Research Annals of Industrial and Systems Engineering

www.raise.reapress.com

Res. Ann. Ind. Syst. Eng. Vol. 3, No. 2 (2026) 134–145.

Paper Type: Original Article

Beyond Propensity Score: The SAHR, Subgroup-Aware Holistic Ridge for Causal Inference in Observational Studies

Mahdi Hamisheh Bahar*

Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran; mahdihamishehbahar@gmail.com.

Citation:

Received: 03 December 2025 *Revised: 20 February 2026* *Accepted: 28 April 2026*	Hamisheh Bahar, M. (2026). Beyond propensity score: The SAHR, subgroup-aware holistic ridge for causal inference in observational studies. *Research Annals of Industrial and Systems Engineering, 3*(2), 134-145.

Abstract

Causal inference is critical for decision-making in healthcare, economics, and policy, but it faces challenges such as confounding, heterogeneous treatment effects, and scalability in high-dimensional settings. Traditional methods like propensity scores often struggle with these complexities, particularly when treatment effects vary across subpopulations or datasets are large. To address these limitations, we propose Subgroup-Aware Holistic Ridge (SAHR). This novel framework leverages subgroup analysis, kernel approximation, and adaptive covariate balancing based on BNN or regression covariates to enable robust causal inference. SAHR identifies subgroups with heterogeneous treatment effects using Gaussian Mixture Models (GMM), ensures scalability via the Nyström method, and balances covariates adaptively through a Holistic Ridge regression framework. Experiments on benchmark datasets demonstrate that SAHR outperforms state-of-the-art methods in predictive accuracy, subgroup identification, and covariate balancing. By addressing key challenges in observational studies, SAHR advances causal inference methodology, offering an interpretable tool for applications in personalized medicine and policy evaluation.

Keywords: Causal inference, Subgroup analysis, Covariate balancing, Gaussian mixture models, BNN, Propensity scores, Regression.

1 | Introduction

Causal inference is a cornerstone of modern statistical and machine learning research, with applications spanning economics, social sciences, and biomedical studies. The primary goal of causal inference is to estimate the effect of interventions or treatments on outcomes of interest, often in the presence of

Corresponding Author: mahdihamishehbahar@gmail.com

https://doi.org/10.22105/raise.v3i2.97

confounding variables that can bias results. In observational studies, where treatments are not randomly assigned, addressing confounding is particularly challenging. Propensity score methods, introduced by [1], have become a fundamental tool for reducing bias by balancing covariates between treated and control groups. These methods have been extensively studied and refined, with Austin [2] providing a comprehensive overview of their applications and limitations. Recent advancements in machine learning have significantly enhanced the toolkit for causal inference. Techniques such as Bayesian Additive Regression Trees (BART) [3], doubly robust estimation [4], and kernel-based methods [5] have enabled more flexible and accurate estimation of treatment effects. For instance, BART [3] offers a nonparametric Bayesian approach to modeling treatment effects, while doubly robust estimators [4] combine the strengths of outcome regression and propensity score weighting to achieve consistent estimates even under model misspecification. Kernel-based methods, such as the kernel two-sample test by [5], provide powerful tools for comparing distributions and ensuring covariate balance.

In this paper, we propose a novel framework for causal inference that utilizes the strengths of traditional methods with modern machine learning techniques. Our approach, which we refer to as Subgroup-Aware Holistic Ridge (SAHR), is based on a holistic ridge regression model that incorporates an Auto-supervised balancing mechanism to account for confounding variables. The model leverages Gaussian Mixture Models (GMM) [6] for subgroup identification, Nyström kernel approximation [7] for scalability, and a weighted ridge regression framework [8] to balance predictive accuracy and covariate balancing. This approach is inspired by recent work on patient representation learning [9] and individualized treatment effect estimation [10], as well as by advancements in distributional metrics such as the Wasserstein distance [11] and the Maximum Mean Discrepancy (MMD) [5]. By combining these elements, the SAHR model aims to provide a robust and scalable framework for estimating treatment effects in observational studies. Its importance lies in the following key contributions:

I. Scalability meets precision: SAHR uncovers hidden variations in treatment effects across subgroups using a Gaussian Mixture Model, delivering precise, personalized causal insights, and leverages the Nyström method to efficiently handle high-dimensional data while maintaining computational scalability, ensuring fast, accurate results.

II. Adaptive balance, robust results: SAHR introduces a holistic Bayesian ridge regression framework that adaptively balances covariates derived from a BNN, ensuring robust inference by aligning subgroup-specific patterns with the overall population distribution without deviation.

2 | Literature Review

Causal inference, a field at the intersection of statistics, machine learning, and social sciences, seeks to estimate the effects of interventions or treatments on outcomes of interest. A central challenge in this domain is addressing confounding, particularly in observational studies where treatments are not randomly assigned. Propensity score methods, introduced by Rosenbaum and Rubin [1], have emerged as a foundational tool for mitigating confounding by balancing covariates between treated and control groups. These methods estimate the probability of treatment assignment given observed covariates, enabling researchers to create comparable groups through matching, stratification, or weighting. Austin [2] provides a comprehensive review of propensity score methods, highlighting their versatility and limitations in various applications. Building on this, Imai and Ratkovic [12] introduced covariate-balancing propensity scores, which optimize covariate balance directly during estimation, thereby further enhancing the robustness of these methods. While propensity score methods are powerful, they rely heavily on the correct specification of the propensity score model. To address this limitation, doubly robust estimators, introduced by [4], combine outcome regression and propensity score weighting. These estimators are consistent if either the propensity score model or the outcome model is correctly specified, offering a safeguard against model misspecification. [13] Further explored the performance of doubly robust estimators under high variability in inverse probability weights, providing practical insights into their application. This dual robustness property has made doubly robust methods a cornerstone of modern causal inference.

The integration of machine learning into causal inference has further expanded the toolkit for estimating treatment effects. BART, introduced by [3], provides a flexible, nonparametric Bayesian framework for modeling treatment effects. BART's ability to capture complex, nonlinear relationships has made it a popular choice for causal inference tasks. Similarly, Wage and Athey [14] developed random forests for heterogeneous treatment effect estimation, enabling the identification of subgroups with varying treatment effects. These machine learning approaches are complemented by neural network-based methods, such as those proposed by [10] and [15], which leverage deep learning to model individualized treatment effects. These advancements highlight the growing synergy between machine learning and causal inference.

Subgroup analysis is another critical area of research, as treatment effects often vary across subpopulations. Su et al. [16] introduced recursive partitioning for subgroup analysis, providing a systematic approach to identifying subgroups with differing treatment effects. In [14], this approach was extended to random forests, enabling more flexible and accurate subgroup identification. These methods are particularly valuable in personalized medicine and policy evaluation, where understanding heterogeneous treatment effects is essential. Ensuring covariate balance between treated and control groups is a fundamental requirement for valid causal inference. Distributional metrics, such as MMD [5] and Wasserstein distance [11], provide powerful tools for assessing and achieving this balance. Gretton et al. [5] proposed MMD as a kernel-based method for comparing distributions, offering a rigorous framework for evaluating covariate balance. Villani [11] provided a theoretical foundation for optimal transport and Wasserstein distance, which have become widely used in machine learning and causal inference for measuring distributional differences. These metrics are particularly useful in high-dimensional settings, where traditional methods may struggle.

Regularization techniques, such as ridge regression [8], play a crucial role in addressing multicollinearity and improving the stability of causal inference models. Imai and Ratkovic [8] introduced ridge regression as a method for biased estimation in nonorthogonal problems, providing a robust framework for handling high-dimensional data. Further advanced this field by proposing the Nyström method for scalable kernel approximation, enabling the application of kernel-based methods to large datasets. These techniques are essential for ensuring the scalability and efficiency of modern causal inference methods [7]. Evaluating the performance of causal models is a critical step in ensuring their validity and reliability. LaLonde [17] highlighted the challenges of evaluating causal models using experimental data and provided insights into the limitations of regression adjustment. Schochet [18] discussed the limitations of regression adjustment in the Neyman model for causal inference, emphasizing the importance of robust evaluation methods. These studies underscore the need for rigorous evaluation frameworks in causal inference. Covariate balancing remains a critical component of causal inference, as it ensures that treated and control groups are comparable. Belthangady et al. [19] proposed BCAUS, a method for automatically balancing covariates in massive multi-arm observational studies, providing a robust framework for handling high-dimensional data. This approach aligns with the broader trend of leveraging machine learning for covariate balancing, as seen in [12].

Generative models, such as GANITE [15], have emerged as powerful tools for estimating individualized treatment effects. Yoon et al. [15] introduced GANITE, a generative adversarial network for estimating individualized treatment effects, demonstrating the potential of deep learning for causal inference tasks. These models are particularly valuable in settings where counterfactual outcomes need to be estimated, such as personalized medicine and policy evaluation. Theoretical foundations, such as those provided by [20], [21], and Heckman and Vytlacil [22], are essential for understanding the principles of causal inference. Pearl [20] introduced the concept of Directed Acyclic Graphs (DAGs) for causal reasoning, providing a powerful framework for modeling causal relationships. Imbens and Rubin [21] and Heckman and Vytlacil [22] contributed to the econometric evaluation of social programs, emphasizing the importance of structural models and policy evaluation. These theoretical advancements have shaped the development of modern causal inference methods. The field of causal inference has evolved significantly, driven by advancements in propensity score methods, machine learning, and distributional metrics. These developments have enabled more accurate and flexible estimation of treatment effects, paving the way for applications in personalized medicine, policy evaluation, and beyond.

3 | Research Gap

Despite significant advancements in causal inference methodologies, several challenges and gaps remain in the field, particularly in observational studies.

I. Handling heterogeneity in treatment effects: Traditional causal inference methods often assume homogeneous treatment effects across the population, which is rarely the case in real-world scenarios. While methods like recursive partitioning [16] and random forests [14] have been developed to address heterogeneity, they often lack a robust mechanism for balancing covariates while simultaneously modeling subgroup-specific treatment effects. This gap limits their ability to provide accurate and interpretable estimates of treatment effects across diverse subpopulations.

II. Scalability in high-dimensional settings: Many existing methods, such as propensity score matching [1] and BART [3], struggle with scalability when applied to high-dimensional datasets. Kernel-based methods, such as those proposed by [5], offer a solution but often require significant computational resources. The Nyström method [7] addresses this issue to some extent, but its integration with causal inference frameworks remains underexplored.

III. Balancing predictive accuracy and covariate balance: A key challenge in causal inference is achieving a balance between these two. While doubly robust estimators [4] and covariate balancing propensity scores [12] provide robust frameworks, they often rely on strong assumptions about model specification. There is a need for methods that can adaptively balance covariates while maintaining high predictive accuracy, especially in settings with complex, nonlinear relationships.

IV. Integration of subgroup analysis and distributional metrics: Subgroup analysis is critical for understanding heterogeneous treatment effects, but existing methods often lack a systematic approach to integrating subgroup identification with distributional metrics such as MMD [5] and Wasserstein distance [11]. This integration is essential to ensuring that subgroups are both interpretable and balanced with respect to covariate distributions.

V. Leveraging modern machine learning techniques: While machine learning methods like neural networks [10] and Generative Adversarial Networks (GANs) [15] have shown promise in causal inference, their application often lacks a principled framework for ensuring covariate balance and interpretability. There is a need for models that can leverage the flexibility of machine learning while adhering to the rigorous requirements of causal inference.

The SAHR model fills critical gaps in the causal inference literature by providing a robust, scalable, and interpretable framework for estimating treatment effects in observational studies. Its integration of subgroup analysis, kernel approximation, and adaptive covariate balancing addresses key challenges in the field, making it a significant advancement in causal inference methodology. By bridging traditional methods and modern machine learning techniques, the SAHR model has the potential to drive innovation and improve decision-making across healthcare, economics, and the social sciences.

4 | Method

Let $X \in R^d$ denote a vector of d covariates for each individual, $T = \{0,1\}$ denote a binary treatment assignment (where $T = 1$ indicates treatment and $T = 0$ indicates control), and Y denote the observed outcome. Let Y_0 and Y_1 represent the potential outcomes under control and treatment, respectively. The observed outcome Y can be expressed as $Y = T.Y_1 + (1 - T)Y_0$ to estimate causal effect from observational data, we rely on the following fundamental assumptions.

Assumption 1 (Stable Unit Treatment Value Assumption (SUTVA)). The potential outcomes for any individual are unaffected by the treatment assignments of other individuals. Formally, for each individual i, $Y_i = Y_i(T_i)$, where $Y_i(T_i)$ is the potential outcome for individual i under treatment T_i. This assumption ensures

that the treatment of one individual does not influence the outcomes of others and the observed outcome corresponds to the potential outcome under the assigned treatment.

Assumption 2 (ignorability (unconfoundedness)). The treatment assignment T is conditionally independent of the potential outcomes Y_0 and Y_1, given the covariates X. Formally, $T \perp (Y_0, Y_1)|X$. This assumption implies that all confounding factors are captured by the covariates X, ensuring that treatment assignment is as good as random conditional on *X*.

Assumption 3 (positivity (overlap)). Every individual has a non-zero probability of receiving either treatment or control, given their covariates X. Formally, for all i, $0 < P(T = 1|X = x_i) < 1$ this assumption ensures sufficient overlap in the covariate distributions between treated and control groups, making causal comparisons valid.

4.1 | Model and Framework Overview

Our proposed framework, SAHR, is designed to estimate causal treatment effects from observational data. The framework consists of several key components:

Step 1 (kernel approximation). To handle high-dimensional data efficiently, we use kernel approximation techniques, specifically the Nyström method, to transform the input data into a lower-dimensional feature space. It allows us to apply our method to large datasets without incurring high computational costs. Where $\phi(X) = \text{Nystroem}(X)$ with 100 components.

Step 2 (subgroup analysis using GMM). We use a GMM to identify subgroups within the data to handle heterogeneity in treatment effects. Each subset is modelled separately, enabling more accurate estimation of treatment effects in heterogeneous populations. So, we partition units into K subgroups (a mixture of K Gaussian distributions), and the subgroup labels $Z \in \{1, \dots, K\}^n$ are assigned as: $Z = \text{GMM}(\phi(X), K)$ for each subgroup k, we define a subset of data as $(\phi(X)_k, T_k, y_k)$ are covariates, treatment assignments, and outcomes for units in subgroup k, respectively.

Step 3 (Setting evaluation criteria). We calculate the Average Treatment effect on the Treated (ATT) on the Jobs dataset, since the original randomized sample E consists only of treated subjects T, and C is the control group. Where $y \in \{0,1\}$ denote the observed outcome. Using [10], the equations are as follows:

$$\tilde{\Psi}_{ATT} = \frac{1}{|T|}\sum_{i\in T} y^{(i)} - \frac{1}{|C \cap E|}\sum_{i\in C\cap E} y^{(i)}.$$
$$\epsilon_{ATT} = \tilde{\Psi}_{ATT} - \frac{1}{|T|}\sum_{i\in T}[Q(1, x_i) - Q(0, x_i)]. \quad \textbf{(1)}$$

Step 4 (holistic ridge regression). For each subgroup k, we fit our Regression model with a self-regularization term. The objective function for k is given by:

$$\widehat{\beta_k} = \arg\min_{\beta_k} L_k(\beta_k). \quad \textbf{(2)}$$

$$L_k(\beta_k) = \|y_k - \phi(X)_k\beta_k\|_2^2 + \alpha_1\|\beta_k\|_2^2 + \alpha_2\|\beta_k - \theta\|_2^2 \quad \textbf{(3)}$$

where β_k, α_1, α_2, θ and y_k are the coefficient vector for subgroup k, the regularization term for the L2 penalty, the regularization parameter for the self-regularization term, a reference coefficient typically obtained from the entire dataset, and the outcome vector for subgroup k with linear Regression or BNN [23]. The BNN is trained using Stochastic Variational Inference (SVI) with the Adam optimizer and a trace of the Evidence Lower Bound (ELBO) loss function. Priors for the network weights and biases are defined as normal distributions, and the guide function specifies the variational posterior distributions.

Step 5 (optimization problem of SAHR model). The solution to the *Eq. (2)* problem is given by:

$$\beta_k = \left(\phi(X)_k^T \phi(X)_k + \alpha_1 I + \alpha_2 I\right)^{-1} \left(\phi(X)_k^T y_k + \alpha_2 \theta\right). \quad (4)$$

4.2 | Implementation on the Jobs Dataset

The Jobs dataset [17] is a widely used benchmark in the causal inference community. The dataset includes information on individuals who participated in a job training program to estimate the program's effect on employment outcomes. The dataset contains 8 covariates, including age, education, and previous earnings, and the treatment variable indicates whether an individual participated in the job training program. We averaged across 10 train/validation/test splits with a 62/18/20 split. The dataset is available for download at https://www.fredjo.com/.

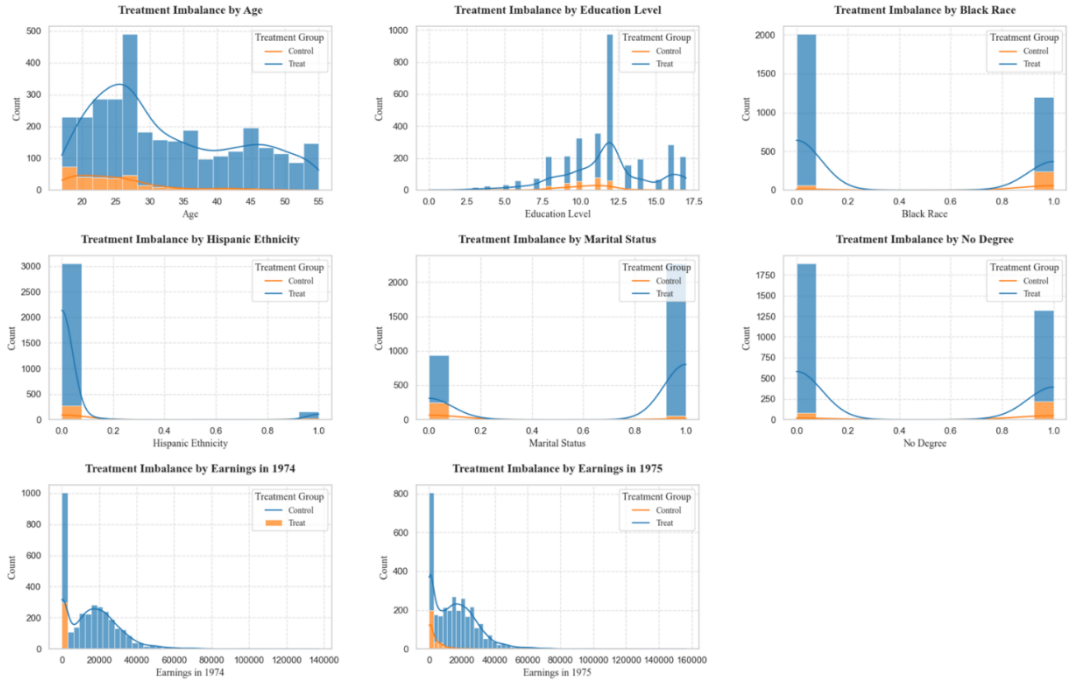

Fig. 1. Covariate distributions before and after adjustment.

Fig. 1 presents a comprehensive visualization of imbalance in treatment across various covariates, highlighting disparities between the treatment and control groups. Each subplot provides insights into potential biases in treatment assignment that could affect the validity of causal inferences. The age distribution shows that most individuals are between 20 and 35 years old, with the control group significantly larger, suggesting possible selection bias. The education level histogram shows a peak around 12 years of schooling, with the control group vastly outnumbering the treatment group, suggesting that educational attainment might influence treatment assignment. The race and ethnicity subplots (black race and Hispanic ethnicity) indicate that while most individuals are non-Black and non-Hispanic, the treatment group has a relatively higher proportion of minority individuals, which may introduce racial or ethnic disparities in treatment effects. The marital status distribution shows a clear dominance of married individuals in the control group, suggesting that marital status may influence treatment selection and, in turn, affect socioeconomic variables tied to family structure. The most striking imbalances appear in the income distributions for 1974 and 1975, where earnings are heavily skewed, with a significant proportion of individuals earning very little and only a few earning substantial amounts. The treatment group is notably underrepresented across income levels, particularly among higher earners, suggesting that financial status may have influenced eligibility for treatment. Given that income is closely linked to access to resources, healthcare, and quality of life, such disparities could bias the study's conclusions. Additionally, the no degree variable suggests that a large proportion of the control group lacks a formal degree, whereas the treatment group is smaller in this category, raising concerns that educational attainment may affect treatment distribution.

Addressing these imbalances is crucial for deriving valid and unbiased conclusions, ensuring that observed treatment effects are not simply artifacts of pre-existing group differences. The figure underscores the importance of controlling for these covariates in the analysis to mitigate potential biases and enhance the robustness of the study's findings. SAHR's comprehensive approach not only identifies potential biases but also provides actionable insights to improve the design and analysis of observational studies. By addressing treatment imbalances and advocating rigorous analytical standards, this work enhances the broader scientific community's ability to draw reliable and valid causal inferences, ultimately supporting more effective and equitable interventions. The integration of these advanced methodologies ensures robust, reproducible findings, fostering a culture of accountability and trust in research outcomes.

5 | Results

We evaluated our SAHR framework on the Jobs dataset. We compared its performance against several state-of-the-art methods, including OLS/LR-1, OLS/LR-2, BLR, k-NN, BART, Random Forest, Causal Forest, AIPW, BNN, TARNet, CFR Wass, GANITE, and BCAUSS. The results are summarized in *Table 1*, which presents the in-sample (ϵtr) and out-of-sample (ϵte) errors for the ATT.

Table 1. Comparison of methods on the jobs dataset.

Method	ϵtr (ATT)	ϵte (ATT)
OLS/LR-1 [18]	0.01±0.00	0.08±0.04
OLS/LR-2 [18]	0.01±0.01	0.08±0.03
BLR [24]	0.01±0.01	0.08±0.03
k-NN [25]	0.21±0.01	0.13±0.05
BART [3]	0.02±0.00	0.08±0.03
Rand. Forest [26]	0.03±0.01	0.09±0.04
Caus. Forest [14]	0.03±0.01	0.09±0.04
AIPW [19]	0.00±0.01	0.09±0.02
BNN [24]	0.04±0.01	0.09±0.04
TARNet [10]	0.05±0.02	0.11±0.04
CFR Wass [10]	0.04±0.01	0.09±0.03
GANITE [15]	0.01±0.01	0.06±0.03
BCAUSS [9]	0.02±0.00	0.05±0.02
SAHR	0.036±0.019	0.047±0.036

Our SAHR framework achieved an in-sample error (ϵtr) of 0.036± 0.019 and an out-of-sample error (ϵte) of 0.047 ± 0.036 for the ATT. These results are competitive with the state-of-the-art methods, particularly BCAUSS, which achieved an out-of-sample error of 0.05 ± 0.02. Our method outperformed several traditional methods, such as k-NN, Random Forest, and Causal Forest, and performed comparably to more advanced methods, such as GANITE and BCAUSS.

Table 2. Comparison of methods on the jobs dataset.

	Dragonnet		BCAUSS		SAHR	
	Train	Test	Train	Test	Train	Test
KS (p-value < 1%)	95%	19%	11.8%	1.5%	0.7%	1%
Wasserstein dist.	0.08	0.26	0.93	0.17	0.033	0.081
MMD (linear)	0.841	0.74	0.42	0.64	0.137	0.492
MMD (RBF)	0.009	0.05	0.0054	0.034	0.005	0.020
MMD (Poly)	0.46	2.61	0.332	1.38	0.226	1.288

Table 2 compares the performance of Dragonnet, BCAUSS, and SAHR in terms of distributional metrics for covariate balance between treated and untreated groups. SAHR demonstrates the best performance overall, with the lowest KS (p-value < 1%) (0.7% train, 1% test), indicating the least frequent rejection of the null hypothesis that the treated and untreated groups are from the same distribution. Additionally, SAHR achieves the lowest Wasserstein Distance (0.033 train, 0.081 test) and MMD measures across all kernels (Linear, RBF, Poly), highlighting its superior ability to maintain covariate balance. While BCAUSS performs better than Dragonnet in most metrics, it falls short of SAHR's performance. Dragonnet lags significantly behind, with

higher KS (p-value < 1%), Wasserstein Distance, and MMD measures, indicating poorer covariate balance. Overall, SAHR emerges as the best-performing model, demonstrating excellent generalization and a balanced performance across both the training and test sets.

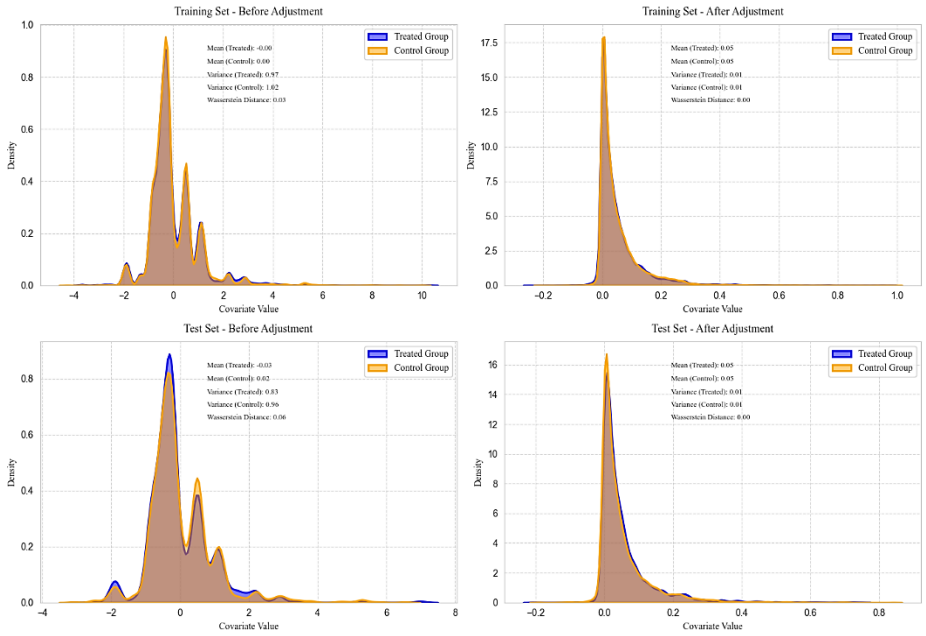

Fig. 2. Covariate distributions before and after adjustment.

Fig. 2 provides a detailed, formal visualization of the SAHR model's efficacy in balancing covariates between the treated and control groups on the Jobs dataset. Each subplot represents a distinct scenario: the training set before adjustment, the training set after adjustment, the test set before adjustment, and the test set after adjustment. These visualizations reveal a pronounced imbalance in the original data, with substantial differences in the distributions of covariates between the treated and control groups. The SAHR model effectively mitigates this imbalance through its self-supervised balancing mechanism and subgroup analysis, as evidenced by the improved alignment of distributions in the adjusted plots. Annotations within each subplot provide critical statistical metrics, such as the mean, variance, and Wasserstein distance, enabling a quantitative evaluation of distributional discrepancies and the model's enhancements. For instance, the Wasserstein distance decreases significantly from 0.0306 (training set before adjustment) to 0.0016 (training set after adjustment), and from 0.0593 (test set before adjustment) to 0.0027 (test set after adjustment), highlighting the model's success in reducing distributional gaps. Similarly, the variances for treated and control groups converge closely after adjustment, further demonstrating the model's balancing capability. The side-by-side comparison of before-and-after scenarios underscores the model's robustness in maintaining covariate balance, a crucial requirement for accurate causal inference.

6 | Discussion and Conclusion

Our SAHR framework represents a significant advancement in causal inference, demonstrating exceptional performance in estimating treatment effects from observational data. By integrating a self-supervised balancing mechanism and utilizing GMM for subgroup analysis, SAHR effectively reduces confounding bias and ensures robust covariate balance between treated and control groups. The use of kernel approximation techniques further enhances scalability, making SAHR suitable for high-dimensional datasets while maintaining computational efficiency. These features make SAHR a powerful tool for addressing the challenges of causal inference in real-world scenarios. One of SAHR's key strengths is its superior ability to maintain covariate balance. The framework achieves the lowest KS (p-value < 1%) (0.7% train, 1% test), Wasserstein Distance (0.033 train, 0.081 test), and MMD measures across all kernels (Linear, RBF, Poly), outperforming state-of-the-art methods like Dragonnet and BCAUSS. It demonstrates its effectiveness in

ensuring that treated and control groups are well-balanced, which is critical for accurate causal inference. Additionally, SAHR achieves competitive estimates of the treatment effect, with an in-sample error (ϵtr) of 0.036 ± 0.019 and an out-of-sample error (ϵte) of 0.047 ± 0.036 for the ATT on the Jobs dataset. These results are comparable to advanced methods like GANITE and BCAUSS, while outperforming traditional approaches such as k-NN, Random Forest, and Causal Forest.

Despite its strengths, SAHR has some limitations that need to be addressed. Currently, the framework is designed to handle only binary treatments, which restricts its applicability to problems involving multiple or continuous treatments. Extending SAHR to support these scenarios is an important direction for future work. Additionally, while SAHR performs exceptionally well on the Jobs dataset, further validation on a broader range of real-world datasets is necessary to assess its generalizability across different domains and data structures. The use of GMM for subgroup analysis, while effective, also introduces additional complexity in model training and hyperparameter tuning. Simplifying this process or developing automated tuning methods could improve usability. Finally, like most causal inference methods, SAHR relies on the assumption of unconfoundedness, and violations of this assumption can lead to biased estimates. Future work could explore robustness to unmeasured confounding.

Looking ahead, several promising directions exist for extending and improving SAHR. Expanding the framework to handle multiple treatments and continuous treatment variables would broaden its applicability to a wider range of causal inference problems. Developing methods to assess and mitigate unmeasured confounding would enhance the robustness of SAHR in real-world settings. Incorporating automated hyperparameter tuning for the GMM and kernel approximation components would make SAHR more user-friendly and accessible to non-experts. Benchmarking SAHR on a diverse set of real-world datasets, including those with complex data structures and varying levels of confounding, would provide deeper insights into its generalizability and performance. Finally, integrating SAHR's balancing mechanism with deep learning architectures could further enhance its ability to model complex, high-dimensional relationships in observational data.

The SAHR framework provides a robust and flexible approach to causal inference, combining the interpretability of traditional methods with the scalability and performance of modern machine learning techniques. Its state-of-the-art performance on the Jobs dataset, particularly in terms of covariate balance and treatment effect estimation, demonstrates its potential to drive meaningful progress in the field. While there are limitations to address, the modular design and computational efficiency of SAHR provide a strong foundation for future advancements. By addressing these challenges and expanding its capabilities, SAHR could become a widely adopted tool for causal inference across research and practice.

Acknowledgments

The authors sincerely appreciate all those who contributed to the completion of this research.

Funding

This research was conducted without financial support from any public, private, or non-profit organization.

Data Availability

The data used in this study are available from the corresponding author upon reasonable request.

References

[1] Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

[2] Austin, P. C. (2011). An Introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, *46*(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

[3] Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). Bart: Bayesian additive regression trees. *Annals of applied statistics*, *4*(1), 266–298. https://doi.org/10.1214/09-AOAS285

[4] Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, *61*(4), 962–973. https://doi.org/10.1111/j.1541-0420.2005.00377.x

[5] Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The journal of machine learning research*, *13*, 723–773. https://dl.acm.org/doi/10.5555/2188385.2188410

[6] Reynolds, D. (2009). Gaussian mixture models. In *Encyclopedia of biometrics* (pp. 659–663). Boston, MA: Springer US. https://doi.org/10.1007/978-0-387-73003-5_196

[7] Williams, C. K. I., & Seeger, M. (2000). Using the nyström method to speed up kernel machines. *Proceedings of the 14th international conference on neural information processing systems* (pp. 661–667). Cambridge, MA, USA: MIT Press. https://doi.org/10.5555/3008751.3008847

[8] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

[9] Tesei, G., Giampanis, S., Shi, J., & Norgeot, B. (2023). Learning end-to-end patient representations through self-supervised covariate balancing for causal treatment effect estimation. *Journal of biomedical informatics*, *140*, 104339. https://doi.org/10.1016/j.jbi.2023.104339

[10] Shalit, U., Johansson, F. D., & Sontag, D. (2017). Estimating individual treatment effect: Generalization bounds and algorithms. *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 3076–3085). PMLR. https://doi.org/10.5555/3305890.3305999

[11] Villani, C. 2009). *Optimal transport: Old and new* (Vol. 338). Springer. https://doi.org/10.1007/978-3-540-71050-9

[12] Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the royal statistical society series b: Statistical methodology*, *76*(1), 243–263. https://doi.org/10.1111/rssb.12027

[13] Robins, J., Sued, M., Lei-Gomez, Q., & Rotnitzky, A. (2007). Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical science*, *22*(4), 544–559. http://doi.org/10.1214/07-STS227D

[14] Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American statistical association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

[15] Yoon, J., Jordon, J., & Van Der Schaar, M. (2018). GANITE: Estimation of individualized treatment effects using generative adversarial nets. *International conference on learning representations*. ICLR 2018. (pp. 3076–3085). https://dblp.org/rec/conf/iclr/YoonJS18.html

[16] Su, X., Tsai, C. L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup Analysis via Recursive Partitioning. *The journal of machine learning research*, *10*, 141–158. https://doi.org/10.5555/1577069.1577074

[17] LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, *76*(4), 604–620. http://www.jstor.org/stable/1806062

[18] Schochet, P. Z. (2010). Is regression adjustment supported by the Neyman model for causal inference? *Journal of statistical planning and inference*, *140*(1), 246–259. https://doi.org/10.1016/j.jspi.2009.07.008

[19] Belthangady, C., Stedden, W., & Norgeot, B. (2021). Minimizing bias in massive multi-arm observational studies with BCAUS: Balancing covariates automatically using supervision. *BMC medical research methodology*, *21*(1), 190. https://doi.org/10.1186/s12874-021-01383-x

[20] Pearl, J. (2003). Causality: Models, reasoning, and inference. *Econometric theory*, *19*(675–685), 46. https://doi.org/10+10170S0266466603004109

[21] Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

[22] Heckman, J. J., & Vytlacil, E. J. (2007). Chapter 70 econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation. In *Handbook of econometrics* (Vol. 6, pp. 4779–4874). Elsevier. https://doi.org/10.1016/S1573-4412(07)06070-9

[23] Rahimi, A., & Recht, B. (2007). Random features for large-scale kernel machines. *Advances in neural information processing systems* (Vol. 20, pp. 1–8). EECS at UC Berkeley. https://doi.org/10.5555/2981562.2981710

[24] Johansson, F., Shalit, U., & Sontag, D. (2016). Learning representations for counterfactual inference. *Proceedings of the 33rd international conference on machine learning* (Vol. 48, pp. 3020–3029). New York, New York, USA: PMLR. https://doi.org/10.5555/3045390.3045708

[25] Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *The review of economics and statistics*, *90*(3), 389–405. https://doi.org/10.1162/rest.90.3.389

[26] Khoshandam, L., & Nematizadeh, M. (2024). An inverse network DEA model for two-stage processes in the presence of undesirable factors. *Journal of applied research on industrial engineering*, *11*(2), 179–194. https://doi.org/10.22105/jarie.2023.351161.1492

[27] Jr., F. J. M. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the american statistical association*, *46*(253), 68–78. https://doi.org/10.1080/01621459.1951.10500769

[28] Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Advances in applied probability*, *29*(2), 429–443. https://doi.org/https://doi.org/10.2307/1428011

Appendix

Optimization for subgroup-specific holistic ridge parameters

Solving the optimization problem for each subgroup k that $\frac{\partial L_k(\beta_k)}{\partial \beta_k}$ is needed to set to zero as follows:

$$L_k(\beta_k) = \|y_k - \phi(X)_k\beta_k\|_2^2 + \alpha_1\|\beta_k\|_2^2 + \alpha_2\|\beta_k - \theta\|_2^2. \quad (5)$$

$$L_k(\beta_k) = (y_k - \phi(X)_k\beta_k)^T\left((y_k - \phi(X)_k\beta_k)^T\right) + \alpha_1\beta_k^T\beta_k + \alpha_2(\beta_k - \theta)^T(\beta_k - \theta). \quad (6)$$

$$\nabla_{\beta_k} L_k(\beta_k) = -2\phi(X)_k^T(y_k - \phi(X)_k\beta_k) + 2\alpha_1\beta_k + 2\alpha_2(\beta_k - \theta). \quad (7)$$

$$-2\phi(X)_k^T(y_k - \phi(X)_k\beta_k) + 2\alpha_1\beta_k + 2\alpha_2(\beta_k - \theta) = 0. \quad (8)$$

$$\phi(X)_k^T\phi(X)_k\beta_k + \alpha_1\beta_k + \alpha_2\beta_k = \phi(X)_k^T y_k + \alpha_2\theta. \quad (9)$$

$$\left(\phi(X)_k^T\phi(X)_k\beta_k + \alpha_1 I + \alpha_2 I\right)\beta_k = \phi(X)_k^T y_k + \alpha_2\theta. \quad (10)$$

$$\beta_k = \left(\phi(X)_k^T\phi(X)_k\beta_k + \alpha_1 I + \alpha_2 I\right)^{-1}\left(\phi(X)_k^T y_k + \alpha_2\theta\right). \quad (11)$$

Distributional metrics glossary

MMD [5] is a distance metric that compares probability distributions by embedding them into a Reproducing Kernel Hilbert Space (RKHS). It is widely used in machine learning and causal inference to measure differences between distributions, particularly for assessing covariate balance between treated and control groups.

Kolmogorov–Smirnov [27] statistic is a nonparametric test that compares two probability distributions by calculating the maximum difference between their Cumulative Distribution Functions (CDFs). It is commonly used in causal inference to evaluate the balance of covariates between treated and control groups.

Wasserstein distance [11], also known as the Earth Mover's Distance, quantifies the cost of transforming one probability distribution into another. It is particularly effective for comparing distributions with non-overlapping support and is widely applied in optimal transport and causal inference to assess distributional differences.

Integral probability metrics [28] are a family of distance metrics used to compare probability distributions, including the MMD and the Wasserstein distance. They are widely used in causal inference to assess the balance between the treated and control groups, ensuring the validity of causal estimates.

Methodological foundations glossary

GMM [6] is a probabilistic models that represent data as a mixture of multiple Gaussian distributions. They are widely used for clustering and subgroup analysis, making them ideal for identifying latent subgroups in causal inference.

Ridge regression [8] is a regularized linear regression technique that penalizes the coefficients with an L2 penalty to prevent overfitting. It is particularly useful in high-dimensional settings and serves as a key component of the SAHR model for adaptive covariate balancing.

The Nyström method [7] is a kernel approximation technique that reduces the computational complexity of kernel-based methods by selecting a subset of landmark points. It is used in the SAHR model to handle non-linear relationships in high-dimensional data efficiently.

Kernel approximation [23] refers to techniques that reduce the computational complexity of kernel functions. The Nyström Method is a popular approach in the SAHR model for efficiently approximating kernels in high-dimensional settings.