Efficient Likelihood Ratio Confidence Intervals using Constrained Optimization*

Gregor Reich
Dept. of Banking and Finance
University of Zurich

Kenneth Judd
Hoover Institution
Stanford University

January 29, 2020

Abstract

Using constrained optimization, we develop a simple, efficient approach (applicable in both unconstrained and constrained maximum-likelihood estimation problems) to computing profile-likelihood confidence intervals. In contrast to Wald-type or score-based inference, the likelihood ratio confidence intervals use all the information encoded in the likelihood function concerning the parameters, which leads to improved statistical properties. In addition, the method does no suffer from the computational burdens inherent in the bootstrap. In an application to Rust's (1987) bus-engine replacement problem, our approach does better than either the Wald or the bootstrap methods, delivering very accurate estimates of the confidence intervals quickly and efficiently. An extensive Monte Carlo study reveals that in small samples, only likelihood ratio confidence intervals yield reasonable coverage properties, while at the same time discriminating implausible values.

1 Introduction

Inherent to most parameter estimation problems is the variation of the estimator arising from the impossibility to "perfectly" (or exhaustively) sample the population from which to recover the parameters. Several approaches to estimation error quantification exist. The most prominent concept is the confidence set: If the sampling from the population were to be repeated, how to construct a subset of the parameter space that would—from an ex-ante point of view—contain the true underlying parameter with a fixed, predetermined coverage probability? In this note, we develop a simple approach to compute one-dimensional profile likelihood confidence intervals—a type of confidence interval known for its advantageous statistical properties, but avoided due to its perceived computational complexity. We show how to efficiently compute it by expressing it as a constrained optimization problem, and compare the approach to other popular types of confidence intervals.

In particular, the following three approaches to the construction of confidence sets are relevant in this note and thus briefly outlined here—presumably in decreasing order of popularity: First, standard-error-based or Wald confidence sets take the asymptotic distribution of an estimator, say $\hat{\theta}$,—under appropriate regularity conditions a normal distribution—and assign to the $\gamma \cdot 100\%$ confidence set all parameter values θ_0, for which $\hat{\theta}$ has a p-value larger than or equal to $1-\gamma$ under $H_0 : \theta = \theta_0$. Since the covariance matrix of the asymptotic distribution

*We thank Philipp Eisenhauer, Robert Erbe, Janos Gabler, Philipp Müller, Harry Paarsch, seminar participants at the University of Zurich, and participants of the Stanford Institute on Theoretical Economics 2018 Asset Pricing session for helpful comments and discussions. Reich gratefully acknowledges financial support from the Swiss National Science foundation under grant P2ZHP1-164978, and from the Fonds zur Förderung des akademischen Nachwuchses (FAN), Zurich.

is estimated consistently by the inverse Hessian of the likelihood function at its maximum, it is straightforward to numerically verify inclusion in the confidence set for any given parameter value, or represent the confidence set implicitly; moreover, the confidence interval for a one-dimensional parameter can be computed explicitly as the $0.5 \cdot (1 \pm \gamma)$ quantiles of the univariate normal distribution. While handy to compute, Wald confidence sets existentially require that the distribution is close to normal; if not, the resulting coverage might be substantially different from γ.

Second, there is the bootstrap which mimics drawing from the model-implied distribution by re-sampling the data set. Running the estimation on each of the so obtained data sets provides a non-parametric estimate of the distribution of the original estimator, from which properties of the distribution like standard errors can be obtained. For a one-dimensional parameter, sample quantiles of the distribution can be taken to approximate the confidence intervals. While the bootstrap does not per se require many assumptions on the underlying distribution of the estimator, it is much more expensive to apply, because usually hundreds or even thousands of full estimation runs have to be performed. Moreover, if the confidence intervals are directly approximated using sample quantiles, we are confronted with a tail sampling problem, which requires either a very high number of samples to yield robust results, or more sophisticated methods which might in turn require stronger assumptions.

Third, likelihood ratio confidence sets lie somewhat in between standard errors and bootstrapping complexity- and requirement-wise. The likelihood ratio test uses the fact that the difference in log-likelihood between a model and a restricted version of it approximately follows a χ^2 distribution, whose number of degrees of freedom is determined by the number of parameters pinned down through the restriction. Similar to the Wald confidence sets, this implies a confidence set through test inversion, where all parameters with a likelihood above the critical value (seen from its maximum) enter the set. The key advantage of this confidence set is that while the critical value is also obtained from an asymptotic result, it uses the information about the parameters given the data encoded in the likelihood not only at its maximum, but in particular close to the boundaries of the confidence interval. Therefore, inference with log-likelihood functions dissimilar to a quadratic function is much more reliable compared to standard error approaches in terms of coverage; moreover, regions with zero likelihood (such as unit roots estimated through unconditional likelihood, non-positive variances, etc.) cannot be part of a likelihood ratio confidence interval. Finally, the extension of confidence intervals for (non-linear) continuous, scalar-valued transformations of the parameter vector is straightforward.

Given the preferable properties of the likelihood ratio confidence sets compared to the Wald sets, and its better numerical efficiency compared to the bootstrap, the lack of popularity is probably due to the fact that its computation requires the inversion of a level set condition (i.e. finding all argument values θ that have likelihood larger or equal to the critical value). This note contributes to the literature an efficient and straightforward way to compute it using standard constrained optimization techniques, in particular in light of the following two additional requirements:

First, confidence of parameters is often assessed on a per-parameter basis. While this ignores important information about the potential correlation in the estimation error across parameters, it is a practical requirement if the results are to be presented in tables alongside with the point estimates, for example. This calls for a need to "marginalize" the confidence sets. The marginalization of Wald style sets with normally distributed estimators is straightforward because the variances of individual components is also normal, and its variances can be read off directly from the diagonal of the joint covariance matrix. Similarly, the non-parametric confidence sets from the bootstrap are typically defined for one-dimensional sub-vectors and are thus inherently marginal. In contrast, the marginalization of likelihood ratio confidence sets requires that the level set condition is formulated in terms of the profile likelihood, which "optimizes out" all but the parameter of interest. Consequently, the level set problem has an inherent bi-level nature,

which can cause—if applied naïvely—severe numerical difficulties. In this note, we show how to reformulate the problem and solve it using standard constrained optimization software.

Second, the use of constrained maximum likelihood estimation is of increasing interest, for example in form of structural estimation. In particular, we relate this note to the seminal papers of Rust (1987) and later Su and Judd (2012), who provided efficient methods to estimate the structural parameters of dynamic economic models whose solution can be represented as a system of non-linear equalities. If these equalities enter the likelihood maximization problem as constraints, and thus the model variables are part of the estimation as it is the case with MPEC estimation (Su and Judd, 2012), standard Wald-type inference becomes more involved. Aitchison and Silvey (1958) provide conditions under which the resulting estimator is (asymptotically) normally distributed, and indicate a numerical procedure to obtain its covariance matrix, which is, however, generally rank-deficient. On the other hand, the bootstrap is not affected by the addition of constraints as it relies on the estimates only. We show that our approach to compute likelihood ratio confidence intervals using constrained optimization naturally incorporates the model constraints as well in an MPEC-like fashion.

To assess the performance of all confidence interval types introduced above in a practical example, we first apply them to the estimation of the well-known bus engine replacement model of Rust (1987), a dynamic discrete choice estimation problem. We replicate the original estimates and standard errors, compute Wald-type, bootstrapping, and (profile) likelihood ratio confidence intervals, and compare the respective timings. We find that all methods deliver plausible results on the original data set, but timings differ by orders of magnitude, with the likelihood ratio constrained optimization method ranking between Wald (fastest) and bootstrapping. Second, we analyze their respective coverage properties in a Monte Carlo study, assessing the question to which extend the nominal confidence level is met. We find that standard error based methods (both Fisher information and bootstrapping based) as well as likelihood ratio intervals yield respectable coverage results even for small data sets, but only likelihood ratio and bootstrapping sample quantiles can discriminate implausible values (because they are both strictly likelihood based). Moreover, only likelihood ratio confidence intervals yield adequate coverage for a non-linear transformation of the parameters—a parameter ratio—on the small data sets.

The remainder of this note is organized as follows: Section 2 introduces the problem of constrained maximum likelihood estimation by taking the example of structural estimation, briefly outlines solution approaches, and finally develops and formally justifies our constrained optimization approach to compute one-dimensional, marginal likelihood ratio confidence intervals.[1]Section 3 applies all three variants of confidence interval computation discussed above to the bus engine replacement model of Rust (1987) and compares their implications and timings; furthermore, we compare their coverage properties using several hundred simulated data sets for the model. Section 4 concludes.

2 Efficient Likelihood Ratio Confidence Intervals using Constrained Optimization

In this note, we consider the estimation of structural economic models of the following structure: Let $\theta \in \Theta \subseteq \mathbb{R}^p$ be a vector of structural parameters; $\sigma \in \Sigma \subseteq \mathbb{R}^m$ a vector of endogenous variables; and $x \in \mathcal{S} \subseteq \mathbb{R}^n$ a vector of state variables.[2] Suppose that the economic model relates

[1]We recently became aware of Wu and Neale (2012), who formulate the idea of treating the likelihood level as an (equality) constraint to obtain likelihood ratio confidence intervals in the context of unconstrained likelihood estimation, without formal justification.

[2]We assume that the model can be expressed in a finite manner. For example, in a dynamic programming model, the state space is either discretized, so that the value function is a discrete function, too, or the value function is approximated using a finite-dimensional representation.

those objects through a system of inequalities:

$$h(x;\sigma,\theta)=0,\ \forall x\in\mathcal{S}. \tag{1}$$

For example, in a dynamic programming model, the parameters θ parametrize the payoff function $\pi(\cdot;\theta)$ and the law of motion of the state, $Pr(x'|x;\sigma,\theta)$; the endogenous variables characterize the value function $V_\theta(x;\sigma)$ (or, usually, its approximation $\hat{V}_\theta$), which itself implies optimal decisions as a function of the state variables; and (1) represents the Bellman equation $V_\theta(x;\sigma)=T(V_\theta)(x;\sigma)$, which has to hold for all states in the state space.

Given data on some[3] of the states and the endogenous variables or any function thereof, $\{\tilde{x}_{t,i},\tilde{\sigma}_{t,i}\}_{t=1,i=1}^{T,I}$, it is common to estimate the vector of structural parameters θ that best explains the data through the model. In this note, we focus on estimation by maximum likelihood: Suppose the model implies a probability distribution of the states and the endogenous variables (or any function thereof) in dependence of the parameter vector, $f(\cdot;\sigma,\theta)$. Then, the likelihood function L is the distribution f—or, usually, the logarithm thereof—for a particular data set, seen as a function of the parameter: $L(\theta;\sigma,\{\tilde{x}_{t,i},\tilde{\sigma}_{t,i}\}_{t=1,i=1}^{T,I})\equiv\log f(\{\tilde{x}_t,\tilde{\sigma}_t\}_{t=1}^{T};\sigma,\theta)$, or, shorter, $L(\theta;\sigma)$. The maximum likelihood estimate (MLE) corresponds to its maximizer, $\hat{\theta}\equiv\arg\max_\theta L(\theta;\sigma)$.

It is important to note that the likelihood and in particular its maximizer are only interpretable if the corresponding endogenous variables σ are chosen such that the model (1) is actually solved. Rust (1987) proposed to solve the model equations for every evaluation of the likelihood function. Consequently, the endogenous variables (and thus the likelihood) can be regarded as an implicit function of the structural parameters,

$$\psi(\theta):\theta\mapsto\sigma\in\{\varsigma\in\Sigma\mid(\forall x\in\mathcal{S})[h(x;\varsigma,\theta)=0]\}, \tag{2}$$

and the estimation problem reads

$$\hat{\theta}=\arg\max_{\theta\in\Theta}\ L(\theta;\psi(\theta)). \tag{3}$$

Because the initial application domain of Rust (1987) was a value function problem, he referred to this as the *nested fixed point algorithm* (NFXP).

Su and Judd (2012) proposed to encode this dependency in a constrained optimization problem:

$$(\hat{\theta},\hat{\sigma})=\arg\max_{\theta\in\Theta,\sigma\in\Sigma}\ L(\theta;\sigma) \tag{4a}$$

$$\text{s.t. } h(x;\sigma,\theta)=0,\ \forall x\in\mathcal{S}. \tag{4b}$$

Since f often represents some kind of optimality or equilibrium conditions, this type of estimation is called *mathematical programming with equilibrium constraints* (MPEC), and guarantees that the MLE both solves the model and maximizes the likelihood at the same time.

It is well-known that efficient confidence sets for maximum likelihood estimators can be obtained from inverting the likelihood ratio test. Rather than to solely rely on the asymptotic distribution of the estimator itself, the idea behind this type of confidence set is to obtain—again through an asymptotic result—a critical value of the likelihood function (more precisely: of a quadratic approximation around its maximum) to quantify the potential variation coming from the finiteness and randomness of the sample. Given such a critical value, we can then invert the likelihood function to obtain all parameter values that have a likelihood equal or larger than this value.

[3] In many instances, only a subset of the states and/or the endogenous variables are observed by the econometrician; for details on how to treat unobserved variables, in particular if they are serially correlated, see, for example, Reich (2018).

For notational brevity, let $\hat{L}_{\gamma,p} \equiv L(\hat{\theta};\hat{\sigma}) - 0.5\chi^2_p(\gamma)$, where $\chi^2_p(\cdot)$ denotes the quantile function of the χ^2 distribution with p degrees of freedom. Then, a (joint) $\gamma \cdot 100\%$ *likelihood ratio confidence set* for the (full) parameter vector θ is given by

$$\widehat{CS}_\gamma \equiv \left\{ \theta \in \Theta \;\middle|\; (\exists \sigma \in \Sigma) \left[L(\theta;\sigma) \geq \hat{L}_{\gamma,p} \wedge (\forall x \in \mathcal{S})[h(x;\sigma,\theta) = 0] \right] \right\}. \tag{5}$$

Obviously, computing $\widehat{CS}_\gamma \subset \mathbb{R}^p$ for $p > 1$ is non-trivial, as the representation of arbitrary sets is a numerical challenge. Moreover, the presentation and interpretation of estimation results often require scalar quantities in order to represent them in tables. Therefore, we focus on the computation of one-dimensional *profile likelihood ratio confidence intervals* in this note: Let $\theta_i \in \Theta_i$ be the ith scalar component of θ, and let $\theta_{-i} \in \Theta_{-i}$ denote the vector of the remaining $p-1$ components. The profile likelihood function of parameter θ_i is the maximum of the likelihood w.r.t. θ_{-i}, seen as a function of the "restricted" parameter θ_i:

$$L_p(\theta_i) \equiv \max_{\theta_{-i} \in \Theta_{-i}} L(\theta_i, \theta_{-i}; \sigma). \tag{6}$$

Recall that the likelihood ratio confidence set is an inverted likelihood ratio test. The general set in (5) jointly tests a restriction of the model on the full parameter. However, we can also test the restricted model for each dimension of the parameter individually, where the model restriction is now expressed through the profile likelihood. This yields the $\gamma \cdot 100\%$ profile likelihood ratio confidence interval (LRCI) for θ_i as

$$\widehat{CI}_{\gamma,i} \equiv \left\{ \theta_i \in \Theta_i \;\middle|\; (\exists \sigma \in \Sigma) \left[\left(\max\nolimits_{\theta_{-i} \in \Theta_{-i}} L(\theta_i, \theta_{-i}; \sigma) \right) \geq \hat{L}_{\gamma,1} \wedge (\forall x \in \mathcal{S})[h(x;\sigma,\theta) = 0] \right] \right\}. \tag{7}$$

For more details on (profile) likelihood ratio confidence intervals and sets, see, for example, the textbooks of Pawitan (2013), or Held and Sabanés Bové (2014).

Next, consider the following relaxation of (7):

$$\widetilde{CI}_{\gamma,i} \equiv \left\{ \theta_i \in \Theta_i \;\middle|\; (\exists \sigma \in \Sigma)(\exists \theta_{-i} \in \Theta_{-i}) \left[L(\theta_i, \theta_{-i}; \sigma) \geq \hat{L}_{\gamma,1} \wedge (\forall x \in \mathcal{S})[h(x;\sigma,\theta) = 0] \right] \right\} \tag{8}$$

We now show that this relaxation is actually tight:

Proposition 1. $\widehat{CI}_{\gamma,i} = \widetilde{CI}_{\gamma,i}$.

Proof. Take any $\theta_i \in \widehat{CI}_{\gamma,i}$; if there exists a $\sigma \in \Sigma$ such that $\max_{\theta_{-i} \in \Theta_{-i}, \sigma \in \Sigma} L(\theta_i, \theta_{-i}; \sigma) \geq \hat{L}_{\gamma,1}$ exists on the restriction $h(x;\sigma;\theta) = 0$, $\forall x \in \mathcal{S}$, then, trivially, there exists a $(\theta_{-i}, \sigma) \in \Theta_{-i} \times \Sigma$ such that $L(\theta_i, \theta_{-i}; \sigma) \geq \hat{L}_{\gamma,1}$ on $h(x;\sigma;\theta) = 0$, $\forall x \in \mathcal{S}$, if the maximum of the (profile) likelihood exists, which we tacitly assume. Thus, $\widehat{CI}_{\gamma,i} \subseteq \widetilde{CI}_{\gamma,i}$.

Conversely, the existence of a $(\theta_{-i}, \sigma) \in \Theta_{-i} \times \Sigma$ such that $L(\theta_i, \theta_{-i}; \sigma) \geq \hat{L}_{\gamma,1}$ on $h(x;\sigma;\theta) = 0$, $\forall x \in \mathcal{S}$ implies that—for the same σ—also $\max_{\theta_{-i} \in \Theta_{-i}} L(\theta_i, \theta_{-i}; \sigma) \geq \hat{L}_{\gamma,1}$ on $h(x;\sigma;\theta) = 0$, $\forall x \in \mathcal{S}$, because we only require greater or equal for the level set condition. Consequently, $\widehat{CI}_{\gamma,i} \supseteq \widetilde{CI}_{\gamma,i}$. □

Given that we cover only one-dimensional confidence intervals in this note, obtaining the boundary of $\widetilde{CI}_{\gamma,i}$ (and the boundary of $\widehat{CI}_{\gamma,i}$) corresponds to the min and max of θ_i in $\widetilde{CI}_{\gamma,i}$ ($\widehat{CI}_{\gamma,i}$), respectively:[4]

$$\widehat{CI}_{\gamma,i} = \widetilde{CI}_{\gamma,i} = \left[\hat{\theta}_{i,l}, \hat{\theta}_{i,u} \right], \tag{9}$$

[4] At this point, we tacitly assume that the likelihood ratio confidence interval really is an interval, and not a set of several disconnected interval, which can arise from non-monotonic likelihood functions. If the case, our confidence interval estimator defines the convex hull of the sub-intervals, and will thus be overly conservative (i.e. it will implement a higher coverage than nominally prescribed by γ).

with

$$\hat{\theta}_{i,l} = \min_{\theta_i \in \Theta_i} \left\{ \theta_i \;\middle|\; (\exists \sigma \in \Sigma)(\exists \theta_{-i} \in \Theta_{-i}) \left[L(\theta_i, \theta_{-i}; \sigma) \geq \hat{L}_{\gamma,1} \wedge (\forall x \in \mathcal{S})[h(x; \sigma, \theta) = 0] \right] \right\}, \tag{10a}$$

$$\hat{\theta}_{i,u} = \max_{\theta_i \in \Theta_i} \left\{ \theta_i \;\middle|\; (\exists \sigma \in \Sigma)(\exists \theta_{-i} \in \Theta_{-i}) \left[L(\theta_i, \theta_{-i}; \sigma) \geq \hat{L}_{\gamma,1} \wedge (\forall x \in \mathcal{S})[h(x; \sigma, \theta) = 0] \right] \right\}. \tag{10b}$$

Note that both boundary elements in (10) constitute constrained optimization problems, similar to the original estimation problem (4), but with the likelihood function now entering through the level set condition as a constraint itself:

$$\hat{\theta}_{i,l} = \min_{\theta_i \in \Theta_i, \theta_{-i} \in \Theta_{-i}, \sigma \in \Sigma} \theta_i \tag{11a}$$

$$\text{s.t. } L(\theta_i, \theta_{-i}; \sigma) \geq \hat{L}_{\gamma,1} \tag{11b}$$

$$h(x; \sigma, \theta) = 0, \; \forall x \in \mathcal{S}, \tag{11c}$$

and $\hat{\theta}_{i,u}$ as the analogous maximization problem. Thus, likelihood ratio confidence intervals for each parameter can be obtained by solving 2 constrained optimization problems similar to the constrained optimization formulation of the original estimation problem (4), implying a total of $2p$ constrained optimization problems.

We conclude this section with two remarks.

Remark 1. Confidence intervals for continuous transformations $\tau : \Theta \to \mathbb{R}$—non necessarily one-to-one—can be obtained as

$$\widetilde{CI}_{\gamma,\tau} \equiv \left\{ t \in \mathbb{R} \;\middle|\; (\exists \sigma \in \Sigma)(\exists \theta \in \Theta) \left[\tau(\theta) = t \wedge L(\theta; \sigma) \geq \hat{L}_{\gamma,1} \wedge (\forall x \in \mathcal{S})[h(x; \sigma, \theta) = 0] \right] \right\} \tag{12}$$

In fact, (8) is a special case of (12) with $\tau : \theta \mapsto \theta_i$.

To obtain the boundaries of $\widetilde{CI}_{\gamma,\tau}$, we solve the equivalent problem to (11), but with objective function $\tau(\theta)$ in (11a).

Remark 2. Although called "likelihood ratio" confidence interval, the logic behind the inversion of the likelihood ratio test and the corresponding asymptotic results carry through for many other extremum estimators, including GMM estimators with appropriate weighting matrices (c.f. Newey and West, 1987), and thus render our constrained optimization approach applicable.

3 Application: The Model of Harold Zurcher (Rust, 1987)

In the bus engine replacement model of Rust (1987), a dynamic discrete choice problem, the manager of a fleet of public transportation buses periodically examines his buses to decide if a bus can stay in service after some regular maintenance work (whose cost are increasing in the mileage traveled by the bus so far), or whether he has to completely overhaul it, in particular replacing its engine (which sets back the odometer to zero, and thus reduces expected future maintenance costs). The cost trade-off the manager faces is estimated from data on his decisions for all buses of the fleet for a particular time window, as well as the corresponding odometer readings when the decision was made.

Formally, the agent faces (immediate) costs per bus and period like

$$u(x, i; \theta_{11}, RC) + \epsilon(i) \equiv \begin{cases} -RC + \epsilon(1) & i = 1 \\ -\theta_{11} \cdot x + \epsilon(0) & i = 0 \end{cases}, \tag{13}$$

where decision $i = 1$ encodes complete overhaul with engine replacement, and $i = 1$ encodes regular maintenance; x is the mileage of respective bus since last replacement. $\epsilon(i)$ forms a

random utility component which is observed by the manager before making the decision; as usual in this literature, they are assumed to be independently and identically extreme value type I (EV1) distributed. Following Rust (1987), we discretize mileage into bins of 5.000 miles each, with a maximum of 90 bins (i.e. 450.000 miles), and regress the parameters on the indices of the bins. Depending on the decision and the mileage today, x and i, a bus' mileage state in the next period, say x', will evolve according to the following law of motion,

$$\theta_{3\Delta} \equiv Pr(x' = (1-i)x + \Delta | x, i), \tag{14}$$

which is estimated alongside with the cost parameters.[5]

Assuming that the agent behaves dynamically optimal, i.e. that he takes into account the fact that replacement today will potentially save him maintenance cost in the future, Rust (1987) imposes the following Bellman equation as a sufficient optimality condition for minimal expected discounted future cost (omitting parameter dependence of the model primitives for brevity from now on):

$$V_\theta(x, \epsilon) = \max_{i \in \{0,1\}} \left\{ u(x, i) + \epsilon(i) + \beta \mathbb{E}\left[V_\theta(x', \epsilon') | x, i \right] \right\} \equiv T[V_\theta](x, \epsilon). \tag{15}$$

Note that this is an equation in the value function V_θ. Since the integration needed to compute the conditional expectation in (15) is over a non-smooth function (due to the presence of the max operator), but has a partial closed-form solution under the EV1 assumption due to Rust (1987), usually the expected Bellman equation is solved instead,

$$\begin{aligned} EV_\theta(x, i) &\equiv \mathbb{E}\left[V_\theta(x', \epsilon') | x, i \right] && \text{(16a)} \\ &= \sum_{\Delta \in \{0,1,2\}} \log \left(\sum_{j \in \{0,1\}} \exp\left(u((1-i)x + \Delta, j) + \beta EV_\theta((1-i)x + \Delta, j) \right) \right) \theta_{3\Delta} && \text{(16b)} \\ &\equiv T[EV_\theta](x, i). && \text{(16c)} \end{aligned}$$

Note that since $x \in \{1, \ldots, 90\}$ and due to the fact that $EV_\theta(x, 1) = EV_\theta(1, 0)$, (16) constitutes a system of 90 equations in 90 unknowns.

It remains to derive the likelihood function of the model. Recall that we use data on mileage and decisions, $\{x_t, i_t\}_{t=1}^T$. For a given bus, the joint (log-)likelihood of the parameter vector $\theta \equiv (RC, \theta_{11}, \theta_{30}, \theta_{31}, \theta_{32})$ can be factored as[6]

$$L(\theta; \{x_t, i_t\}_{t=1}^T) = \log \left(\prod_{t=1}^T Pr(i_t | x_t) Pr(x_t | x_{t-1}, i_{t-1}) \right). \tag{17}$$

While the mileage transition probabilities $Pr(x_t | x_{t-1}, i_{t-1})$ are parameters themselves (see equation 14), in the presence of EV1 errors the choice probabilities $Pr(i_t | x_t)$ can be computed directly using

$$Pr(i|x) = \frac{\exp(u(x, i) + \beta EV_\theta(x, i))}{\sum_{j \in \{0,1\}} \exp(u(x, j) + \beta EV_\theta(x, j))}, \tag{18}$$

a result also due to Rust (1987).

Since the function EV_θ enters the choice probabilities, we need to ensure that the (expected) Bellman equation (16) is solved for any parameter value of interest. For our analysis, this includes the maximum likelihood estimate itself as well as all confidence interval boundaries. For the estimation problem, this can be implemented using NFXP (3) or MPEC (4), with the constraint function h now being $EV(x) - T[EV_\theta](x) = 0, \ \forall x$, and σ representing the actual

[5] Since mileage increases by two bins per period at maximum in the data, only θ_{30} and θ_{31} are estimated; θ_{32} follows as $1 - \theta_{30} - \theta_{31}$.

[6] Following Rust (1987), we fix the discount factor at $\beta = 0.9999$.

function values $EV(x)$, $\forall x$. The definitions for the corresponding likelihood ratio confidence interval boundaries (10) for all parameters is analogous.

In addition to the estimates of the original parameters, we also report a non-linear transformation of the two cost parameters, RC/θ_{11}. This transformation is informative in the following sense: Dynamic logit models usually identify the natural level of the utility only up to a multiplicative constant because the variance of the residuals (i.e. the EV1 shocks) is normalized.[7] If, however, an alternative specification with a different residual variance is estimated and compared against the original one, the parameter estimates might vary significantly even if the underlying trade-off did not change much. For example, Reich (2018) estimates a version of the bus engine model where the residuals are potentially serially correlated, finding that the cost parameter estimates almost triple, but their ratio changes by merely two percent, leaving the agent with essentially the same trade-off as in the original specification.

Table 1 presents the main results:[8] The first section reproduces the original estimates from Rust (1987) together with their standard errors, and computes the 95% Wald confidence intervals implied by the 0.025 and 0.975 quantiles of the univariate normal distribution with standard deviation equal to Rust's standard errors centered around the estimates. In the second section (MPEC/LRCI) we report our maximum likelihood estimate of the Rust (1987) model and data using MPEC, and the associated likelihood ratio confidence intervals using the constrained optimization approach developed in Section 2. The third section (MPEC/Wald) reports Wald-type confidence intervals based on the standard errors for constrained maximum likelihood problems developed in Aitchison and Silvey (1958) around our MPEC estimate (see Appendix A); the standard errors for the non-linear transformation of the parameters are obtained using the delta method. The last section (MPEC/Bootstrapping) reports standard errors and sample quantiles from bootstrapping the original data set 1000 times; additionally, we report Wald-type confidence intervals around the MPEC estimate using the bootstrap standard error estimates.

Assessing the quality of the confidence intervals in terms of coverage is not possible without knowledge of the true data generating process parameters; therefore, we assess the difference in coverage in a simulation study below. The comparison of running times already reveals that—as expected—Wald-style confidence intervals are very cheap to compute. At the same time, our constrained optimization approach to likelihood ratio confidence intervals is slower, but far from being prohibitively expensive; indeed, given its (ex-ante) desirable statistical properties, it appears actually quite appealing in terms of cost-benefit ratio. Finally—again as expected—, bootstrapping turns out to be very expensive, in particular given the fact that we are at the lower end of the number of data sets needed (at least for the truly non-parametric sample quantiles).

To obtain evidence on whether or not the confidence intervals are reliable and actually contain the true parameter value in proportion to the confidence level—i.e. if they match the nominal coverage level—, we simulate the model to obtain 400 data sets based on the estimate by Rust (1987) reported in Table 1; we then re-estimate the parameters for each data set and compute the corresponding confidence intervals. (Note that for bootstrapping, this requires us to re-sample each simulated data set 400 times.) This enables us to check for each confidence interval whether the true (and by now known) parameter value is contained or not. Ideally, a fraction equal to the confidence level γ does contain it, whereas a fraction of $1-\gamma$ misses it out. We carry out this study once with data sets of size comparable to Rust (1987) (8'112 observations), and once with much smaller data sets of roughly 10% of the original size (780 observations).[9]

[7] Moreover, the fact that only differences in utility matter leaves another additive constant unidentified.

[8] All experiments have been carried out using MATLAB and the constrained non-linear solver KNITRO interfaced through CasADi (Andersson et al., 2018) for automatic differentiation on a 2012 MacBook Pro with a 2.6GHz Intel Core i7 processor and 16GB 1600MHz RAM.

[9] While less than one thousand observations might appear small in the context of DDCMs, it is actually very realistic in the estimation of dynamic macro models based on monthly post-war data; in fact, the use of quarterly or even annual data can lead to even smaller data sets.

Table 2 reports the main results: Each horizontal section contains one type of confidence interval; the left pane reports the results for the large data sets, the results for the small data sets are on the right. For each combination, we report actual coverage for all parameters and the transformation RC/θ_{11} as well as the smallest lower bound and the largest upper bound over all data sets. The nominal confidence level is 95%. For the large data sets, we find that all types of confidence intervals roughly implement the desired confidence level, even for the non-linear transformation; moreover, also the confidence intervals all span a reasonable range. For the small data sets, the picture is, however, different: Both likelihood ratio and Wald confidence intervals, and, to a lesser extend, also standard error based bootstrapping intervals implement very precise coverage, except for the non-linear transformation for which only likelihood ratio gets adequate coverage. However, both standard error based methods yield minimal values that imply implausible values, such as negative costs or negative probabilities. This is a direct consequence of the fact that they rely on purely local information at the maximum of the likelihood function, namely the Fisher information matrix.[10]

Furthermore, the results are visualized in Figures 1 and 2, which plot the confidence intervals for the replacement cost parameter RC for all data sets stacked on top of each other, sorted according to the left interval boundary. The true parameter value ($RC = 9.7558$) is denoted by a black vertical line. Intervals that contain the true value are colored blue, those that do not contain it are colored in cyan; finally, magenta-colored intervals contain values implying negative costs. For the large data sets, we find that all methods covered in this paper perform roughly equal, with the bootstrapping intervals being slightly looser on average. For the small data sets, however, only the likelihood ratio and the Wald confidence intervals have reasonable coverage and are, at the same time, substantially tighter than the bootstrapping intervals. However, a substantial share of both standard error based intervals contain unreasonable values. Finally, sample quantiles from bootstrapping are very loose and do not achieve good coverage.

We conclude that the likelihood ratio confidence intervals show the best overall performance, and thus appear—given the computational feasibility demonstrated above—an excellent choice for practical purposes. While we are aware that our findings and their interpretations stem from a single model instance, we conjecture them to be quite general properties of the respective methods: First, many models feature natural parameter bounds which cannot be enforced by purely local methods based on standard errors. Second, many bootstrapping implementations are in essence Monte Carlo methods and thus inherit their general applicability—but also their slow convergence. Moreover, for the truly non-parametric version using sample quantiles, we actually face a tail sampling problem, which is inherently inefficient as the majority of samples is drawn in regions of non-interest (i.e. the center of the distribution), or extremely volatile for fat-tailed distributions.

4 Conclusions

In this note, we cast computing the likelihood-ratio confidence interval as a constrained optimization problem. Our approach naturally incorporates inference in models whose solution can be expressed by a system of equations, which include additional constraints to the interval computation problem. Using data from the well-known dynamic programming model of Rust (1987) as a laboratory, we compared our approach to Wald-type methods as well as the bootstrap—using sample quantiles from bootstrapping, too. We found that our method is quite competitive in terms of computational speed. In an extensive Monte Carlo study simulating Rust's (1987)

[10] A popular remedy for this issue is to apply the delta method to estimate a one-to-one transformation of the model whose inverse respects the domain, and then re-transform the corresponding confidence interval boundaries. However, this procedure can be very unstable if the estimates are close to the boundary. In our setup, we regularly observed diverging interval boundaries and rather poor coverage for θ_{11} and RC/θ_{11} for those data sets where domain violations are a problem.

model to estimate the realized coverage of the different interval types, we found that all methods have comparable coverage properties. In terms of computing times, Wald-type intervals are cheapest (on the order of seconds), likelihood-ratio in the middle (minutes), and bootstrapping the most expensive (hours). When we reduced the sample size to about ten percent of the original sample size, we found that while both likelihood ratio and Wald intervals provide good coverage for the parameter estimates, a substantial share of all standard error-based methods contain unreasonable values, such as negative costs; moreover, only likelihood ratio confidence intervals yield adequately coverage for the non-linear transformation of parameters. At the same time, truly non-parametric bootstrapping intervals (sample quantiles) naturally do not contain such values, but yield poor coverage. Even though we have demonstrated these results for one particular example, we believe they extend to other methods and applications.

	Rust (1987)			
	$\hat{\theta}$	SE	—	normal quantiles
RC	9.7558	1.227		[7.351, 12.16]
θ_{11}	2.6275	0.618		[1.416, 3.839]
θ_{30}	0.3489	0.0052		[0.339, 0.359]
θ_{31}	0.6394	0.0053		[0.629, 0.650]
RC/θ_{11}	3.7130	—		—
L	-6,055.250			
time (sec)			—	
	MPEC/LRCI			
	$\hat{\theta}$	—	LRCI	—
RC	9.7584		[8.200, 11.76]	
θ_{11}	2.6275		[1.810, 3.669]	
θ_{30}	0.3489		[0.337, 0.359]	
θ_{31}	0.6394		[0.629, 0.650]	
RC/θ_{11}	3.7139		[3.103, 4.656]	
L	-6,055.250			
time (sec)			288	
	MPEC/Wald			
	$\hat{\theta}$	SE	—	normal quantiles
RC	9.7584	0.8781		[8.038, 11.48]
θ_{11}	2.6275	0.4551		[1.736, 3.520]
θ_{30}	0.3489	0.0053		[0.339, 0.359]
θ_{31}	0.6394	0.0053		[0.629, 0.650]
RC/θ_{11}	3.7139	0.3686		[2.992, 4.436]
L	-6,055.250			
time (sec)			12	
	MPEC/Bootstrapping			
	$\hat{\theta}$	SE	sample quantiles	normal quantiles
RC	9.7584	0.7567	[8.680, 11.84]	[8.230, 11.29]
θ_{11}	2.6275	0.4632	[1.931, 3.844]	[1.695, 3.561]
θ_{30}	0.3489	0.0071	[0.334, 0.362]	[0.335, 0.363]
θ_{31}	0.6394	0.0070	[0.626, 0.654]	[0.625, 0.653]
RC/θ_{11}	3.7139	0.4010	[2.938, 4.586]	[2.928, 4.500]
L	-6,055.250			
time (sec)			6,305	

Table 1: Maximum likelihood estimates, standard errors, and 95% confidence intervals for the model of Rust (1987). Top section (Rust, 1987): original estimates ($\hat{\theta}$), standard errors (SE), and implied Wald confidence intervals (N); second section (MPEC/LRCI): replicated estimates using MPEC and likelihood ratio confidence intervals as derived in Section 2; third section (MPEC/Wald): standard errors for constrained MLE as in Aitchison and Silvey (1958) and implied Wald confidence intervals; bottom section (MPEC/Bootstrapping): standard errors and sample quantiles from bootstrapping and implied Wald confidence intervals. Implied Wald intervals are the 0.025 and 0.975 quantiles of the normal distribution with mean $\hat{\theta}$ and standard deviation equal to the reported standard errors SE; bootstrapping was carried out on 1,000 datasets; running times are serial and measured in seconds.

	LRCI					
	Sample size: 8'112			Sample size: 780		
	coverage	min	max	coverage	min	max
RC	0.961	6.465	21.77	0.958	4.333	153.7
θ_{11}	0.953	0.558	7.888	0.938	7e-16	73.33
θ_{30}	0.955	0.321	0.376	0.958	0.280	0.433
θ_{31}	0.952	0.610	0.667	0.960	0.556	0.712
θ_{32}	0.945	0.006	0.019	0.944	4e-04	0.035
RC/θ_{11}	0.942	2.348	12.07	0.911	1.305	4e07
	Wald					
	Sample size: 8'112			Sample size: 780		
	coverage	min	max	coverage	min	max
RC	0.952	6.367	20.85	0.955	-42.53	132.8
θ_{11}	0.928	0.450	7.404	0.935	-22.60	61.00
θ_{30}	0.955	0.321	0.376	0.952	0.270	0.433
θ_{31}	0.954	0.610	0.667	0.957	0.555	0.726
θ_{32}	0.947	0.006	0.019	0.935	-0.001	0.034
RC/θ_{11}	0.962	2.212	10.30	0.791	-8e04	8e04
	Bootstrapping (sample quantiles)					
	Sample size: 8'112			Sample size: 780		
	coverage	min	max	coverage	min	max
RC	0.928	5.736	20.56	0.675	4.709	350.0
θ_{11}	0.939	0.273	7.723	0.813	1e-12	167.4
θ_{30}	0.955	0.322	0.376	0.918	0.271	0.438
θ_{31}	0.956	0.611	0.666	0.905	0.549	0.727
θ_{32}	0.949	0.006	0.019	0.868	1e-16	0.037
RC/θ_{11}	0.939	2.231	11.11	0.880	1.181	5e12
	Bootstrapping (normal quantiles)					
	Sample size: 8'112			Sample size: 780		
	coverage	min	max	coverage	min	max
RC	0.949	6.563	22.13	0.950	-167.0	382.1
θ_{11}	0.934	0.539	8.659	0.953	-83.99	180.8
θ_{30}	0.954	0.322	0.376	0.930	0.272	0.441
θ_{31}	0.953	0.610	0.667	0.918	0.548	0.725
θ_{32}	0.945	0.006	0.019	0.868	-0.004	0.036
RC/θ_{11}	0.958	2.294	12.45	0.838	-5e13	5e13

Table 2: Realized coverage as well as minimums and maximums of the confidence intervals (95% confidence level) for 1,000 simulated data sets, using the estimates of Rust (1987) as true population parameters. Top section (LRCI): likelihood ratio confidence intervals as derived in Section 2; second section (Wald): Wald confidence intervals based on standard errors for constrained MLE as in Aitchison and Silvey (1958); third section (Bootstrapping, sample quantiles): sample quantiles from bootstrapping; bottom section (Bootstrapping, normal quantiles): Wald confidence intervals based on standard errors from bootstrapping. Bootstrapping was carried out on 1,000 datasets.

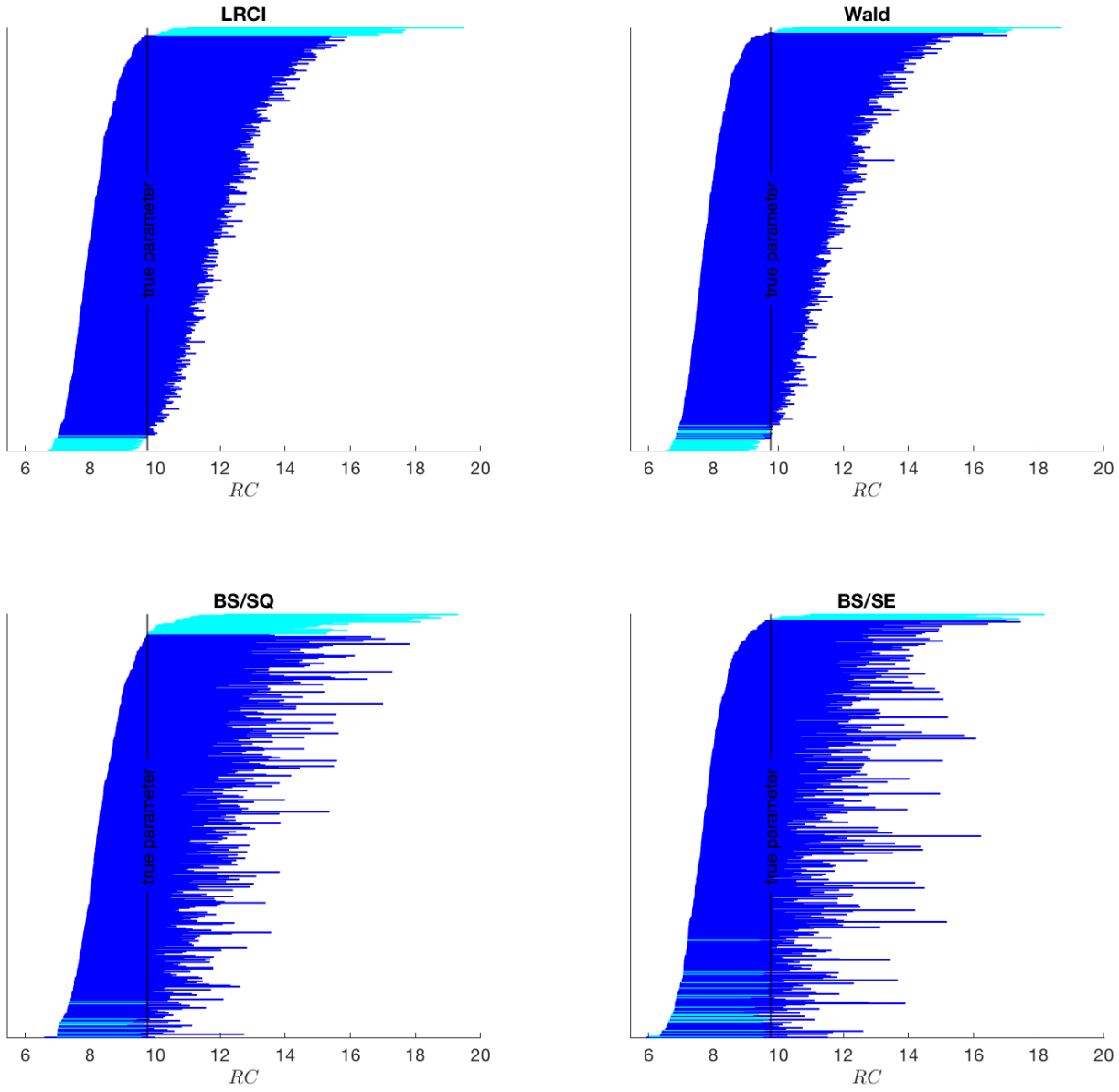

Figure 1: Confidence intervals (stacked; 95% confidence level) for the replacement cost parameter RC for 400 simulated data sets of 8'112 observations each, using the estimate of Rust (1987). Top left: Likelihood ratio CI; top right: Wald CI (normal quantiles, standard errors as in Aitchison and Silvey, 1958); bottom left: Bootstrapping (sample quantiles); bottom right: Bootstrapping (normal quantiles, standard errors from bootstrapping). Blue intervals contain the true parameter (denoted by black vertical line), cyan colored intervals do not. Intervals are sorted by the value of their left boundary.

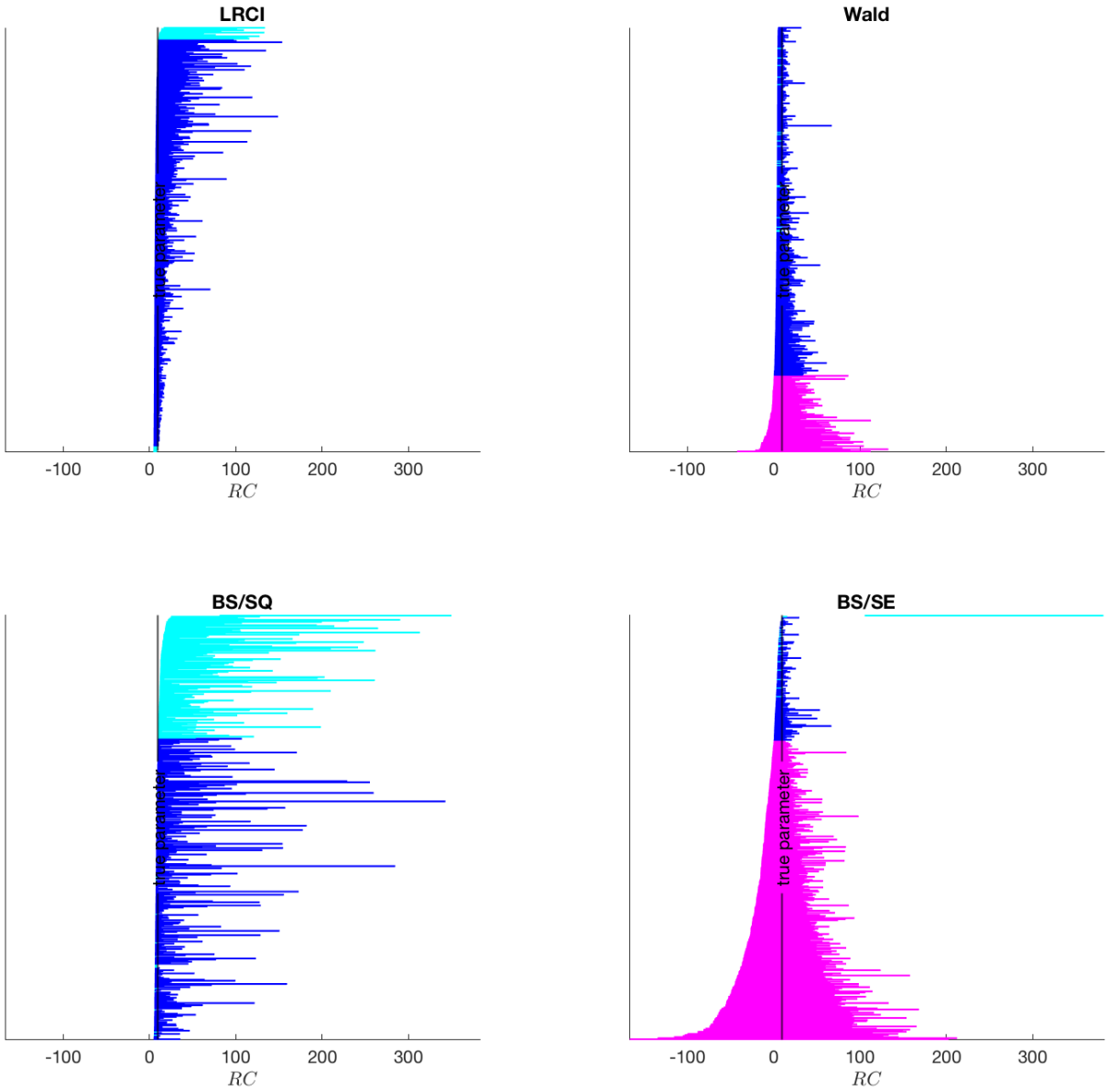

Figure 2: Confidence intervals (stacked; 95% confidence level) for the replacement cost parameter RC for 400 simulated data sets of 780 observations each, using the estimate of Rust (1987). Top left: Likelihood ratio CI; top right: Wald CI (normal quantiles, standard errors as in Aitchison and Silvey, 1958); bottom left: Bootstrapping (sample quantiles); bottom right: Bootstrapping (normal quantiles, standard errors from bootstrapping). Blue intervals contain the true parameter (denoted by black vertical line), cyan colored intervals do not; intervals which cover negative replacement cost values are colored in magenta. Intervals are sorted by the value of their left boundary.

References

Aitchison, J. and Silvey, S. D. (1958). Maximum-Likelihood Estimation of Parameters Subject to Restraints. *The Annals of Mathematical Statistics*, 29(3):813–828.

Andersson, J. A. E., Gillis, J., Horn, G., Rawlings, J. B., and Diehl, M. (2018). CasADi: a software framework for nonlinear optimization and optimal control. *forthcoming in: Mathematical Programming Computation*, 20(3):1–36.

Held, L. and Sabanés Bové, D. (2014). *Applied Statistical Inference*. Springer, Berlin, Heidelberg.

Newey, W. K. and West, K. D. (1987). Hypothesis Testing with Efficient Method of Moments Estimation. *International Economic Review*, 28(3):777–787.

Pawitan, Y. (2013). *In All Likelihood*. Statistical Modelling and Inference Using Likelihood. Oxford University Press.

Reich, G. (2018). Divide and Conquer: Recursive Likelihood Function Integration for Hidden Markov Models with Continuous Latent Variables. *Operations Research*, 66(6):1457–1470.

Rust, J. (1987). Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher. *Econometrica: Journal of the Econometric Society*, 55(5):999–1033.

Su, C.-L. and Judd, K. L. (2012). Constrained Optimization Approaches to Estimation of Structural Models. *Econometrica: Journal of the Econometric Society*, 80(5):2213–2230.

Wu, H. and Neale, M. C. (2012). Adjusted Confidence Intervals for a Bounded Parameter. *Behavior Genetics*, 42(6):886–898.

A Standard Errors for Equality-Constrained Maximum Likelihood Estimation

Aitchison and Silvey (1958) derive the basic mathematical theory for existence and asymptotic normality of maximum likelihood estimators under constraints (or "restraints" as they phrase it). In particular, their result on the asymptotic normality of the estimator (Theorem 2) states the variance-covariance matrix of the distribution. The paper gives an explicit equation for it (p. 823), which, however, relies on an unknown object that is close to, but not exactly equal to the (unrestricted) Hessian of the likelihood function. In fact, the formula cannot be used as-is if this Hessian is singular, which can easily happen. (Consider, for example, a discrete state setting, where some—observable—states in the state space are never reached in the data).

However, when developing an implementable algorithm for solving the constrained problem, Aitchison and Silvey (1958) propose[11] to invert the "augmented" Hessian (p. 826)

$$\begin{bmatrix} P(\theta,\sigma) & Q(\theta,\sigma) \\ Q^T(\theta,\sigma) & R(\theta,\sigma) \end{bmatrix} = \begin{bmatrix} -D^2_{\theta,\sigma}L(\theta,\sigma) & -D_{\theta,\sigma}h(\cdot;\theta,\sigma) \\ -D^T_{\theta,\sigma}h(\cdot;\theta,\sigma) & 0 \end{bmatrix}^{-1}. \tag{19}$$

We take $P(\hat{\theta},\hat{\sigma})$ obtained at the solution to the maximum likelihood problem as an estimator of the variance-covariance matrix of the free parameters. Note, however, that P is generally not of full rank, due to the fact that only a subset of the parameters is free (the block R corresponds to the variance-covariance matrix of the Lagrange multipliers, which are also asymptotically normal).

[11] Aitchison and Silvey (1958) credit other authors for this, but are unable to provide an explicit source.