbf{u}_{m_t}) - \widetilde{\Delta}_m(\mathbf{u}_{m_t})\Big]^2 \right\}. \qquad \text{(A.7)}
\end{aligned}$$

For the first summation on the right-hand side of (A.7), by the Mean Value Theorem for integrals and the fact that every cell $\mathcal{U}_m$ has volume $1/\widetilde{n}_2$, we have

$$\sum_{m=1}^{\widetilde{n}_2} \int_{\mathcal{U}_m} \Big[\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big]^2 dP_{\mathbf{U}} = \frac{1}{\widetilde{n}_2} \sum_{m=1}^{\widetilde{n}_2} f(\mathbf{u}_m^*)\Big[\Delta(\mathbf{u}_m^*) - \Delta(\mathbf{u}_{m_t})\Big]^2,$$

where $\mathbf{u}_m^*$ is some point in the hypercube $\mathcal{U}_m$ and $f(\mathbf{u}_m^*)$ represents the design density at this point. The smoothness conditions on f_t and f_c imply that Δ satisfies a Lipschitz condition

with Lipschitz constant $\sqrt{p}L$. Thus for any m, since the distance between $\mathbf{u}_m^*$ and $\mathbf{u}_{m_t}$ is at most $\sqrt{p}h$, $\Delta(\mathbf{u}_m^*) - \Delta(\mathbf{u}_{m_t}) \leq pLh$. Thus we have

$$\sum_{m=1}^{\widetilde{n}_2} \int_{\mathcal{U}_m} \Big[\Delta(\mathbf{u}) - \Delta(\mathbf{u}_{m_t})\Big]^2 dP_{\mathbf{U}} \leq \overline{c}(pLh)^2. \tag{A.8}$$

Combining (A.6), (A.7), and (A.8), we have established

$$\begin{aligned}
&\mathbb{E}_\pi\left[\|\Delta - \widetilde{\widehat{\Delta}}_\pi\|_2^2 \Big| D_\pi\right] \\
&\quad \leq 2\overline{c}(pLh)^2 + 2\overline{c}B_2\left(\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\Big[f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\Big]^2\right. \\
&\qquad + \inf_j\left\{\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + \frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\left(\Big[\sigma_t(\mathbf{u}_{m_t}) - \hat{\sigma}_{t,n_1,j}(\mathbf{u}_{m_t})\Big]^2 + \Big[\sigma_c(\mathbf{u}_{m_c}) - \hat{\sigma}_{c,n_1,j}(\mathbf{u}_{m_c})\Big]^2\right)\right. \\
&\qquad\quad \left.\left. + \frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big]^2\right\}\right).
\end{aligned} \tag{A.9}$$

Next we relate the global risk of each $\widehat{\Delta}_{n_1,j}$ to its average risk at the design points. Again using the Mean Value Theorem for integrals and conditioning on D_π, we have for any $j \geq 1$,

$$\begin{aligned}
&\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big]^2 - \mathbb{E}_\pi\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 \\
&\qquad \leq \frac{c^*}{\widetilde{n}_2}\mathbb{E}_\pi\sum_{m=1}^{\widetilde{n}_2}\left\{\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big]^2 - \Big[\Delta(\mathbf{u}_m^*) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_m^*)\Big]^2\right\},
\end{aligned}$$

where c^* is a constant bounded by $\max(1/\underline{c}, \overline{c})$ that exists by the boundedness of $P_{\mathbf{U}}$. The difference in the squared differences after the summation can be bounded for each m by the smoothness of Δ and $\widehat{\Delta}_{n_1,j}$.

Indeed, for each m we have

$$\begin{aligned}
&\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big]^2 - \Big[\Delta(\mathbf{u}_m^*) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_m^*)\Big]^2 \\
&= \Big\{\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big] + \Big[\Delta(\mathbf{u}_m^*) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_m^*)\Big]\Big\} \\
&\qquad \times \Big\{\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big] - \Big[\Delta(\mathbf{u}_m^*) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_m^*)\Big]\Big\}.
\end{aligned}$$

Since Δ and $\widehat{\Delta}_{n_1,j}$ both are bounded between $-2A$ and $2A$,

$$\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big] + \Big[\Delta(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m)\Big] \leq 4A.$$

Meanwhile, the smoothness of Δ and $\widehat{\Delta}_{n_1,j}$ ensure that both satisfy a Lipschitz condition with Lipschitz constant $\sqrt{p}L$. Thus for any m, since each $\mathcal{U}_m$ has diameter $\sqrt{p}h$,

$$\begin{aligned}&\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big] - \Big[\Delta(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m)\Big] \\ &= \Big[\Delta(\mathbf{u}_{m_t}) - \Delta(\mathbf{u}^*_m)\Big] + \Big[\widehat{\Delta}_{n_1,j}(\mathbf{u}^*_m) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big] \leq 2pLh.\end{aligned}$$

Therefore, conditional on D_π,

$$\frac{1}{\widetilde{n}_2}\mathbb{E}_\pi \sum_{m=1}^{\widetilde{n}_2}\Big[\Delta(\mathbf{u}_{m_t}) - \widehat{\Delta}_{n_1,j}(\mathbf{u}_{m_t})\Big]^2 \leq \mathbb{E}_\pi\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 + 8c^*ApLh. \tag{A.10}$$

Because σ_t and σ_c and their corresponding estimators also are bounded and smooth (conditions 4 and 5), we can apply similar arguments to obtain, conditional on D_π,

$$\frac{1}{\widetilde{n}_2}\mathbb{E}_\pi \sum_{m=1}^{\widetilde{n}_2}[\sigma_t(\mathbf{u}_{m_t}) - \widehat{\sigma}_{t,n_1,j}(\mathbf{u}_{m_t})]^2 \leq \mathbb{E}_\pi\|\sigma_t - \widehat{\sigma}_{t,n_1,j}\|_2^2 + 4c^*\overline{\sigma}pLh \tag{A.11}$$

and

$$\frac{1}{\widetilde{n}_2}\mathbb{E}_\pi \sum_{m=1}^{\widetilde{n}_2}[\sigma_c(\mathbf{u}_{m_c}) - \widehat{\sigma}_{c,n_1,j}(\mathbf{u}_{m_c})]^2 \leq \mathbb{E}_\pi\|\sigma_c - \widehat{\sigma}_{c,n_1,j}\|_2^2 + 4c^*\overline{\sigma}pLh. \tag{A.12}$$

Thus combining (A.10), (A.11) and (A.12) with (A.9), we have established that

$$\begin{aligned}&\mathbb{E}_\pi\left[\|\Delta - \widetilde{\widehat{\Delta}}_\pi\|_2^2\middle| D_\pi\right] \\ &\quad\leq 8c^*pLh(A+\overline{\sigma}) + \overline{c}(pLh)^2 + B_2\Bigg\{\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\Big[f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\Big]^2 \\ &\qquad + \inf_j\left[\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + \mathbb{E}_\pi\|\sigma_t - \hat{\sigma}_{t,n_1,j}\|_2^2 + \mathbb{E}_\pi\|\sigma_c - \hat{\sigma}_{c,n_1,j}\|_2^2 + \mathbb{E}_\pi\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right]\Bigg\}.\end{aligned}$$

Using the Lipschitz condition for f_c within each cell, in a similar fashion as before, we can show that

$$\frac{1}{\widetilde{n}_2}\sum_{m=1}^{\widetilde{n}_2}\mathbb{E}_\pi\Big[f_c(\mathbf{u}_{m_t}) - f_c(\mathbf{u}_{m_c})\Big]^2 \leq (pLh)^2.$$

Thus we have

$$\mathbb{E}_\pi\left[\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi\right] \leq 8c^* pLh(A+\overline{\sigma}) + B_3\Bigg\{(pLh)^2 + \inf_j\left[\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + \mathbb{E}_\pi\|\sigma_t - \hat{\sigma}_{t,n_1,j}\|_2^2 + \mathbb{E}_\pi\|\sigma_c - \hat{\sigma}_{c,n_1,j}\|_2^2 + \mathbb{E}_\pi\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right]\Bigg\}, \tag{A.13}$$

for a constant B_3 depending on $\underline{\sigma}$, $\overline{\sigma}$, A, and $\overline{c}$.

Now,

$$\mathbb{E}_\pi\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \leq \mathbb{E}_\pi\left[\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi\right] + \mathbb{E}_\pi\left[\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi^c\right] \times P(D_\pi^c). \tag{A.14}$$

By the boundedness of Δ and $\widetilde{\widetilde{\Delta}}_\pi$ between $-2A$ and $2A$,

$$\mathbb{E}_\pi\left[\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \middle| D_\pi^c\right] \leq 16A^2. \tag{A.15}$$

To use (A.14), we need to bound $P(D_\pi^c)$. Denote the event that all cells in our partition contain at least one observation from the treatment group by $D_{\pi,t}$, and let $D_{\pi,c}$ denote the corresponding event for the control group. Since $D_\pi = D_{\pi,t} \cap D_{\pi,c}$, $P(D_\pi^c) \leq P(D_{\pi,t}^c) + P(D_{\pi,c}^c)$.

Let $\mathcal{U}_g$ denote an arbitrary cell in the partition. By the first regularity condition, the probability that any observation from the treatment group falls into $\mathcal{U}_g$ is at least $\underline{c}h^p$. Since the covariate values of the n_{t_2} treatment observations are i.i.d., the probability that $\mathcal{U}_g$ contains no treatment observations from $\mathbf{Z}^{(2)}$ is at most

$$(1-\underline{c}h^p)^{n_{t_2}} = e^{n_{t_2}\log(1-\underline{c}h^p)} \leq e^{-n_{t_2}\underline{c}h^p},$$

where the last inequality results from the fact that $\log x \leq x - 1$.

Since $\mathcal{U}_g$ is arbitrary and there are $(1/h)^p$ such cells in the partition of $\mathcal{U}$, the probability that any of them contain no treatment observations is at most

$$(1/h)^p e^{-n_{t_2}\underline{c}h^p} = \exp[-n_{t_2}\underline{c}h^p + p\log(1/h)].$$

By the choice of h in Step 2 of the TEEM$_\text{A}$ algorithm, $h \geq [2\log(n_2^*)/\underline{c}n_2^*]^{1/p}$. Therefore,

$$\begin{aligned}
&- n_{t_2}\underline{c}h^p + p\log(1/h) \\
&\quad \leq \frac{-2n_{t_2}\log(n_2^*)}{n_2^*} + \log\left(\frac{\underline{c}n_2^*}{2\log n_2^*}\right) \\
&\quad \leq \log\left(\frac{\underline{c}}{2n_2^*\log n_2^*}\right) \\
&\quad \leq \log\left(\frac{\underline{c}}{2\widetilde{n}_2\log\widetilde{n}_2}\right).
\end{aligned}$$

The second inequality in the above expression results from $n_{t_2} \geq n_2^*$. Thus

$$P(D_{\pi,t}^c) \leq \exp\left[\log\left(\frac{\underline{c}}{2n_2^*\log n_2^*}\right)\right] = \left(\frac{\underline{c}}{2n_2^*\log n_2^*}\right).$$

The same bound may be established for $P(D_{\pi,c}^c)$; therefore,

$$P(D_\pi^c) \leq \frac{\underline{c}}{n_2^*\log n_2^*}. \tag{A.16}$$

Using (A.14) together with (A.13), (A.15), and (A.16), and using the fact that $h = B_4\{\log(n_2^*)/n_2^*\}^{1/p}$ for some B_4 depending on $\underline{c}$ and p, we have

$$\begin{aligned}
&\mathbb{E}_\pi\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \\
&\quad \leq 8c^*pL(A+\overline{\sigma})B_4\left(\frac{\log n_2^*}{n_2^*}\right)^{1/p} + B_3(B_4pL)^2\left(\frac{\log n_2^*}{n_2^*}\right)^{2/p} + 16A^2\underline{c}\left(\frac{1}{n_2^*\log n_2^*}\right) \\
&\qquad + B_3\inf_j\left[\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + \mathbb{E}_\pi\|\sigma_t - \hat{\sigma}_{t,n_1,j}\|_2^2 + \mathbb{E}_\pi\|\sigma_c - \hat{\sigma}_{c,n_1,j}\|_2^2 + \mathbb{E}_\pi\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right].
\end{aligned} \tag{A.17}$$

With the exception of small n_2^*,

$$\frac{1}{n_2^*\log n_2^*} \leq \left(\frac{\log n_2^*}{n_2^*}\right)^{2/p} \leq \left(\frac{\log n_2^*}{n_2^*}\right)^{1/p},$$

so we can rewrite expression (A.17) as

$$\begin{aligned}
\mathbb{E}_\pi\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \leq B_5\Bigg\{&\left(\frac{\log n_2^*}{n_2^*}\right)^{1/p} \\
&+ \inf_j\left[\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + \mathbb{E}_\pi\|\sigma_t - \hat{\sigma}_{t,n_1,j}\|_2^2 + \mathbb{E}_\pi\|\sigma_c - \hat{\sigma}_{c,n_1,j}\|_2^2 + \mathbb{E}_\pi\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right]\Bigg\},
\end{aligned}$$

for a constant B_5 depending on $\underline{c}$, $\overline{c}$, $\underline{\sigma}$, $\overline{\sigma}$, A, p, and L.

Now n_2^* and $\widetilde{n}_2$, which heretofore we have treated as fixed, are random variables determined by the values of $(\mathbf{U}_i, T_i)_{i=1}^n$ and the permutation π. By the iterated expectation law, unconditional on the permutation π,

$$\begin{aligned}\mathbb{E}\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 &= \mathbb{E}\left(\mathbb{E}_\pi\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2\right) \\ &\leq B_5\Bigg\{\mathbb{E}\left[\left(\frac{\log n_2^*}{n_2^*}\right)^{1/p}\right] \\ &\qquad + \inf_j\left[\mathbb{E}\frac{1}{\widetilde{n}_2}\log\frac{1}{\omega_j} + \mathbb{E}\|\sigma_t - \hat{\sigma}_{t,n_1,j}\|_2^2 + \mathbb{E}\|\sigma_c - \hat{\sigma}_{c,n_1,j}\|_2^2 + \mathbb{E}\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2\right]\Bigg\}.\end{aligned} \tag{A.18}$$

Let $\alpha \in (0,1)$ be a fixed constant and let $H_{\alpha,\pi}$ denote the event that $P(n_2^* \geq \alpha n_2)$. Since $(\log n_2^*/n_2^*)^{1/p} \leq 1$, we have

$$\begin{aligned}\mathbb{E}\left[\left(\frac{\log n_2^*}{n_2^*}\right)^{1/p}\right] &\leq \mathbb{E}\left[\left(\frac{\log n_2^*}{n_2^*}\right)^{1/p}\middle| H_{\alpha,\pi}\right] + P(H_{\alpha,\pi}^c) \\ &\leq \alpha^{-1/p}\left(\frac{\log n_2}{n_2}\right)^{1/p} + P(H_{\alpha,\pi}^c).\end{aligned}$$

For $P(H_{\alpha,\pi}^c)$, the exponential bound on the upper tail probability of the hypergeometric distribution established by Chvátal (1979) can be used to show that we can find $\alpha \in (0,1)$ depending on a and b from the second regularity condition such that

$$P(H_{\alpha,\pi}^c) \leq B_6 e^{-n_2},$$

for a constant B_6 depending on a and b. Thus

$$\mathbb{E}\left[\left(\frac{\log n_2^*}{n_2^*}\right)^{1/p}\right] \leq B_7\left(\frac{\log n_2}{n_2}\right)^{1/p}, \tag{A.19}$$

for B_7 depending on a and b.

For $\mathbb{E}(1/\widetilde{n}_2)$, conditional on D_π,

$$\frac{1}{\widetilde{n}_2} = h^p = \left\{\left\lfloor\left(\frac{\underline{c}n_2^*}{2\log n_2^*}\right)^{1/p}\right\rfloor\right\}^{-p} \leq B_8\left(\frac{\log n_2^*}{n_2^*}\right) \leq B_7 B_8\left(\frac{\log n_2}{n_2}\right), \tag{A.20}$$

for a constant B_8 depending on $\underline{c}$. As established earlier in this proof, $P(D_\pi^c)$ converges faster than $O(1/n_2^*) = O(1/n_2)$.

Using (A.19) and (A.20) to replace the random variables in (A.18) with fixed constants, we obtain a bound for the risk of $\widetilde{\widetilde{\Delta}}_\pi$:

$$\begin{aligned}
&\mathbb{E}\|\Delta - \widetilde{\widetilde{\Delta}}_\pi\|_2^2 \\
&\quad \leq B_9 \Bigg\{ \left(\frac{\log n_2}{n_2}\right)^{1/p} \\
&\qquad + \inf_j \left[\left(\frac{\log n_2}{n_2}\right) \log \frac{1}{\omega_j} + \mathbb{E}\|\sigma_t - \hat{\sigma}_{t,n_1,j}\|_2^2 + \mathbb{E}\|\sigma_c - \hat{\sigma}_{c,n_1,j}\|_2^2 + \mathbb{E}\|\Delta - \widehat{\Delta}_{n_1,j}\|_2^2 \right] \Bigg\},
\end{aligned} \tag{A.21}$$

for a constant B_9 depending on a, b, $\underline{c}$, $\overline{c}$, $\underline{\sigma}$, $\overline{\sigma}$, A, p, and L.

For $P > 1$, the estimator $\overline{\widetilde{\Delta}}$ from Step 8 of the algorithm is the average (over the set of P permutations) of $\widetilde{\widetilde{\Delta}}_{\pi_p}$. Therefore, by the convexity of the L_2 loss, an application of Jensen's inequality gives us

$$\mathbb{E}\|\Delta - \overline{\widetilde{\Delta}}\|_2^2 \leq \frac{1}{P}\sum_{p=1}^{P} \mathbb{E}\|\Delta - \widetilde{\widetilde{\Delta}}_{\pi_p}\|_2^2. \tag{A.22}$$

The permutation π used to establish the bound in (A.21) was arbitrary; therefore, by (A.22), the bound in (A.21) also holds for $\mathbb{E}\|\Delta - \overline{\widetilde{\Delta}}\|_2^2$. This completes the proof of the theorem. ■

REFERENCES

Barron, A. R. (1987) Are Bayes rules consistent in information? In T. M. Cover & B. Gopinath (eds.), *Open Problems in Communication and Computation*, pp. 85–91. Springer-Verlag.

Chvátal, V. (1979) The tail of the hypergeometric distribution. *Discrete Mathematics* 25, 285–287.

Rice, J. (1984) Bandwidth choice for nonparametric regression. *The Annals of Statistics* 12, 1215–1230.

Rolling, C. A. & Y. Yang (2014) Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76, 749–769.

Yang, Y. (2001) Adaptive regression by mixing. *Journal of the American Statistical Association* 96, 574–588.

Yang, Y. (2007) Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 2450–2473.

Zhang, Y. & Y. Yang (2015) Cross-validation for selecting a model selection procedure. *Journal of Econometrics* 187, 95–112.