The Annals of Applied Statistics
2026, Vol. 20, No. 1, 408–433
https://doi.org/10.1214/25-AOAS2129

ASYMPTOTICALLY EFFICIENT DATA-ADAPTIVE PENALIZED SHRINKAGE ESTIMATION WITH APPLICATION TO CAUSAL INFERENCE

BY HERBERT P. SUSMANN[1,a], YITING LI[2,d], MARA A. MCADAMS-DEMARCO[2,e], WENBO WU[1,b] AND IVÁN DÍAZ[1,c]

[1]*Division of Biostatistics, Department of Population Health, NYU Grossman School of Medicine,* [a]*susmah01@nyu.edu,* [b]*wenbowu@jhu.edu,* [c]*ivan.diaz@nyu.edu*

[2]*Department of Surgery, NYU Grossman School of Medicine,* [d]*yiting.li@nyulangone.org,* [e]*mara.mcadamsdemarco@nyulangone.org*

A rich literature exists on constructing nonparametric estimators with optimal asymptotic properties. In addition to asymptotic guarantees, it is often of interest to design estimators with desirable finite-sample properties, such as reduced mean-squared error of a large set of parameters. We provide examples drawn from causal inference where this may be the case, such as estimating a large number of group-specific treatment effects. We show how finite-sample properties of nonparametric estimators, particularly their variance, can be improved by careful application of *penalization*. Given a target parameter of interest, we derive a novel penalized parameter defined as the solution to an optimization problem that balances fidelity to the original parameter against a penalty term. By deriving the nonparametric efficiency bound for the penalized parameter, we are able to propose simple data-adaptive choices for the L_1 and L_2 tuning parameters designed to minimize finite-sample mean-squared error while preserving optimal asymptotic properties. The L_1 and L_2 penalization amounts to an adjustment that can be performed as a postprocessing step applied to any asymptotically normal and efficient estimator. We show in extensive simulations that this adjustment yields estimators with lower MSE than the unpenalized estimators. Finally, we apply our approach to estimate provider quality measures of kidney dialysis providers within a causal inference framework.

1. Introduction. In many settings it is of interest to define and estimate a large set of related statistical parameters. This is often the case in causal inference, where one may wish to estimate a large set of related treatment effects. For example, in studies of an intervention applied in multiple sites, one may wish to estimate both the average effect of the intervention marginally across all sites as well as the average effect within each site; here, there are as many statistical parameters as there are sites. When there are many sites, estimating the site-specific effects may be challenging; this is especially true when there are sites with few data. Another salient example arises in healthcare provider profiling applications in which many healthcare providers are evaluated based on their patient outcomes. A more general example is determining the importance of a large number of variables in a prediction model, which may involve estimating a large number of variable importance measures (Williamson et al. (2021)).

When estimating a set of statistical parameters in real-world scenarios, there is not typically sufficient mechanistic knowledge to justify the use of parametric models. Nonparametric, data-adaptive approaches are instead warranted. For example, the relationship between patient health outcomes, patient characteristics, and healthcare provider characteristics is highly complex, and cannot be accurately described by a simple (e.g., linear) relationship

Received September 2025.
Key words and phrases. Causal inference, doubly robust estimation, penalization, shrinkage estimator.

between variables. In order to avoid such strong assumptions, we prefer to work within a nonparametric framework in which we seek to estimate low-dimensional statistical summaries, such as a set of treatment effects, of an infinite-dimensional nuisance parameter, such as the set of all probability laws defined on the support of the data.

We guide the development of our estimators using semiparametric efficiency theory, which characterizes lower bounds on the asymptotic performance of nonparametric estimators. Based on foundational work by Hájek (1969/70, 1972) and Le Cam (1972) and further developed by Pfanzagl and Wefelmeyer (1985), van der Vaart (1992), Bickel et al. (1998), among others (see van der Vaart (1998), Chapter 25) for an overview), this theory extends classical efficiency results for finite-dimensional parameters of smooth parametric models to the functionals of nonparametric, infinite-dimensional nuisance parameters. A key result is the convolution theorem, which establishes that the optimal limiting distribution for regular nonparametric estimators is Gaussian with covariance determined by the *efficient influence function* (EIF) of the functional. The EIF plays a similar role as the Fisher information for parametric models, which characterizes the parametric efficiency bound through the Cramér–Rao theorem. Thus, characterizing the form of the EIF for a statistical functional is a key task, as it characterizes the efficiency bound for estimating the functional in a nonparametric model.

Remarkably, several nonparametric estimation strategies have been developed for constructing nonparametric estimators that achieve the semiparametric efficiency bound; these include one-step estimation, targeted maximum likelihood estimation, and estimating equations, among others (Pfanzagl and Wefelmeyer (1985), Bickel et al. (1998), Tsiatis (2006), van der Laan and Rubin (2006); see Kennedy (2024) for an accessible review). These estimators are typically built using the form of the EIF for the target statistical functional. Thus, deriving the EIF is useful for another reason: it both characterizes the efficiency bound and provides a path toward constructing estimators that achieve this bound.

Semiparametric efficiency theory, including the convolution theorem, provides an asymptotic theory of optimality for nonparametric estimators. However, we may wish to design estimators with additional finite-sample properties. For example, it may be desirable to find an estimator for a set of parameters for which each individual estimator may be *biased* in finite samples, yet the *mean-squared error* defined jointly over the set of parameters is lower. A related goal may be to find an estimator that has lower joint finite-sample mean-squared error and simultaneously summarizes the parameters in a useful way, for example, by introducing *sparsity*. That is, it is often desirable to have estimates that are not "meaningfully far from zero" shrunk identically to zero (where what it means to be "meaningfully far from zero" requires careful formalization). Ideally, an estimator would have these finite-sample properties while still achieving the asymptotic optimality given by the convolution theorem in which the limiting distribution is gaussian with variance given by the variance of the EIF.

In this paper we investigate how penalization can be used to construct alternative estimators with useful finite-sample properties, such as improved finite-sample variance and sparsity, while nonetheless having optimal asymptotic properties. First, we propose a general theoretical framework for defining penalized parameters. Our framework defines a penalized parameter as the solution to an optimization problem that balances fidelity to the original parameter (as measured via an arbitrary loss function) and an arbitrary penalization term. Our framework, therefore, encompasses penalized parameters defined using squared-error loss functions and L_2 and L_1 penalties, aping Ridge and Lasso regression, respectively. In practice, we allow the degree of penalization to depend on the sample size, with the goal that as sample size goes to infinity the penalized parameter converges to the original parameter. The penalized estimator, therefore, inherits the favorable asymptotic properties of the original estimator. We provide three examples to illustrate our proposals. First, we examine

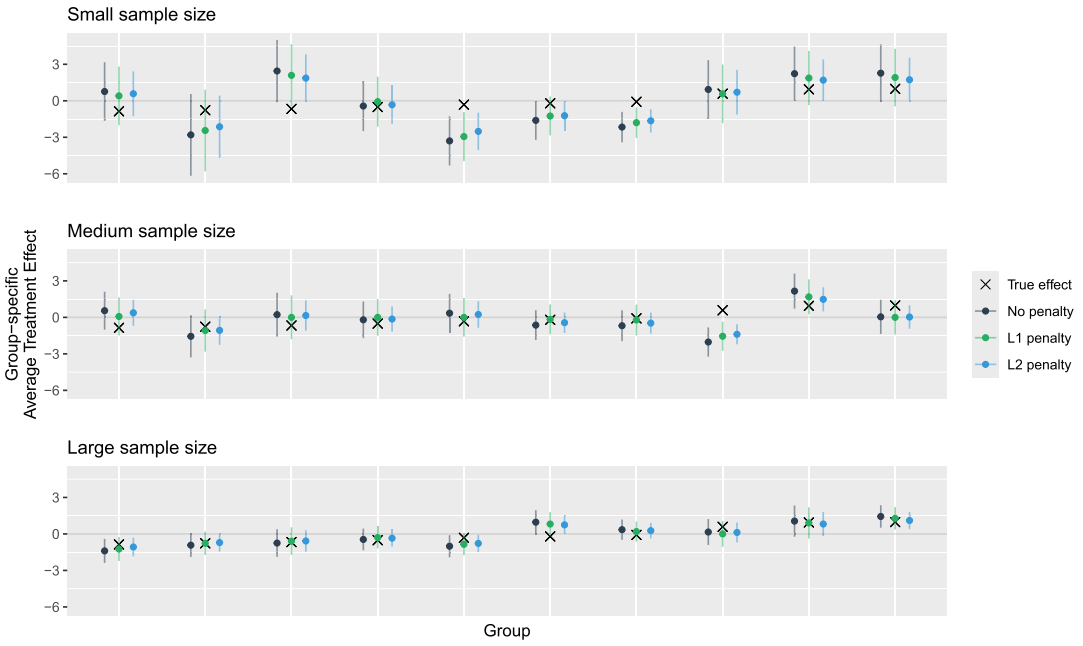

FIG. 1. *Example based on simulated data illustrating our penalized estimators applied to estimating group-specific average treatment effects. The true group-specific average treatment effects in each group are shown by the crosses. The estimated effects (point estimates and 95% confidence intervals) for each group are shown based on a double-robust targeted estimator (first point in each group), an L_1 penalized estimator (second point), and an L_2 penalized estimator (third point) for small, medium, and large sample sizes.*

a nonparametric linear association parameter with which we directly compare our approach to traditional penalized regression methods. Second, we use as further examples two causal inference parameters: group-specific average treatment effects and indirectly standardized outcomes.

Next, we apply tools from semiparametric efficiency theory to derive a general form for the efficient influence function (EIF) of the penalized parameter. The EIF characterizes the efficiency bound of semiparametric estimators of the penalized parameters. Knowledge of the efficiency bound allows us to derive data-adaptive choices of the penalization tuning parameters in the L_2 and L_1 cases. Under these data-adaptive choices, for which the degree of penalization depends on the sample size, we show that, as sample size increases, the EIF of the penalized parameter converges to the EIF of the original parameter. Thus, asymptotically our estimator recover the same limiting properties of nonpenalized nonparametric estimator. Furthermore, the asymptotic results lead to construction of asymptotically valid statistical inference on the original target parameter of interest, including the construction of confidence intervals. As such, our method amounts to a finite-sample correction of the point estimate designed to yield lower variance at the cost of introducing (finite-sample) bias. Practically speaking, we show that this penalization procedure can be applied as a postprocessing step to the estimates yielded by any asymptotically normal and efficient estimator of the target parameter. This makes our methods easily applicable to the outputs provided by standard statistical software.

Our approach is illustrated using simulated data in Figure 1. Data are simulated for a trial of an intervention applied in multiple groups; for example, it could be a treatment intervention in multiple hospitals. The simulation is designed such that the true treatment effect for each group is uniformly distributed between -1 and 1. For three simulated datasets with increasing sample sizes, we first applied a doubly robust targeted causal inference estimator separately

within each group to estimate the group-specific treatment effects (referred to as *no penalty* in the figure). We then applied our proposed methods to estimate L_1 and L_2-penalized group-specific treatment effects. At the smallest sample sizes, the penalized estimates are shrunk toward zero, which improves the quality of the estimates. The L_1-penalized estimates are in some cases shrunk exactly to zero. Due to the data-adaptive choice of the penalization parameter, as the sample size increases, the unpenalized and penalized point estimates (and confidence intervals) converge to each other; as such, the penalized estimates inherit the optimal asymptotic properties of the doubly robust unpenalized estimator. Simulations based on a similar data-generating process are investigated in more depth in Section 6.

Motivating application. While our approach is applicable to a general class of nonparametric parameters, the motivating application for this work is to estimate the performance of healthcare providers in terms of patient outcomes, a task referred to as *provider profiling* (Normand, Glickman and Gatsonis (1997)). A key challenge in provider profiling is to adjust for differing case-mix across providers, which is also known as risk adjustment (DeLong et al. (1997)). Indirect standardization is a standard adjustment method (Keiding and Clayton (2014)) and can be described succinctly when framed as a causal inference problem (Daignault and Saarela (2017)). In causal terms, indirect standardization compares the observed outcomes of patients from a provider to their counterfactual outcomes had they been randomly reassigned to another provider with a similar case-mix. Common methods for indirect standardization leverage generalized linear models with either fixed or random provider effects, with some debate over which to prefer (Kalbfleisch and Wolfe (2013), Yang et al. (2014), Kalbfleisch and He (2018)).

There are two areas of improvement for provider profiling estimators that we address in this work. First, the use of parametric models with linear functional form assumptions may result in biased or underpowered estimators when such assumptions do not hold (Varewyck et al. (2014), Susmann et al. (2025)), such as in scenarios where there are nonlinear relationships between covariates and patient outcomes. This concern motivates our adoption of a nonparametric estimation framework leveraging flexible machine-learning algorithms for nuisance estimation that do not require unrealistic functional form assumptions. Second, the presence of providers with few observations poses challenges for estimation (Brakenhoff et al. (2018)). In the generalized linear model framework, a way to stabilize estimation for small providers is through the use of random, rather than fixed, provider effects (MacKenzie et al. (2015)). The contribution of our work is to show how shrinkage estimation can be applied within a nonparametric framework to estimate provider profiling metrics and compare the results to standard generalized linear modeling based approaches via simulations.

Prior work. The utility of shrinkage estimators that trade bias for variance is well known through the famous example of the James–Stein estimator, which demonstrates that in a certain normal mean model an estimator that scales unbiased initial estimates toward zero dominates the unbiased estimator in terms of joint mean squared error (MSE; Stein (1956), Efron and Morris (1977)). Similarly, James–Stein inspired estimators have also been derived in other contexts, such as simultaneous equations and two-stage least squares (Maasoumi (1978), Hansen (2017)). The original James–Stein estimator can also be motivated by empirical Bayes arguments (Efron (2024)). Indeed, shrinkage estimators are a major topic in empirical Bayes methodology (Armstrong, Kolesár and Plagborg-Møller (2022)); we draw on such arguments to justify a simple modification of our L_2 penalization method to allow for adaptive shrinkage that depends on the precision of the individual parameter estimates. Our overall project is distinct from empirical Bayes methods, however, in that we define our parameters via penalization.

In another context, estimating regression coefficients with penalization in linear models was popularized by various regularized regression methods including the Lasso, Ridge, and

Elastic-Net, to name only several examples in a vast literature (Tibshirani (1996), Hoerl and Kennard (1970), Zou and Hastie (2005)). These regression penalization methods yield estimators that trade bias for reduced variance, with a focus on improving predictive performance. Depending on the penalization term, the estimates of the regression coefficients can also be sparse, as is the case for the Lasso (using L_1 penalization). A variety of methods have been proposed for choosing the key tuning parameter controlling the strength of penalization of the coefficients, including cross-validation and diverse in-sample estimators (see, e.g., Table 4 of Imdad Ullah, Aslam and Altaf (2018) which details 20 such options). For example, an early tuning-free method for ridge regression derives a simple form for the optimal degree of penalization by minimizing MSE (Hoerl, Kannard and Baldwin (1975)). Our work has a similar goal in that we seek to derive estimators of a penalization parameter that minimizes MSE, but we are concerned with the MSE of the estimator of a *vector-valued parameter* rather than that of *regression model predictions*. Our work also diverges from the penalized regression literature in that we make no modeling assumptions, working rather in a fully nonparametric framework. In addition, our focus is on inference for statistical functionals, rather than on predictions, and our asymptotic results lead to straightforward constructions of confidence intervals. On the other hand, statistical inference for penalized regression coefficients typically depends on postselection inference techniques (Lee et al. (2016)).

In applied Bayesian methodology, shrinkage of parameter estimates is ubiquitous through the application of priors. Bayesian shrinkage methods are appealing in that inference is available automatically via standard Bayesian arguments; for example, a common approach is to shrink parameter estimates in linear mixed models by placing hierarchical priors on the model coefficients. For treatment effect estimates in particular, Feller and Gelman (2015) advocate for shrinking multiple effect estimates (such as group-specific effects) toward a common mean and propose a parametric Bayesian modeling approach to that end. Our work has a similar goal, although we approach the problem in a frequentist nonparametric framework.

In the context of causal inference, penalization has been previously investigated for estimating nuisance parameters that are involved in forming the final estimates of the causal target parameters of interest (Smucler, Rotnitzky and Robins (2019), Shortreed and Ertefaie (2017), Benkeser and van der Laan (2016)). However, achieving a desired bias-variance trade-off for the nuisance parameters does not necessarily imply that the subsequent estimates of the causal effects will share the same desirable properties. For example, using sparse regression methods for the nuisance parameters will not necessarily imply that the causal effect estimates are sparse. Our work takes a different approach by defining a new target parameter that incorporates the penalization. Nuisance parameters can be estimated using diverse methods and are not limited to regularized regression methods, for example.

Our work can be seen as a specific form of nonparametric marginal structural model (MSM) proposed in the context of causal inference. Nonparametric MSMs summarize a possibly high-dimensional set of target parameters by projecting them onto a lower-dimensional working model. Such approaches have also been referred to as projection learners (McClean, Branson and Kennedy (2024)). Semiparametric theory for a general class of MSMs is reviewed in Susmann and Chambaz (2023). Also closely related to our work is that of Bahamyirou et al. (2022), who developed a penalized method for discovery of conditional average treatment effect (CATE) modifiers. The principal differences in our approaches lie in that ours is fully general and applicable to a large class of parameters beyond the CATE, and our results eschew any modeling assumptions, such as the linear marginal structural model, that their work imposes.

Outline. The rest of the manuscript is organized as follows. In Section 2 we introduce a general class of penalized parameters. In Section 3 we derive the semiparametric efficiency

properties of this general parameter class. In Sections 4 and 5, we apply the results to parameters defined with L_2 and L_1 penalties, respectively. Simulation studies are included in Sections 6 and 7, and an application to estimating the performance of kidney dialysis providers is presented in Section 8. We conclude in Section 9 with a discussion.

2. Framework for penalized parameters. Suppose we observe n i.i.d. draws $O_1, \ldots, O_n$ of the generic variable $O \in \mathcal{O}$ from a law P_0. We assume only that P_0 falls in the nonparametric model $\mathcal{M}$ (i.e., $\mathcal{M}$ is the set of all probability laws defined on the support of O). Let $\psi : \mathcal{M} \to \mathbb{R}^{|\mathcal{D}|}$ be a vector-valued parameter defined by $\psi(P) = (\psi_d(P) : d \in \mathcal{D})$, where $\psi_d : \mathcal{M} \to \mathbb{R}$ is a statistical functional indexed by $d \in \mathcal{D}$. We assume throughout that the ψ_d are sufficiently smooth so as to be *pathwise differentiable*, a concept introduced in the next section.

Notation. Whenever there is a set or vector $\mathcal{D}$, we will use the subscript "d" to denote the dth element of the set, as in ψ_d for the dth element of the vector $\psi(P)$. When we make a statement concerning "the ψ_d" we are applying the statement to all ψ_d for $d \in \mathcal{D}$. For convenience we will use the subscript "0" to denote a parameter evaluated at P_0, for example, $\psi_0 = \psi(P_0)$. We will also use the subscript "n" to signal dependence on n; for example, we will write ψ_n to denote an estimator of ψ_0. For a function f and $P \in \mathcal{M}$ we write the expectation of f with respect to P as either $\mathsf{E}_P[f]$ or $Pf = \int f \, dP$. We may write expectation with respect to the empirical measure as $P_n f = n^{-1} \sum_{i=1}^n f(O_i)$. A reference table listing key notation is provided in Section 1 of the Supplementary Material (Susmann et al. (2026)).

Now we introduce three examples of vector-valued statistical parameters. We will use these parameters later to evaluate our proposed methods in simulation studies.

EXAMPLE 1 (Nonparametric linear association). Let $O = (X, Y)$, where $X = (X_1, \ldots, X_D)$ is a D-dimensional vector of covariates and $Y \in \mathbb{R}$ is a continuous outcome. Denote by $X_{(-d)}$ the vector containing all but the dth element of X. For each $d \in \mathcal{D} = \{1, \ldots, D\}$, define

$$\psi_d(P) = \mathsf{E}_P\big[\mathsf{Cov}_P(Y, X_d | X_{(-d)})\big],$$

where $\mathsf{Cov}_P(Y, X_d \mid X_{-d})$ denotes the conditional covariance of Y and X_d given covariates X_{-d} under the distribution $P \in \mathcal{M}$. Collecting these into a vector yields the parameter $\psi(P) = \{\psi_d(P) : d \in \mathcal{D}\}$.

Note that this parameter has the useful property that it can be estimated using linear regression in that in a main-terms linear regression of Y on X the coefficient estimate $\hat{\beta}_d$ converges to $\psi_d(P)/\mathsf{E}_P[\mathsf{Var}_P(Y|X_d)]$. This property will allow us to compare our methods directly to penalized generalized linear models in simulations.

EXAMPLE 2 (Group-specific treatment effects). Let X be a vector of covariates, $G \in \{1, \ldots, D\}$ a variable indexing assignment to a group, and $A \in \{0, 1\}$ a binary treatment. Let $Y(0), Y(1) \in \mathbb{R}$ be potential outcomes corresponding to treatment assignments $A = 0$ and $A = 1$, respectively, and let $Y = AY(1) + (1 - A)Y(0)$ be the observed outcome. The observed data are, therefore, $O = (X, G, A, Y)$. The causal parameter of interest is the group-specific average treatment effect, denoted in terms of potential outcomes as, for $d \in \mathcal{D} = \{1, \ldots, D\}$,

$$\psi_d^*(P) = \mathsf{E}_P\big[Y(1) - Y(0) | G = d\big].$$

Let $\mu_P(a, d, X) = \mathsf{E}_P[Y \mid A = a, G = d, X]$. Then under standard causal assumptions (conditional ignorability and positivity), the parameter $\psi_d(P)$ is identified in terms of only observable variables as, for $d \in \mathcal{D}$,

$$\psi_d(P) = \mathsf{E}_P\big[\mu_P(1, d, X) - \mu_P(0, d, X) \mid G = d\big].$$

EXAMPLE 3 (Indirectly standardized outcomes). Let X be a vector of covariates and $A \in \mathcal{D} = \{1, \ldots, D\}$ a categorical treatment indicator. Let $\{Y(a) : a \in \mathcal{D}\}$ be a set of potential outcomes corresponding to each of the treatment assignments and $Y = Y(A)$ be the outcome under the observed treatment assignment. The observed data comprise $O = (X, A, Y)$.

Let $Z \sim P_{A|X}$ be a random draw from the conditional distribution of the treatment assignment, given covariates. Let $Y(Z)$ be the potential outcome under the stochastic intervention in which the individual was reassigned to treatment Z (which possibly differs from the observed treatment assignment A). The target causal parameter is defined as

$$\psi_d^*(P) = \mathsf{E}_P\big[Y(Z)|A = d\big].$$

That is, $\psi_d^*(P)$ is the expected outcome if, possibly contrary to fact, all observations from group d were randomly reassigned to an alternative treatment Z.

Let $\mu_P(X) = \mathsf{E}_P[Y \mid X]$. Then $\psi^*(P)$ is identified using only observable variables as

$$\psi_d(P) = \mathsf{E}_P\big[\mu(X)|A = d\big].$$

The parameter ψ_d is sometimes referred to as an *indirectly standardized outcome*. As discussed in the Introduction, one application of this parameter is in provider profiling, where the observations are patients with baseline characteristics X who were treated at healthcare provider A and experienced the outcome Y. One way of evaluating the performance of a provider is to ask what would have happened if the population of patients who were treated by that provider had instead been randomly reassigned for treatment to another provider that tends to treat a similar patient population. This counterfactual parameter can be estimated by comparing $\psi_d(P)$ to the mean outcome of patients treated at the provider, for example, through the difference $\psi_d(P) - \mathsf{E}_P[Y \mid A = d]$ (Daignault and Saarela (2017), Díaz (2024), Susmann et al. (2025)).

Penalized parameter. We now define a novel *penalized parameter* defined in terms of the original parameter ψ. For any $P \in \mathcal{M}$, define the penalized parameter $\tilde{\psi}_\lambda \in \mathbb{R}^{|\mathcal{D}|}$ as the solution to the following optimization problem:

$$\tilde{\psi}_\lambda(P) = \underset{\tilde{\psi} \in \mathbb{R}^{|\mathcal{D}|}}{\arg\min}\, U_\lambda\big(\psi(P), \tilde{\psi}\big), \tag{1}$$

where the optimization objective $U_\lambda : \mathbb{R}^{|\mathcal{D}|} \times \mathbb{R}^{|\mathcal{D}|} \to \mathbb{R}$ is the map

$$(x, \tilde{x}) \mapsto U_\lambda(x, \tilde{x}) = L(x, \tilde{x}) + V_\lambda(\tilde{x}).$$

The loss function $L : \mathbb{R}^{|\mathcal{D}|} \times \mathbb{R}^{|\mathcal{D}|} \to \mathbb{R}$ measures the fidelity of the penalized parameter to the original parameter, and $V_\lambda : \mathbb{R}^{|\mathcal{D}|} \to \mathbb{R}$ is a penalization term. The tuning parameter $\lambda \in \Lambda$ controls the strength of the penalization. Typically, $\Lambda = \mathbb{R}_{>0}$; that is, λ is a positive number, with higher values of λ implying stronger penalization. Further assumptions may be necessary to assure that the optimization problem in (1) has a unique solution and that subsequently $\tilde{\psi}(P)$ is well defined. We first consider the case where the penalization parameter λ is fixed and user-defined. After developing theory for the case of fixed λ, we apply the results to suggest optimal data-adaptive methods for choosing λ.

3. General results. In this section we review foundations from semiparametric efficiency theory, which we then apply to derive the semiparametric efficiency bound for estimating the penalized parameter $\tilde{\psi}_\lambda$ at the true data-generating distribution P_0 under sufficiently smooth choices of loss function and penalty term. A general estimator based on one-step estimation that achieves the efficiency bound is presented in Section 2 of the Supplementary Material. Accessible and high-quality reviews of the relevant semiparametric theory, with an emphasis on applications to causal inference, can be found in Kennedy (2016, 2024). Other key references include van der Vaart and Wellner (1996), van der Vaart (1998), Bickel et al. (1998).

3.1. *Semiparametric efficiency theory and one-step estimation.* For the purposes of introducing the principal concepts, consider a generic statistical functional $\phi : \mathcal{M} \to \mathbb{R}^p$ (for $p \geq 1$). We focus on functionals that are sufficiently smooth so as to be *pathwise differentiable*, as this is a crucial property that allows for the derivation of nonparametric efficiency bounds. To introduce pathwise differentiability, for every $P \in \mathcal{M}$ and $s \in L_0^2(P)$, s bounded and not identically zero, define a parametric submodel $\mathcal{P}_s = \{P_{s,\epsilon} : \epsilon \in \mathbb{R}^p, \|\epsilon\|_\infty < \|s\|_\infty^{-1}\} \subset \mathcal{M}$, where $dP_{s,\epsilon} = (1 + \epsilon^\top s)\, dP$. Note that $\mathcal{P}_s$ is a fluctuation of P in the direction s, in the sense that $P_{s,\epsilon} = P$ at $\epsilon = 0$ and the score of $P_{\epsilon,s}$ at $\epsilon = 0$ is s. We call ϕ pathwise differentiable at P if there exists a functional $D_\phi^*(P) : \mathcal{O} \to \mathbb{R}^p$ with mean zero and finite variance referred to as an *influence curve* such that, for every s, the following derivative exists and can be expressed as

$$\frac{\partial}{\partial \epsilon}\phi(P_{s,\epsilon})|_{\epsilon=0} = \mathsf{E}_P\big[s(O)D_\phi^*(P)(O)^\top\big].$$

Because every $s \in L_0^2(P)$ induces a fluctuation model $\mathcal{P}_s$, if the derivative exists, then $D_\phi^*(P)$ is unique and is referred to as the *efficient influence function* of ϕ at P. A key result of semiparametric efficiency theory is that the asymptotic covariance of any regular estimator of $\phi(P)$ is lower bounded by the variance of the efficient influence function,

$$\sigma_\phi^2(P) = \mathsf{E}_P\big[D_\phi^*(P)(O)D_\phi^*(P)(O)^\top\big].$$

When a parameter is pathwise differentiable, the influence curve serves as the first-order term of a type of distributional Taylor expansion of the parameter. Formally, for any $P_1, P_2 \in \mathcal{M}$, write

$$\phi(P_1) - \phi(P_2) = (P_1 - P_2)D_\phi(P_1) + R(P_1, P_2), \tag{2}$$

for an influence function $D_\phi(P_1) : \mathcal{O} \to \mathbb{R}^p$ of ϕ at P and second-order remainder term $R : \mathcal{M} \times \mathcal{M} \to \mathbb{R}^p$. The remainder term is called second order because R is is a function only of squares or products of differences in its arguments. This expansion is sometimes referred to as the *von Mises* expansion of the parameter (von Mises (1947)).

Our analyses of the semiparametric efficiency properties of the proposed penalized parameters, therefore, proceeds in two steps: first, we establish whether the parameter is pathwise differentiable, and, if so, derive the form of its efficient influence function and the associated second-order remainder term. By characterizing the form the EIF and the remainder term, we can propose estimators and, subsequently, establish conditions under which that estimator is consistent, efficient, and asymptotically normal.

In this work we focus on penalized parameters defined with respect to an underling parameter ψ that is pathwise differentiable and admits a von Mises expansion of the form (2). For the three example target parameters, we give below the form of their associated efficient influence functions and the remainder term of the von Mises expansion.

EXAMPLE 1 (Nonparametric regression coefficient, continued). Let $\pi_P(X_{(-d)}) = \mathsf{E}_P[X_d \mid X_{(-d)}]$ and $\mu_P(X_{(-d)}) = \mathsf{E}_P[Y \mid X_{(-d)}]$. The parameter ψ_d is pathwise differentiable with efficient influence function $D_{\psi_d}^*$ at P characterized by

$$D_{\psi_d}^*(P)(O) = \big\{X_d - \pi_P(X_{(-d)})\big\}\big\{Y - \mu(X_{(-d)})]\big\}.$$

Furthermore, ψ_d satisfies a von Mises expansion with remainder term R_d for any $P, P_0 \in \mathcal{M}$ characterized by

$$R_d(P_0, P) = \mathsf{E}_{P_0}\big[\big\{\pi_P(X_{(-d)}) - \pi_0(X_{(-d)})\big\}\big\{\mu_P(X_{(-d)}) - \mu_0(X_{(-d)})\big\}\big].$$

EXAMPLE 2 (Group-specific treatment effects, continued). Fix $d \in \mathcal{D}$. Let $\pi_P(d, a, X) = P(A = a \mid G = d, X)$ and $\mu_P(d, a, X) = \mathsf{E}_P[Y \mid A = a, G = d, X]$. The parameter ψ_d is pathwise differentiable with efficient influence function $D^*_{\psi_d}$ at any $P \in \mathcal{M}$ characterized by

$$D^*_{\psi_d}(P)(O) = \frac{\mathbb{I}[G = d]}{P(G = d)}\Big[\frac{2A - 1}{\pi_P(d, A, Y)}\big(Y - \mu_P(G, A, X)\big) + \mu(d, 1, X) - \mu(d, 0, X) - \psi_d(P)\Big].$$

The parameter ψ_d satisfies a von Mises expansion with remainder term R_d for any P, $P_0 \in \mathcal{M}$ characterized by

$$R_d(P_0, P) = \sum_{a \in \{0,1\}} \frac{2a - 1}{P(G = d)} \mathsf{E}_{P_0}\Big[\mathbb{I}[A = d]\Big\{\frac{1}{\pi_P(d, a, X)} - \frac{1}{\pi_0(d, a, X)}\Big\} \times \big\{\mu_0(d, a, X) - \mu_P(d, a, X)\big\}\pi_0(d, a, X)\Big].$$

EXAMPLE 3 (Indirectly standardized outcomes). Fix $d \in \mathcal{D}$. Let $\pi_P(a, X) = P(A = a \mid X)$ and $\mu_P(X) = \mathsf{E}_P[Y \mid X]$. The indirectly standardized outcome parameter ψ_d is pathwise differentiable (Susmann et al. (2025)) with efficient influence function $D^*_{\psi_d}$ at any $P \in \mathcal{M}$ characterized by

$$D^*_{\psi_d}(P)(O) = \frac{1}{P(A = d)}\big\{\pi_P(d, X)\big(Y - \mu_P(X)\big) + \mathbb{I}[A = d]\big(\mu_P(X) - \psi_d(P)\big)\big\}.$$

The parameter ψ_d satisfies a von Mises expansion with remainder term R for any P, $P_0 \in \mathcal{M}$ characterized by

$$R_d(P_0, P) = \mathsf{E}_{P_0}\Big[\frac{1}{P(A = d)}\big(\pi_P(d, X) - \pi_0(d, X)\big)\big(\mu_0(X) - \mu_P(X)\big)\Big].$$

3.2. *Pathwise differentiability of general penalized parameters.* In the following theorem, we provide conditions under which $\tilde{\psi}_\lambda$ is pathwise differentiable and provide the form of its EIF when the penalization tuning parameter λ is fixed. Theory for the fixed λ scenario is useful for two reasons. First, doing so leads to strategies for choosing λ data-adaptively. Second, as we show in the next section, when λ is itself estimated from the data and applied to form a penalized parameter $\tilde{\psi}_\lambda$, the uncertainty arising from estimating λ is asymptotically negligible; in other words, under mild conditions the estimated λ can be treated as fixed, and the results proved here for fixed λ can be applied.

The following theorem and its conditions are an adaption of Susmann and Chambaz ((2023), Theorem 1). The proof is a straightforward application of the proof of that theorem and is, therefore, omitted.

THEOREM 3.1 (Efficient influence function of $\tilde{\psi}_\lambda$ for fixed λ). *Fix $\lambda \in \Lambda$. Assumptions*:

1. *The parameter ψ is pathwise differentiable at any $P \in \mathcal{M}$ with EIF $D^*_\psi(P) : \mathcal{O} \to \mathbb{R}^{|\mathcal{D}|}$.*
2. *For every $x \in \mathbb{R}^{|\mathcal{D}|}$, the following conditions are met*:
   (a) *$\tilde{x} \mapsto U_\lambda(x, \tilde{x})$ is differentiable at every $\tilde{x}$ with derivative $\dot{U}_\lambda(x, \tilde{x}) \in \mathbb{R}^{|\mathcal{D}|}$.*
   (b) *$\tilde{x} \mapsto \dot{U}_\lambda(x, \tilde{x})$ is differentiable at every $\tilde{x}$ with derivative $\ddot{U}_\lambda(x, \tilde{x}) \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}|}$.*

   In addition, for every $\tilde{x} \in \mathbb{R}^{|\mathcal{D}|}$, it holds that:
   (a) *$x \mapsto \dot{U}_\lambda(x, \tilde{x})$ is differentiable at every $x \in \mathbb{R}^{|\mathcal{D}|}$ with derivative $\nabla \dot{U}_\lambda(x, \tilde{x}) \in \mathbb{R}^{|\mathcal{D}|}$, and $\nabla \dot{U}_\lambda(x, \tilde{x})$ is invertible.*

Then the functional $P \mapsto \tilde{\psi}_\lambda(P)$ is pathwise differentiable at every $P \in \mathcal{M}$ with an efficient influence function $D^*_{\tilde{\psi}_\lambda}(P)$ at P given by

$$O \mapsto D^*_{\tilde{\psi}_\lambda}(P)(O) = M^{-1}\big[\nabla \dot{U}_\lambda\big(\psi(P), \tilde{\psi}(P)\big) \times D^*_\psi(P)(O) + \dot{U}_\lambda\big(\psi(P), \tilde{\psi}(P)\big)\big],$$

where the normalizing matrix M is given by

$$M = -\ddot{U}_\lambda\big(\psi(P), \tilde{\psi}(P)\big).$$

The required pathwise differentiability of ψ (Assumption 1) must be verified separately for the specific choice of underlying parameter, as we have done for the three examples. Assumption 2, requiring that various derivatives of objective function exist, must be verified for each choice of loss function and penalty term.

In Section 2 of the Supplementary Material, we describe a one-step estimator for $\tilde{\psi}_\lambda$ when λ is fixed that, under mild conditions, is consistent, asymptotically normal and achieves the nonparametric efficiency bound implied by the form of the EIF given in Theorem 3.1. In the following we focus on the scenario in which the tuning parameter is chosen data-adaptively.

4. L_2 penalty. In many real-world scenarios, we wish to choose the penalization tuning parameter data-adaptively in order to yield an estimator with desirable finite-sample properties. In this section we consider the choice of tuning parameter when using the L_2-norm penalty. We start with the L_2-norm because its infinite differentiability leads to particularly tidy results. Throughout, we use a squared-error loss function $L(x, \tilde{x}) = \|x - \tilde{x}\|_2^2$. For the penalty term, let $V_2(\tilde{x}) = \lambda\|\tilde{x}\|_2^2$. Begin by fixing a $\lambda \geq 0$. The objective function is then

$$U_\lambda(x, \tilde{x}) = \|x - \tilde{x}\|_2^2 + \lambda\|\tilde{x}\|_2^2,$$

and the optimization problem (1) has the solution, for any $P \in \mathcal{M}$,

$$\tilde{\psi}_\lambda(P) = \frac{1}{1+\lambda}\psi(P).$$

Applying Theorem 3.1 (Assumption 2 thereof easily verified due to the infinite differentiability of the L_2-norm) shows that the EIF of $\tilde{\psi}_\lambda$ is simply the scaled EIF of ψ,

$$D^*_{\tilde{\psi}_\lambda}(P)(O) = \frac{1}{1+\lambda} D^*_\psi(P)(O). \tag{3}$$

Indeed, the machinery of Theorem 3.1 isn't necessary to derive the above EIF, as it follows straightforwardly from the fact that $\tilde{\psi}_\lambda$ is simply a scaled version of λ.

In practice, we often do not have a value of λ fixed a priori; rather, we wish to choose λ data-adaptively. We propose choosing λ by minimizing the following criterion as a function of λ, which we denote Crit:

$$\mathsf{Crit}\big(\lambda, \psi(P), \sigma^2_\psi(P), n\big) = \frac{\lambda^2}{(1+\lambda)^2}\big\|\psi(P)\big\|_2^2 + \frac{1}{n(1+\lambda)^2}\mathrm{tr}\big(\sigma^2_\psi(P)\big),$$

where $n \geq 0$. The data-adaptive choice of λ is then given by

$$\lambda^* = \arg\min_{\lambda \geq 0} \mathsf{Crit}\big(\lambda, \psi(P), \sigma^2_\psi(P), n\big).$$

We argue that this is a useful way to choose λ because the criterion can be understood as an asymptotically valid approximation of the mean-squared error of an estimator of $\tilde{\psi}_\lambda$ relative to the true parameter value ψ. To illustrate this, for any $P \in \mathcal{M}$, define for the MSE of an estimator $\tilde{\psi}_{\lambda,n}$ of $\tilde{\psi}_\lambda(P)$ relative to $\psi(P)$ as

$$\mathsf{MSE}\big(\tilde{\psi}_{\lambda,n}, \psi(P)\big) = \mathsf{Bias}\big(\tilde{\psi}_{\lambda,n}, \psi(P)\big)^2 + \mathsf{Variance}(\tilde{\psi}_{\lambda,n}), \tag{4}$$

where $\mathsf{Bias}(\tilde{\psi}_{\lambda,n}, \psi(P))^2 = \|\mathsf{E}_P[\tilde{\psi}_{\lambda,n}] - \psi(P)\|_2^2$, $\mathsf{Variance}(\tilde{\psi}_{\lambda,n}) = \mathrm{tr}(\mathsf{Var}[\tilde{\psi}_{\lambda,n}])$, and tr is the matrix trace operator. An asymptotically normal and efficient estimator $\tilde{\psi}_{\lambda,n}$ of $\tilde{\psi}(P)$ satisfies

$$\sqrt{n}\big(\tilde{\psi}_{\lambda,n} - \tilde{\psi}_\lambda(P)\big) \xrightarrow{d} N\big(0, \sigma^2_{\tilde{\psi}_\lambda}(P)\big).$$

Therefore, an asymptotically valid estimate of the variance of $\tilde{\psi}_{\lambda,n}$ is $\sigma^2_{\tilde{\psi}_\lambda}(P)/n$. Using this as an estimate of the variance yields a simple form for the MSE (4),

$$\frac{\lambda^2}{(1+\lambda)^2}\|\psi(P)\|_2^2 + \frac{1}{n(1+\lambda)^2}\mathrm{tr}\big(\sigma^2_{\tilde{\psi}_\lambda}(P)\big) = \mathsf{Crit}\big(\lambda, \psi(P), \sigma^2_\psi(P), n\big). \tag{5}$$

Therefore, minimizing Crit as a function of λ can be seen as minimizing an asymptotic approximation of the MSE of the penalized estimator relative to the true parameter. The major caveat with this choice is that it depends on an asymptotic approximation of the variance of the estimator.

Conveniently, there is a closed-form solution for the value of λ that minimizes Crit. To express the closed form solution succinctly, first define, for any $P \in \mathcal{M}$ such that $\|\psi(P)\|_2^2 > 0$, the parameter $\gamma : \mathcal{M} \to \mathbb{R}$ as

$$P \mapsto \gamma(P) = \frac{\mathrm{tr}(\sigma^2_\psi(P))}{\|\psi(P)\|_2^2}.$$

The parameter γ is interesting in its own right as a summary of the efficiency bound of ψ relative to the overall scale of ψ, and its squared root is often referred to as the coefficient of variation. In addition, it is useful because the value of λ that minimizes the MSE given in (5) is a simple function of $\gamma(P)$,

$$\lambda^*\big(\gamma(P), n\big) = \frac{1}{n} \times \gamma(P).$$

For intuition, $\lambda^*(\gamma(P), n)$ has a simple interpretation as the ratio of the sum of the (approximate) variance of the estimator of each parameter divided by sum of squares of the parameters. Thus, when the variance is low relative to the magnitude of the parameter, less shrinkage is applied and vice versa when the variance is high.

We continue by studying the semiparametric efficiency properties of the parameter γ. Because γ is a differentiable function of ψ and σ^2_ψ, it follows that it will be pathwise differentiable so long as the same holds for ψ and σ^2_ψ. The following theorem formalizes this result.

LEMMA 4.1 (Efficient influence function of γ). *For all $d = 1, \ldots, D$, assume that $\sigma^2_{\psi_d}$ is pathwise differentiable at any $P \in \mathcal{M}$ with EIF $D^*_{\sigma^2_{\psi_d}}(P) : \mathcal{O} \to \mathbb{R}$. Then the parameter γ is pathwise differentiable with EIF $D^*_\gamma(P) : \mathcal{O} \to \mathbb{R}$ at $P \in \mathcal{M}$ characterized by*

$$O \mapsto D^*_\gamma(P)(O) = -2 \times \frac{\mathrm{tr}(\sigma^2_\phi(P))}{\|\psi(P)\|_2^3} \sum_{d=1}^{D} D^*_{\psi,d}(P)(O) + \frac{\sum_{d=1}^{D} D^*_{\sigma^2_{\phi_d}}(P)(O)}{\|\psi(P)\|_2^2}.$$

We can now go one step further and derive the EIF of the penalized parameter $\tilde{\psi}_{\lambda^*}$, the penalized parameter where the optimizer $\lambda^*(\gamma(P), n)$ is chosen as the penalization parameter.

THEOREM 4.2 (Efficient influence function of $\tilde{\psi}_{\lambda^*}$). *Fix $n > 0$. For any $P \in \mathcal{M}$, set $\lambda^* = \frac{1}{n}\gamma(P)$. The parameter $\tilde{\psi}_{\lambda^*}$ is pathwise differentiable at P with EIF $D^*_{\tilde{\psi}_{\lambda^*}}(P) : \mathcal{O} \to \mathbb{R}^{|\mathcal{D}|}$ characterized by*

$$D^*_{\tilde{\psi}_{\lambda^*}}(P)(O) = \frac{1}{1+\lambda^*} D^*_{\psi}(P)(O) - \frac{1}{n} \times \frac{\psi(P)}{(1+\lambda^*)^2} D^*_{\gamma}(P)(O).$$

The first term of the EIF for $\tilde{\psi}_{\lambda^*}$ is simply the EIF of the original parameter scaled by λ^*; this term can be interpreted as representing uncertainty in estimating $\tilde{\psi}_{\lambda^*}$ when λ^* as fixed, as in (3). The second term represents uncertainty in estimating λ^*. Notably, this term is scaled by $1/n$. This suggests that the second term of the EIF will be negligible as n increases.

An estimator of $\tilde{\psi}_{\lambda^*}$ could be constructed using the full EIF of $\tilde{\psi}_{\lambda^*}$ given in Theorem 4.2 (e.g., using the one-step approach described in Section 2 of the Supplementary Material). However, doing so would require estimating D^*_{γ}, which may be difficult or involve estimating additional nuisance parameters beyond those required for estimating ψ, D^*_{ψ} and λ^*. Therefore, we propose forming a simpler estimator that disregards the D^*_{γ} term. We subsequently prove that ignoring this term is justified in an asymptotic analysis.

To form the estimator, suppose that we have an asymptotically normal and efficient estimator ψ_n of ψ_0 and a consistent estimator γ_n of γ_0. We propose setting the penalty term to $\lambda^*_n = \frac{1}{n}\gamma_n$ and estimating $\tilde{\psi}_{\lambda^*_n}$ by simply scaling ψ_n by the estimated shrinkage factor,

$$\tilde{\psi}_{\lambda^*_n, n} = \frac{1}{1+\lambda^*_n}\psi_n.$$

To justify this simplified estimator, we prove the following alternative decomposition of the penalized parameter that shows, if the original parameter admits a von Mises expansion, then the penalized parameter satisfies a similar expansion that differs only by terms related to λ^*. The proof is provided in Section 3.1 of the Supplementary Material.

THEOREM 4.3. *Suppose that ψ satisfies a von Mises expansion of the form (2) with EIF D^*_{ψ} and second-order remainder R. Fix $n > 0$, and let $\lambda^* = \frac{1}{n}\gamma(P)$. Let $\tilde{\psi}_{\lambda^*} = \frac{1}{1+\lambda^*}\psi(P)$, and assume that $\tilde{\psi}_{\lambda}$ satisfies a von Mises expansion with EIF $D^*_{\tilde{\psi}}$ and second-order remainder $R_{\tilde{\psi}}$. Then the parameter $\tilde{\psi}_{\lambda^*}$ satisfies the following expansion*:

$$\begin{aligned}\tilde{\psi}_{\lambda^*}(P_1) - \tilde{\psi}_{\lambda^*}(P_2) = -P_2\left[\frac{1}{1+\lambda^*(P_1)} D^*_{\psi}(P_1)\right] + \left\{\frac{1}{1+\lambda^*(P_1)} - \frac{1}{1+\lambda^*(P_2)}\right\}\psi(P_2) \\ + \frac{1}{1+\lambda^*(P_1)} R(P_1, P_2).\end{aligned}$$

This result is notable because, as $n \to \infty$, the decomposition converges to

$$\tilde{\psi}_{\lambda^*}(P_1) - \tilde{\psi}_{\lambda^*}(P_2) = -P_2\big[D^*_{\psi}(P_1)\big] + R(P_1, P_2).$$

The proof is given in Section 3.1 of the Supplementary Material. Asymptotic consistency, normality and efficiency, therefore, follows for $\tilde{\psi}_{\lambda^*_n}$ under the same conditions necessary for the original parameter ψ, with the only other condition necessary being that an estimator γ_n of γ_0 does not diverge This is formalized in the following theorem, which establishes conditions under which $\tilde{\psi}_{\lambda^*}$ is asymptotically normal and efficient estimator of ψ_0.

THEOREM 4.4 (Asymptotic normality and efficiency of $\tilde{\psi}_{\lambda^*_n}$ for L_2-penalization). *Let ψ_n and γ_n be estimators of ψ_0 and γ_n, respectively. Let $\lambda^*_n = \frac{1}{n} \times \gamma_n$. Assume each of the following*:

1. *The estimator ψ_n is asymptotically normal and efficient*,

$$\sqrt{n}(\psi_n - \psi_0) \xrightarrow{d} N(0, \sigma^2_{\psi,0}).$$

2. *The estimator γ_n converges*: *there exists a γ_∞ with $\infty < \gamma_\infty < \infty$ such that $\gamma_n - \gamma_\infty = o_P(1)$.*

Then $\tilde{\psi}_{\lambda^*_n,n} = \frac{1}{1+\lambda^*_n}\psi_n$ is an asymptotically normal and efficient estimator of ψ_0,

$$\sqrt{n}(\tilde{\psi}_{\lambda^*_n} - \psi_0) \xrightarrow{d} N(0, \sigma^2_{\psi,0}).$$

PROOF. By assumption, $\gamma_n - \gamma_\infty = o_P(1)$. Therefore, $\lambda^*_n = o_P(1)$, and furthermore, the shrinkage factor $1/(1+\lambda^*_n) = 1 + o_P(1)$. Thus, Slutsky's theorem and the fact that the estimator of ψ is asymptotically normal and efficient implies the stated result. □

Establishing conditions under which Assumption 1 holds depends on the underlying parameter of interest. Typically, convergence of γ_n, required by Assumption 2, will hold under weak assumptions; indeed, γ_n will typically be a consistent estimator of γ_0 under the same assumptions necessary for Assumption 1. In the interest of generality, Theorem 4.4 is stated in terms of a generic asymptotically efficient estimator ψ_n of ψ_0. Alternatively, one could use the expansion in Theorem 4.3 to construct an estimator of $\tilde{\psi}_0$, for example, by using a one-step estimation strategy.

Based on the asymptotic normality result of Theorem 4.4, a straightforward and asymptotically valid $(1-\alpha)\times 100\%$ confidence interval for ψ can be formed using the estimated variance of the unpenalized parameter estimate,

$$C_{1-\alpha}(\tilde{\psi}_{\lambda^*_n}) = \left(\tilde{\psi}_{\lambda^*_n} - q_{1-\alpha}\sqrt{\frac{\sigma^2_{d,n}}{n}}, \tilde{\psi}_{\lambda^*_n} + q_{1-\alpha}\sqrt{\frac{\sigma^2_{d,n}}{n}}\right), \tag{6}$$

where $\sigma^2_{d,n}$ is an estimate of the efficiency bound of ψ_d. Assuming that we have access to an asymptotically normal and efficient estimator of ψ_d, then such an estimate of the efficiency bound is typically available through the estimator's reported standard error. This is similar to the recent proposal in Kaplan and Liu (2024) for forming confidence intervals of biased parameters that are centered on the biased parameter estimate but use the standard error of the original (unbiased) estimator to determine the confidence interval width.

The above confidence interval is asymptotically valid, but not entirely satisfying, as it has the same width as a confidence interval for the unpenalized parameter. As an alternative, we can form a narrower confidence interval based on the estimated shrinkage factor,

$$C'_{1-\alpha} = \left(\psi_n - \frac{q_{1-\alpha}}{1+\lambda^*_n}\sqrt{\frac{\sigma^2_{d,n}}{n}}, \psi_n + \frac{q_{1-\alpha}}{1+\lambda^*_n}\sqrt{\frac{\sigma^2_{d,n}}{n}}\right).$$

The asymptotic validity of the confidence interval follows from the same logic as the proof of Theorem 4.4.

In some applications the fact that the penalized estimator $\tilde{\psi}_{\lambda^*_n}$ shrinks all estimates by the same factor $1/(1+\lambda^*_n)$ may not be desirable. Instead, we may wish to shrink each estimate in a manner proportional to the precision of the estimate. To propose such an estimator, note that we can rewrite the penalized parameter in the following form:

$$\begin{aligned}\tilde{\psi}_{\lambda^*_n}(P) &= \frac{1}{1+\lambda^*_n(P)}\psi(P) \\ &= \frac{\frac{1}{D}\|\psi(P)\|_2^2}{\frac{1}{D}\|\psi(P)\|_2^2 + \frac{1}{D}\sum_{d'=1}^{D}\frac{1}{n}P[D^*_{\psi,d'}(P)^2]}\psi(P).\end{aligned}$$

In this form the shrinkage is recognizable as the ratio involving the variance of the original parameter ψ around zero and the mean of the approximate estimator variances. This form also suggests a simple modification to allow for variable shrinkage. For a parameter ψ_d ($d \in \mathcal{D}$), estimate the shrinkage using the approximate estimator variance of only ψ_d,

$$\tilde{\psi}_d^{\mathsf{eb}}(P) = \frac{\frac{1}{D}\|\psi(P)\|_2^2}{\frac{1}{D}\|\psi(P)\|_2^2 + \frac{1}{n}P[D_{\psi,d}^*(P)^2]}\psi(P).$$

This estimator has a natural connection to empirical Bayes, as it can be interpreted as the posterior mean of ψ_d under a normal observation model with $\psi_{d,n} \sim N(\psi_d, P[D_{\psi,d}^*(P)]^2)$ and prior $\theta_d \sim N(0, \tau^2)$. In practice, given an asymptotically normal and efficient estimator ψ_n of ψ_0 with estimated standard errors σ_n^2, we form the empirical Bayes estimator

$$\tilde{\psi}_{d,n}^{\mathsf{eb}} = \frac{\frac{1}{D-1}\|\psi_n\|_2^2}{\frac{1}{D-1}\|\psi_n\| + \sigma_{d,n}d^2}\psi_n.$$

Confidence intervals can be formed as before, but plugging in the d-specific shrinkage factors such that their length adapts to the precision of the estimates of the parameters.

5. L_1 penalty. In this section we consider penalized parameter defined with an L_1 penalty term. As before, we combine the penalty term with the squared-error loss function $L(x, \tilde{x}) = \|x - \tilde{x}\|_2^2$. Let $V_1(\tilde{x}) = \lambda\|\tilde{x}\|_1^1$ where $\lambda \geq 0$ is fixed. The objective function is then

$$U(x, \tilde{x}) = \|x - \tilde{x}\|_2^2 + \lambda\|\tilde{x}\|_1^1.$$

That the objective is not differentiable everywhere means we cannot apply Theorem 3.1 to find an EIF for $\tilde{\psi}$, which precludes the type of analysis we were able to conduct in the previous section for the L_2 penalty. We proceed instead by noting that the penalized parameter has a closed form solution

$$\tilde{\psi}_d(P) = S_\lambda\big(\psi_d(P)\big),$$

where $S_\lambda : \mathbb{R} \to \mathbb{R}$ is the soft-thresholding operator

$$x \mapsto S_\lambda(x) = \begin{cases} x + \lambda, & x < -\lambda, \\ 0, & |x| \leq \lambda, \\ x - \lambda, & x > \lambda. \end{cases}$$

When applied to a vector (i.e., for $S_\lambda : \mathbb{R}^d \to \mathbb{R}^d$) the soft-thresholding operator is to be understood as applying elementwise. This solution shows that the penalized parameter simply shifts the original parameter toward zero by the amount λ, unless the original parameter is already within λ of zero, in which case it is shrunk identically to zero.

As in the L_2 case, we propose a data-driven approach for choosing λ. Our goal is to pick a λ that reduces the finite-sample variance of the penalized parameter with respect to the original parameter. In addition, the L_1 penalty may induce a parameter that is sparse, in the sense that it may contain more zeros than the original parameter. We seek an estimator that converges asymptotically to the original parameter by choosing λ data-adaptively such that λ converges to zero with sample size.

Our method for choosing λ involves approximating the finite-sample bias and variance of an estimator of $\tilde{\psi}_\lambda$ depending on the choice of λ. The nonpathwise differentiability of $\tilde{\psi}_\lambda$ in this context precludes the approach we took for the L_2-penalized parameter; accordingly, we

need to make a bolder approximation. An asymptotically normal and efficient estimator $\psi_{d,n}$ of $\psi_d(P)$, for $d \in \mathcal{D}$, has a limiting distribution given by

$$\sqrt{n}\big(\psi_n - \psi(P)\big) \xrightarrow{d} N\big(0, \sigma^2_{\psi,d}(P)\big).$$

Based on this, we approximate the finite-sample distribution of $\psi_{d,n}$ by the normal distribution,

$$Z_d \sim N\left(\psi_d(P), \frac{1}{n}\sigma^2_{\psi,d}(P)\right).$$

Suppose that we apply the soft-thresholding operator S_λ to Z_d, yielding a transformed random variable $S_\lambda(Z_d)$. In Section 4 of the Supplementary Material, we give closed forms for the mean and variance of $S_\lambda(Z_d)$ as a function of λ and the mean and variance of $S_\lambda(Z_d)$, which we denote $\mu_\lambda(\psi_d(P), \sigma^2_{\psi,d}, n)]$ and $\sigma^2_\lambda(\psi_d(P), \sigma^2_{\psi,d}, n)$. We propose setting the tuning parameter λ to the value λ^*_n that minimizes the following criterion:

$$\mathsf{Crit}\big(\lambda, \psi(P), \sigma^2_\psi(P), n\big) = \sum_{d=1}^{D}\big[\big(\mu_\lambda\big(\psi_d(P), \sigma^2_{\psi,d}, n\big) - \psi_d(P)\big)^2 + \sigma^2_\lambda\big(\psi_d(P), \sigma^2_{\psi,d}, n\big)\big].$$

The tuning parameter λ is then set to be the minimizer of the above criterion,

$$\lambda^*\big(\psi(P), \sigma^2_\psi(P), n\big) = \arg\min_{\lambda \geq 0} \mathsf{Crit}\big(\lambda, \psi(P), \sigma^2_\psi(P), n\big). \tag{7}$$

The criterion can be interpreted as an approximation of the mean-squared error of the soft-thresholded estimator relative to the original parameter. The minimizer of the above optimization problem does not have a closed form solution; in practice, we solve it numerically.

We propose estimating λ^* by the plug-in estimator $\lambda^*_n = \lambda^*_n(\psi_n, \sigma^2_{\psi,n}, n)$ based on estimates ψ_n and $\sigma^2_{\psi,n}$ of ψ_0 and $\sigma^2_{\psi,0}$. The estimated λ^*_n can then be applied to soft-threshold the initial estimates of ψ_n,

$$\tilde{\psi}_{\lambda^*_n} = S_{\lambda^*_n}(\psi_n). \tag{8}$$

The following theorem establishes the asymptotic normality and efficiency of the proposed estimator.

THEOREM 5.1 (Asymptotic normality and efficiency of $\tilde{\psi}_{\lambda^*_n}$ for L_1-penalization). *Let ψ_n and $\sigma^2_{\psi,n}$ be estimators of ψ_0 and $\sigma^2_{\psi,0}$, respectively. Let λ^*_n and $\tilde{\psi}_{\lambda^*_n}$ be defined as in (7) and (8). Assume each of the following*:

1. *There exists at least one nonzero $\psi_{d,0}$: $\|\psi_0\|_\infty > 0$.*
2. *The estimator ψ_n is an asymptotically normal and efficient*,

$$\sqrt{n}(\psi_n - \psi_0) \xrightarrow{d} N\big(0, \sigma^2_{\psi,0}\big).$$

3. *The estimator $\sigma^2_{\psi,n}$ is consistent*: $\|\sigma^2_{\psi,n} - \sigma^2_{\psi,0}\|_\infty = o_P(1)$.
4. *The estimators λ^*_n nearly minimize the minimization criterion, in the sense that*

$$\mathsf{Crit}\big(\lambda^*_n, \psi_n, \sigma^2_{\psi,n}, n\big) \leq \inf_{\lambda \geq 0} \mathsf{Crit}\big(\lambda, \psi_n, \sigma^2_{\psi,n}, n\big) + o_P(1).$$

Then it follows that $\tilde{\psi}_{\lambda^*_n}$ is an asymptotically normal and efficient estimator of ψ_0,

$$\sqrt{n}(\tilde{\psi}_{\lambda^*_n} - \psi_0) \xrightarrow{d} N\big(0, \sigma^2_{\psi,0}\big).$$

The proof is given in Section 3.2 of the Supplementary Material. Assumption 1 is necessary only to ensure that the limiting criterion function has a unique minimizer. Otherwise, if all the $\psi_{d,0}$ are zero, then the limiting criterion function is constant, and any $\lambda \geq 0$ is a minimizer. This assumption could be removed by modifying the criterion to penalize large values of λ. Assumptions 2 and 3 are equivalent to the assumptions for Theorem 4.4. Assumption 4 is a weak assumption that we expect to hold in practice.

Asymptotically valid confidence intervals for the soft-thresholded estimator can be formed using the estimated standard errors for the unpenalized parameter, as in (6).

6. Simulation studies. We investigated the finite-sample performance of the proposed L_1 and L_2 penalized estimators for each of the parameters from Examples 1, 2, and 3. In this section we include Simulation Study 1, for the nonparametric linear association parameter of Example 1, and Simulation Study 3, for the third example concerning indirectly standardized outcomes and motivated by the provider profiling application. Simulation Study 2, for the group-specific average treatment effect example, is included in Section 5 of the Supplementary Material (Susmann et al. (2026)). Reproduction materials for the simulation studies are available at https://github.com/herbps10/efficient_penalized_estimation_paper and in the Supplementary Material.

6.1. *Simulation study* 1: *Nonparametric linear association.* In this simulation we directly compare our proposed approach to penalized regression methods. The target parameter is the scaled nonparametric regression coefficient of Example 1, where for each $d \in \mathcal{D}$, the parameter is $\mathsf{E}_P[\mathsf{Cov}_P(Y, X_d \mid X_{(-d)})/\mathsf{E}_P[\mathsf{Var}_P(Y \mid X_d)]$. The scaling by the expected variance is introduced such that the parameter is equal to the coefficient $\hat{\beta}_d$ of a main-terms linear regression of Y on X, allowing us to directly compare our approach to traditional penalized linear regression estimators.

The simulation setup is a sparse linear regression scenario. Let $X = (X_1, \ldots, X_{100})^T$ be a row vector of covariates, where $X_k \sim \text{Binomial}(0.5)$ for $k = 1, \ldots, 100$. Let $\beta \in \mathbb{R}^K$ be a vector of coefficients, and draw $Y = \beta X + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$. The regression coefficients are fixed at the beginning of each simulation by drawing $\beta_k \sim \text{Binomial}(\theta)$ with $\theta = 30\%$. The simulation study tested all combinations of sample size $N \in \{50, 100, 250, 500\}$ and noise standard deviation $\sigma \in \{0.5, 1, 3\}$.

To implement the penalized estimators, we need a nonparametric estimator of the nonparametric linear association that is asymptotically normal and efficient. Section 2 of the Supplementary Material describes such an estimator based on one-step estimation. The nuisance parameters are estimated using L_1-regularized generalized linear regressions with tuning parameters chosen via cross-validation, using the implementation in the `glmnet` `R` package (Friedman, Tibshirani and Hastie (2010), Tay, Narasimhan and Hastie (2023)). The unpenalized estimator is then adjusted using the proposed penalization methods to form L_1- and L_2-regularized estimators of $\tilde{\psi}_d$.

As a benchmark, we estimated the linear association parameters by fitting L_1- and L_2-regularized main-terms linear models of Y with respect to covariates X and an intercept term, and take the estimated coefficient $\hat{\beta}_d$ as an estimate of the corresponding linear association parameter $\psi_{d,0}$. The tuning parameters were chosen by the default cross-validation method implemented in `glmnet`. We expect this benchmark estimator to be a consistent estimator of $\psi_{d,0}$), as the simulation data-generating process is a linear model. We compare our approach to the benchmark in terms of the estimates mean error (ME), variance (Var), mean squared error (MSE), and 95% empirical coverage. The comparison method `glmnet` does not report confidence intervals by default, so we do not compare our method to `glmnet` in terms of empirical coverage.

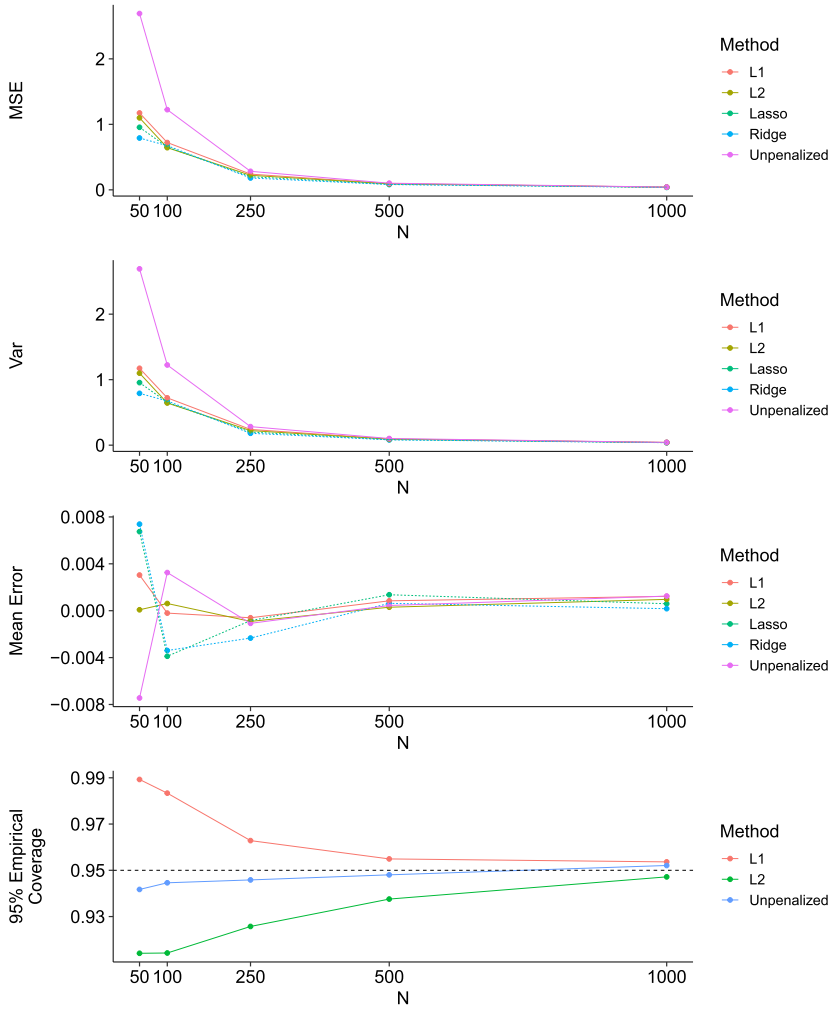

FIG. 2. *Subset of results from Simulation Study* 1 *for the nonparametric linear association parameter plotting MSE for all methods with the data-generating process noise size* $\sigma = 3$.

A subset of the results corresponding to simulations with noise $\sigma = 3$ are shown in Figure 2; a complete table of the results is available as Table 3 in the Supplementary Material. Our proposed L_1 and L_2 penalized estimators match or outperform the unpenalized one-step estimator for all sample sizes and noise levels. The benchmark penalized regressions tended to achieve slightly lower MSE. The better performance of the benchmark in this setting is probably because these methods are tuned using cross-validation, which likely provides better finite-sample approximations of variance than our method, which chooses the strength of penalization parameter λ based on an asymptotic approximation.

7. Simulation study 3: Indirectly-standardized outcomes. For the third simulation study, we adopted a data-generating process described based on the second simulation study of (Susmann et al. (2025)), which we describe briefly here. We draw simulated datasets of N i.i.d. patient observations (X, A, Y) where $X \in [0, 1]^5$ is a set of five patient covariates drawn from the joint uniform distribution, $A \in \{1, \ldots, 50\}$ indexes the provider, and $Y \in \{0, 1\}$ is a binary outcome. Provider covariates $\{\beta_a : a = 1, \ldots, 50\}$ are drawn as $\beta_a \sim \text{Bernoulli}(0.5)$. Then, conditional on patient covariates, patients are assigned to a provider with probability

$P(A = d \mid X)$, given by

$$\operatorname{logit}(\beta_d, X)) \propto 1 + 20\beta_d\big(1.5\mathbb{I}[X_1 > 0.4] - 0.5\mathbb{I}[X_1 > 0.4]\big) + \mathbb{I}[0.4 < X_2 < 0.7] - 0.5X_5,$$

suitably normalized such that $\sum_a g(\beta_a, X) = 1$. Then the outcome is drawn from the law

$$Y \mid X, A, \beta_A \sim \text{Bernoulli}\big(\text{logit}^{-1}\big(0.7\mathbb{I}[W_1 > 0.7] + 3\mathbb{I}[0.4 < X_1 < 0.7] + X_2^2 + (W_4 + 1)X_5 + 2\big(-1 + 2\mathbb{I}[X_1 > 0.7]\big)\beta_A\big)\big).$$

This simulation design incorporates selective assignment of patients to providers according to their covariates, mimicking differential case-mix between providers, as well as nonlinear provider assignment and outcome processes, mimicking complex real-world relationships.

The above data-generating process was sampled 250 times for each sample size in $N \in \{3000, 5000, 9000\}$. For each simulated dataset, the doubly-robust TMLE method described in (Susmann et al. (2025)) was applied to estimate the indirectly standardized outcome ratios for each simulated provider with at least 10 observations. Nuisance parameters were estimated using `lightgbm` with 200, 100, and 50 iterations, `ranger`, `glm`, and `gam` learners. These unpenalized estimates were then adjusted using our proposed $L2$ and $L1$ penalization approach with data-adaptive choice of tuning parameter. As a comparison, we estimated the indirectly standardized outcome ratios using generalized linear models with either fixed- or random-effect provider effects using implementations available in the `R` package `pprof`, based on computationally efficient algorithms described in Wu et al. (2022). We also applied the empirical Bayes adjustment described in Section 4. The results were compared by their MSE, mean error (ME), and empirical coverage of the 95% confidence intervals. We computed these metrics both overall, for all simulated providers, and in providers stratified by sample size in order to compare how shrinkage effects perform for small vs. large providers. For each overall sample size N, we computed the tertiles of the sample sizes for each simulated provider; we refer to these as "small, 'medium," and "large" providers. The metrics were then calculated independently among providers in each group.

The nonstratified results are shown in Table 1, and several overall trends are noticeable. First, as expected, the mean error of the L_1, L_2, and empirical Bayes adjusted estimates is higher than the unadjusted estimate. At the lowest sample size, the adjusted estimates trade the higher mean error for lower MSE than the unadjusted estimator, and the L_2 method has lower MSE at all sample sizes than both the unadjusted TMLE estimator and the comparison fixed-effects estimator. Second, the adjusted estimators (particularly the L_2 estimator) outperform the comparison random-effects estimator, the state-of-the-art for applying shrinkage in estimating indirect standardization ratios, by achieving lower MSE, lower mean error, and better coverage. However, except for the L_1 method, both L_2 and empirical Bayes do not achieve the nominal 95% empirical coverage of the unadjusted TMLE.

Next, we turn to the stratified results in which providers are grouped by their observed sample size (Table 2). By viewing the results stratified by sample size, we can investigate how the penalization methods impact providers of different sizes. Starting with mean error, the unpenalized TMLE method and the comparison fixed-effects method have similar error across all provider sizes. The empirical Bayes method, on the other hand, has the largest error for small providers and the smallest error for large providers. This matches what we might expect, given that the empirical Bayes method significantly shrinks low-precision estimates while applying little shrinkage to high-precision estimates. The same can be seen for the random-effects comparison method, which partially pools information between providers resulting in point estimates that are shrunk proportional to provider size. The L_2 method, on the other hand, applies the same shrinkage factor to all providers. Accordingly, error is spread somewhat more evenly across provider sizes than the empirical Bayes method. A

TABLE 1

Results from Simulation Study 3 comparing a fixed effects model, random effects model, the unpenalized TMLE estimator, L_1-regularized estimator, L_2-regularized estimator, and empirical Bayes (EB) shrinkage estimator

N	TMLE	L_1	L_2	EB	Fixed effects	Random effects
			Mean Squared Error $\times 100$			
1500	9.3	7.2	5.2	5.7	12.9	6.8
3000	5.0	5.3	4.2	4.8	6.1	5.6
4500	3.4	3.7	3.2	3.7	4.1	4.6
6000	2.6	2.9	2.6	3.0	3.2	4.0
			Mean Error $\times 100$			
1500	1.1	8.1	7.9	12.7	−1.9	14.4
3000	3.8	7.2	8.8	12.6	1.8	14.6
4500	3.6	4.9	7.6	11.0	2.9	13.5
6000	3.4	4.1	6.7	9.7	3.9	13.0
			95% Empirical Coverage			
1500	95.6%	99.6%	87.5%	83.1%	92.9%	66.2%
3000	94.8%	98.3%	85.2%	79.3%	92.2%	76.4%
4500	94.2%	97.0%	87.0%	81.6%	90.2%	79.4%
6000	94.1%	95.9%	86.8%	82.7%	86.4%	77.3%

similar trend holds for the L_1 method. In terms of MSE, applying penalization adjustments improved MSE relative to the unpenalized TMLE and comparison methods. For coverage, the L_1 method tended to be anticonservative for all provider sizes, while the L_2 method was conservative. The empirical Bayes method had near-optimal empirical coverage for large providers, which is explained by the point estimates and confidence intervals avoiding shrinkage; on the other hand, the small providers had lower coverage rates, due to more aggressive shrinkage.

8. Application. In this section we illustrate the real-world utility of our penalization methods through a healthcare provider profiling application, estimating the standardized readmission ratios (SRR) for kidney dialysis providers. Briefly, the observed data are a set of baseline patient covariates X, a treatment variable $A \in \{1, \ldots, D\} = \mathcal{D}$ that indexes the dialysis provider seen by each patient, and an outcome variable $Y \in \{0, 1\}$ which indicates all-cause unplanned hospital readmission within 30 days of discharge ($Y = 1$ indicates unplanned readmission, which is considered a negative outcome). Define the indirectly-standardized outcome ψ_d for a provider $d \in \mathcal{D}$ as in Example 3. That is, ψ_d is (under causal assumptions) the mean unplanned readmission rate if the population of patients treated by provider d had rather been randomly assigned to another provider according to the observed provider-assignment mechanism. We then define the centered standardized readmission ratio (SRR) as the ratio of ψ_d to the observed readmission rates for patients treated by provider d, centered at zero,

$$\mathsf{SRR}_d(P) := \frac{\psi_d(P)}{\mathsf{E}_P[Y \mid A = d]} - 1.$$

A positive SRR means that the unplanned readmission rate would have been higher if patients had been randomly assigned to a provider that treated a similar patient mix; this can be seen as evidence of better performance of provider d relative to its peers treating a similar population. Similarly, a negative SRR suggests that the unplanned readmission rate would have been lower if patients were randomly reassigned to another provider.

TABLE 2

Results from Simulation Study 3 comparing a fixed effects model, random effects model, the unpenalized TMLE estimator, L_1-regularized estimator, L_2-regularized estimator, and empirical Bayes (EB) shrinkage estimator. Results are stratified by the tertile of the sample size of each provider with respect to the sample sizes of all simulated providers at the same overall N

Provider sample size	N	TMLE	L_1	L_2	EB	Fixed effects	Random effects
			Mean Squared Error $\times 100$				
Small	1500	20.3	15.6	11.2	11.9	29.0	14.6
	3000	10.4	10.8	8.5	9.8	12.9	11.5
	4500	6.9	7.5	6.5	7.7	8.4	9.4
	6000	5.2	5.8	5.3	6.3	6.5	8.3
Medium	1500	5.3	4.6	3.3	3.7	6.8	4.4
	3000	3.8	4.4	3.4	3.8	4.6	4.5
	4500	2.7	3.1	2.7	3.1	3.4	3.9
	6000	2.1	2.5	2.2	2.5	2.7	3.4
Large	1500	0.7	0.4	0.3	0.6	0.7	0.3
	3000	0.3	0.3	0.2	0.3	0.4	0.2
	4500	0.2	0.3	0.2	0.2	0.3	0.2
	6000	0.2	0.3	0.1	0.2	0.3	0.2
			Mean Error $\times 100$				
Small	1500	2.5	23.6	20.1	29.0	−5.7	34.7
	3000	7.9	19.7	20.0	26.9	2.1	30.2
	4500	7.6	14.6	17.2	23.5	3.7	26.3
	6000	7.3	12.0	15.2	20.8	5.7	24.5
Medium	1500	0.9	4.7	4.9	7.5	−0.9	8.5
	3000	3.7	6.6	8.1	10.5	1.5	12.4
	4500	3.4	4.9	7.2	9.5	2.2	12.1
	6000	3.1	4.1	6.3	8.4	3.0	11.6
Large	1500	−0.5	−6.0	−3.0	−0.8	1.5	−2.7
	3000	−0.6	−5.8	−2.5	−0.8	1.9	−0.3
	4500	−0.4	−5.1	−2.0	−0.6	2.6	1.5
	6000	−0.5	−4.3	−1.8	−0.6	3.0	2.2
			95% Empirical Coverage				
Small	1500	96.0%	99.1%	85.3%	66.5%	90.9%	40.6%
	3000	94.4%	97.1%	82.1%	60.8%	91.0%	54.8%
	4500	93.4%	94.5%	84.2%	66.4%	89.6%	62.1%
	6000	93.3%	92.3%	83.8%	70.1%	86.2%	61.3%
Medium	1500	95.5%	99.6%	88.4%	89.7%	93.6%	74.7%
	3000	94.2%	97.7%	84.9%	83.0%	91.6%	78.6%
	4500	93.9%	96.6%	86.5%	84.0%	91.6%	80.7%
	6000	94.6%	95.9%	87.1%	84.5%	89.4%	78.3%
Large	1500	95.2%	100.0%	89.0%	95.2%	94.3%	86.9%
	3000	95.8%	100.0%	88.8%	95.7%	94.1%	97.6%
	4500	95.2%	99.9%	90.5%	95.1%	89.3%	96.2%
	6000	94.3%	99.8%	89.6%	94.1%	83.6%	93.1%

Estimating the above SRR parameter may be difficult, especially for providers with few patients. In addition, there are typically high policy stakes involved in provider profiling, as the results may be used to identify underperforming providers for remedial action. Thus, there is often interest in having any estimates be conservative by shrinking high-variance estimates toward zero (Normand, Glickman and Gatsonis (1997)). This approach avoids unfairly penal-

izing small providers who, for example, purely by chance happened to have treated patients who had a unusually high number of unplanned readmissions.

A popular approach for estimating provider profiling measures with shrinkage is via generalized mixed models with a provider-specific random effect that is shrunk toward zero. However, as explored in simulations in Susmann et al. (2025), generalized linear models introduce parametric assumptions on the data-generating process that can lead to biased estimates when the model is misspecified. In addition, we argue that shrinking the actual parameter of interest, the SRR, toward zero is more interpretable than shrinking the provider-specific random effects of a generalized linear model, which have a complex interpretation. In addition, the results of Simulation Study 3 show that linear generalized linear models with provider-specific random effects can exhibit bias for the SRR that does not show evidence of converging to zero with sample size.

We analyze data from a Medicare claims dataset from the United States Renal Data System (USRDS) consisting in dialysis provider treatment records for patients with end-stage renal disease (ESRD) (U.S. Renal Data System (2022)). These data were previously analyzed in Susmann et al. (2025) in which nonpenalized SRRs were estimated using doubly robust and asymptotically consistent estimators. Our analysis dataset comprises all dialysis providers in New York State with at least 20 observations in the year 2020 (this enlarges our previous analysis of the same data, which used only those providers with at least 100 observations). First, we computed nonpenalized estimates of the nonpenalized SRR using the nonparametric targeted minimum loss-based estimation (TMLE) method described in Susmann et al. (2025), using cross-fitted ensembles of generalized linear models, regularized generalized linear models, and gradient boosting trees for nuisance parameter estimation. We then estimated the L_2-penalized SRR with penalization parameter λ chosen using the data-driven criterion proposed in Section 4. We also applied the empirical Bayes shrinkage derived in Section 4 that adaptively shrinks estimates as a function of the standard error. As a comparison, we estimated the SRR using a generalized linear model with provider-specific random effects.

Results for the nonparametric estimates with L_2 penalization and empirical Bayes shrinkage adjustments are shown in Figure 3. The results are displayed as funnel plots, which plot the precision of the unpenalized SRR estimator vs. the SRR point estimates, before and after adjustment. A notable difference in the estimates adjusted by L_2 penalization vs. empirical Bayes shrinkage is in the high-precision estimates. As expected, the L_2 penalization is based on a single penalization parameter λ, which causes all parameters to be shrunk toward one, including the high-precision estimates. This is not true of the empirical Bayes estimates, which are shrunk less for high-precision estimates. Each of the methods, therefore, has tradeoffs, and the choice of which to use depends on the goals of the analysis. If the goal is to minimize false positives when identifying outlying providers, then empirical Bayes may be a good choice due to its more aggressive shrinkage of low-precision estimates. On the other hand, if the goal is to increase power, the less extreme shrinkage of the L_2 method, applied uniformly across all providers, may be preferred.

Figure 4 compares estimates of the SRR based on a generalized linear model with provider-specific random effects vs. the nonparametric estimates adjusted with our proposed L_2 and empirical Bayes shrinkage method. Overall, the estimated SRRs are similar across the methods. However, compared to both adjustment methods, the random effects model yields attenuated SRR estimates that are more conservative than the nonparametric approach. In particular, several providers that have the highest estimated values using TMLE and L_2 penalization have lower estimates under the random effects model. These comparisons can be interpreted through our theoretical and empirical simulation results. From the theoretical standpoint, the nonparametric TMLE estimates rely on fewer assumptions than the random effects

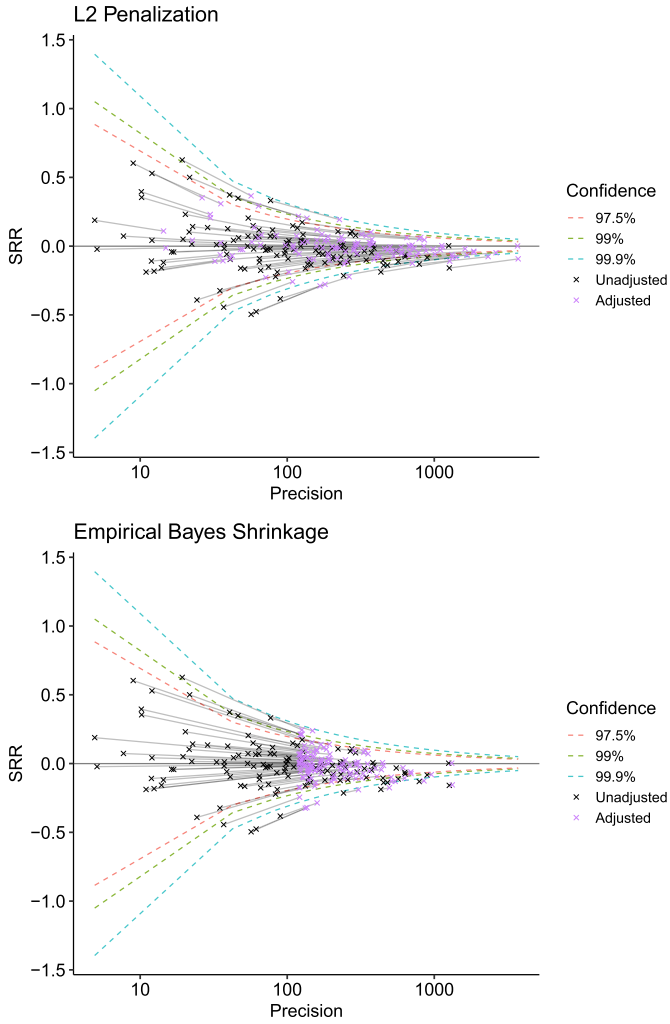

FIG. 3. *Funnel plots of standardized readmission ratios (SRR, Section 8) for New York State dialysis providers in the analysis dataset. In both the top and bottom plots the nonadjusted estimates are derived from a nonparametric targeted minimum loss-based estimation (TMLE) approach. Then in the top plot, SRR estimates are adjusted by L_2 penalization using the data-adaptive choice of penalization hyperparameter λ proposed in Section 4 and shrinkage standard errors. In the bottom plot, SRR estimates are shrunk using the empirical Bayes based method described in Section 4. Vertical lines connect dialysis provider SRR estimates before and after adjustment.*

model. Importantly, the nonparametric estimates do not require linearity assumptions. In this provider profiling example, we find it reasonable that there may be complex interactions and nonlinear relationships between patient covariates and outcomes, and we, therefore, find the plausibility of the assumptions necessary for the nonparametric estimation strategy to be more reasonable than requiring linearity. This conclusion is also supported by the empirical results from Simulation Study 3, which suggests that the flexibility of the nonparametric estimation strategy combined with penalization or shrinkage does not incur a trade-off in terms of bias or variance in a practical scenario.

9. Discussion. Estimating a large set of statistical parameters introduces challenges beyond those of estimating a single parameter. To improve estimation, it may be of interest to trade bias in one of the constituent parameters in favor of controlling the overall variance across all estimates. In addition, to aid interpretation or communication it may also be of

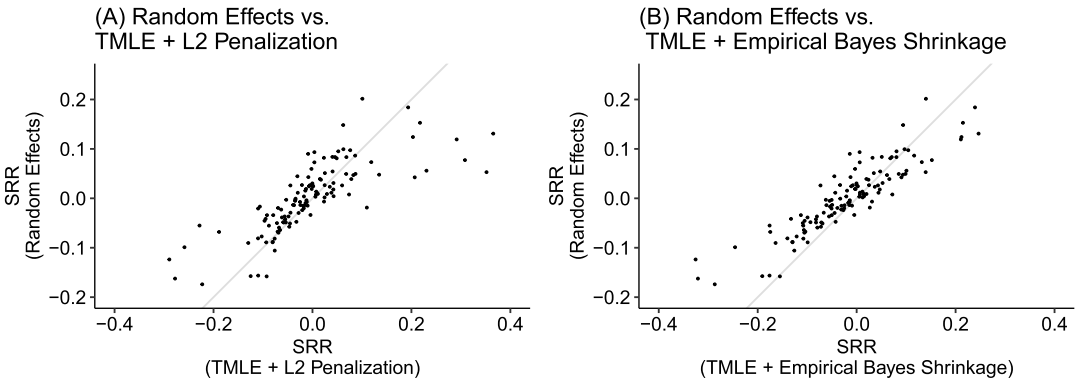

FIG. 4. *Comparison of standardized readmission tatio* (*SRR*, *Section* 8) *estimates for New York State dialysis providers in the analysis dataset from a linear model with provider random effects vs. estimates from the nonparametric targeted minimum loss-based estimation* (*TMLE*) *approach adjusted via the proposed L2 penalization and empirical Bayes shrinkage methods.*

interest to find a set of point estimates that are *sparse*, in that estimates statistically indistinguishable from zero are shrunk identically to zero. To address these concerns, we introduced a novel framework for defining regularized statistical parameters via penalization. This framework maintains the substantive focus on the original parameter of interest, and penalized parameters are introduced as a way to derive estimators with desirable finite properties such as lower variance and sparsity.

The framework we proposed for regularized estimation was motivated by problems in healthcare provider profiling, particularly for cases where provider quality metrics are estimated for providers with limited observations. Simulation Study 3 was designed to reproduce patterns seen in provider profiling applications, including a range of provider sizes and complex, nonlinear relationships between variables. The simulation results demonstrate that the proposed penalized methods can achieve lower MSE than both the unadjusted doubly-robust estimator and fixed- and random-effects comparison methods, especially at small sample sizes. Whether a method that trades increased bias for lower variance, particularly with higher bias for small providers, will depend on the goals of the specific profiling application. In general, the use of shrinkage estimators may be desirable when the reduction of false positives in detecting outlying providers is preferred. Compared to the dominant extant approach of fixed- or random-effects modeling, the benefit of our approach is that it allows for the use of penalization and shrinkage within a fully nonparametric framework that does not rely on strong functional form assumptions.

The penalized parameters we propose are formulated in a completely nonparametric framework, and our results are, therefore, applicable in very general settings. One particular area where they are relevant is causal inference, where the target parameters of interest are typically a (possibly large) set of low-dimensional summaries of counterfactual quantities, such as treatment effects. While existing methods such as penalized regression can be applied to estimate the nuisance parameters required for forming efficient and doubly-robust causal effect estimators, it is less clear how to apply penalization directly to the causal effect estimates themselves. Our research fills this gap.

We explored two important examples of penalized parameters that fall within our framework: those defined with L_2 and L_1 penalization terms. Many other options are available; considering L_p-norm penalties in more generality would be an immediate extension, or penalties such as the Elastic-Net penalty or the Huber loss function. Going further, our framework could be expanded to capture functional parameters (such as those in a Banach or Hilbert Space) regularized with functional norms.

Within the L_2 and L_1 examples we investigated, we proposed data-adaptive approaches for choosing the penalization hyperparameter λ. We propose choosing λ to go to zero as sample size increases so that the penalized estimates converge to the unpenalized estimates. This reflects the fact that the target parameter of interest remains the unpenalized parameter, with penalization applied only to improve finite-sample performance of the estimator. The data-driven choices for λ are based on an asymptotically valid approximation of the finite-sample variance of the unpenalized estimators. For the L_2 penalty example, for example, we use asymptotically-justified variance approximations with the goal of forming an estimator with better *finite-sample* performance. The reliance on asymptotic approximations is due to the generality of our approach in which the key restriction is the pathwise differentiability of the original parameter, the property that leads to the existence of an EIF for the target parameter characterizing its efficiency bound. We then use this asymptotic efficiency bound to approximate finite-sample variance. However, other methods for choosing λ may perform better than our approach, in particular when cross-validation can be applied. Indeed, results from the first simulation study show that penalized linear regression tuned with cross-validation can yield lower MSEs than our proposed estimators. However, the applicability of cross-validation in that context hinges on the fact that the parameter of interest is identified as a linear regression coefficient. Cross-validation can then be applied to find an optimal degree of penalization based on the model's predictive performance. However, for the other causal target parameters we investigated, it is not clear how cross-validation could be so straightforwardly applied as the target parameters are low-dimensional summaries of counterfactuals and are not predictive. The strength of our approach, then, is its general applicability to low-dimensional target parameters, such as those of interest in causal inference that are typically defined in terms of counterfactual quantities.

Acknowledgments. Correspondence may be directed to Herbert P. Susmann (susmah01@nyu.edu).

We would like to thank Antoine Chambaz and Alec McClean for helpful discussions. The computational requirements for this work were supported in part by the NYU Langone High Performance Computing (HPC) Core's resources and personnel. The data reported here have been supplied by the United States Renal Data System (USRDS). The interpretation and reporting of these data are the responsibility of the authors and in no way should be seen as an official policy or interpretation of the U.S. government.

SUPPLEMENTARY MATERIAL

Supplement: Derivations and additional results (DOI: 10.1214/25-AOAS2129SUPPA; .pdf). Notation guide, proofs of theoretical results presented in the manuscript, and additional simulation study results.

Supplement: Analysis code (DOI: 10.1214/25-AOAS2129SUPPB; .zip). Collection of `R` scripts for reproducing the simulation studies in the paper.

REFERENCES

ARMSTRONG, T. B., KOLESÁR, M. and PLAGBORG-MØLLER, M. (2022). Robust empirical Bayes confidence intervals. *Econometrica* **90** 2567–2602. MR4524894 https://doi.org/10.3982/ecta18597

BAHAMYIROU, A., SCHNITZER, M. E., KENNEDY, E. H., BLAIS, L. and YANG, Y. (2022). Doubly robust adaptive LASSO for effect modifier discovery. *The International Journal of Biostatistics* **18** 307–327. https://doi.org/10.1515/ijb-2020-0073

BENKESER, D. and VAN DER LAAN, M. (2016). The highly adaptive lasso estimator. In 2016 *IEEE International Conference on Data Science and Advanced Analytics* (*DSAA*) 689–696. https://doi.org/10.1109/DSAA.2016.93

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. and WELLNER, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer, New York. MR1623559

BRAKENHOFF, T. B., MOONS, K. G., KLUIN, J. and GROENWOLD, R. H. (2018). Investigating risk adjustment methods for health care provider profiling when observations are scarce or events rare. *Health Services Insights* **11** 1178632918785133. https://doi.org/10.1177/1178632918785133

DAIGNAULT, K. and SAARELA, O. (2017). Doubly robust estimator for indirectly standardized mortality ratios. *Epidemiologic Methods* **6** 20160016. https://doi.org/10.1515/em-2016-0016

DELONG, E. R., PETERSON, E. D., DELONG, D. M., MUHLBAIER, L. H., HACKETT, S. and MARK, D. B. (1997). Comparing risk-adjustment methods for provider profiling. *Stat. Med.* **16** 2645–2664. https://doi.org/10.1002/(SICI)1097-0258(19971215)16:23<2645::AID-SIM696>3.0.CO;2-D

DÍAZ, I. (2024). Non-agency interventions for causal mediation in the presence of intermediate confounding. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **86** 435–460. MR4754091 https://doi.org/10.1093/jrsssb/qkad130

EFRON, B. (2024). Empirical Bayes: Concepts and methods. In *Handbook of Bayesian, Fiducial, and Frequentist Inference*. Chapman & Hall, London.

EFRON, B. and MORRIS, C. (1977). Stein's paradox in statistics. *Scientific American* **236** 119–127.

FELLER, A. and GELMAN, A. (2015). Hierarchical models for causal effects. In *Emerging Trends in the Social and Behavioral Sciences* 1-16 Wiley, New York. https://doi.org/10.1002/9781118900772.etrds0160

FRIEDMAN, J., TIBSHIRANI, R. and HASTIE, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33** 1–22. https://doi.org/10.18637/jss.v033.i01

HÁJEK, J. (1969/70). A characterization of limiting distributions of regular estimates. *Z. Wahrsch. Verw. Gebiete* **14** 323–330. MR0283911 https://doi.org/10.1007/BF00533669

HÁJEK, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I: Theory of Statistics* 175–194. Univ. California Press, Berkeley, CA. MR0400513

HANSEN, B. E. (2017). Stein-like 2SLS estimator. *Econometric Rev.* **36** 840–852. MR3680746 https://doi.org/10.1080/07474938.2017.1307579

HOERL, A. E., KANNARD, R. W. and BALDWIN, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics* **4** 105–123. https://doi.org/10.1080/03610927508827232

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67. https://doi.org/10.1080/00401706.1970.10488634

IMDAD ULLAH, M., ASLAM, M. and ALTAF, S. (2018). Lmridge: A comprehensive R package for ridge regression. *The R Journal* **10** 326–346. https://doi.org/10.32614/RJ-2018-060

KALBFLEISCH, J. D. and HE, K. (2018). Discussion on "Time-dynamic profiling with application to hospital readmission among patients on dialysis," by Jason P. Estes, Danh V. Nguyen, Yanjun Chen, Lorien S. Dalrymple, Connie M. Rhee, Kamyar Kalantar-Zadeh, and Damla Senturk. *Biometrics* **74** 1401–1403.

KALBFLEISCH, J. D. and WOLFE, R. A. (2013). On monitoring outcomes of medical providers. *Statistics in Biosciences* **5** 286–302. https://doi.org/10.1007/s12561-013-9093-x

KAPLAN, D. M. and LIU, X. (2024). Confidence intervals for intentionally biased estimators. *Econometric Rev.* **43** 197–214. MR4720046 https://doi.org/10.1080/07474938.2024.2312288

KEIDING, N. and CLAYTON, D. (2014). Standardization and control for confounding in observational studies: A historical perspective. *Statist. Sci.* **29** 529–558. MR3300358 https://doi.org/10.1214/13-STS453

KENNEDY, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In *Statistical Causal Inferences and Their Applications in Public Health Research. ICSA Book Ser. Stat.* 141–167. Springer, Cham. MR3617956

KENNEDY, E. H. (2024). Semiparametric doubly robust targeted double machine learning: A review. In *Handbook of Statistical Methods for Precision Medicine* 10 CRC Press, Boca Raton.

LE CAM, L. (1972). Limits of experiments. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (*Univ. California, Berkeley, Calif.*, 1970/1971), *Vol. I: Theory of Statistics* 245–261. Univ. California Press, Berkeley, CA. MR0415819

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948 https://doi.org/10.1214/15-AOS1371

MAASOUMI, E. (1978). A modified Stein-like estimator for the reduced form coefficients of simultaneous equations. *Econometrica* **46** 695–703. MR0483203 https://doi.org/10.2307/1914241

MACKENZIE, T. A., GRUNKEMEIER, G. L., GRUNWALD, G. K., O'MALLEY, A. J., BOHN, C., WU, Y. and MALENKA, D. J. (2015). A primer on using shrinkage to compare in-hospital mortality between centers. *The Annals of Thoracic Surgery* **99** 757–761. https://doi.org/10.1016/j.athoracsur.2014.11.039

MCCLEAN, A., BRANSON, Z. and KENNEDY, E. H. (2024). Nonparametric estimation of conditional incremental effects. *J. Causal Inference* **12** Paper No. 20230024, 42. MR4736178 https://doi.org/10.1515/jci-2023-0024

NORMAND, S.-L. T., GLICKMAN, M. E. and GATSONIS, C. A. (1997). Statistical methods for profiling providers of medical care: Issues and applications. *J. Amer. Statist. Assoc.* **92** 803–814. https://doi.org/10.1080/01621459.1997.10474036

PFANZAGL, J. and WEFELMEYER, W. (1985). Contributions to a general asymptotic statistical theory. *Statistics & Risk Modeling* **3** 379–388.

SHORTREED, S. M. and ERTEFAIE, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics* **73** 1111–1122. MR3744525 https://doi.org/10.1111/biom.12679

SMUCLER, E., ROTNITZKY, A. and ROBINS, J. M. (2019). A unifying approach for doubly-robust ℓ_1 regularized estimation of causal contrasts.

STEIN, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, 1954–1955, *Vol. I* 197–206. Univ. California Press, Berkeley, CA. MR0084922

SUSMANN, H. and CHAMBAZ, A. (2023). Inference in Marginal Structural Models by Automatic Targeted Bayesian and Minimum Loss-Based Estimation.

SUSMANN, H., LI, Y., MCADAMS-DEMARCO, M. A., DÍAZ, I. and WU, W. (2025). Doubly Robust Nonparametric Efficient Estimation for Provider Evaluation. *J. R. Stat. Soc. Ser. A Stat. Soc.* MR2137327 https://doi.org/10.1093/jrsssa/qnaf145

SUSMANN, H. P, LI, Y., MCADAMS-DEMARCO, M. A, WU, W. and DÍAZ, I. (2026). Supplement to "Asymptotically efficient data-adaptive penalized shrinkage estimation with application to causal inference." https://doi.org/10.1214/25-AOAS2129SUPPA, https://doi.org/10.1214/25-AOAS2129SUPPB

TAY, J. K., NARASIMHAN, B. and HASTIE, T. (2023). Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software* **106** 1–31. https://doi.org/10.18637/jss.v106.i01

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TSIATIS, A. A. (2006). *Semiparametric Theory and Missing Data. Springer Series in Statistics*. Springer, New York. MR2233926

U.S. Renal Data System (2022). 2022 USRDS annual data report: Epidemiology of kidney disease in the United States National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD.

VON MISES, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Ann. Math. Statistics* **18** 309–348. MR0022330 https://doi.org/10.1214/aoms/1177730385

VAN DER LAAN, M. J. and RUBIN, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostat.* **2** Art. 11, 40. MR2306500 https://doi.org/10.2202/1557-4679.1043

VAN DER VAART, A. W. (1992). Asymptotic linearity of minimax estimators. *Statist. Neerlandica* **46** 179–194. MR1178478 https://doi.org/10.1111/j.1467-9574.1992.tb01336.x

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247 https://doi.org/10.1017/CBO9780511802256

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

VAREWYCK, M., GOETGHEBEUR, E., ERIKSSON, M. and VANSTEELANDT, S. (2014). On shrinkage and model extrapolation in the evaluation of clinical center performance. *Biostatistics* **15** 651–664. https://doi.org/10.1093/biostatistics/kxu019

WILLIAMSON, B. D., GILBERT, P. B., CARONE, M. and SIMON, N. (2021). Nonparametric variable importance assessment using machine learning techniques. *Biometrics* **77** 9–22. MR4229718 https://doi.org/10.1111/biom.13392

WU, W., YANG, Y., KANG, J. and HE, K. (2022). Improving large-scale estimation and inference for profiling health care providers. *Stat. Med.* **41** 2840–2853. MR4441590 https://doi.org/10.1002/sim.9387

YANG, X., PENG, B., CHEN, R., ZHANG, Q., ZHU, D., ZHANG, Q. J., XUE, F. and QI, L. (2014). Statistical profiling methods with hierarchical logistic regression for healthcare providers with binary outcomes. *J. Appl. Stat.* **41** 46–59. MR3291199 https://doi.org/10.1080/02664763.2013.830086

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67** 301–320. MR2137327 https://doi.org/10.1111/j.1467-9868.2005.00503.x