Advance Access publication February 12, 2008

Political Analysis (2009) 17:89–106
doi:10.1093/pan/mpm032

A Comment on Diagnostic Tools for Counterfactual Inference

Nicholas Sambanis
Department of Political Science, Yale University,
PO Box 208301, New Haven, CT 06520
e-mail: nicholas.sambanis@yale.edu (corresponding author)

Alexander Michaelides
London School of Economics, Department of Economics,
Houghton Street, London WC2A 2AE, UK
e-mail: a.michaelides@lse.ac.uk

We evaluate two diagnostic tools used to determine if counterfactual analysis requires extrapolation. Counterfactuals based on extrapolation are model dependent and might not support empirically valid inferences. The diagnostics help researchers identify those counterfactual "what if" questions that are empirically plausible. We show, through simple Monte Carlo experiments, that these diagnostics will often detect extrapolation, suggesting that there is a risk of biased counterfactual inference when there is no such risk of extrapolation bias in the data. This is because the diagnostics are affected by what we call the *n/k* problem: as the number of data points relative to the number of explanatory variables decreases, the diagnostics are more likely to detect the risk of extrapolation bias even when such risk does not exist. We conclude that the diagnostics provide too severe a test for many data sets used in political science.

1 Introduction

Counterfactual questions of the sort "What would happen to Y if we changed X while keeping everything else constant?" are the subject of much empirical work in political science. Answers to such "what if" questions are frequently extrapolations that go beyond the range of the observed data. The further from the data we take a counterfactual, the greater the risk of "extrapolation bias" and the more model dependent are the conclusions. Model-dependent extrapolation can lead to bias when the support of the distribution of the covariates in a regression model differs between the treatment and control groups.

King and Zeng (2006) develop two diagnostic tools to help researchers avoid analyzing data sets that cannot support empirically plausible causal inferences. These tools promise to help by identifying data points that are "too far" from the counterfactuals. Such data points should be discarded since they have no empirical content to support reasonable

Author's note: We thank Komei Fukuda, Don Green, Alan Gerber, and Jasjeet Sekhon for their generous help, Mike Kane for assistance with R programming, and five anonymous referees for constructive comments.

causal inferences. The diagnostics require no assumptions about the model and apply to any dependent variable and any type of data.

We argue that when the number of observations relative to the number of variables is "small," the diagnostics will find the counterfactuals "far" from the data, even for data that are simulated ex ante not to generate concern over extrapolation bias. By "small" data sets we mean sample sizes that one usually finds in applied cross-country analyses in international relations and other fields of political science (100 or even 1000 observations over, say, 10 variables). Our argument is supported by evidence from Monte Carlo simulations. Our Monte Carlo evidence suggests that researchers will not be able to use these diagnostics as reliable indicators that their causal inferences from a data set will be biased due to extrapolation.

2 The Convex Hull and Gower Distance Tests

We first describe the two diagnostic tests for completeness. Consider a data matrix $\boldsymbol{X}$ with dimensions $n \times k$, where n is the number of data points and k the number of variables. One of the k variables is a binary treatment. We are interested in identifying the causal effects of the treatment on some dependent variable Y. A "counterfactual" is defined as 1 minus the treatment, leaving all the other variables unchanged. How should we evaluate such a counterfactual? The diagnostics tell us which counterfactuals are empirically plausible and which require extrapolation. The diagnostics therefore provide measures of the degree of model dependence in any counterfactual analysis.

The first diagnostic proposes a definition of extrapolation based on the concept of the convex hull. The convex hull is the smallest geometric shape that includes all the points in a given set of data (the points lie either in the interior or on the vertices of the hull). If a counterfactual point is outside of the convex hull of $\boldsymbol{X}_{n \times k}$, then analyzing that point requires extrapolation (if it is inside, then inferences are based on *inter*polation).

The first diagnostic tool provides a computationally feasible way to check whether a particular observation lies inside the convex hull. The method defines the check of whether a query point is inside the convex hull of a set of data as a linear programming problem. The innovation of King and Zeng is the following argument: if $x_{1 \times k}$[1] is in the convex hull of $\boldsymbol{X}_{n \times k}$, then $x_{1 \times k}$ can be expressed as a convex combination of all points in S, where S is the set of the vertices of the convex hull of $\boldsymbol{X}_{n \times k}$ containing all boundary points in $\boldsymbol{X}_{n \times k}$. All points in $\boldsymbol{X}_{n \times k}$ are also convex combinations of points in S.

So, the condition is equivalent to x being a convex combination of all points in $\boldsymbol{X}$ (King and Zeng 2006, 154–155):

> "if x can be expressed as a convex combination of points in X, then there exists a vector of coefficients $\eta_{n \times 1}$, constrained to the simplex so that $X'\eta = x'$. This last equation contains k linear constraints, each stating that an element (variable) of x is a convex combination of the corresponding elements of rows in X. Combining this with the constraint that the elements of η sum to one, we have a total of $k + 1$ linear constraints in the form of $A'\eta = B'$, where A' and B' are X' and x' with a row of ones added, respectively." Thus, to check if a data point (x) is in the convex hull of a set of data X, they "check for the existence of a feasible solution to the following standard form linear programming problem:
>
> $$\min C'\eta$$
>
> $$\text{s. t. } A'\eta = B'$$
>
> $$\eta \geq 0,$$

[1]Note that $x_{1 \times k}$ is not necessarily the first row of $\boldsymbol{X}_{n \times k}$.

where C is a vector of zeros (so there is no objective function to minimize). Checking whether there is a feasible solution to [this] problem is what all standard LP software does in phase I, and it can be done very efficiently for large k and n" (p. 155).

The second diagnostic test measures how "far" a counterfactual is from the data.[2] Specifically, the test measures the fraction of observations (rows) in $\boldsymbol{X}$ "near" the counterfactual x. This test is adapted from Gower's (1966, 1971) coefficient of similarity and is defined as $G_{ij}^2 = \frac{1}{K}\sum_{i=1}^{k} \frac{|x_{ik} - x_{jk}|}{r_k}$, where x_i and x_j are two points (rows) in $\boldsymbol{X}$ and r_k is the range of the kth explanatory variable.[3] "If G^2 is 0, the [query] point and the row in question of X are identical" (p. 137). The larger the G^2, the more different the two rows are and the further away from the data we are extrapolating. Gower's measure identifies extrapolation by measuring the distances between the query point (the counterfactual) and each row in $\boldsymbol{X}$.

King and Zeng's (2006, 138) rule of thumb "in defining observations that are sufficiently close to the counterfactual to make for reasonable inferences is to use the fraction (or number) of observations in the data with distances (values of G^2) less than the 'geometric variability' (GV) of X—which is roughly the average distance among all pairs of observations in the data." Data further away than one GV from the counterfactual are "too far" for reasonable inferences and should be discarded.

3 Should We Rely on the Convex Hull Test to Evaluate Counterfactuals?

Angrist and Krueger (2001, 14) explain why we should care about extrapolation: "Since the sample size and range of variability in many empirical studies are quite limited, extrapolation to other populations is naturally somewhat speculative and often relies heavily on theory and common sense. (A fertilizer that helps corn to grow in Iowa will probably have a beneficial effect in California as well, though one can't be sure.)" This problem is clear, but what is not clear is how the diagnostics described in the previous section are related to what we actually care about: identifying the region of common support in the data so we can make reasonable inferences.

Counterfactuals should be outside the convex hull of the factuals if the support of the distribution of the covariates for the treatment group differs from that of the control group. With nonoverlapping densities causal inferences will involve extrapolation. The key question here is how we should determine if we are making inferences off the common support. We argue that the diagnostics that we evaluate do not reliably identify the region of common support in many reasonable applied situations. We show in subsequent sections that the convex hull test will often find counterfactuals outside the hull even in cases where there is overlapping support in the distributions of the treatment and control groups and each treated observation could be matched to an observation in the control group. The convex hull test is too conservative and will lead researchers to discard data that would have made estimation more efficient without necessarily increasing the risk of extrapolation bias.

Our claim is supported by evidence from Monte Carlo experiments. We use a deliberately simple setup, simulating data from multivariate distributions and applying the diagnostics to those data. The simplicity of this setup allows us to avoid the criticism that our

[2] The convex hull test is a "simplified," "dichotomous criterion" of the measure of distance between two points in $\boldsymbol{X}_{n \times k}$. Thus, if a point is outside the convex hull of $\boldsymbol{X}_{n \times k}$, it is judged to be "too far" from the data for valid counterfactual inference.

[3] So, the measure gives "the average absolute distance between the elements of the two points, divided by the range of the data" King and Zeng (2006, 137). The range "$r_k = \max(X_{.k}) - \min(X_{.k})$, and the min and max functions return the smallest and largest elements, respectively, in the set including the kth element of the explanatory variables X" (p. 137).

conclusions are too specific or that they depend on the design of our experiments. The data are simulated so that we know a priori that there is no risk of extrapolation bias, given a sufficient number of observations. Furthermore, our simple setup allows us to avoid complications that might arise from selection, omitted variables, or endogeneity bias in real observational data. For illustrative purposes, we also provide a simple example of our argument using real data from a study on civil war occurrence. But our argument rests on our simulated data, which isolate the question of extrapolation detection and allow us to see if the diagnostics are equally reliable for different n and k.

Our evidence leads us to the conclusion that as the number of variables (k) grows and/or the number of observations (n) declines, the diagnostics will lead the researcher to infer that the data are likely to suffer from a high risk of extrapolation bias although this does not have to be true (in our setup, by construction, this is not true). We call this the "n/k problem." We also present a theoretical probabilistic argument that explains why the n/k problem affects the convex hull test.

3.1 *Relationship between the Convex Hull Test and Common Support*

We design simple Monte Carlo experiments to investigate how the convex hull diagnostic performs in situations where we know ex ante that the data come from the same distribution support. The basis of the claim that the convex hull test can detect the risk of extrapolation bias in causal inference is that if there is no common support in the distributions of the treatment and control group, then the convex hull test will find counterfactuals outside the hull of the factuals. To investigate this further, we construct a data set that should not raise concern over extrapolation bias. The data set has 11 variables and 1000 observations. The first 10 variables are drawn randomly from a multivariate standard normal distribution and are uncorrelated. The 11th variable is a zero-one dichotomous variable drawn randomly from a binomial distribution with probability $p = .5$ of drawing a 1 (we later investigate the robustness of our conclusions by experimenting with different values for p). A reasonable simulation, for example, results in 490 ones and 510 zeros.

The fact that the data are randomly generated from the same distribution means that they are on common support. We can check in this specific simulated data set if there is balance by estimating the propensity score for the treatment and plotting the distributions of the propensity score for the treatment and control group, checking the degree of overlap in the distributions of the treatment and control groups. The propensity score can be estimated via logistic regression, and the balancing property is satisfied.[4] We show in Table 1 the inferior bound of the propensity score (that bound is identified by the matching program), the number of treated observations, and the number of controls in each block of the propensity score.

Since we have drawn the data randomly, we know that there are no specification problems in the propensity score equation, so there is no concern that the balancing test is unreliable.[5] The usefulness of this approach is that the propensity score summarizes

[4]The propensity score and balancing test are based on Becker and Ichino's (2002) program in *Stata* 8.0. The density plots are generated in *Gauss* 5.0.

[5]Our replication file includes balance statistics on the covariates, obtained using GenMatch (Sekhon 2007). The bootstrapped p value for the univariate Kolmogorov-Smirnov (KS) test is .02 for one covariate (1000 bootstraps). For all other variables in the raw data, we cannot reject the null hypothesis of no significant difference between treated and controls. Having imperfect balance in one covariate does not affect the convex hull test, however. In another $n = 1000$, $k = 11$ example generated the same way, all covariates are balanced (minimum bootstrapped p value for the KS test is .21) and results from the convex hull test are consistent with the ones presented in the text (14.6% of counterfactuals inside the hull; see replication file).

Table 1 Propensity score using our simulated data

	Treatment		
Inferior of block of p score	*0*	*1*	*Total data points*
.2	30	18	48
.4	293	245	538
.5	179	219	398
.6	8	8	16
Total	510	490	1000

relevant aspects of the multidimensional covariate space. Thus, in Fig. 1 we can see that there is overlap in the densities of the treatment and control group (the x axis of Fig. 1 gives the probability of assignment to treatment, given the set of covariates Z). This, in turn, means that there should be no concern over extrapolation bias in these data since extrapolation is a problem that arises "when there are certain values of [the covariates] that some members of one group take on with positive probability but no members of the other group possess" (King and Zeng 2006, 149).[6] By construction, we are not concerned with that problem in this data set.

Since Fig. 1 suggests that counterfactual inferences from this data set should be empirically plausible, it follows that counterfactuals from these data should be inside the convex hull of the factuals if the convex hull is intended to provide an evaluation of the plausibility of counterfactuals. Nevertheless, even for this fairly sizable data set (n = 1000), the convex hull test finds only 13.6% of the counterfactuals inside the hull and only an average of 2.7% of the data are "nearby" the counterfactuals.[7] The implication is that counterfactual inferences from these data are risky (model-dependent) extrapolations.

Drawing 100 observations from these factuals and repeating the convex hull test for the new set of counterfactuals, we find that only 1% of the counterfactuals are inside the hull and 2.9% of the data are nearby the counterfactuals.[8] By contrast, if we limit the number of covariates to 1 plus the treatment, we find 96% of the counterfactuals in the hull with n = 100. This example illustrates the n/k problem. We investigate this finding further and reinforce it with results from Monte Carlo experiments that we report next.

3.1.1 Monte Carlo results on the effects of the n/k problem

The experiments investigate how the diagnostics behave as we vary the number of observations (n) and number of variables (k) with randomly generated data. For each combination of n and k, we simulate 100 different data sets and report the average statistics for the diagnostics. In the first set of experiments, we report the percent of the counterfactuals that are inside the hull. The data sets are randomly generated: all k variables except the last one are drawn from the multivariate standard normal distribution (with no correlation

[6]In practical applications, the propensity score equation may be misspecified, which could invalidate the balance test. This problem does not arise in our setup since we have generated the data, so we can argue that we know the true propensity score model.

[7]Results presented here are based on the "What If" R-based software (version 1.4-6) of King and Zeng. In our replication folder, the factuals are "a1_1000_fact.txt." The treatment is the 11th variable. The counterfactual data are "a1_1000_count.txt." The first 10 variables are identical to the factual data. The counterfactual treatment is defined as 1 − var11 in the factual data.

[8]In our replication folder, factual data are "a1_100_fact.txt" and counterfactual data are "a1_100_count.txt."

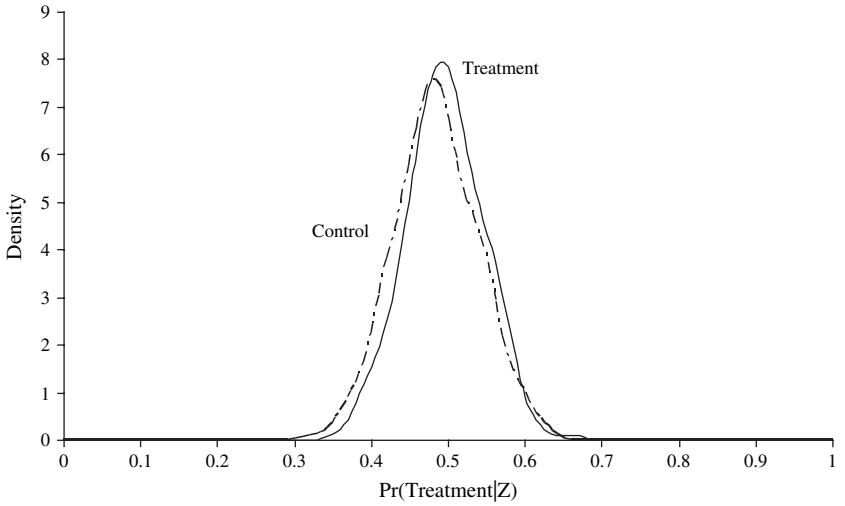

Fig. 1 Propensity score density estimates for treatment and control groups.

between them). The last variable is a binary treatment that is drawn from the binomial distribution with probability .5 of drawing a 1. As before, counterfactuals are defined as 1 minus the treatment, leaving all other variables unchanged. Figure 2 presents the results from 100 different Monte Carlo experiments.

The *y* axis gives the percentage of counterfactuals inside the hull and the *x* axis gives the number of observations. We report results for $k = 2, 3, 4, 6, 9$, and 11. The average percentage of counterfactuals inside the hull over 100 experiments is clearly very small (near zero) when the number of variables is large relative to the number of data points (see results for $k = 9$ and 11 when $n = 100$). As n gets larger, more counterfactuals will be found inside the hull for all k, but the percentage of counterfactuals inside the hull increases at different rates for different k (compare the slopes of the lines for $k = 2$, 6, and 11, e.g.).

Given that these mistaken inferences are made based on our simulated data that are superior to real-world political science data, these results demonstrate that if we were to apply the convex hull test to actual data sets in political science, we could find counterfactuals outside the hull simply because *of the small number of data points relative to the large number of variables even if there is no reason to be concerned with bias due to extrapolation or any other source of inferential bias*. Our simple Monte Carlo experiments ensure that extrapolation *bias* is not a risk in the data by construction, yet the convex hull diagnostic flags this only for cases that, arguably, are empirically uninteresting ($n > 500$ and $k < 5$).

The results in Fig. 2 are due to the sparseness of high-dimensional space, which makes the convex hull test too conservative a test of common support. This is true especially when all variables are continuous, as in our examples above. However, the *n*/*k* problem also applies to binary data. Although both binary and continuous data demonstrate our point, the continuous data example is more relevant for our argument since it is more likely to find mixed or continuous variable data sets in political science than it is to find data sets with all binary variables.[9] When all the data are binary, the *n*/*k* problem should be less severe because the shape of the convex hull is much less complex.

[9]The *n*/*k* problem is more pronounced for data with a greater mix of binary and continuous variables as compared to data sets with only binary variables, though the problem also applies to binary data. See Appendix, Figs. A1 and A2.

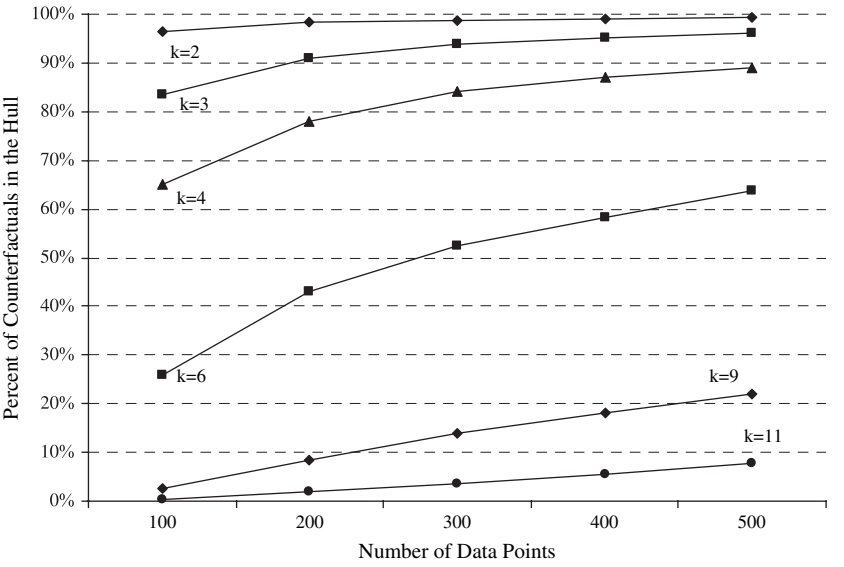

Fig. 2 Counterfactuals in the hull for different k (variables) and n.

A simple example in two-dimensional space illustrates the effects we discuss. In Fig. 3a–3f, we generate two variables drawn from the Uniform distribution on the unit square, making the latter the domain of the distribution. We plot the convex hull as we increase N from 10 to 40 and increase the correlation (rho) between the two variables from 0 to 0.8 to 1.[10] It is clear that the convex hull area becomes smaller relative to the area of support as N is reduced from 40 to 10 (compare Fig. 3a and 3b). Similarly, the convex hull is smaller when the two variables are correlated (contrast Fig. 3b to 3d or Fig. 3a to 3c). In the extreme case where correlation is 1 (Fig. 3e and 3f), the area of the convex hull is the area of the line connecting the points on the two variables. We probe the effect of correlation among the $\boldsymbol{X}$s further in the next section with a new set of Monte Carlo experiments.

3.1.2 Monte Carlo results on the effects of correlation among the $\boldsymbol{X}$s

The data used in the previous set of Monte Carlo experiments were constructed without any correlation. In most practical applications in social science, however, analysts work with explanatory variables that are correlated. Correlation between the treatment and the rest of the data will have profound effects on the degree of detected extrapolation risk. This should be obvious, but what is not obvious is how the n/k problem affects the detection of extrapolation in the presence of correlated variables.

To explore this question, we conduct another set of Monte Carlo experiments. We draw $k = 7$ variables: the first five are drawn from a multivariate normal distribution and the last two are binary variables drawn randomly each with probability $p = .5$ to draw a 1. We impose an arbitrary correlation (rho) between the binary variables. One of the binaries is the treatment, and the counterfactual is again defined as 1 minus the treatment. In Fig. 4, we plot the average fraction of counterfactuals that are inside the convex hull for n from 100 to 300 with different degrees of correlation (rho varies from 0 to 0.8).[11]

[10] We simulate correlated standard uniform variables by first simulating correlated standard normal variables and then using the cumulative distribution function of the standard normal to get a number in the [0,1] interval.

[11] In unreported simulations, we find that correlation among the first five variables does not affect the results of the diagnostics.

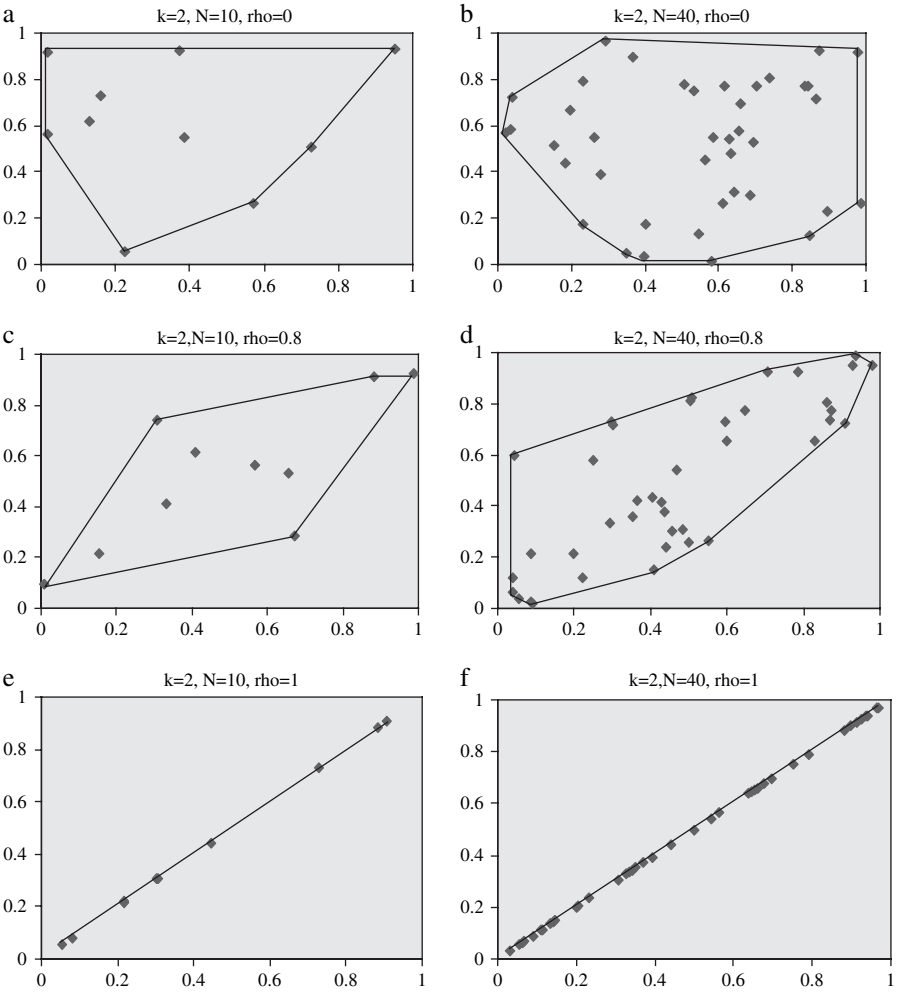

Fig. 3 Convex hull area relative to the area of common support.

It is clear from the downward slope of these curves that correlation between the treatment and another variable reduces the number of counterfactuals in the hull. But, for $n = 100$, high correlation (rho $= 0.8$) does not have a big effect on the results of the convex hull test because the *n/k* problem dominates the results (we find very few counterfactuals in the hull even for rho $= 0$ when $n = 100$). The effect of correlation is much more pronounced as the *n/k* problem becomes smaller (see results for $n = 300$ in Fig. 4).

We also see that simple correlation is sufficient to take counterfactuals outside the hull even in data sets that are not affected by the *n/k* problem. This implies that there is a trade-off between extrapolation bias and omitted variable bias. If researchers add controls to a regression model to reduce the risk of omitted variable bias in estimates of causal effects, they will inevitably increase the risk of extrapolation bias. This trade-off between extrapolation and omitted variable bias limits the usefulness of the diagnostic in practical applications since, in the absence of a theory that claims to present the right model, most scholars will add controls to check the robustness of their findings. Yet, doing so will inevitably result in more extrapolation. A counterargument might be that control variable

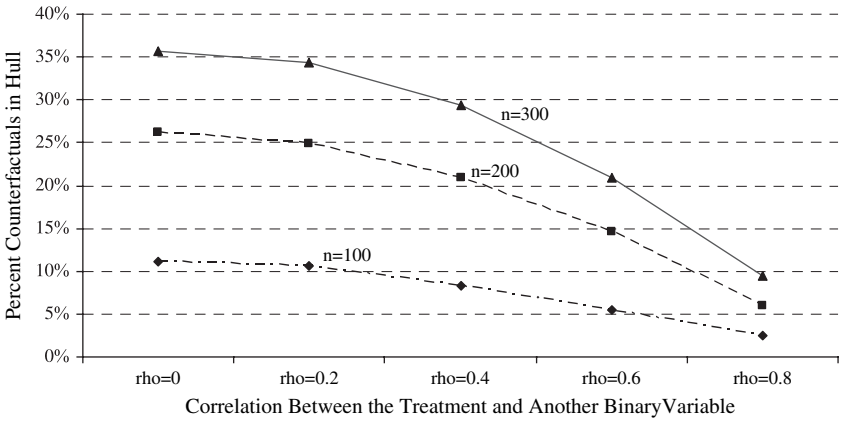

Fig. 4 Average percentage of counterfactuals in the hull for different n and rho ($k = 7$).

selection must be done before the convex hull test is applied since the diagnostic can only be valid for a given set of covariates $\boldsymbol{X}$. However, in applied settings, researchers will not be certain that they know the true model, so they will face this trade-off as they try different controls, which will change $\boldsymbol{X}$. If researchers knew the true model, they would not be concerned with extrapolation in the first place.

The intuition that underlies the results from our simulations also applies to real data. As we add standard controls to a regression, we will find more counterfactuals outside the hull partly because of the n/k problem. We illustrate how the n/k problem affects real data analysis using a cross-country data set on civil war occurrence. Data for our example come from a study of the effects of ethnic polarization on the risk of civil war occurrence (Montalvo and Reynal-Querol 2005). The authors use a data set that codes all civil wars that took place in all countries of the world at any time during the period from 1960 to 1999. There are 90 observations with nonmissing data in the cross-sectional data set. The authors' favorite model regresses civil war occurrence on per capita income, population size, primary commodity exports, mountainous terrain, a dummy variable for countries with large noncontiguous territories, ethnic polarization, and a dummy variable for democratic regimes (results are presented in Model 1, Table 6 in Montalvo and Reynal-Querol 2005, 812).

We use this data set to consider counterfactuals about democracy. The counterfactual question is what would happen to civil war risk if all democracies became nondemocracies and vice versa. We check how adding one control variable to the model affects results from the convex hull test. Using the data from the original model, we find that nine counterfactuals (10%) about democracy are inside the hull of the observed data. When we add a control variable—a regional dummy for the seven East Asian countries in the data set—we see a 44% drop in the number of counterfactuals in the hull.

This result could be due to the n/k problem but also to the properties of the data, if East Asian countries are significantly different from other countries with respect to the other covariates. We could add other controls for possibly unobserved regional effects, but important regional differences across some of the covariates are likely to emerge, thereby complicating the interpretation of the convex hull test results. East Asian countries were only different from the rest with respect to the presence of noncontiguous territories.[12]

[12]Equality of means tests shows significant differences between East Asian and other countries with respect to noncontiguity but not other variables at standard significance levels (the lowest p value is .11 for the log of population size; it is .06 with a one-tailed test; p values are much higher for the other covariates).

Countries with large territories that are separated from the mainland may face higher risks of rebellion in their far-flung regions at least in theory. But noncontiguity is not significant in any of the regressions, so we could drop it and rerun the test. We now find 11 out of 90 counterfactuals in the hull, and adding the East Asian dummy again results in a 45% drop in the number of counterfactuals in the hull (six democracy counterfactuals are inside the hull).

This example shows that as we add covariates to reduce the risk of omitted variable bias or to check the robustness of our results, the convex hull test will lead us to the conclusion that a significant number of cases ought to be dropped. Dropping these cases would deprive analysts of valuable information that could improve their estimates. We show this in the next section, where we study how dropping cases outside of the convex hull affects estimates of average treatment effects (ATEs).

3.1.3 Relationship between the convex hull test and unbiasedness of ATEs

Our discussion of trade-offs that applied researchers are likely to face leads us to consider the relationship between membership in the convex hull and the unbiasedness of ATEs, which is the quantity of interest in causal inference. The convex hull test is intended to help researchers identify regions of common support so as to improve estimates of ATEs. For instance, Ho et al. (2007, 27) propose dropping all observations that are outside of the convex hull before estimating the model.[13]

However, we showed in previous sections that the convex hull test is simply too conservative and will lead researchers to discard data that are actually on common support. We now show that dropping observations outside of the hull might actually impede causal inference because being outside the hull does not necessarily bias causal inferences and dropping counterfactuals that are outside the convex hull of the observed data will inevitably increase the uncertainty around ATE estimates.

We demonstrate this by extending our previous Monte Carlo experiments. We now draw a dependent variable and estimate ATEs for a binary treatment variable. Keeping the analysis intentionally simple to avoid any model-dependence criticisms, we generate a dependent variable by a simple ordinary least squares routine based on our generated control and treatment variables. Specifically, we generate Y_i as $Y_i = x_i\beta + \gamma T_i + \sigma\varepsilon_i$, where x is a $1 \times k$ vector of continuous independent variables, T is the binary treatment variable, ε_i comes from the standard normal distribution, and σ is chosen to determine the fit implied by the model ($\sigma = 0$ implies a perfect fit).[14] We set for simplicity all coefficients (β and γ) equal to 1 and investigate how inference changes when we perform the estimation based on all the data as compared to only those observations that are found to be inside the convex hull of the simulated data. Table 2 reports results from 500 Monte Carlo experiments.

For $n = 200$, $k = 3$, and a model R^2 equal to 0.5, an average of 91% of counterfactuals are inside the hull (i.e., 91% of the data are on common support according to the in-hull criterion).[15] Based on these in-hull observations, we estimate γ to equal 0.9999 with an SE

[13] King and Zeng (2006, 153) write that to "reliably estimate [a model] without high levels of model dependence, we would also want to drop treated units that fall outside the convex hull of the control units."

[14] We pick σ to generate an "intended" adjusted R^2 equal to a specified number. In our baseline example in Table 2, this is set equal to 0.5, but we experiment with different σ in our replication file. For every Monte Carlo experiment, we pick the slightly different σ that will generate the intended adjusted R^2 through a bisection method (there is a one-to-one relationship between the adjusted R^2 and σ and, given that we want to perform experiments with different number of continuous independent variables (k), we must ensure that the adjusted R^2 stays constant as we change k). See our replication file for results with adjusted R^2 of 0.2 and 0.8.

[15] We do not report results with $n = 100$ because for $k = 3$ or lower or 7 and higher, over many Monte Carlo draws, we get fewer counterfactual data points in the hull than k and cannot run the regression. We therefore increase n and report averages over Monte Carlo experiments where this feature of the n/k problem does not arise.

Table 2 Relationship between convex hull membership and bias in ATE estimates (results based on 500 Monte Carlo experiments)

	All data	*In-hull*	*Outside-hull*	*All data*	*In-hull*	*Outside-hull*
	200 points, 3 variables[a]			*500 points, 3 variables*		
	Average percentage of counterfactuals in hull = 91%			*Average percentage of counterfactuals in hull = 96%*		
Coefficient estimate	1.0007	0.9999	1.0159	1.0004	0.9995	1.0351
SE	0.1531	0.1602	0.6145	0.0956	0.0975	0.5495
MSE	0.0263	0.0278	0.4174	0.0086	0.0089	0.3606
Model adjusted-R^2	0.5000	0.4472	0.7539	0.5000	0.4708	0.7965
	200 points, 5 variables			*500 points, 5 variables*		
	Average percentage of counterfactuals in hull = 61%			*Average percentage of counterfactuals in hull = 77%*		
Coefficient estimate	0.9985	0.9935	1.0117	1.0001	1.0011	0.9971
SE	0.2114	0.2741	0.3499	0.1316	0.1501	0.2847
MSE	0.0496	0.0907	0.1357	0.0166	0.0234	0.0823
Model adjusted-R^2	0.5000	0.3797	0.6182	0.4999	0.4193	0.6644
	200 points, 7 variables			*500 points, 7 variables*		
	Average percentage of counterfactuals in hull = 27%			*Average percentage of counterfactuals in hull = 48%*		
Coefficient estimate	1.0002	1.0055	1.0034	1.0005	1.0033	1.0013
SE	0.2584	0.5229	0.3059	0.1604	0.2344	0.2244
MSE	0.0725	0.2885	0.1045	0.0242	0.0535	0.0500
Model adjusted-R^2	0.5000	0.3130	0.5491	0.5000	0.3681	0.5805

[a]Results for $n = 200$ and $k = 3$ are based on 435 Monte Carlo experiments.

equal to 0.1602, whereas the estimate based on the whole data set equals 1.0007 with an SE equal to 0.1531. The difference between the two coefficients is not statistically significant, whereas the SE from the smaller sample is slightly larger, resulting in a slightly higher mean squared error (MSE) for the in-hull model. But since most of the counterfactual data points are in the hull because of the favorable n/k ratio, we would not expect much of a difference in the model MSEs.

For $n = 200$ and $k = 5$, more interesting differences emerge because now only 61% of the data set is found to be in the region of the common support. Hence, we must discard a larger proportion of the data. Based on this much smaller data set, we find that the mean estimate of γ based on the in-hull observations (0.9935) is again not very different from the estimate using the whole data set (0.9985) or the out-of-hull data set (1.0117, with an SE of 0.3499). What is different is the SE of the coefficient, which now rises to 0.2741 (compared to 0.2114 for the whole data set). This difference is as it should be and arises simply because we are using a smaller data set.[16] Accordingly, the MSE is higher now (0.0907 versus 0.0496) and is driven by the increase in the variance of the estimator due to the

[16]According to $\sqrt{n}$ asymptotics, with approximately half the sample we should be getting SEs that are approximately $\sqrt{2}$ higher than the ones based on the original sample.

smaller sample size. The coefficient estimate is not statistically different for data points inside and outside the hull. The difference in the model MSEs gets even larger for $n = 200$ and $k = 7$ with 27% of counterfactuals in the hull, increasing from 0.0725 for the whole data set to 0.2885 for the in-hull only data and is actually lower for the model estimated only on data outside of the hull (0.1045). The same picture emerges for larger n. With 500 data points and 7 variables (and 48% of counterfactuals outside the hull), the estimate of γ is 1 for all three regressions (using all the data, in-hull data only, and outside-the-hull data only), but the SE of the estimate is actually higher, resulting in a higher MSE for the in-hull model.

In all our simulations, the MSE of the model estimated on all the data is always smaller than the MSE of the model estimated on in-hull data only, so we would be better off using all the data to estimate causal effects. The results are robust to experiments with different adjusted R^2 in the original regression.[17] We conclude that for the causal effect of interest to applied researchers, using the convex hull to eliminate observations from the analysis can substantially decrease estimator efficiency simply because the convex hull criterion is too conservative a way to ascertain the region of common support in a data set.[18] Moreover, making inferences outside the hull does not necessarily cause bias in regression estimates.

3.2 *Theoretical Motivation for the n/k Problem*

The experiments presented above suggest that the probability that a data point is inside the hull of a set of points is always extremely small and that the number of data points (n) must increase exponentially in the number of dimensions (k) for that probability to be reasonably high. This is due to the complex geometry of multidimensional space and is unrelated to nonrandom assignment, selection, functional form, or other modeling assumptions.

This argument relies on a general result by Elekes (1986) who showed that one cannot approximate the volume of any convex body in the k-dimensional Euclidean space by the convex hull of a "small" number of points.[19] One interpretation of Elekes's proof is that the probability that a uniform random point in a unit k-dimensional hypercube will hit the convex hull of fixed n points in dimension k is the minimum of 1 and $n/(2^k)$.[20] To have a reasonably high probability that a random point in unit k-dimensional hypercube will hit the convex hull of n extreme points of the hypercube, the number of points n must be exponential in k. Given the above, it follows that it is not reasonable to apply the convex hull test to a small data set because the query point will almost certainly be outside the hull and we will not know if this is simply because of the small number of data points relative to the number of variables or other features of the data that might bias counterfactual inferences. Moreover, if extrapolation is defined as making inferences about points that are outside the convex hull of the factuals, then any method that is applied to high-dimensional

[17] As the adjusted R^2 gets larger, all estimates (in-hull and out-of-hull) are closer to the true γ. But the MSE using all the data is always smaller than the MSE using in-hull data only. See our replication file, Table A1, for results for adjusted $R^2 = 0.8$ and Table A2 for results for adjusted $R^2 = 0.2$.

[18] In political science, the quantity of interest is typically the base effect of the treatment. In some applications, the kth-order interaction effect of the treatment may also be of interest and it may be the case that the convex hull test is informative for that effect. This could be investigated with a modification of our code.

[19] Professor Komei Fukuda pointed this out. Our argument is closely related to a more general argument about the likelihood that a random point in a given convex body will hit the convex hull of n points from the given convex body. We can consider the given convex body as the space of all possible candidates of factual and counterfactual data sets. Elekes's result shows that the probability in question is very small unless n is exponentially large in k.

[20] For completeness, we state the actual context of Elekes's assumptions and theoretical result. Let S be a ball in the k-dimensional Euclidean space with volume 1. Choose any n points $\{P_1, P_2, \ldots, P_n\}$ in S. Denote by C_n the convex hull of $\{P_i; 1 \leq i \leq n\}$. What is the upper bound of the volume C_n maximized over all possible set of points? The upper bound of this volume is $n/(2^k)$. Note that this is a conservative upper bound and that, when this is less than 1, this volume can be interpreted as the probability that a random point may lie in C_n.

data sets will have to rely on extrapolation to some degree even if we have 1000 or more observations.

As a consequence of Elekes's theorem, if $n = 500$, $k = 11$, and an equal number of points get 1 and 0 in the treatment variable, then the *upper bound* of the probability that the counterfactuals will be inside the factuals' convex hull is 0.244, whereas for the same number of dimensions k and only 100 points, the upper bound is 0.0488. Equivalently, note that if $n = 500$ and $k = 2$, the upper bound of the probability is given by the minimum of 1 and 125; in this case, a researcher knows with probability 1 if a point is in the convex hull in the two-dimensional case with 500 points. In fact, the intuitive result arises that for $k = 1$, a researcher needs only two points to reach the probability's upper bound: a given point can be determined to be inside or outside the line linking two points with probability 1, whereas with $k = 2$, the same upper bound is reached with $n = 4$. In the simulations that we presented above drawing data from the normal distribution, we found that only about 13.6% of the counterfactuals were inside the hull with $n = 1000$ and $k = 11$, whereas the number dropped to 1% when $n = 100$.

These Monte Carlo results are consistent with Elekes's general result because they show that, for a given k, the probability of hitting the convex hull decreases exponentially as n increases, and this arises simply from the mathematics of n-dimensional geometry.[21] To see how Elekes's theorem applies to our simulated data, suppose that we have two sets of factual data, the first set with 0 in the first component and the second set with 1 in the first component:

$$(0, a_1), (0, a_2), \ldots, (0, a_M),$$

$$(1, b_1), (1, b_2), \ldots, (1, b_N),$$

where a_is and b_js are vectors of dimension 1 less than the dimension of factuals and M and N indicate the number of 0's and 1's in the data (the number of rows) and a and b have the same number of columns. The convex hull test takes, for example, $(1, a_1)$, which is the counterfactual of point $(0, a_1)$, and asks if it is in the convex hull of all the factuals. Such a test is equivalent to testing whether a_1 is in the convex hull of $b_1, b_2, \ldots, b_N$, simply because no $(0, a_i)$s can be used to represent $(1, a_1)$ as a convex combination.

Proof:[22] Let $(1, a_1)$ be in the convex hull of $(0, a_i)$s and $(1, b_j)$s, namely,

$$(1, a_1) = p_1(0, a_1) + \cdots + p_M(0, a_M) + q_1(1, b_1) + \cdots + q_N(1, b_N), \tag{1}$$

for some nonnegative p_is and q_js such that $p_1 + \cdots + p_M + q_1 + \cdots + q_N = 1$.

The first component of equation (1) says

$$1 = q_1 + \cdots + q_N. \tag{2}$$

This together with $p_1 + \cdots + p_M + q_1 + \cdots + q_N = 1$ implies

$$p_1 + \cdots + p_M = 0. \tag{3}$$

Since all p_is are nonnegative, this implies all p_is are zero. This proves that no $(0, a_i)$s can be used to represent $(1, a_1)$ as a convex combination.

[21] In Fig. 2, the ratio of points in the hull over n is less than $n/(2^k)$ (where k is dimension 11 in this case). It is in fact much less than that since the threshold we gave is an upper bound.

[22] We thank Komei Fukuda for providing us with this proof.

If 0s and 1s are values of a binary "treatment" variable and a_i and b_j are vectors with continuous data (as in the data we have used in our simulations), then our proof shows that the convex hull test amounts to checking if vectors a_is are in the convex hull of the b_js. Elekes's result therefore applies (since a_i and b_j are continuous data) and suggests that the probability that a randomly chosen a_i point is in the convex hull of b_js would be very small if a_is are independent of b_js (i.e., if there is no correlation between them).[23]

3.3 *Monte Carlo Results on the Gower Distance Test*

Because a data point may lie just outside the hull, but closer to a region of the data than another point that is inside the hull but near an empty region of the data, a second diagnostic may complement the convex hull test. The Gower distance test (G^2) reports "the distance between the two points as a proportion of the distance across the data, $\boldsymbol{X}$.... A distance between two points of $G^2 = 0.3$ means that to get from one point to the other, we need to go the equivalent of 30% of the way across the range of the data set" (King and Zeng 2006, 140).

But how can we tell how far is too far to take a counterfactual inference? The proposed rule of thumb is to use the data set's GV as the threshold of nearness for reasonable empirically based inferences. If most of the data lie beyond one GV from the counterfactual point, this is going too far. Yet, there is no inherent reason why one GV is the right threshold to use in all cases (e.g., higher thresholds can be used if we have more confidence in the model), nor is there a clear way to establish how far is too far in counterfactual inference because the Gower distance test is also affected by the n/k problem. We use Monte Carlo experiments to demonstrate this point and highlight one argument: the distance test is affected by the distribution of 1's and 0's in the treatment and this is unrelated to modeling assumptions or other characteristics of the data that could cause inferential bias (including extrapolation bias). Our experiments investigate the sensitivity of the Gower distance test to the skewness (p) of the distribution of the treatment variable while keeping n and k constant.

Figure 5 shows the results of 100 runs of a Monte Carlo experiment with $n = 100$ and $k = 6$. On the x axis, we plot the percentage of observations assigned to the treatment in our simulated data, and on the y axis we plot the average percentage of the data that are nearby the *counterfactuals* or the other *factuals*.[24] For any size n, as we increase k, there will always be fewer data nearby the counterfactuals.[25] With an even chance of being assigned to treatment, Fig. 5 shows an average of 10% of the data nearby the counterfactuals. Is this enough for valid counterfactual inference? When the treatment is very skewed (0.9), the fraction of data nearby the counterfactuals declines to about 2%. Although the Gower distance tells us the fraction of the data nearby the counterfactuals, that measure in and of itself cannot tell us much since we would need to know ex ante how many data points must be near the counterfactuals to make empirically valid inferences.

In Fig. 5, we also show the fraction of the data nearby other factuals. It should be clear that, in sparse data sets, that fraction can be low. What do we make of that? Another rule of thumb might be that "good" counterfactuals should have an amount of data nearby that are not too far from that for the factuals. Good counterfactuals would be those that do not require extrapolation.

[23]No correlation implies that a_is and b_js are chosen independently and the probability that any given vector $\boldsymbol{x}$ is chosen as a_i is equal to that of x being chosen as b_j, for any i, j. This would be consistent, for example, with a random assignment of the treatment to a_is and b_js drawn from the same distribution.

[24]If we keep p constant at .5 and vary n from 100 to 500, the average percentage of the data that are near the factuals or the counterfactuals is fairly constant regardless of the size of k. The average percentage of data that are nearby either the factuals or the counterfactuals is several times higher for each n when k is low. See our replication folder for the output from several experiments.

[25]See replication file for results from $n = 100$ to 500.

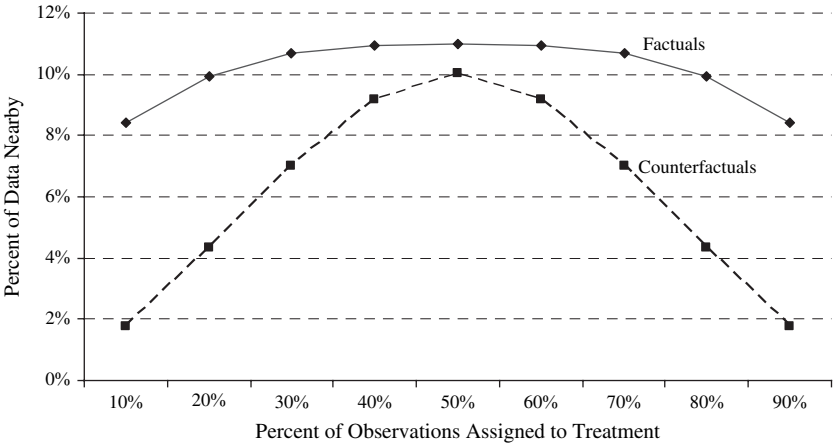

Fig. 5 Percent of the data "nearby" as we vary the skewness of the treatment ($n = 100$, $k = 6$).

The average fraction of the data nearby the factuals will be near the fraction of data nearby the counterfactuals for different n and k as long as there is an even split between 0's and 1's in the treatment (for $n = 500$, $k = 11$, and $p = .5$, that fraction is 2.8%). It is apparent from Fig. 5 that as the distribution of the treatment becomes more skewed, there is divergence between these fractions.[26] The divergence gets larger if there is correlation between the treatment and another explanatory variable, even if the skew of the treatment remains unchanged.[27]

This result is also sensitive to the *n*/*k* problem. With the same highly skewed treatment ($p = .9$) and same n (100), the ratio of the average fraction of the data nearby the other factuals over the average fraction of the data nearby the counterfactuals *decreases* from 4.8 to 1.7 as we increase k from 6 to 11 (there is less divergence, but we should be making the problem of extrapolation worse by adding covariates). The divergence *increases* from 1.7 to 3.5 as we keep k constant at 11 and increase n from 100 to 500. With a skew of 0.9 and a high k (11), the fraction of data near the counterfactuals is lower for large n (0.005 for 500 points as compared to 0.013 for 100 points), even though we saw earlier that extrapolation becomes less problematic with larger n. If the treatment is not skewed at all, the fraction is the same for $n = 100$ and $n = 500$. Thus, the problem is the skewness of the treatment. But any systematic differences between the treatment and control groups should become *smaller* with random data as n gets larger.[28] The "good counterfactual" rule of thumb based on the comparison of the fraction of data nearby the factuals and counterfactuals points to a different direction than the convex hull test as n increases from 100 to 500. The results discussed here suggest that, in any given data set, there is a mechanical "upper bound" for the percent of the data that can be nearby the factuals relative to the counterfactuals and that upper bound is a function of the skewness of the treatment.

[26]The results of the convex hull test are also affected by the skewness of the treatment (with $n = 100$ and $k = 6$, five times more counterfactuals are inside the hull when $p = .5$ as compared to $p = .1$).

[27]Results not shown. With a correlation rho = 0.8, $n = 100$, $k = 7$, and no skew of the treatment ($p = .5$), there are 3.9 times more data nearby the factuals than there are nearby the counterfactuals. With the same rho, p, and k, if $n = 300$, that ratio increases to 4.5. See replication folder for these results.

[28]We show in the replication file (Fig. A3) that, with $n = 1000$, there is near-complete overlap in the distributions of the propensity score for the treatment and control groups in one simulated data set for a highly skewed, randomly assigned treatment ($p = .8$).

We use another Monte Carlo experiment to evaluate this new rule of thumb. Increase n to 1000 observations, set $k = 2$ (one independent normal variable and a randomly drawn binary), and set the skewness of the binary variable to 0.1. A priori we expect that the convex hull test will not detect the risk of extrapolation bias in these data because of the high n/k ratio. Indeed, after 100 runs of the Monte Carlo experiment, the average number of counterfactuals inside the hull is 98.4%. Since there is no extrapolation, these should therefore be good counterfactuals, yet the average percent of data nearby the counterfactuals is 10.6%, whereas it is 47.8% for the factuals.

These simple examples, combined with results presented in previous sections, illustrate that the results from the two diagnostics are contradictory if by good counterfactuals we mean counterfactuals with about the same amount of data nearby as that for the factuals. There is no straightforward way to connect results from the convex hull test and the Gower distance test and that we cannot use the two diagnostics to determine how "far" we can safely take our counterfactual inferences. Before any determination is made about how far we can take a counterfactual inference, we would need to know how theoretically sound our model is and what fraction of the data should be nearby to make reasonable inferences that are not sensitive to specification assumptions. Neither the Gower distance test nor the convex hull test provides us with that information, and interpretations of the two diagnostics are at times contradictory.

4 Conclusion

Inferences that involve extrapolation are likely to be model dependent and, therefore, risky. The diagnostic tools that we have used and evaluated in this paper identify such inferences by determining if counterfactual points are outside the convex hull of the observed data. If they are, the proposed rule of thumb is to drop these cases before estimation to reduce the risk of extrapolation bias.

We make two arguments based on Monte Carlo simulations. First, the convex hull test will find all or most of the counterfactuals outside the hull of the observed data in data sets with a small (but reasonable) number of observations and/or a large (but also reasonable) number of explanatory variables. This is independent of data characteristics that might generate concern over extrapolation bias. The convex hull test is too conservative and will lead researchers to discard valuable data points, reducing their ability to address many important questions in political science.

Second, even though a counterfactual may lie outside the hull, it can still lie within a reasonable distance of the data, but there is no way to know a priori how "far" is "too far" to make extrapolations based on a statistical model. This is because the results of the second diagnostic—the Gower distance test—are affected by, among other things, the distribution of treatments to nontreatments in the data. A skewed distribution will lead the test to detect the risk of extrapolation bias, even when the distribution of treatments to nontreatments is random and skewness in the distribution of the treatment does not diminish the plausibility of a counterfactual comparison (as would be the case with a large n).

Both problems are likely to arise frequently in applied settings and thus limit the usefulness of the diagnostics in detecting the risk of extrapolation bias in statistical inference. Moreover, using the diagnostics to preprocess the data can actually hurt researchers rather than help them. We have shown using simulated data that one can derive unbiased estimates and make unbiased extrapolations even when the convex hull test suggests that there should be significant extrapolation bias. Thus, failing these diagnostic tests is neither a necessary nor a sufficient case for bias in causal inference and the

suggestion to discard data outside of the hull may not only be too conservative but also create unnecessary uncertainty in estimates of causal effects.

The question of interest underlying this analysis is how we can improve causal inference. We have argued that relying on the convex hull to identify and prevent extrapolation is not the best way to improve causal inference. With observational data, more promise may lie in other methods, such as covariate matching as the way to identify observations on common support. Another fruitful road might be to estimate structural models of political behavior and political outcomes.

Appendix

Figure A1 supplements the results reported in Fig. 2 by showing how the *n*/*k* problem affects the convex hull test for binary data. In this experiment, all variables are randomly drawn from the binomial distribution with a .5 probability of getting a 1. We report results from 100 Monte Carlo experiments for 3, 6, 9, and 11 variables (*k*). The *n*/*k* problem applies to these data, too, but is less severe than with the data in Fig. 2.

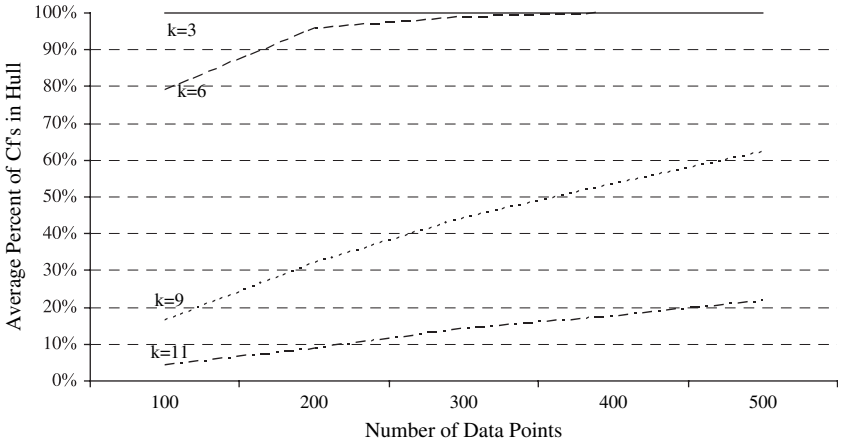

Fig. A1 Convex hull test results for binary data.

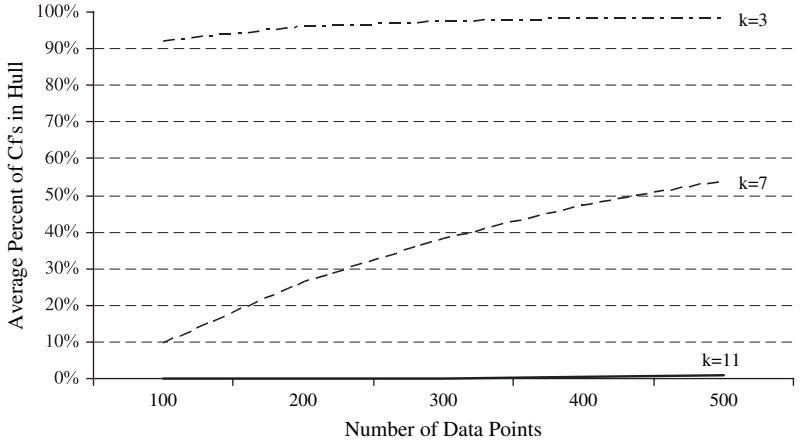

Fig. A2 Convex hull test results for mixed data.

Next, we report results for the same experiments using "mixed" data. Figure A2 supplements the results reported in Fig. 2 by showing how the n/k problem affects the convex hull test for a combination of continuous and binary variables. In this experiment, all binary variables are randomly drawn from the binomial distribution with a .5 probability of getting a 1. There are $k + 1$ such variables including the treatment. There are also k continuous variables, randomly drawn from the multivariate normal distribution. We report results from 100 Monte Carlo experiments for 3, 7, and 11 variables (k) for n ranging from 100 to 500. The n/k problem is more severe with a greater mix of binary and continuous variables.

References

Angrist, Joshua D., and Alan B. Krueger. 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. Working Paper #455. Princeton University Industrial Relations Section.

Becker, Sasha O., and Andrea Ichino. 2002. Estimation of average treatment effects based on propensity scores. *The Stata Journal* 2:358–77.

Elekes, Gyorgi. 1986. A geometric inequality and the complexity of computing volume. *Discrete and Computational Geometry* 1:289–92.

Gower, J. C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* 53:325–88.

———. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27:857–72.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15:199–236.

King, Gary, and Langche Zeng. 2006. The dangers of extreme counterfactuals. *Political Analysis* 14:131–59.

Montalvo, Jose G., and Marta Reynal-Querol. 2005. Ethnic polarization, potential conflict, and civil wars. *American Economic Review* 95:796–816.

Sekhon, Jasjeet S. 2007. *Multivariate and propensity score matching software with automated balance optimization: The matching package for R.* http://sekhon.berkeley.edu/matching.