

# A Rational Approach to Inference on Multiple Parameters

Brigham R. Frandsen

September 17, 2018

## Abstract

The presence of multiple parameters complicates statistical inference. Ignoring the presence of multiple parameters can result in misleading inferences while existing remedies often produce inferences too imprecise to be economically meaningful. This article proposes an approach to inference on multiple parameters that produces valid simultaneous inferences on multiple parameters while maintaining precision on the parameter or parameters of greatest interest. The approach allows inference to reflect differing preferences for precision that the researcher may have across parameters, resulting in hypothesis tests that are more powerful and confidence regions with shorter projections on the parameters that the researcher cares more about, while remaining jointly valid across all parameters. A researcher using the procedure specifies in advance non-negative weights that correspond to the relative preference for precision across parameters. The proposed procedure chooses a confidence region to minimize the weighted sum of the projections on the parameter dimensions. A decision theoretic framework presents axioms for researcher preferences under which the proposed procedure is optimal. An empirical example from a field experiment on charitable giving shows the method offers substantial improvements in real-world settings.

## 1 Introduction

Nearly all empirical studies involve inference on multiple parameters. For example, an experiment may investigate treatment effects on several outcomes or in several subgroups. Sometimes the presence of multiple parameters can be safely ignored for inference purposes and researchers can perform hypothesis tests and construct

confidence intervals in the usual way for one parameter at a time. This will be the case when questions to be answered or conclusions to be drawn from the research depend on only one parameter. This situation is likely rare in economics. More commonly, the research seeks to answer a question whose answer depends on several parameters. For example, whether a particular health or education intervention is cost effective depends on its benefits—often measured in terms of effects on several outcomes—and its costs. Distributional concerns depend jointly on effects on different subgroups or at different points of the outcome distribution. In situations like these, researchers should perform inference jointly over all parameters upon which the question to be answered depends.

The presence of multiple parameters can complicate inference, however. There are two common approaches. The first, and most common, ignores the presence of multiple parameters and performs inference one parameter at a time. Any resulting conclusions that depend on more than one parameter will be misleading more frequently than intended. For example, suppose the purpose of a study is to assess whether some education intervention is beneficial, where benefits might be measured by achievement test scores, graduation, college enrollment, or long-term earnings. Ineffective interventions will be wrongly touted as beneficial more often than intended when outcomes are examined one at a time.

The second common approach recognizes the faults of ignoring the presence of multiple parameters and adjusts significance levels or uncertainty intervals to ensure that joint inference on the several parameters of interest will have certain prescribed properties. For example, [Bonferroni \(1935\)](#) corrections to significance levels ensure that the probability of rejecting any one of a set of true hypotheses stays below some nominal level. This approach has its own set of drawbacks. Often it leads to inferences that are too imprecise to be useful for drawing conclusions or guiding policy decisions.

Precision about any one parameter is sacrificed by simultaneously examining several parameters. The loss of precision involved with accounting for multiple inferences can have an even worse side effect: it increases the temptation to selectively omit results from reported findings, a tendency which has contributed to the ongoing “replication crisis” in social sciences ([Open Science Collaboraton, 2015](#)).

The pitfalls of inference on multiple parameters are not limited to the classical null hypothesis significance testing (NHST) paradigm or frequentist confidence intervals. Inference based on Bayesian testing procedures or credible regions also suffers from the drawbacks associated with either performing inference on one parameter at a time or simultaneously.

I propose an approach to inference on multiple parameters that produces valid simultaneous inferences on multiple parameters while maintaining precision on the parameter or parameters of greatest interest. The approach allows inference to reflect differing preferences for precision that the researcher may have across parameters, resulting in hypothesis tests that are more powerful and confidence regions with shorter projections on the parameters that the researcher cares more about, while remaining jointly valid across all parameters. Under this approach, the researcher specifies in advance weights that correspond to the relative value the researcher places on precision on the parameters. The shape of the confidence region is then chosen to minimize the weighted sum of the projections on the parameter dimensions, subject to the constraint that the confidence region has the required coverage. This paper focuses on confidence intervals, which are often of greatest interest to economists, but the confidence region projections correspond to critical values for each parameter dimension, resulting in tests that are the most powerful for parameters with the highest weights, while remaining jointly valid.

The approach generalizes current standard methods for inference on multiple pa-

rameters, nesting them as special cases. Specifying a positive weight on one parameter and zero on the others produces conventional confidence intervals and hypothesis tests on a single parameter. Specifying equal weights on all parameters is equivalent to a Bonferroni correction if the parameter estimates are uncorrelated.

## 1.1 Relationship to existing methods for multiple inference

The method proposed here builds on a large literature on multiple inference. The familiar [Bonferroni \(1935\)](#) adjustment is one of the simplest and earliest approaches, and provides the basis for many subsequent methods. It is most commonly applied to hypothesis tests, but also lends itself easily to constructing jointly valid confidence intervals. The Bonferroni approach adopts worst-case bounds, and so is quite conservative when estimates are positively correlated. As in the current paper, different parameters can be assigned different weights to direct power to particular hypotheses in what [Miller \(1966\)](#) refers to as *allocation*.

In a refinement of the Bonferroni method, [Holm \(1979\)](#) developed a sequential step-down procedure that yields more powerful tests than Bonferroni, although it, too, adopts worst case bounds and can therefore be quite conservative. Confidence intervals based on the Holm procedure, derived by [Strassburger and Bretz \(2008\)](#) and [Guilbaud \(2008\)](#), are not uniformly shorter than Bonferroni intervals, however, and have other undesirable properties. [Simes \(1986\)](#) and [Hochberg \(1988\)](#) have a related sequential step-up procedure that is even more powerful, but controls size only under certain dependence conditions, and does not lend itself to confidence intervals, nor does it take advantage of correlation among tests to increase power.

An early method that does take advantage of correlation among the tests was developed by [Dunnett \(1955\)](#) for the special case of a comparing several treatment

groups to a control under homoskedasticity normality. [White \(2000\)](#) developed a bootstrap procedure that relaxes normality and homoskedasticity. [Westfall and Young \(1993\)](#), [Romano and Wolf \(2005\)](#) and [List et al. \(2016\)](#) propose resampling-based sequential step-down procedures that refine the Dunnett and White approach in much the same way that Holm refines Bonferroni. It allows more powerful testing of remaining hypotheses if the most significant test rejects, and allows for heteroskedasticity. A drawback, however, is that it does not lend itself to confidence intervals or weighting of parameters.

The method proposed in this paper contributes to the literature by proposing simultaneous confidence intervals, rather than only hypothesis tests, that account for correlation among individual estimates and allow precision to be preserved on the parameters of greatest interest.

The methods described above, including the method proposed in this paper, control the familywise error rate, meaning the probability of incorrectly rejecting a null hypothesis concerning any one of the parameters or that any confidence interval fails to include the corresponding parameter will be at most the specified  $\alpha$  level. [Benjamini and Hochberg \(1995\)](#) propose an adjustment to test statistics that controls instead the false discovery rate, the expected proportion of the number of rejected true null hypotheses to the number of rejected hypotheses (defined as zero if no hypotheses are rejected), under the assumption of independence. While this approach can be more powerful than methods that control the familywise error rate, it is not appropriate when conclusions to be drawn depend on several parameters simultaneously or when confidence intervals are desired.

## 2 Econometric framework

Consider inference about a parameter  $\beta$  of finite dimension  $J < \infty$ , for which a consistent and asymptotically normal estimate,  $\hat{b}$ , is available, satisfying

$$\sqrt{n} \left( \hat{b} - \beta \right) \xrightarrow{d} N(0, \Omega),$$

where we can consistently estimate  $\hat{\Omega} \xrightarrow{p} \Omega$ , which we assume is positive definite.

This general framework includes as a special case the common scenario in randomized controlled trials (RCTs) where  $\beta$  consists of the average treatment effects of a single treatment variable on a set of  $J$  outcomes,  $y_{1i}, \dots, y_{Ji}$ . The empirical example in Section 5 illustrates this scenario. In this situation, and assuming independent observations, the asymptotic variance covariance matrix is given by

$$\Omega^{RCT} = \frac{\Sigma_1}{p} + \frac{\Sigma_0}{1-p}, \tag{1}$$

where  $\Sigma_1$  is the covariance matrix of the  $J$  outcomes in the treated group,  $\Sigma_0$  is the same in the untreated group, and  $p$  is the probability of being assigned to the treated group. In another common special case of this general framework,  $\beta$  consists of coefficients on a vector of regressors  $x_i$  in a regression specification  $y_i = x_i' \beta + \varepsilon_i$ , where  $E[x_i \varepsilon_i] = 0$ . In this case a possible estimator is  $\hat{b} = (\sum_{i=1}^n x_i x_i')^{-1} \sum_{i=1}^n x_i y_i$  and (given independent observations) the asymptotic covariance matrix is

$$\Omega^{regression} = E[x_i x_i']^{-1} E[x_i x_i' \varepsilon_i^2] E[x_i x_i']^{-1}.$$

Many other settings fit into this general framework, including quantile regression, maximum likelihood estimation of structural models, or GMM estimates.

## 2.1 Conventional confidence regions

A level  $(1 - \alpha)$  confidence region for one dimension, say,  $\beta_j$ , can be constructed as

$$CI_{1-\alpha}^j = \left[ \hat{b}_j \pm C^{\text{simple}} \sqrt{\hat{\Omega}_{j,j}/n} \right], \quad (2)$$

where  $C^{\text{simple}}$  is the  $(1 - \alpha/2)$ -quantile of the standard normal distribution.

A naive confidence region for the parameter vector  $\beta$  is simply the product of the intervals for each element:

$$CR_{1-\alpha}^{\text{naive}} = \prod_j CI_{1-\alpha}^j.$$

Unfortunately, this naive confidence region covers the true parameter with lower than nominal probability:

$$\lim_{n \rightarrow \infty} \Pr(\beta \in CR_{1-\alpha}^{\text{naive}}) < 1 - \alpha.$$

Conclusions based on the naive confidence region that depend on more than one parameter will therefore be misleading a higher-than-expected fraction of the time. For example, even when a treatment has no (true) effect on any outcomes under consideration, the naive joint confidence region will exclude zero for some outcome with probability greater than  $\alpha$ , and perhaps much greater, depending on the number of outcomes considered and their correlation.

One common solution is a Bonferroni-like adjustment to the critical values ensuring that the joint coverage is at least nominal:

$$CR_{1-\alpha}^{\text{Bonferroni}} = \prod_j \left[ \hat{b}_j \pm C^{\text{Bonferroni}} \sqrt{\hat{\Omega}_{j,j}/n} \right],$$

where the Bonferroni critical value,  $C^{\text{Bonferroni}}$ , is equal to the  $1 - \alpha/(2J)$  quantile of the normal distribution. Note that  $C^{\text{Bonferroni}} > C^{\text{simple}}$  for  $J > 1$ , resulting in longer

projections on each dimension. The Bonferroni region is valid in that it will cover the truth with limiting probability greater than  $1 - \alpha$ :

$$\lim_{n \rightarrow \infty} \Pr(\beta \in CR_{1-\alpha}^{\text{Bonferroni}}) > 1 - \alpha.$$

A drawback of the Bonferroni region is that it is conservative, and thus leads to unnecessarily imprecise inferences. It is conservative for two reasons. First, by construction it produces confidence regions with at least the nominal coverage under the worst-case scenario when the estimates  $\{\hat{b}_j\}$  are independent. When the estimates are correlated, the confidence region will be conservative. Second, even if the estimates were independent, the critical value  $C^{\text{Bonferroni}}$  is a conservative first-order approximation (in  $\alpha$ ) to the actual critical value necessary to obtain correct coverage,  $C^{\text{independent}}$ , equal to the  $(1 + (1 - \alpha)^{1/J})/2$  quantile of the standard normal distribution, which is always less than  $C^{\text{Bonferroni}}$  for any  $\alpha \in (0, 1)$ . The degree of conservativeness due to the first-order approximation is typically quite small compared to that due to the assumption of independence when the estimates  $\{\hat{b}_j\}$  are correlated.

A confidence region that takes into account correlation among the estimates can be constructed by inverting Wald  $\chi^2$  tests, that is, by collecting all values for the parameter that would not be rejected by a joint Wald test:

$$CR^{\text{inverse}} = \left\{ b : n (\hat{b} - b)' \hat{\Omega}^{-1} (\hat{b} - b) \leq \chi_{1-\alpha}^2(J) \right\},$$

where  $\chi_{1-\alpha}^2(J)$  is the  $(1 - \alpha)$ -quantile of the  $\chi^2$  distribution with  $J$  degrees of freedom. Confidence regions formed in this way in general define an ellipsoid centered on  $\hat{b}$ . The shape and volume of the ellipsoid depend on the correlation among the estimates. Projecting the ellipsoid on each dimension produces a set of jointly valid



confidence intervals. Remarkably, however, the widths of the projected confidence intervals do not depend on the correlation among the estimates, as the following proposition shows.

**Proposition 0** *The projection of  $CR^{inverse}$  onto the  $j$ -th coordinate axis is*

$$CI_j^{inverse} = \left[ \hat{b}_j \pm C^{inverse} \sqrt{\hat{\Omega}_{j,j}/n} \right],$$

where  $C^{inverse} = \sqrt{\chi_{1-\alpha}^2(J)}$ .

**Proof.** Change variables  $v = \hat{b} - b$ , and define  $P = n\hat{\Omega}^{-1}$ , and  $k = \sqrt{\chi_{1-\alpha}^2(J)}$ . Note that  $CR^{inverse}$  corresponds to the ellipsoid  $\{v : v'Pv \leq k^2\}$ . Since  $P$  is symmetric and positive definite, Lemma 6 in the appendix applies, whose result states that the projection of  $\{v : v'Pv \leq k^2\}$  on the  $j$ -th coordinate axis is  $[\pm k\sqrt{p_{j,j}}]$ , where  $p_{j,j}$  is the  $j$ -th diagonal element of  $P^{-1}$ . Changing variables back yields the result. ■

Proposition 0 means that although the ellipsoidal confidence region  $CR^{inverse}$  takes into account correlation among the estimates, the confidence intervals obtained by projecting that region onto each parameter dimension do not: even when estimates are highly correlated, projected confidence intervals will be just as wide as when estimates are independent. For  $J > 1$  we have  $C^{inverse} > C^{Bonferroni}$ , so inverting  $\chi^2$  tests produces longer confidence intervals for each dimension than the Bonferroni method, whether estimates are correlated or not.

## 2.2 Weighted confidence region

Confidence intervals derived from Bonferroni adjustments or inverting  $\chi^2$  tests are valid in the sense that in large samples their coverage will be at least nominal, but they each have drawbacks that hinder their usefulness in many economic settings.

Bonferroni adjustments do not take into account correlation among the estimates and thus can yield quite conservative confidence intervals. Confidence regions constructed by inverting joint  $\chi^2$  tests account for correlation, but the projections on each dimension remain even wider than Bonferroni intervals. Both of these methods sacrifice precision in any one dimension to achieve joint validity across all dimensions. This is especially harmful when the researcher cares more about precision in some dimensions than others.

Here I propose a method for constructing confidence regions that remedies the drawbacks of conventional methods. In particular, the method takes into account correlation (thus improving on Bonferroni) in a way that shortens projected confidence intervals on each dimension (improving on inverting  $\chi^2$  tests) and it allows the researcher to preserve precision along the dimensions of most interest while maintaining joint validity (improving on both).

The idea is to choose a confidence region with projected confidence intervals as short as possible subject to maintaining the required coverage. This leads to rectangular confidence regions centered on the estimate. Define such a confidence region by its vector of normalized half-widths,  $h := (h_1, \dots, h_J)'$ , so that the projected confidence interval for  $\beta_j$  is given by  $\hat{b}_j \pm h_j \sqrt{\hat{\Omega}_{j,j}/n}$ . The normalized half-widths  $\{h_j\}$  have the interpretation of effective critical values applying to each dimension. Given an asymptotic variance matrix  $\Omega$ , the coverage associated with any choice of  $h$  is given by

$$\Gamma(h, \Omega) = \int \cdots \int_{-h}^h \phi(\mathbf{x}; \mathbf{R}(\Omega)) \, d\mathbf{x},$$

where  $\phi$  is the multivariate normal density with correlation matrix

$$\mathbf{R}(\Omega) := \text{diag}(\Omega)^{-1/2} \Omega \text{diag}(\Omega)^{-1/2},$$

and  $\text{diag}(\Omega)^{-1/2}$  is a diagonal matrix containing the reciprocals of the square roots of the diagonal values of  $\Omega$ .

Minimizing the projected confidence interval lengths subject to maintaining required coverage involves a tradeoff over parameter dimensions, which the researcher specifies by choosing ex ante a nonnegative  $J$ -vector of preference weights  $\omega$ , summing to one. The proposed confidence region then minimizes the weighted sum of the half-widths subject to maintaining correct coverage:

$$\begin{aligned} \hat{h}(\omega, \mathbf{R}, \alpha) &= \arg \min_h \omega' h & (3) \\ \text{subject to } \Gamma(h, \mathbf{R}) &\geq 1 - \alpha. \end{aligned}$$

The projected confidence interval for, say,  $\beta_j$ , is then

$$CI_j^{\text{weighted}} = \left[ \hat{b}_j \pm \hat{h}_j \sqrt{\hat{\Omega}_{j,j}/n} \right].$$

The resulting confidence intervals respond naturally to different choices of weights  $\omega$ : the higher the weight on one dimension relative to the others, the shorter the confidence interval for that dimension. Any dimension for which the specified weight is zero will have a confidence interval of infinite length. The shortest possible confidence interval for a given dimension is obtained when all weight is placed on that dimension, and zero on the others.

The weighted confidence intervals include existing methods as special cases. In the special case where the estimates are independent (i.e.,  $\Omega$  is diagonal) and all weights are equal (i.e.,  $\omega_j = 1/J$  for all  $j$ ), the weighted confidence intervals coincide with the Bonferroni intervals (to first order in  $\alpha$ ). The special case where  $\omega$  places weight one on one dimension and zero on the others reduces to simple inference on one

parameter, with half-width equal to the  $(1 - \alpha/2)$ -quantile of the standard normal distribution.

### 3 Decision Theoretic Motivation

The weighted confidence intervals proposed above are appealing because they offer researchers a middle ground between the two extremes of performing inference on one parameter only, while ignoring all others, and performing symmetric joint inference on all parameters, sacrificing precision on the subset of parameters the researcher cares most about. Nevertheless, the procedure may seem ad hoc. This section motivates the procedure above in a decision theoretic framework, presenting axioms for a researcher's preferences under which the above procedure is optimal.

To accomplish this we will consider a researcher's preferences over confidence regions, and conditions under which those preferences can be represented by a linear combination of the form  $\omega'h$ , or more generally  $\sum_{j=1}^J \omega_j u(h_j)$ , where  $h$  is the vector of projected confidence interval lengths associated with the confidence region.

For our purposes, we will define confidence regions standardized relative to the location and scale of the estimates  $\hat{b}$ . Thus, a confidence region is a measurable subset of  $\bar{\mathbb{R}}^J := \{\mathbb{R} \cup \{-\infty, \infty\}\}^J$ , where the origin corresponds to the estimate  $\hat{b}$ , and the units are in the standard errors of each dimension of  $\hat{b}$ , that is,  $\sqrt{\hat{\Omega}_{j,j}/n}$ . A researcher chooses the confidence region prior to observing  $\hat{b}$ .

The axioms below will formally capture the intuitive assumptions that the researcher's preferences over confidence regions depend on the lengths of the projected intervals for each dimension, that shorter confidence intervals are more preferred, and that the tradeoffs the researcher is willing to make on precision among different dimensions are scale invariant. Naturally the researcher also wants confidence regions

to have certain coverage properties, but this is incorporated as a constraint in the decision problem, rather than a characterization of preferences, much as a budget constraint is not ordinarily included in a characterization of a consumer's preferences over bundles of goods.

Formally, let the researcher's preferences over confidence regions be denoted by  $\succeq^*$ , a complete, transitive, and reflexive binary relation on  $\mathcal{S}$ , the set of all measurable subsets of  $\bar{\mathbb{R}}^J$ . We define the corresponding indifference relation  $\sim^*$  and strict preference relation  $\succ^*$  in the usual way. Define  $h : \mathcal{S} \rightarrow \bar{\mathbb{R}}_+^J$ , where  $\bar{\mathbb{R}}_+^J$  is the nonnegative orthant of  $\bar{\mathbb{R}}^J$ , to be a mapping giving the vector of projected confidence interval half-widths associated with a confidence region.

The first axiom we postulate is that a researcher's preferences over confidence regions depend only on the projected confidence intervals.

**Axiom 1 (Sufficiency of projections)**  $h(r) = h(s)$  implies  $r \sim^* s$  for all  $r, s \in \mathcal{S}$ .

This axiom means the researcher is indifferent between two confidence regions if they yield the same vector of projected confidence intervals. It is consistent with empirical practice of reporting only marginal confidence intervals, including projections of joint confidence regions (Bugni et al., 2017; Ciliberto and Tamer, 2009; Grieco, 2014). It nevertheless may rule out some reasonable preferences. For example, in a two-parameter setting with highly correlated estimates an elliptical confidence region may have the same projected confidence intervals as a rectangular region that is much larger in area. The elliptical confidence region rules out many more linear combinations of the two parameters, such as their difference, than the rectangular region, and thus it would be reasonable for a researcher to prefer it. If it is the difference between two parameters that is of interest, however, then the estimation problem can

be parameterized so that the difference becomes its own parameter, and the axiom may apply.

The consequence of Axiom 1 is that we can equivalently consider preferences over vectors of confidence interval projection lengths. We take  $\succeq$  to be a binary relation on  $\bar{\mathbb{R}}_+^J$  defined by

$$f \succeq g : \iff r \succeq^* s, \tag{4}$$

for some  $r, s \in \mathcal{S}$  such that  $h(r) = f$  and  $h(s) = g$ .

Lemma 7 in the Appendix shows that  $\succeq^*$  (defined over confidence regions) and  $\succeq$  (defined over vectors of projected confidence interval half-lengths) are equivalent in the sense that  $r \succeq^* s \iff h(r) \succeq h(s)$ . A representation of  $\succeq$  is therefore also a representation of  $\succeq^*$ .

The task then becomes to find axioms satisfied by  $\succeq$ , a complete, transitive, and reflexive binary relation on  $\bar{\mathbb{R}}_+^J$ , such that  $h \succeq h'$  if and only if  $\sum_{j=1}^J \omega_j u(h_j) \leq \sum_{j=1}^J \omega_j u(h'_j)$ . Mathematically this setting is analogous to a subjective expected utility setting analyzed in Hens (1992), where the parameter dimensions take on the role of states, the precision preference weights  $\{\omega_i\}$  take on the role of subjective probabilities, and the vectors of projected interval lengths take on the role of acts, in the terminology of Savage (1972). Because of the finite dimensionality of the parameter, classic results such as Savage do not apply; Hens (1992) extends the basic result to situations with a finite state space, as we have here. The following therefore adapts Hens's (1992) model to this setting.

Following Hens, we adopt a smoothness condition à la Debreu (1972) that indifference sets defined by  $\succeq$  are continuously differentiable manifolds. Our next axiom, also assumed in Hens (1992), captures the intuitive notion that shorter confidence

intervals are strictly preferred:

**Axiom 2 (Monotonicity)** *Consider  $h, h' \in \bar{\mathbb{R}}^{+J}$ . Then  $h < h'$  implies  $h \succ h'$ .*

In other words, researchers prefer more precise inferences. Here we use the vector inequality  $h < h'$  to mean  $h_j \leq h'_j$  for all  $j = 1, \dots, J$ , and  $h \neq h'$ . Monotonicity should be reasonable in nearly every setting.

Given completeness, transitivity, reflexivity, smoothness, and monotonicity, standard results (Mas-Colell, 1985) imply that  $\succeq$  can be represented by a continuously differentiable function  $V : \bar{\mathbb{R}}_+^J \rightarrow \mathbb{R}$  such that

$$r \succeq s \iff V(r) \leq V(s) \text{ for all } r, s \in \bar{\mathbb{R}}_+^J.$$

The next axiom captures the notion that when comparing two confidence regions with identical projected confidence intervals on some subset of dimensions, attention can be restricted to dimensions on which they differ.

**Axiom 3 (Dimension separability)** *Consider a subset of dimensions  $\Delta \subset \{1, \dots, J\}$  and confidence regions  $x, x', y, y' \in \bar{\mathbb{R}}^{+J}$  with  $x_j = y_j$  for  $j \in \Delta$  and  $x_j = x'_j$  for  $j \in \{1, \dots, J\} \setminus \Delta$ ,  $x'_j = y'_j$  for  $j \in \Delta$  and  $y_j = y'_j$  for  $j \in \{1, \dots, J\} \setminus \Delta$ . Then  $y \succeq y'$  implies  $x \succeq x'$ .*

This axiom means there is a kind of separability between the parameter dimensions in that a researcher's evaluation of a subset of confidence intervals is not affected by the other dimensions. The axiom is likely reasonable in most settings, but would be violated if, for example, a researcher's relative preference for precision among a subset of dimensions changes depending on precision in another set dimensions.

Axiom 3 means that the researcher's preferences can be represented by an additively separable function with dimension-specific utility functions  $u_j$  (or, perhaps

more aptly, disutility functions) such that

$$r \succeq s \iff \sum_{j=1}^J u_j(r_j) \leq \sum_{j=1}^J u_j(s_j) \text{ for all } r, s \in \bar{\mathbb{R}}_+^J,$$

and the dimension-specific utility functions  $u_j$  are identical up to monotone increasing transformations.

The final axiom, also adapted from [Hens \(1992\)](#), captures the intuition that a researcher's willingness to trade off precision in one dimension for precision in another is scale invariant:

**Axiom 4 (Scale-invariant precision tradeoff)** *For all pairs of dimensions  $j, j' \in \{1, \dots, J\}$ , the marginal rate of substitution between  $j$  and  $j'$  is constant along the equal-precision line, i.e.,*

$$\frac{V_j(\bar{r}, \dots, \bar{r})}{V_{j'}(\bar{r}, \dots, \bar{r})} = c_{j,j'} \text{ for all } \bar{r} > 0,$$

where

$$V_j(r) := \frac{\partial V(r)}{\partial r_j} \text{ for } r \in \bar{\mathbb{R}}_+^J.$$

This assumption means that from a starting point of a symmetrically precise (square) confidence region, a researcher's willingness to trade off precision in one dimension for another does not depend on the overall precision of the confidence region. This axiom should be reasonable in many settings, but would be violated if, for example, a researcher's priorities for precision across parameter dimensions were different in settings with a high degree of precision (e.g., large sample size) and settings with less precision (e.g., small sample size).

Axioms 1-4 together imply that a researcher's preferences over confidence regions can be represented by a weighted sum of projected confidence interval lengths, as the



following proposition states formally.

**Proposition 5 (Hens’s (1992) Representation)** *Given Axioms 1 and 2, and smoothness, the preference relation  $\succeq$  satisfies Axioms 3-4 if and only if it has the following representation:*

$$r \succeq s \iff \sum_{j=1}^J \omega_j u(r_j) \leq \sum_{j=1}^J \omega_j u(s_j) \text{ for all } r, s \in \bar{\mathbb{R}}_+^J, \quad (5)$$

where  $\{\omega_j\}$  are nonnegative and sum to one. Moreover,  $u$  is unique up to positive affine transformations, and the weights  $\{\omega_j\}$  are unique.

Proposition 5, which follows from Hens’s (1992) subjective expected utility representation, formally justifies the weighted confidence region procedure (3). A researcher with preferences satisfying Axioms 1-4 and facing a constraint that the region maintains adequate coverage will choose a rectangular confidence region centered on the origin with half-widths that minimize  $\sum_{j=1}^J \omega_j u(h_j)$ , subject to the region maintaining adequate coverage. Making the simplification  $u(h) = h$ , the researcher’s choice satisfies (3).

## 4 Monte Carlo Simulations

This section illustrates via numerical simulations the how the weighted inference procedure developed above compares to conventional methods, and how its performance depends on factors that are likely to vary across empirical settings, including the number of parameters and the degree of correlation among estimates of those parameters.

The simulations mimic a study examining the effect of a treatment on several outcomes with a sample size chosen such that separate outcome-wise size  $\alpha = .05$

tests of the null hypothesis of zero effect will each have 80 percent power against an alternative of an effect of .2 standard deviation of the untreated outcome, a relatively common scenario (e.g. DellaVigna and Pope, 2018). The simulations will illustrate how the number of parameters, the degree of correlation among the estimates, and the specification of preference weights affect the performance of the proposed weighted inference procedure and its comparison with conventional methods.

The simulated data generating process is as follows. Outcome  $j \in \{1, \dots, J\}$  for observation  $i$  is generated as

$$Y_{ji} = \gamma_{j0} + \delta_j D_i + \varepsilon_{ji},$$

where  $\gamma_{j0} = 0$  and  $\delta_j = .2$  for all  $j \in \{1, \dots, J\}$ ,  $D_i$  is an iid binary random variable with  $\Pr(D_i = 1) = .5$ , and  $(\varepsilon_{1i}, \dots, \varepsilon_{Ji})'$  is a random vector of error terms, independent across observations, with distribution

$$\begin{pmatrix} \varepsilon_{1i} \\ \vdots \\ \varepsilon_{Ji} \end{pmatrix} \sim N \left( 0, \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix} \right).$$

The simulations vary the correlation parameter  $\rho$  across values between zero and one. To achieve 80 percent power against an effect size of .2, the simulations use a sample size of  $n = 785$ .<sup>1</sup>

---

<sup>1</sup>Assuming homoskedasticity for simplicity, to achieve power of  $1 - \beta$  against an effect size of  $\delta$  in terms of standard deviations of the outcome using a test of size  $\alpha$ , the sample size needs to satisfy:

$$1 - \beta = \Phi(\Phi^{-1}(\alpha/2) - \sqrt{n}\delta/2) + 1 - \Phi(\Phi^{-1}(1 - \alpha/2) - \sqrt{n}\delta/2).$$

For  $1 - \beta = .8$ ,  $\delta = .2$ , and  $\alpha = .05$  this implies  $n = 785$ .

The parameter vector in this setting is the vector of treatment effects:

$$\beta = (\delta_1, \dots, \delta_J)'$$

The natural (and efficient) estimator for the effect on the  $j$ -th outcome is the difference in means:

$$\hat{b}_j = \frac{\sum_{i=1}^n Y_{ji} D_i}{\sum_{i=1}^n D_i} - \frac{\sum_{i=1}^n Y_{ji} (1 - D_i)}{\sum_{i=1}^n (1 - D_i)}.$$

The vector of estimators,  $\hat{b} := (\hat{b}_1, \dots, \hat{b}_J)$ , has the distribution:

$$\sqrt{n} (\hat{b} - \beta) \sim N \left( 0, 4 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \vdots \\ \vdots & & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix} \right).$$

The standard error of an individual estimate will therefore be approximately  $2/\sqrt{785} \approx 0.071$ . A 95-percent confidence interval for a single parameter will on average therefore have upper and lower bounds (.06, .34).

The first set of simulations illustrates how the number of parameters affects each inference method's performance. Figure 6 plots the mean confidence intervals produced by the naive method (which ignores the presence of multiple parameters), Bonferroni correction, inverting  $\chi^2$  tests, and two versions of the weighted procedure developed here: one where all parameters receive equal weights, and one where one parameter receives 80 percent of the weight, with the remaining weight split evenly among the other parameters. The simulations set  $\rho = .8$  and vary the number of outcomes  $J$  from 2 to 13.

The widest band in Figure 6 corresponds to the inverse  $\chi^2$ -test method. With

only two outcomes, this method is only modestly less precise than the others, but its disadvantage grows quickly as the number of outcomes increases. With as few as four outcomes the (mean) confidence interval includes zero. The second widest band corresponds to the Bonferroni correction, which is substantially more precise than the inverse  $\chi^2$  method, but still loses precision as the number of outcomes increases, and begins to include zero when there are 10 or more outcomes. The narrowest band in the figure corresponds to the naive method which takes no account of multiple parameters, and remains constant as the number of parameters grows. The band just wider than the naive band corresponds to the weighted method, where 80 percent of the weight is concentrated on the first outcome (whose confidence intervals are plotted), while the other outcomes split the remaining 20 percent. The resulting confidence intervals are only slightly wider than the naive intervals, and do not widen as the number of parameters grows, since the weight concentrated on the first parameter remains constant. This precision comes at the cost of less precision on the remaining parameters, however. This cost may be worth it if precision on one parameter is more important to the researcher than on others. The band just wider than the weighted procedure with concentrated weights corresponds to the weighted procedure where weights on all outcomes are equal. Like the Bonferroni, it treats all outcomes symmetrically, and so gets less precise as the number of outcomes grows, but unlike the Bonferroni, it takes into account correlation among the estimates, and so is more precise, excluding zero over the range of outcomes in the simulations.

Another way to see how precision depends on the number of parameters is to consider the probability with which the confidence intervals exclude zero, or in other words, the power of the test of the null hypothesis of zero effect. Figure 6 plots these probabilities for each of the methods described above. The power of the naive test is by construction .8 no matter how many outcomes. The figure shows that the power

associated with the weighted procedure where 80 percent of the weight is concentrated on one parameter of interest is only slightly below 80 percent no matter how many outcomes. When the weights are equally distributed over the outcomes the power decreases somewhat as the number of outcomes increases, but remains higher than the Bonferroni method's power and much higher than the inverse  $\chi^2$  method's power.

The next set of simulations illustrates how correlation among the estimates affects the performance of the proposed confidence intervals. Figure 6 plots confidence interval half-widths for each of the methods described above as a function of the correlation  $\rho$  among the estimators. The simulations set the number of parameters at  $J = 2$  and vary  $\rho$  from zero to .99. The naive method produces the smallest half-width, at 1.96 for all levels of  $\rho$ . The inverse  $\chi^2$  method and Bonferroni method produce the widest intervals at 2.45 and 2.24, respectively, across all levels of  $\rho$ , since the projected confidence intervals from neither of these methods takes advantage of correlation among the estimates. Both the weighted method with concentrated weights (80 percent on the first parameter) and the method with equal weights converge to the naive width as the correlation increases, with the concentrated weights yielding narrower intervals for all levels of  $\rho$ . The method with equal weights is equivalent to Bonferroni for  $\rho = 0$ , but improves in precision relative to Bonferroni as  $\rho$  increases.

The final set of simulations shows how the distribution of weights affects the weighted confidence interval procedure. Figure 6 plots confidence interval half-widths of the weighted confidence intervals as a function of the weight given to the highest-weighted parameter, in a scenario with two parameters. The  $x$ -axis runs from .5 (equal weights) to 1 (all weight concentrated on one parameter). Panel A, which corresponds to independent estimates, shows that when the weights are equal the confidence intervals for both parameters are equivalent to the Bonferroni-corrected interval. As the weight on the first parameter increases, its confidence interval narrows

and converges to the naive width (1.96) while the confidence interval on the second interval widens. Panel B, which corresponds to a correlation between the estimates of .8, shows that when the parameters are equally weighted, both CIs are narrower than the Bonferroni-corrected interval, and remain so when the weights are close to evenly distributed (around 60/40). As the weights become more unequal the confidence interval for the less-weighted parameter becomes wider than the Bonferroni-corrected interval.

## 5 Empirical example: Charitable donations and matching

This section illustrates the proposed inference method and compares it to existing methods using data from [Karlan and List's \(2007\)](#) field experiment, re-analyzed by [List et al. \(2016\)](#). studying the effects of matching grants on charitable donations. Experimental subjects, each of whom had donated in the past to a certain U.S.-based nonprofit organization, were sent solicitation letters inviting them to donate. Subjects randomly assigned to the treatment group were told their donations would be matched. Control subjects' donations were not matched. The parameters of interest we focus on are the effects of the match offer on the following four outcomes: an indicator for whether the subject gave any positive amount, the amount the subject gave, the amount given plus the match (i.e., total revenue raised as a result of the subject's donation), and the change in the amount given compared to the subject's previous donation. [Table 1](#) reports the means and standard deviations of these outcomes among the 16,687 subjects assigned to the control group.

The presence of multiple parameters matters for inference in this setting since the

study examines several outcomes to assess the effect of the match offer on giving. For example, naively performing inference one outcome at a time without accounting for multiple parameters can yield misleading conclusions regarding whether matching affects any of the outcomes of interest. Yet conservative adjustments such as Bonferroni may preclude drawing any conclusions at all. We will use this setting to compare conventional approaches to inference on multiple parameters with the weighted procedure proposed in this paper.

Applying the weighted inference approach requires first choosing weights for each parameter that reflect the preference for precision. This choice will surely differ from case to case and should depend on the purpose of the study. For the sake of illustration, we will choose a weight of 0.1 for the effect of treatment on the indicator for giving, a weight of 1 for the effect on the amount given, a weight of 0.1 for the effect on the amount given plus the match, and a weight of zero on the change in the amount given, as shown in the right-hand column of Table 1. Note that the scale of weights does not matter, only their relative sizes. These weights reflect the preferences of a researcher who care much more about precise inference on the effect of treatment on the amount given than the other outcomes. The zero weight for the effect on the change in the amount given seems natural given that the effect on the change in amount given must be equal to the effect on the amount given, since treatment cannot affect pre-treatment variables. Likewise, it seems reasonable to assign a lower weight to the effect on the amount given plus the match, since the effect on the match is mechanically positive due to the experimental design which offered no match to the control group.

The proposed weighted procedure also depends on the correlation among the parameter estimates. With the exception of the estimate of the effect on the change in amount given, the estimates are highly correlated. Table 2 reports the correlations

among the estimates of the treatment effects on the four outcomes of interest. The estimated effects on “gave” and “amount” have a correlation of .73. The estimated effects on “gave” and “amount plus match” have a correlation of .63. The most highly correlated estimates are those on “amount” and “amount plus match,” with a correlation of .85. Estimates on “change in amount” have minimal correlation with the other estimates, due to the large variance in pre-treatment donation amounts. The relatively high correlation among many of the estimates suggests that the proposed weighted procedure may be particularly advantageous in this setting compared to conventional methods, which do not take advantage of correlation to yield tighter projected confidence intervals.

In this example the weighted confidence region approach allows substantially more precise inferences on the parameter of greatest interest than conventional approaches. Table 3 reports point estimates, standard errors, and projected 90-percent confidence intervals calculated via conventional methods and the weighted procedure proposed here. Figure 6 shows the confidence intervals graphically. The top row of Table 3 shows that the estimated effects on the measures of giving are all positive, although the precision of the estimates varies. The estimated effect on the probability of giving is 0.004 (on a mean of 0.02). The estimated effect on amount given is \$0.15. The estimated effect on the amount given plus the match is \$2.09. The estimated effect on the change in amount given compared to the pre-treatment donation is \$6.33, although the standard error of that estimate is \$13.43.

The row labeled “Naive CI” shows that taken one at a time, the effects on each outcome other than the change in amount given would be considered statistically significantly different from zero at the 10-percent level. Considered jointly, however, the naive confidence intervals will fail to include the true vector of parameters more than 10 percent of the time. The next row reports Bonferroni-adjusted confidence



intervals that have joint coverage of at least 90 percent. Not surprisingly, the individual confidence intervals are substantially less precise than the the naive confidence intervals. In particular, the confidence interval for amount given includes negative effects. The confidence intervals obtained by inverting  $\chi^2$  tests are even less precise than the Bonferroni intervals.

The weighted confidence intervals obtained by the procedure proposed above restore precision to the parameter of greatest interest. The weighted confidence interval for the effect on amount given excludes zero and is only slightly wider than the naive confidence interval. The precision on this parameter comes at the cost of slightly less precise intervals for the effect on the probability of giving and amount given plus the match. The zero weight placed on the effect on change in amount given means the confidence interval is unbounded on that parameter. Given that by virtue of the experimental design the effect on the change in amount is equal to effect on amount given—about which we do have precise inference—this does not entail a meaningful loss. The bottom row of the table shows the optimal critical values used in the weighted procedure, determined by the minimization problem (3): 2.44 for the effect on the probability of giving, 1.69 for the effect on amount given, 2.29 for the effect on the amount plus match, and infinity for the effect on the change in amount given.

## 6 Conclusion

Nearly all empirical studies involve multiple parameters, which can complicate inference. Conventional methods for accounting for multiple parameters often sacrifice precision on the parameters of greatest interest. This article proposed a method that results in jointly valid confidence intervals across all parameters while maintaining precision on the parameters of greatest interest. The procedure gains its advantage

by accounting for correlations among the estimates, unlike Bonferroni-like methods, and by allowing the researcher to specify preferences for precision across the parameters; giving up precision on parameters of less interest allows more precise inference on parameters of greatest interest. The procedure was motivated in a decision theoretic framework specifying axioms for a researcher's preferences under which the proposed procedure is optimal.

The empirical example based on a field experiment on the effect of matching on donations illustrated the advantage of the weighted approach in situations where one or more parameters are of more interest than others, but the researcher would like to perform jointly valid inference on all parameters. Conventional methods for conducting valid joint inference sacrifices precision on the parameter of most interest. The weighted procedure maintains precision on the parameter of greatest interest while remaining jointly valid.

A word of caution is warranted regarding the proposed weighted inference procedure. A researcher may be tempted to specify preference weights strategically after observing point estimates in order to achieve statistically significant results. This practice would of course lead to invalid inferences. To ensure valid inferences the weights should be specified in advance. Pre-registered studies may consider including the specified weights in the analysis plan. Studies that do not pre-register should give the ex ante rationale for the chosen weights.

## References

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

- C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60, 1935.
- Federico A. Bugni, Ivan A. Canay, and Xiaoxia Shi. Inference for subvectors and other functions of partially identified parameters in moment inequality models. *Quantitative Economics*, 8(1):1–38, 2017.
- Federico Ciliberto and Elie Tamer. Market structure and multiple equilibria in airline markets. *Econometrica*, 77(6):1791–1828, 2009.
- Gerard Debreu. Smooth preferences. *Econometrica*, 40(4):603–615, 1972.
- Stefano DellaVigna and Devin Pope. What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069, 2018.
- Charles W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272):1096–1121, 1955.
- Paul L. E. Grieco. Discrete games with flexible information structures: an application to local grocery markets. *The RAND Journal of Economics*, 45(2):303–340, 2014.
- Olivier Guilbaud. Simultaneous confidence regions corresponding to holm’s step-down procedure and other closed-testing procedures. *Biometrical Journal*, 50(5):678–692, 2008.
- Thorsten Hens. A note on savage’s theorem with a finite number of states. *Journal of Risk and Uncertainty*, 5(1):63–71, Feb 1992.
- Yosef Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

- Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.
- Dean Karlan and John A. List. Does price matter in charitable giving? evidence from a large-scale natural field experiment. *American Economic Review*, 97(5):1774–1793, December 2007.
- John A. List, Azeem M. Shaikh, and Yang Xu. Multiple hypothesis testing in experimental economics. Working Paper 21875, National Bureau of Economic Research, January 2016.
- Andreu Mas-Colell. *The theory of general economic equilibrium : a differentiable approach*. Cambridge [Cambridgeshire] ; New York : Cambridge University Press, 1985. Includes indexes.
- Rupert G. Miller, Jr. *Simultaneous Statistical Inference*. McGraw-Hill, 1966.
- Open Science Collaboraton. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.
- Joseph P. Romano and Michael Wolf. Stepwise multiple testing as formalized data snooping. *Econometrica*, 73(4):1237–1282, 2005.
- Leonard J. Savage. *The Foundations of Statistics*. Dover Books on Mathematics Series. Dover Publications, 1972.
- R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.
- Klaus Strassburger and Frank Bretz. Compatible simultaneous lower confidence bounds for the holm procedure and other bonferroni-based closed tests. *Statistics in Medicine*, 27(24):4914–4927, 2008.

Peter H. Westfall and S. Stanley Young. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. John Wiley & Sons, Inc., 1993.

Halbert White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

## Appendix

**Lemma 6** *Let  $P$  be a real-valued  $J \times J$  symmetric, positive definite matrix and let  $k \geq 0$ . Then the projection of the ellipsoid defined by  $\{v \in \mathbb{R}^J | v'Pv = k^2\}$  on coordinate axis  $i \in \{1, \dots, J\}$  is the interval  $[-k\sqrt{p_{ii}}, k\sqrt{p_{ii}}]$ , where  $p_{ii}$  denotes the  $i$ -th diagonal element of  $P^{-1}$ .*

**Proof.** We prove the result for  $i = 1$ , which by symmetry is without loss of generality. Denote the  $(i, j)$ -th element of  $P^{-1}$  by  $p_{ij}$ . Let  $P'_1$  be the first row of  $P$  and  $P'_{-1}$  be all rows but the first of  $P$ . Similarly, let  $P^{-1}_1$  be the first column of  $P^{-1}$ . The ellipsoid is defined by the equation

$$v'Pv = k^2. \tag{6}$$

Equation (6) defines  $v_1$  as an implicit function of  $(v_2, \dots, v_J)$ , and the endpoints of the projection correspond to the maximum and minimum of that function. Implicitly differentiating  $v_1$  with respect to  $(v_2, \dots, v_J)$  in equation (6) and setting the derivative equal to zero yields the following first-order conditions satisfied by the points  $\tilde{v}$  on the ellipsoid where  $v_1$  obtains its maximum and minimum:

$$P'_{-1}\tilde{v} = 0.$$

Together with equation (6), these equations form a system of  $J$  equations that define

the points on the ellipsoid corresponding to the projection. Noting that  $\tilde{v} = P^{-1}P\tilde{v}$ , we have

$$\begin{aligned}
\tilde{v} &= P^{-1}P\tilde{v} \\
&= P^{-1} \begin{pmatrix} P'_1\tilde{v} \\ P'_{-1}\tilde{v} \end{pmatrix} \\
&= P^{-1} \begin{pmatrix} P'_1\tilde{v} \\ 0 \end{pmatrix} \\
&= \begin{pmatrix} p_{11}P'_1\tilde{v} \\ p_{21}P'_1\tilde{v} \\ \vdots \\ p_{J1}P'_1\tilde{v} \end{pmatrix} \\
&= P_1^{-1}P'_1\tilde{v},
\end{aligned}$$

where the third equality follows from the first-order condition  $P'_{-1}\tilde{v} = 0$ . Note that the fourth equality implies that the projection endpoints  $\tilde{v}_1$  satisfy  $\tilde{v}_1 = p_{11}P'_1\tilde{v}$ . Substituting  $\tilde{v} = P_1^{-1}P'_1\tilde{v}$  into equation (6) gives

$$\begin{aligned}
k^2 &= (P_1^{-1}P'_1\tilde{v})' P P_1^{-1}P'_1\tilde{v} \\
&= P_1^{-1'} P P_1^{-1} (P'_1\tilde{v})^2.
\end{aligned}$$

But  $P_1^{-1'}P$  is  $(1, 0, \dots, 0)$ , so  $P_1^{-1'} P P_1^{-1}$  is simply  $p_{11}$ . Recalling that  $\tilde{v}_1 = p_{11}P'_1\tilde{v}$ , we have

$$\begin{aligned}
k^2 &= p_{11} (P'_1\tilde{v})^2 \\
&= \frac{\tilde{v}_1^2}{p_{11}},
\end{aligned}$$

or

$$\tilde{v}_1 = \pm k\sqrt{p_{11}},$$

which completes the proof. ■

**Lemma 7** *Suppose  $\succeq$  satisfies completeness, transitivity, reflexivity, and Axiom 1. Then for  $\succeq$  defined in (4) we have  $r \succeq^* s \iff h(r) \succeq h(s)$  for all  $r, s \in \mathcal{S}$ .*

**Proof.** That  $r \succeq^* s$  implies  $h(r) \succeq h(s)$  is true by the definition of  $\succeq$ . Now we show  $h(r) \succeq h(s)$  implies  $r \succeq^* s$ .  $h(r) \succeq h(s)$  by definition means there exist  $r', s' \in \mathcal{S}$  such that  $h(r') = h(r)$ ,  $h(s') = h(s)$ , and  $r' \succeq^* s'$ . By Axiom 1  $r \sim^* r'$  and  $s \sim^* s'$ . By transitivity, therefore  $r \succeq^* s$ . ■

**Proof of Proposition 5.** Given Axiom 1 the decision setting here is a special case of the setting in Hens (1992) where our number of “goods” is  $L = 1$  and where the number of “states”  $S$  is the number of parameter dimensions,  $J$ . Hens’s (1992) representation result requires smoothness, monotonicity, and three axioms: (1) Sure Thing Principle, (2) state independence, and (3) weak homotheticity. We assume smoothness in the premise to Proposition 5 and monotonicity in Axiom 2. The Sure Thing Principle is equivalent to our Axiom 3. State independence in our setting with only one “good” is implied by monotonicity. Weak homotheticity is equivalent to our Axiom 4. Thus our setting satisfies Hens’s (1992) conditions, yielding the result. ■

## Tables and Figures

Table 1: Summary statistics in the control group (N = 16,687)

Variable	Control mean	Control std. dev.	Preference weight
Gave	0.02	0.13	0.1
Amount (\$)	0.81	8.18	1
Amount plus match (\$)	0.81	8.18	0.1
Change in amount (\$)	-56.89	1551.91	0

Notes: The table shows the mean and standard deviation of the variables indicated in the left column among the subjects assigned to the control group. The right-hand column shows the preference weight specified for each outcome. A larger weight means a greater preference for precision on the effects on the corresponding outcome. Data from List, Shaikh, and Xu (2016).

Table 2: Correlations among treatment effect estimates

	Gave	Amount (\$)	Amount plus match (\$)	Change in amount (\$)
Gave	1.000	0.730	0.633	0.005
Amount (\$)	0.730	1.000	0.854	0.005
Amount plus match (\$)	0.633	0.854	1.000	0.004
Change in amount (\$)	0.005	0.005	0.004	1.000

Notes: Each cell in the table reports the estimated correlation between the point estimates of effects on the outcomes corresponding to the given row and column.



Table 3: Comparison of conventional and weighted simultaneous confidence intervals

Outcome: Preference weight:	Gave 0.1	Amount (\$) 1	Amount plus match (\$) 0.1	Change in amount (\$) 0
Treatment effect	0.004 (0.001)	0.154 (0.080)	2.088 (0.161)	6.331 (13.432)
Naïve CI	[.002,.006]	[.022,.285]	[1.82,2.35]	[-15.8,28.4]
Bonferroni CI	[.001,.007]	[-.026,.333]	[1.73,2.45]	[-23.8,36.4]
Inverse chi-squared CI	[.0005,.0078]	[-.070,.377]	[1.64,2.54]	[-31.1,43.8]
Weighted CI	[.001,.007]	[.019,.289]	[1.72,2.46]	$[-\infty, +\infty]$
optimal critical value	2.44	1.69	2.29	$\infty$

Notes: The table shows point estimates, heteroskedasticity-consistent standard errors, and simultaneous 90-percent confidence intervals for the effects of a match offer on the outcome indicated in the top row. The naïve confidence intervals use a critical value equal to the .95-quantile of the standard normal distribution. The Bonferroni confidence intervals use the  $1-.1/(2 \times 4) = .9875$ -quantile of the standard normal distribution. The inverse chi-squared CI uses the square root of the .9-quantile of the chi-squared distribution with four degrees of freedom. The weighted confidence intervals use the preference weights given at the top of each column to compute the optimal critical values shown in the bottom row.

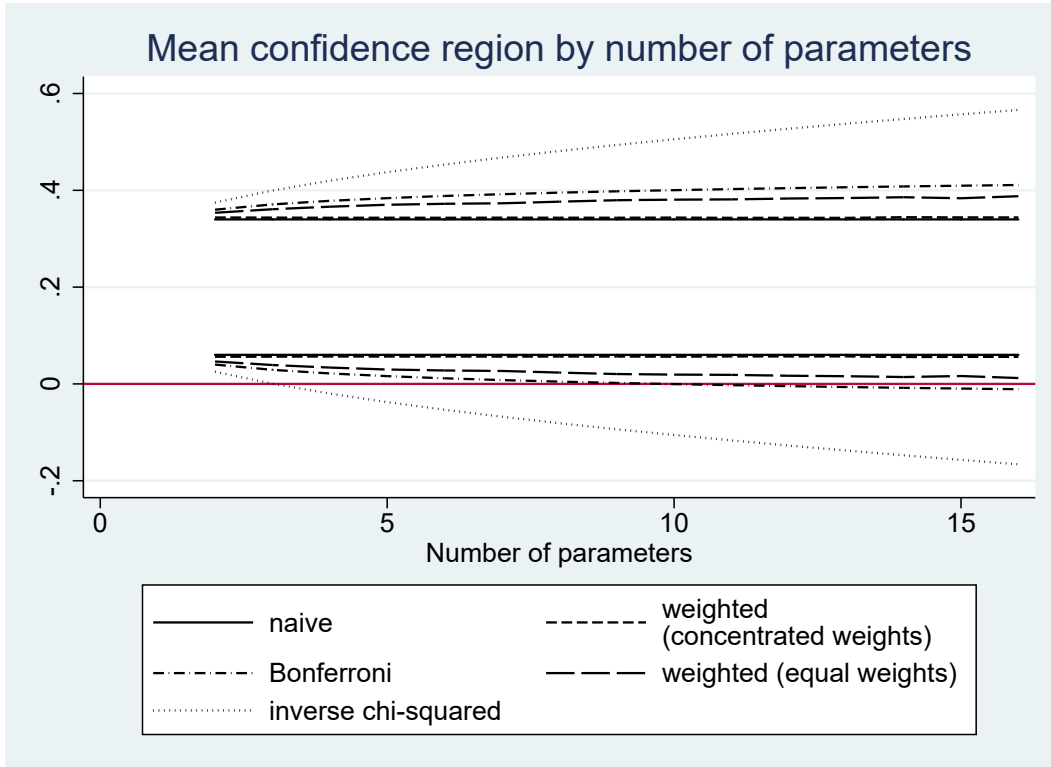


Figure 1: Simulated mean confidence intervals by method and by number of outcomes. Simulations set the correlation among outcomes at  $\rho = .8$ . The intervals labeled “weighted (concentrated weights)” correspond to the weighted procedure developed in the text with a weight of .8 given to the first parameter (whose confidence intervals are plotted) and the remain .2 weight split evenly among the other outcomes.

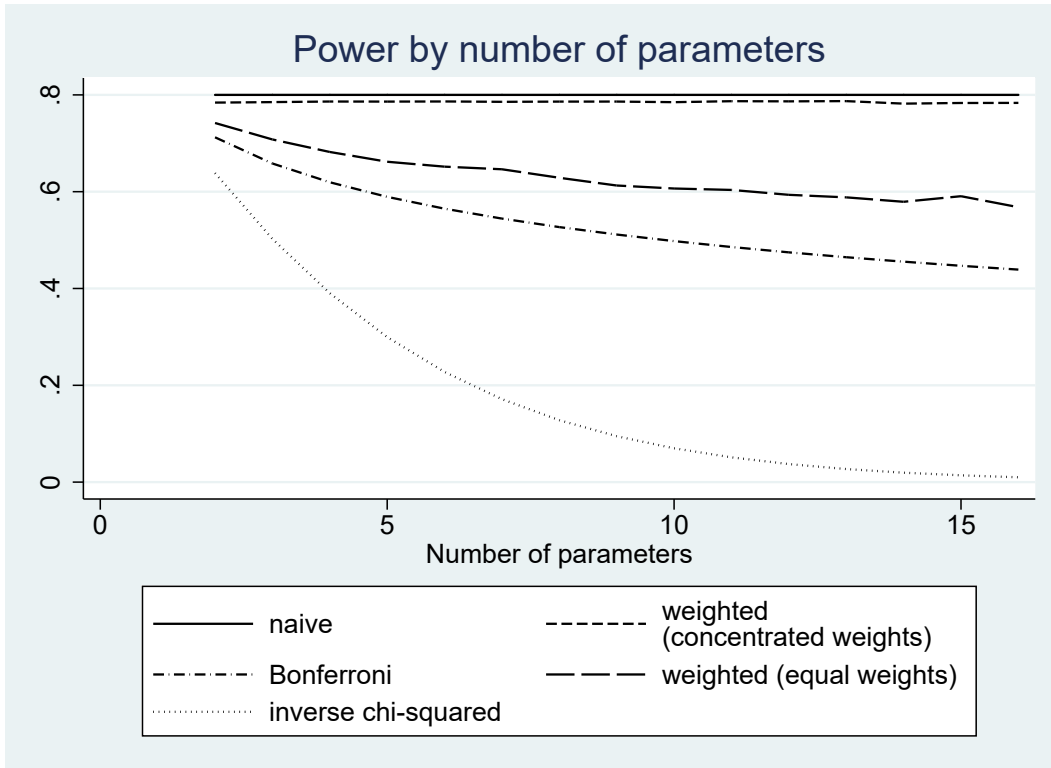


Figure 2: Simulated rejection rates by method and by number of outcomes. Simulations set the correlation among outcomes at  $\rho = .8$ . The rejection rates refer to the probabilities that confidence intervals of the indicated method will exclude zero. The true effect size is  $\delta = .2$ . The curve labeled “weighted (concentrated weights)” corresponds to the weighted procedure developed in the text with a weight of .8 given to the first parameter (whose confidence intervals are plotted) and the remain .2 weight split evenly among the other outcomes.

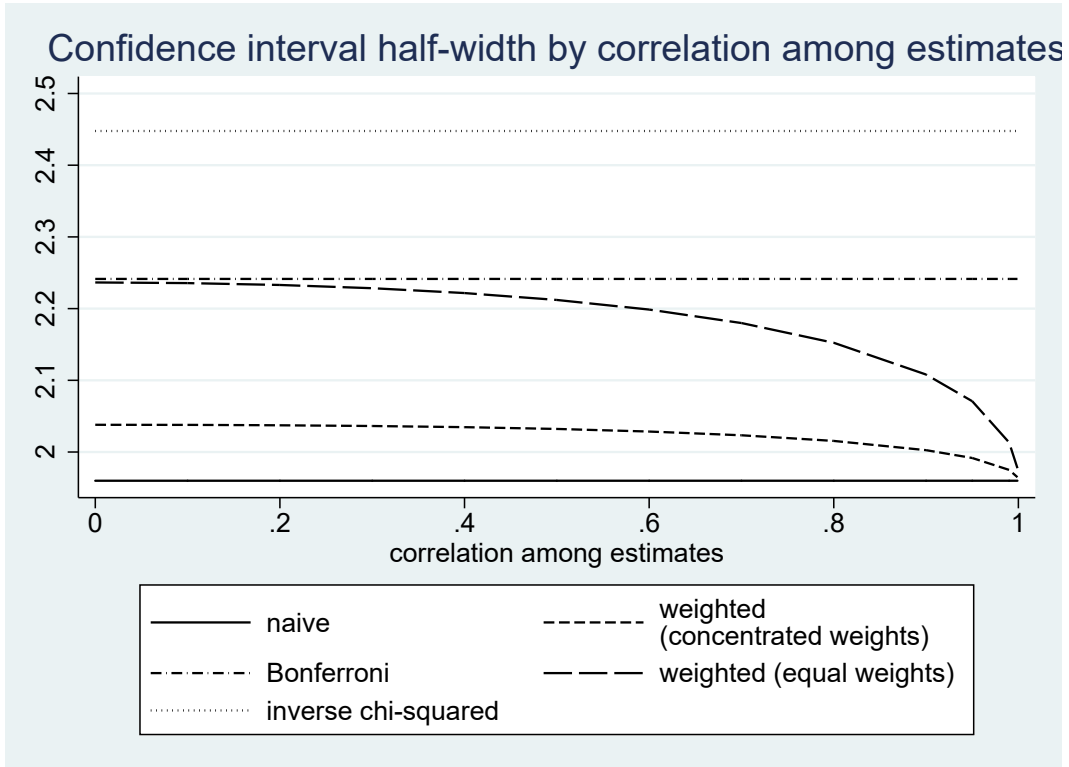
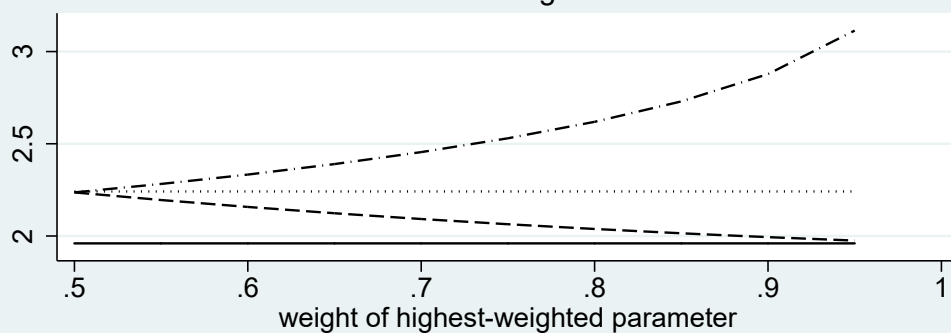


Figure 3: Simulated confidence interval half-widths by method and by number of outcomes. Simulations set the number of parameters at  $J = 2$ . The curve labeled “weighted (concentrated weights)” corresponds to the weighted procedure developed in the text with a weight of .8 given to the first parameter (whose confidence intervals are plotted) and the remain .2 weight split evenly among the other outcomes.

## Confidence interval half-width by weight distribution

A. Correlation among estimates = 0



B. Correlation among estimates = 0.8

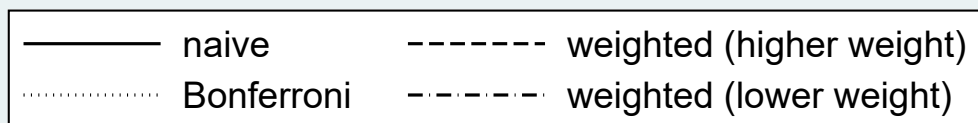
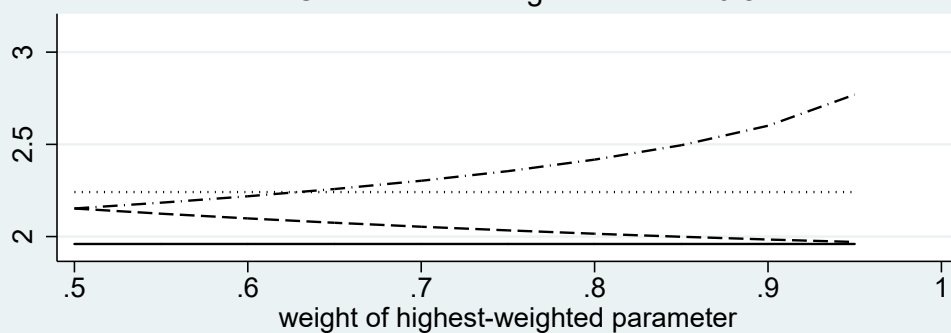


Figure 4: Simulated confidence interval half-widths by method and by weight distribution. Simulations set the number of parameters at  $J = 2$ .

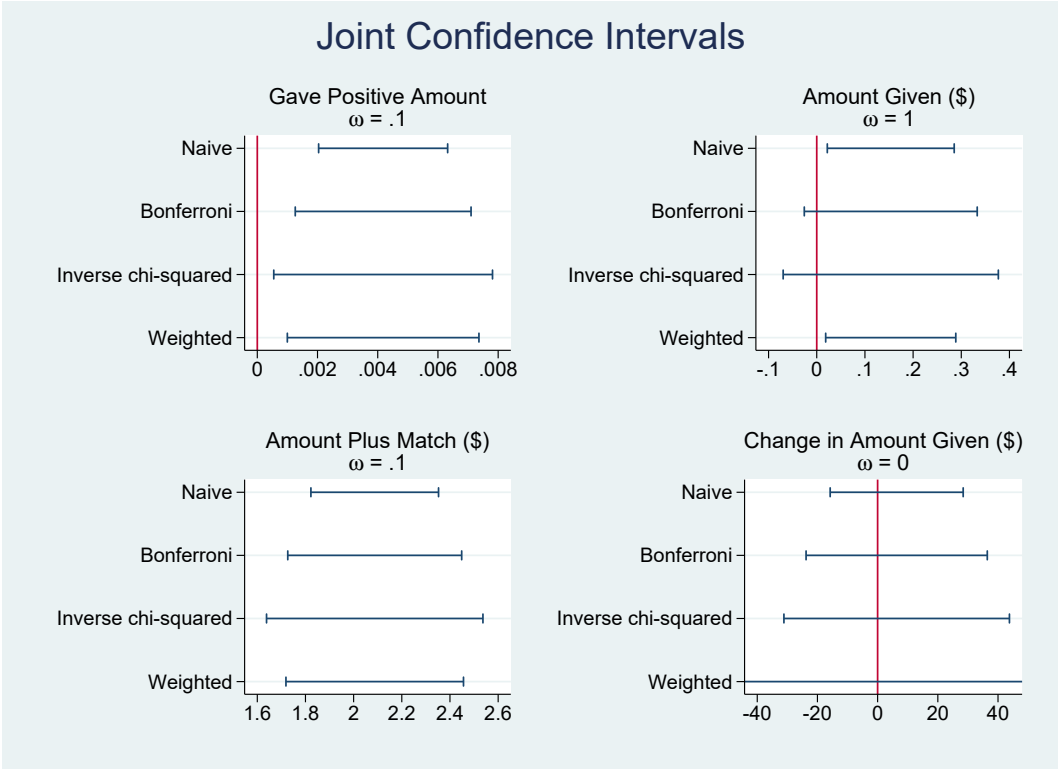


Figure 5: Conventional and weighted simultaneous confidence intervals