

Multimodality Adjusted p^{**} -Formula and Confidence Regions

Kees Jan Van Garderen *
University of Amsterdam

Fallaw Sowell
Carnegie Mellon University

November 7, 2016

Abstract

Barndorff-Nielsen's celebrated p^* -formula and variations thereof have amongst their various attractions the ability to approximate bimodal distributions. In this paper we show that in general this requires a crucial adjustment to the basic formula. The adjustment is based on a simple idea and straightforward to implement, yet delivers important improvements. It is based on recognizing that certain outcomes are theoretically impossible and the density of the MLE should then equal zero, rather than the positive density that a straight application of p^* would suggest. This has implications for inference and we show how to use the adjusted p^{**} -formula to construct improved confidence regions. These can be disjoint as a consequence of the bimodality. The degree of bimodality depends heavily on the value of an approximate ancillary statistic and conditioning on the observed value of this statistic is therefore desirable. The p^{**} -formula naturally delivers the relevant conditional distribution. We illustrate these results in small and large samples using a simple nonlinear regression model and an errors in variables model where the measurement errors in dependent and explanatory variables are correlated and allow for weak proxies.

*Corresponding author. K.J.vanGarderen@uva.nl, Dept of Economics and Econometrics, University of Amsterdam, Roetersstraat 11, P.O. Box 15867, 1001 NJ, Amsterdam, The Netherlands.

We thank Grant Hillier, Peter Phillips, Richard Smith and participants at the Cambridge conference in his honour, participants at the NESG conference, the Co-editors, and especially two referees for earlier comments that substantially improved the paper.

1 Introduction

The p^* -formula was developed and investigated by Barndorff-Nielsen (1980, 1983) and later work, building on the original result in Fisher (1934) and related work including Fraser (1968), Daniels (1954, 1980), Barndorff-Nielsen & Cox (1979), Hinkley (1980), Cox (1980), and Durbin (1980). It provides an approximation to the conditional distribution of the maximum likelihood estimator (MLE) $\hat{\theta}$ given an exact or approximate ancillary \mathbf{a} when the true parameter value is θ . The basic formula for the approximate density of $\hat{\theta}$ at $\hat{\theta} = q$ given $\mathbf{a} = a$ is simply:

$$p_{\hat{\theta}}^*(q|a, \theta) = c(\theta, a) |j(q, a)|^{1/2} \exp\{\ell(\theta; y) - \ell(q; y)\}, \quad (1)$$

where $\ell(\theta; y)$ is the log-likelihood function for θ given a sample of observations y , assumed to be regular and two times continuously differentiable, $|j(q, a)|$ is the absolute value of the determinant of the observed information, i.e. minus the second derivative of the log-likelihood, evaluated at the MLE. The norming constant $c(\theta, a)$ should ensure that the density integrates to 1 and in general depends on θ and a .

In some cases p^* is exact (e.g. inverse Gaussian and transformation models, see e.g. Barndorff-Nielsen, 1980, and Daniels, 1980) and in other cases, despite its apparent simplicity, provides a powerful approximation that is not restricted to a small deviation region. It has a relative rather than an absolute error and can capture asymmetry and bimodality of the true distribution.

Care ought to be taken however, when implementing the formula, since not all combinations of q and a are theoretically possible. These values can nevertheless be substituted in the formula and this mechanistic substitution leads to positive values of the approximate density when it should evidently be zero. The first contribution of the paper is making this observation and to identify these points for which the density is logically zero on theoretical grounds. This does not seem to have received any attention in the literature. Asymptotically this problem is negligible in a neighborhood of the true θ . In finite samples however, or cases with weak instruments and little information, this can be crucial as we will demonstrate.

The issue is related to the underlying assumption that the likelihood has a single stationary point and has a unique global maximum. The formula needs adapting when the likelihood is multimodal and has multiple stationary points. In the one parameter case at least one of these stationary points is a local minimum with a positive second derivative of the log-likelihood. There will also be points of inflexion where the observed information is singular and hence $|j(q, a)| = 0$ and the p^* -approximation (1) equals 0, but with a higher value of the likelihood than at the minimum. This implies that the basic p^* -approximation (1) has a higher value in a local minimum of the likelihood than at the points of inflection, regardless of how much higher the likelihood is in those points. This logical inconsistency is easily resolved by setting the density to zero when the observed information is negative, or in the multivariate case has negative eigenvalues. This, however, is not sufficient. There may be other combinations of (q, a) that are theoretically impossible, despite having a positive definite observed information

matrix. We propose a simple adjusted version of the p^* -formula that is zero when the density is theoretically zero.

A second reason for additional care is that for bimodal distributions the normalizing constant c tends to vary much more with a and θ than in standard, unimodal cases. In fact, as stated by Barndorff-Nielsen (1983, p.348), for large enough sample size c can often be approximated well by a constant c_0 independent of a and θ , and in view of the asymptotic normality frequently taken as $(2\pi)^{-d/2}$, with d the dimension of θ . The reason this does not hold in multimodal cases is that $1/c(\theta, a)$ is obtained by an integration explained below. Combinations of (q, a) that are impossible make zero contribution to this integral. This heavily depends on the value of a and the integral can be much smaller, leading in some cases to extremely large values of $c(\theta, a)$, even if the sample size is large.

The asymptotic theory for p^* and related saddlepoint approximations has been well established. Under standard regularity conditions only one unique mode of the likelihood near the true θ_0 remains as the sample size increases beyond all bounds. Other modes will disappear and the MLE is consistent. The adjusted p^{**} -approximation introduced below will then coincide with the standard p^* -formula. This has the obvious advantage that asymptotic theory already established simply carries over from the original p^* -approximation to the adjusted p^{**} . On the other hand, superiority of the adjusted formula can therefore not be shown by asymptotic techniques. We use logical arguments why the density must be theoretically zero in certain regions. In those regions our p^{**} is strictly dominating other approximations that are not zero.

For outcomes with positive density, we show in Appendix 3 that the method of proof in Barndorff-Nielsen (1980) also applies to p^{**} . The proof is based on a transformation of variables which is valid regardless of the sample size and not dependent on asymptotic arguments. The Jacobian of the transformation is relevant for multimodal cases and other situations where the value of the MLE is not necessarily in a neighborhood of the true θ_0 . The method of proof is related to Hillier and Armstrong (1999) who derive the exact distribution of the MLE even when no explicit formula for the MLE is available. They apply this to exponential regression and show how the p^* -formula follows from their result when there is no conditioning.

We will use a number of stylized econometric models that, despite their simplicity, have interesting features and show that the p^{**} -formula is extremely accurate. The nonlinear regression model illustrates large possible differences with the p^* -formula. In weak instrument type asymptotics, when information does not accrue at the usual rate, techniques that are relevant in small samples, like our p^{**} -approximation may still be relevant. We show this to be the case for an errors in variables type model with correlated measurement errors in the dependent and independent variables and weak proxies.

The examples are members of the general class of curved exponential models introduced by Efron (1975). This class has received particular attention in the development of p^* and other saddlepoint related approximations (reviewed in e.g. Barndorff-Nielsen and Cox, 1994). They serve as a paradigm for smooth models where the dimension of the minimal sufficient statistic

is larger than the dimension of the parameter, similar to an overidentified GMM setting. Van Garderen (1997) gave a number of prominent examples in Econometrics and in this paper we will use this class of models and examples to make points that are relevant more generally outside the class of exponential models.

The second main contribution of the paper is to show how the adjusted p^{**} -formula can be applied successfully to construct confidence regions. In multimodal cases smallest prediction and confidence sets can consist of unconnected regions. The adjusted p^{**} -formula can be extremely useful in improving over standard first order techniques when constructing confidence regions. In particular when the distribution of the MLE is seriously bimodal, this cannot be captured by techniques based on first or even higher order Edgeworth type approximations. We will use p^{**} to construct smallest prediction- or acceptance regions and invert them to obtain reliable confidence regions for the parameter of interest. These confidence regions can be disjoint, but this will depend heavily on the value of the ancillary statistic. It is therefore very important to condition on the actual sample value of the ancillary statistic.

The modified confidence regions have conditionally accurate coverage levels, and are therefore also unconditionally accurate. The standard confidence intervals suffer from two defects. First, the overall (marginal) coverage rate can be far too low. We give an example where coverage is only 80% for a nominal 95% confidence interval. Second, the coverage rate varies significantly with the value of the ancillary. In the example it drops below 50% for values of the ancillary that are not extreme. The reason is that if the density of the MLE has a substantial second mode this is ignored by standard methods. Our adjusted p^{**} -based confidence regions are less than 1% point from its nominal value.

These results are relevant from a theoretical and practical point of view in econometrics. Many models of interest in econometrics have multiple solutions to the first order conditions of the criterion function, be it likelihood, GMM, or other methods, and multimodality is of increasing concern. Our inference procedure provides a solution to the apprehension one might experience when the global optimum differs only slightly from another local optimum.

The paper is organized as follows. Section 2 briefly discusses curved exponential models, introduces the two leading examples, and provides arguments for the adjustment of p^* . Section 3 considers a partitioning of the sample space to formalize the sets where the density is identically zero and used to adjust the p^* -formula defined and illustrated in Section 5, after having introduced the affine ancillary used for conditioning in Section 4. Section 6 deploys the adjusted p^{**} -formula to obtain conditional confidence intervals and shows the superiority over standard methods based on the Normal approximation of the MLE. Section 7 concludes. Proofs and further results are relegated to the Appendix.

2 Curved Exponential Models

Curved exponential models were introduced by Efron (1975) and instrumental in the development of differential geometry in statistics and the derivation of the p^* -formula by Barndorff-Nielsen (1980). He also proposed an ancillary to serve as conditioning statistic in his p^* -formula. We will also use and discuss this score based ancillary statistic here, because it allows adapting his method of proof to multi-modal likelihoods. A Curved Exponential Model of order (k, d) , or CEM(k, d) for short, can be characterized by the property that the number of parameters in the model, d , is smaller than the dimension of the minimal sufficient statistic, k , as was shown by Van Garderen (1997) in an extension of the familiar theorem of Pitman-Koopman-Darmois to CEMs. He further showed that a variety of leading econometric models, including nonlinear regression models, the Single Structural Equations (SSE) model, the Seemingly Unrelated Regressions (SUR) model, Vector Autoregressive (VAR) models (and the cointegrated VAR investigated by Mavroeidis and Van Garderen, 2006), are all examples of CEMs.

An exponential family is a family $\mathcal{P} = \{\mathcal{P}_\eta | \eta \in \mathcal{N}\}$ of distributions with densities that can be written as:

$$p_Y(y; \eta) = \exp\{\langle \eta, t(y) \rangle - \kappa(\eta)\} h(y), \quad (2)$$

with respect to a common σ -finite dominating measure on the sample space of Y . The parameter η is a vector with values in a k -dimensional vector space, t is a k -dimensional vector function of y or of Y when $T = T(Y) \equiv t(Y)$ is a random variable taking values in a sample space \mathbb{T} , and $\langle \cdot, \cdot \rangle : \mathcal{N} \times \mathbb{T} \rightarrow \mathbb{R}$ a non-degenerate bilinear product, which in our derivations will be of the form $\langle \eta, t \rangle = \eta' t$. The representation is minimal if k is the smallest integer such that (2) holds, and (2) is a Full Exponential Model of order k , or FEM(k) for short, assuming η is genuinely k dimensional in the sense that there is a k -dimensional cube in \mathcal{N} in which η can take any value. If η is restricted to lie on a smooth subset \mathcal{M} of \mathcal{N} such that $\eta = \eta(\theta)$ is a smooth continuously differentiable function of a d -dimensional parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ with $d < k$, then the model is a CEM(k, d), and the density can be expressed in terms of θ as:

$$p_Y(t(y); \eta(\theta)) = \exp\{\eta(\theta)' t(y) - \kappa(\theta)\} \tilde{h}(t(y)), \quad (3)$$

with $\tilde{h}(t(y))$ such that the density integrates to 1 and $\kappa(\theta) \equiv \kappa(\eta(\theta))$, a notation used to avoid separate functions for the FEM and embedded CEM. Whenever the parameter θ is used it obviously refers to the CEM. For the log-likelihood we similarly write $\ell(\eta; t)$ and $\ell(\theta; t)$. Further discussion and results can be found in for instance Efron (1978) and Barndorff-Nielsen and Cox (1989, Sect.6.4, 1994, p.61-72). Our FEMs are assumed to be regular and therefore steep.

Moments and cumulants of t can be obtained using the embedding FEM with parameter η and differentiating the identity $\int p(t; \eta) dt = 1$ w.r.t. η , assuming that interchanging integration and differentiation is permitted:

$$E_\eta[T] = \frac{\partial \kappa(\eta)}{\partial \eta} \equiv \tau(\eta); \quad Var_\eta[T] = \frac{\partial^2 \kappa(\eta)}{\partial \eta \partial \eta'} \equiv \Sigma(\eta). \quad (4)$$

The score of the FEM is: $s(\eta; t) = \partial \ell(\eta; t) / \partial \theta = t - \tau(\eta)$, which has mean 0 and variance $\Sigma(\eta)$. The MLE $\hat{\eta}$ solves $t - \tau(\hat{\eta}) = 0$ and is unique for a regular FEM.

For the CEM, mean and variance of T can be obtained by evaluating (4) at $\eta = \eta(\theta)$. This gives the so called expectation manifold $E_\theta [T] = \tau(\theta) = \tau(\eta(\theta))$ and $Var_\theta(T) = \Sigma(\theta) = \Sigma(\eta(\theta))$ as a function of θ . For the derivatives of $\eta(\theta)$ and $\tau(\theta)$ we write:

$$B(\theta) = \frac{\partial \eta(\theta)}{\partial \theta'} \text{ and } C(\theta) = \frac{\partial \tau(\theta)}{\partial \theta'} = \frac{\partial^2 \kappa(\eta)}{\partial \eta \partial \eta'} \frac{\partial \eta(\theta)}{\partial \theta'} = \Sigma(\theta) B(\theta). \quad (5)$$

The score of the CEM, $\partial \ell(\theta; t) / \partial \theta$ can be written using $B(\theta)$ as:

$$s(\theta; t) = B(\theta)' (t - \tau(\theta)), \quad (6)$$

which is $d \times 1$ and not equal to $s(\eta(\theta); t)$ because of the differentiation involved. The expected Fisher information matrix $i(\theta) = E_\theta [s(\theta; T) s(\theta; T)']$ equals:

$$i(\theta) = B(\theta)' \Sigma(\theta) B(\theta),$$

and, since $\Sigma(\theta) B(\theta) = C(\theta)$, could be written as $i(\theta) = B(\theta)' C(\theta)$ which does not require explicit knowledge of the $k \times k$ matrix $\Sigma(\theta)$. The observed information evaluated at the MLE $\hat{\theta} = q$ equals:

$$j(q; t) = - \frac{\partial^2 \ell(\theta; t)}{\partial \theta \partial \theta'} \Big|_{\theta=q} = i(q) - \sum_{l=1}^k (t - \tau(q))_l - \frac{\partial^2 \eta_l(\theta)}{\partial \theta \partial \theta'} \Big|_{\theta=q}. \quad (7)$$

It is easy to see by induction that all higher order derivatives of the log-likelihood will also be linear (affine) in $(t - \tau(\theta))$. The expectation of the corresponding term will be 0 and the difference in expected and observed quantities, like the information, will therefore be linear in $(t - \tau(\theta))$ and the affine ancillary \mathbf{a} explained in Section 4 will be based on a sample version of this term: $\mathbf{a} = A(\hat{\theta})' [T - \tau(\hat{\theta})]$ with an appropriate choice of $A(\hat{\theta})$.

2.1 Examples

The examples we will use to illustrate the limitations of p^* and the improvements of the adjusted version are based on restricted versions of the simultaneous equations model and the non-linear regression model.

Single Structural Equation Model (SSEM)

In the context of exponential families, Van Garderen (1997) considered the SSEM:

$$y_1 = Y_2 \beta + Z_1 \gamma_1 + u_1, \\ Y_2 = (Z_1 : Z_2) \begin{bmatrix} \Pi_{12} \\ \Pi_{22} \end{bmatrix} + V_2,$$

with $y_1, u_1 : n \times 1, Y_2, Z_2 : n \times G_2, (Z_1 : Z_2) : n \times (K_1 + K_2)$ and the disturbances $(u_1 : V_2)$ are *i.i.d.* normally distributed. This SSEM constitutes a CEM(k, d) with $k = (K_1 + K_2)(G_2 + 1) +$

$(G_2 + 1)(G_2 + 2)/2$, and $d = k - (K_2 - G_2)$. So $k - d = K_2 - G_2$ (degree of overidentification in Phillips (1983) terminology) and under the exogeneity restriction ($Y_2 \perp u_1$) the model reduces to a FEM of lower dimension. Basic versions of this model have been used to analyze the effects of weak instruments and bimodality.

Woglom (2001) and Hillier (2006) start with the stripped-down exactly identified model with one explanatory variable:

$$y_i = \beta x_i + u_i, \tag{8}$$

$$x_i = \gamma z_i + v_i, \tag{9}$$

and show bimodality of the MLE. The overidentified version is also analyzed by Hillier (2006) as does Forchini (2006) to investigate and explain the bimodality. Interestingly, Hillier (2006) finds that under certain conditioning, the conditional distribution is unimodal, but unconditionally it is bimodal.

Bergstrom (1962) considered a simple Keynesian model and derived the exact *marginal* distribution of the MLE for β , the propensity to consume parameter in the model with y consumption, x income, z (non-random) investment, and instead of (9) had an identity for x_i with $\gamma = 1$ in the the equation

$$x_i = y_i + \gamma z_i \tag{10}$$

Phillips (2006) used this model in the weak instrument context and showed that bimodality persists even asymptotically. He assumed a known instrument strength parameter γ and noted the restriction that u_i is the single disturbance. The model is formally equivalent to the model by Nelson and Startz (1990) as observed by Maddala and Jeong (1992), see Phillips (2006, Footnote 1). Ariza and Van Garderen (2010) used the p^* -formula for this model and showed that it is very accurate for the chosen parameter values and captures the bimodality of the conditional density, but missed the adjustments of the present paper. It should also be noted that the log-likelihood goes to minus infinity when the parameter β goes to one for any sample from this model. So the density should always equal zero at one, even without any conditioning.

Errors in Variables Model

We consider another restricted version of this SSEM when explanatory variables are not observable. If x in (8) and (9) is measured with error, or is a latent, unobservable variable, then z could be used as proxy or mis-measured version of x . The measurement error in x may actually be correlated with the measurement error in y in this version of the model. The likelihood is obtained in Section 5 below by substituting (9) in (8) and will depend only on y and z , but x will not enter the likelihood since it is not observed. This model is a CEM(2, 1) used later to illustrate the quality of the adapted p^{**} -formula and improved confidence regions.

The errors in variable model has a very long tradition in Econometrics, see e.g. Durbin (1954) and its references to earlier work, and is still actively researched today, see De Nadai and Lewbel (2016) for a recent contribution that also considers correlated measurement errors

as we do. We use a version that is closely related to the SSEM analyzed in some of the weak instrument literature. We may anticipate problematic inference when (i) the variance of v_i (measurement error in x_i) is large or (ii) v_i highly correlated with u_i (the measurement errors in y_i and x_i are highly correlated) (iii) the sample variation in z_i is low or (vi) γ is small (z_i is a weak or a poor proxy), and a key quantity will be $\gamma^2 \sum_{i=1}^n z_i^2$, as in the weak instrument literature (see Stock, Wright, and Yogo, 2012, for a review). We keep $\sigma_u^2 = \sigma_v^2 = 1$ and $\rho = 0.75$ fixed, but show that problems persist even when the sample size n is increased from 25 to 1000 and γ reduced from 1 to 0.1.

Nonlinear Regression

The second model that we will expound is the classic linear regression model with nonlinear restrictions¹. Consider n independent observations y_i with explanatory variables x_i and *i.i.* $N(0, \sigma^2)$ disturbances v_i :

$$y = X\beta + v, \tag{11}$$

$y, v : n \times 1, X : n \times k$. The $k \times 1$ vector of parameters β is assumed to satisfy nonlinear restrictions such that $\beta = \beta(\theta)$ with θ a $d \times 1$ vector $d < k$.²

We expand a little on this model here to illustrate the CEM nature, the relevant sample space, and the issues involved with the p^* -formula. We will show our improvements and confidence regions in later sections. The simplest case would be $k = 2$, $\beta_1 = \theta$, $\beta_2 = \theta^2$, and σ^2 known. The parameter of interest θ is one-dimensional, but the two-dimensional OLS estimator $\hat{\beta}$ for the unrestricted β is a minimal sufficient statistic and satisfies:

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \sim N \left(\begin{pmatrix} \theta \\ \theta^2 \end{pmatrix}, \sigma^2 (X'X)^{-1} \right). \tag{12}$$

For concreteness and explicating the role of the sample size, set $\sigma^2 = 10$ and let X consist of two complementary dummies such that $\sigma^2 (X'X)^{-1} = \text{diag}(10/n_1, 10/n_2)$ with n_1 and n_2 the number of observations in each category, e.g. the number of male and female observations. The log-likelihood is a simple restricted version of a bivariate Normal log-likelihood and for a specific observed value $b = (b_1, b_2)'$ of the estimator $\hat{\beta}$ can be written as:

$$\ell(\theta; b) = b_1 \theta \frac{n_1}{10} + b_2 \theta^2 \frac{n_2}{10} - \frac{1}{20} (\theta^2 n_1 + \theta^4 n_2), \tag{13}$$

¹It should be noted in this context that Spady (1991) developed saddlepoint expansions for regression models and considered an example like (11) where β is unrestricted but the density of the disturbance v is itself bimodal.

²For example if β satisfies r non-linear restrictions $h(\beta) = 0$, that are sufficiently smooth and assuming constant rank of the derivative $\partial h(\beta) / \partial \beta'$, we have by the implicit function theorem that r parameters can be solved and expressed in terms of the $k - r$ remaining parameters. These parameters of interest can be collected in a vector θ and by reordering and redefinition we may write without loss of generality $h((\theta : g(\theta))') = 0$ and $\beta = \beta(\theta)$.

As a referee highlighted, one could start out with a more general model and use a technique to reduce the dimension of the model. The Frisch-Waugh Theorem for example if additional regressors Z have associated parameters not involved in the restrictions and freely varying. y and X would then be the residuals after regression on Z .

with all irrelevant terms that do not depend on parameters dropped. It is immediate from Theorem 1 in Van Garderen (1997) that the model is a CEM(2, 1) with canonical parameter $\eta(\theta) = (\theta n_1/10, \theta^2 n_1/10)'$, canonical statistic b , and cumulant function $\kappa(\theta) = \frac{1}{20}(\theta^2 n_1 + \theta^4 n_2)$.

Figure 1 shows the sample space with the expectation manifold as a function of θ : $\tau(\theta) = E_\theta[\hat{\beta}]$ equal to $(\theta, \theta^2)'$ here. It also shows three possible realizations of the sufficient statistic $\hat{\beta}$ from a sample of 50 observations with $n_1 = 10$ in the first category and $n_2 = 40$ in the second. These three realizations b^+ , b^- and b^\ominus all satisfy the first order conditions that the score $s(\theta; b) = \partial \ell(\theta; b) / \partial \theta$ equals zero at 1.5: $s(1.5; b) = B(1.5)'(b - \tau(1.5)) = 0$. For b^+ and b^- , 1.5 is indeed the MLE, but for b^\ominus the likelihood is globally maximized when $\theta = -1.54$. The likelihood only attains a local maximum for 1.5. In between -1.54 and 1.5 the likelihood has a local minimum at $\theta = 0.04$. The observed information $j(\theta; b^\ominus)$, evaluated at $\theta = 1.5$ and $\theta = -1.54$ is positive, but is negative for $\theta = 0.04$.

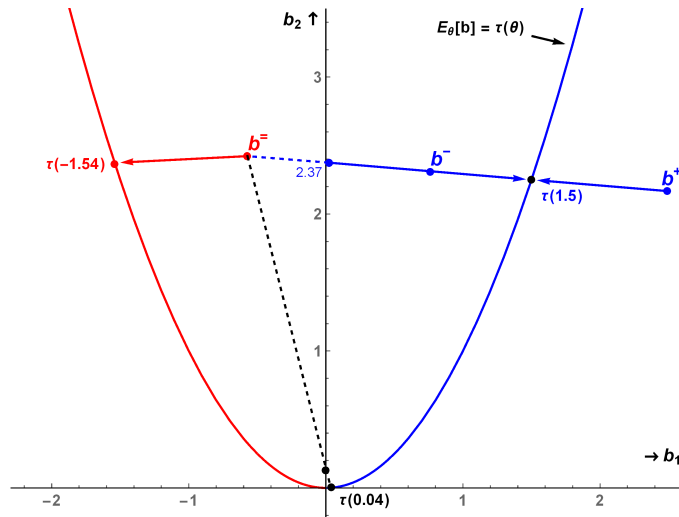


Figure 1: Sample space for the minimal sufficient statistic b with various outcomes b^\ominus , b^- , b^+ are shown with ML estimates -1.54 , 1.5 , and 1.5 respectively. They all satisfy the first order orthogonality condition $B(1.5) \perp (b - \tau(1.5))$ with $\tau(\theta) = E_\theta[b] = (\theta, \theta^2)'$ the expectation of b as a function of θ , the parabola, is also illustrated. The points $\tau(1.5)$, $\tau(-1.54)$, and $\tau(0.04)$ are also shown. Sample size $n = 50$ with $n_1 = 10$, $n_2 = 40$ and variance $\sigma^2 = 10$.

The graph illustrates a number of important issues relevant for the adjustment of p^* . Even if θ is much larger than 0, we can have sample realizations b that lead to negative values of $\hat{\theta}$. This provides a graphical explanation for bimodality in the distribution of $\hat{\theta}$. A second possible reason for bimodality is that the observed information (as defined below) can change from positive to negative. The score $s(0.04, b) = 0$ for all b on the dashed line, but as we observe values of b from $\tau(0.04)$ closer to b^\ominus , the observed information decreases and will change to being negative. Where it is zero and negative, the Jacobian matrix used in Barndorff-Nielsen's (1980) derivation of p^* is no longer appropriate since the basic theorem for the transformation

of variables breaks down and requires a decomposition of the sample space in sets where the Jacobian matrix is of constant rank.

The observed information $j(q, a)$ used in the p^* -formula (1) depends on the value a of an (approximate) ancillary statistic \mathbf{a} . Barndorff-Nielsen (1980) introduced his affine ancillary statistic defined in (15) below for this purpose. Details of \mathbf{a} for the example here are given in Appendix A.4. With $q = 0.04$ and with values of affine ancillary a smaller than -0.26 , the observed information is negative and $\theta = 0.04$ *minimizes* the likelihood locally instead of maximizing it. The outcome b° also shows a *global* reason for adapting p^* and its proof. Although $q = 1.5$ locally maximizes the likelihood and has positive observed information, $q = -1.54$ is the true global maximizer of the likelihood.

This leads to the most important point of the graph since it explains the shortcomings of p^* and suggests the adjustment to rectify it. We cannot jointly have $\hat{\theta} = 0.04$ and negative values of $\mathbf{a} < -0.26$. Such combinations cannot occur and the density must be 0. There are two different reasons (i) locally q could *minimize* the likelihood and q would not be the MLE with this value of a . This can be checked by the calculation of the observed information (ii) globally q might not maximize the likelihood. This requires comparing the likelihood values for different stationary points. Checking positive definiteness of the observed information is not sufficient.

Finally, it may be possible that b has two different values for θ with the same value for the likelihood. The point $b = (0, 2.37)'$ has the two values 1.5 and -1.5 , that maximize the likelihood. The MLE is not unique in that case, or for any other point on the b_2 axis with $b_2 > 1/8$.

3 Partitioning the Sample Space

In this section we will discuss a partitioning of the sample space for estimation, closely related to Efron (1978) and Amari (1985). It is a natural estimation counterpart to the classic Neyman-Pearson approach to partitioning the sample space for hypothesis testing. The decomposition will be based on the MLE $\hat{\theta}$ and an ancillary statistic \mathbf{a} that complements $\hat{\theta}$ and is used as conditioning variable in the (adjusted) p^* -formula.

Let Θ denote the parameter space and let y be the sample outcome of the random vector Y , $Y, y \in \mathbb{Y} \subseteq \mathbb{R}^n$. Let $T = T(Y)$ be a minimal sufficient statistic, which we will assume to be of finite dimension k , although this could be generalized to grow with the sample size. The relevant sample space of T is of dimension k and denoted $\mathbb{T} \subseteq \mathbb{R}^k$. The expectation of T will vary with θ and if $\tau(\theta) = E_\theta[T]$ is continuously differentiable, it constitutes a manifold inside \mathbb{T} and hence the name expectation manifold.

We define the following sets in order to characterize points in the sample space that satisfy the first order conditions (F), and have positive observed information (\tilde{F}), and points that have a global solution (C^+), or have multiple solutions (\tilde{M}) to the maximum likelihood problem.

The score vector $s(\theta; t) = \partial \ell(\theta; t) / \partial \theta$ will be of use to us in three ways. First, for given t ,

fixed at a sample value, it provides the first order conditions $s(q; t) = 0$ for the MLE. Second, for fixed θ_0 the score is a measurable function and $s(\theta_0; T)$ is a statistic with a distribution. Third, with $\hat{\theta}$ and T both random variables, $s(\hat{\theta}; T) = 0$ provides a relation between T and $\hat{\theta}$. When $\hat{\theta}$ is augmented by an ancillary \mathbf{a} such that a bijection between $(\hat{\theta}, \mathbf{a})$ and T is established, then the distribution of $(\hat{\theta}, \mathbf{a})$ follows from the distribution of T by a transformation of variables.

We define the following score related sets.

Definition 1 For fixed $q \in \Theta$ let:

$$F_q = \{t \in \mathbb{T} \mid s(q; t) = 0\};$$

$$\tilde{F}_q = \{t \in F_q \mid j(q; t) \text{ is positive definite}\};$$

For a particular value of the estimator $\hat{\theta} = q$, the set F_q is what Efron (1978) called the inverted MLE and Amari (1982,1985) called it the ancillary space associated with the estimator $\hat{\theta}$. This q maximizes the likelihood for given $t \in F_q$, at least locally, if $j(q; t)$ is positive definite. For local uniqueness of the MLE it is relevant whether the observed information matrix is positive definite or has eigenvalues smaller than or equal to 0. It will be expedient to refer to $j(q; t) = -\partial^2 \ell(q; t) / \partial \theta \partial \theta'$ as the observed information evaluated at $q \in \Theta$ also when q is not the true (global) maximizer of the likelihood MLE, for instance when evaluated at q with $t \in F_q \setminus \tilde{F}_q$ and q minimizes the likelihood locally. With this convention we can have eigenvalues of the observed information evaluated at q that are zero or even negative.

The aforementioned authors considered cases where the MLE is unique, but in curved models the sets $F_{q(1)}$ and $F_{q(2)}$ may intersect for two different points in $\Theta : q^{(1)} \neq q^{(2)}$. This means that there are points t in the sample space with vanishing scores for two different values of q . This is not uncommon in econometric models, yet only one point truly maximizes the likelihood in general. We associate with each fixed q a set in the sample space for which q maximizes the likelihood as follows:

Definition 2 $C_q^+ = \{t \in \tilde{F}_q \mid q = \arg \max_{\theta \in \Theta} \ell(\theta; t)\}$.

If q is the true maximizer of the log-likelihood then C_q^+ contains all points t for which the log-likelihood $\ell(\theta; t)$ is maximized by $\theta = q$.³ The set C_q^+ could be defined more generally with $t \in \mathbb{T}$ but under our smoothness conditions q will satisfy the first order conditions and has $j(q; t)$ positive definite. The set C_q^+ can be empty. In the simple Keynesian model (see 10) the likelihood always goes to minus infinity when β goes to 1 and there is no t that would ever be mapped onto $\hat{\theta} = 1$. This holds whether one conditions on an ancillary statistic or not.

For the nonlinear regression example with MLE $\hat{\theta} = 1.5$, $F_{1.5}$ is the complete line through b^- and b^+ , $C_{1.5}^+$ is the half line through b^- and b^+ starting at, but not including the point on the b_2 axis. $\tilde{F}_{1.5}$ is a half line (not fully shown) through b^+ , b^- , b^- and starting at, but not including the point $(-54, 6.9)$ because $|j(1.5; (-54, 6.9))| = 0$. The point b^- is an element of

³In the econometrics literature Hillier and Armstrong (1999) refer to the sets F and C^+ as S and \hat{S} , but assume the MLE is unique and $rank(j(q; t)) = d$ for all $t \in F_q$.

the three sets $F_{1.5}$, $F_{0.04}$ and $F_{-1.54}$, but $\ell(\theta; b^=)$ is globally maximized for $\theta = -1.54$. Hence $b^=$ is only an element of $C_{-1.54}^+$ but not of $C_{1.5}^+$, $C_{0.04}^+$, or any other $C_{q \neq -1.54}^+$. The point $b^=$ is also an element of both $\tilde{F}_{1.5}$ and $\tilde{F}_{-1.54}$, but not of $\tilde{F}_{0.04}$ because the observed information $j(0.04; b^=)$ is negative.

Note that F_q , \tilde{F}_q , and C_q^+ are in increasing order of difficulty of verification and that by definition $C_q^+ \subseteq \tilde{F}_q \subseteq F_q$. The complement $F_q \setminus C_q^+$, if non-empty, contains points t that satisfy the first order conditions but result in a value of the MLE different from q , or has $|j(q; t)| = 0$. The definition of C_q^+ does not exclude the possibility that there exist points $t \in \mathbb{T}$ that lie on two different $C_{q^{(1)}}^+$ and $C_{q^{(2)}}^+$ with $q^{(1)} \neq q^{(2)}$. For such points t the MLE is not unique. Non-uniqueness can also occur locally, when the determinant of the observed information is zero. Points t in F_q with $|j(q; t)| = 0$, are on the boundary of \tilde{F}_q and the closure of \tilde{F}_q , denoted $cl(\tilde{F}_q)$, includes these points with singular observed information. The set of all points for which the MLE is not unique is defined as

Definition 3 $\tilde{M} = \left\{ t \in \mathbb{T} \mid \exists q^{(1)}, q^{(2)} \in \Theta : q^{(1)} \neq q^{(2)} : t \in cl\left(C_{q^{(1)}}^+\right) \cap cl\left(C_{q^{(2)}}^+\right) \right\}$.

This definition covers global non-uniqueness, when two possibly very different $q^{(1)}$ and $q^{(2)}$ have the same value for the likelihood (otherwise t would be an element of only one C_q^+), and possible local non-uniqueness when the observed information matrix has an eigenvalue equal to zero and $q^{(2)}$ approaches $q^{(1)}$ arbitrarily close in the direction of the associated eigenvector.⁴

In the nonlinear regression example \tilde{M} in Figure 1 consists of the part of the b_2 axis with $b_2 \geq 1/8$. Note that $j(0; (0, 1/8)) = 0$ and the likelihood is locally flat.

The sets C_q^+ and \tilde{M} define a partition of the sample space. Each t is either in \tilde{M} or in one, and only one $C_q^+ \setminus \tilde{M}$. Each $t \in C_q^+ \setminus \tilde{M}$ maps uniquely to the MLE value q when maximizing the likelihood function. We can therefore decompose the sample space using two types of coordinates: the MLE $\hat{\theta}$ and the complementary coordinate (vector) \mathbf{a} defined in (15) below, that provides the location of T in the associated space $C_{\hat{\theta}}^+$. Apart from points $T \in \tilde{M}$, this decomposition is unique. This means that we have a bijective relation.

$$T \leftrightarrow (\hat{\theta}, \mathbf{a}). \quad (14)$$

Even under stringent regularity conditions, including smoothness of the log-likelihood function in both θ and y , this mapping is not continuous in general. There can be regions in the

⁴A singular Hessian is a necessary condition for non-uniqueness, but is not sufficient. Nevertheless, if $t \in cl\left(C_{q^{(1)}}^+\right) \cap cl\left(C_{q^{(2)}}^+\right)$ and v is a unit eigenvector associated with a zero eigenvalue of $\frac{\partial^2 \ell(\theta; t)}{\partial \theta \partial \theta'} \Big|_{\theta=q^{(1)}}$ and $q^{(2)}$ approaches $q^{(1)}$ in the direction v such that $(q^{(2)} - q^{(1)}) = \lambda v$ then a second order Taylor expansion $\ell(q^{(2)}; t) = \ell(q^{(1)}; t) + \frac{\partial \ell(\theta; t)}{\partial \theta'} \Big|_{\theta=q^{(1)}} (q^{(2)} - q^{(1)}) + \frac{1}{2} (q^{(2)} - q^{(1)})' \frac{\partial^2 \ell(\theta; t)}{\partial \theta \partial \theta'} \Big|_{\theta=q^{(1)}} (q^{(2)} - q^{(1)}) + \left\| q^{(2)} - q^{(1)} \right\|^3$ has the second r.h.s. term vanishing since $t \in cl\left(\tilde{F}_{q^{(1)}}\right)$ and the third r.h.s. term vanishes since $(q^{(2)} - q^{(1)})$ is the direction associated with the zero eigenvalue. So the difference in likelihoods is $\ell(q^{(2)}; t) - \ell(q^{(1)}; t) = O\left(\left\| q^{(2)} - q^{(1)} \right\|^3\right)$. For the actual implementation local non-uniqueness is no great concern since both the original p^* and our p^{**} set the density equal to zero if the observed information matrix is singular.

sample space where an arbitrarily small change in T leads to a discrete change in $\hat{\theta}$. This discrete change can be large, as in the nonlinear regression model with $b_2 = 2.37$ and b_1 around 0. A arbitrarily small difference in outcome b_1 , e.g. from (very small) ε to $-\varepsilon$ leads to a jump in the estimate from 1.5 to -1.5 . This discontinuity occurs when crossing \widetilde{M} . For t bounded away from \widetilde{M} there will be a neighborhood of t and (q, a) such that the mapping is continuously differentiable and the Jacobian of the transformation is properly defined. This will be crucial in the derivation of p^* and its adaptation.

The p^* -formula gives an approximation of the conditional distribution of the MLE given an ancillary \mathbf{a} . We will require that \mathbf{a} is (i) a function of the minimal sufficient statistic T , (ii) maximal, i.e. of dimension $k - d$ and such that the mapping from T to $(\hat{\theta}, \mathbf{a})$ is invertible, and (iii) (approximately) ancillary. If the distribution of \mathbf{a} does not depend on parameters, then \mathbf{a} by itself will not contain information on the parameters, but can nevertheless contain important information about the distribution of the MLE. Since we can compute its value using the actual sample, we can condition on the observed value.

Usually, focus is on the conditional variance. An important example of this is Efron and Hinkley (1978) who show that the inverse of the observed information is preferred over the expected information as a measure of variance, given an appropriate ancillary. The usefulness of the variance is reduced, however, if the density has two modes that are seriously apart. This paper turns attention away from the variance and focuses on multimodality instead. We show that it is still appropriate and desirable to condition on \mathbf{a} because it contains crucial information on the degree of bimodality.

4 The Affine Ancillary

Various approximate ancillary statistics are available and a choice has to be made, see for instance the discussion in Barndorff-Nielsen and Cox (1994, 7.2). One could ask which statistic is most informative about the bimodality, without losing its ancillarity properties. Many of these ancillary statistics might not be appropriate since they are generally based on the local geometry of the model around a particular θ , but in contrast bimodality derives from the global topological structure of the model. One might be able to derive new statistics based on the global structure, but ancillarity will be difficult to establish, even approximately, and this will not be investigated here. Instead we will use the affine ancillary which is of the right dimension and approximately ancillary. There are three main reasons for choosing the affine ancillary. Despite being a local ancillary, it still contains valuable information on the bimodality of the distribution of the MLE as we will show. Second, it plays an important role in the derivation of p^* and our adjustment. Third, it is easy to calculate \mathbf{a} and to invert $(\hat{\theta}, \mathbf{a})$, and obtain $t(q, a)$. As a final comment it should be noted that the C_q^+ partitioning of the sample space does not depend on which ancillary is chosen and is an intrinsic property of the model and the chosen estimator (MLE in our setting).

The affine ancillary was introduced by Barndorff-Nielsen (1980) in the context of CEMs as an affine function of the minimal sufficient statistic T such that, for fixed $\hat{\theta}$, it is approximately ancillary with mean 0 and variance I_{k-d} . Although not stated explicitly, \mathbf{a} should also be independent of, or at least approximately uncorrelated with $\hat{\theta}$. We can motivate this using Basu's Theorem, (Basu, 1955), which states that any ancillary statistic must be independent of a complete sufficient statistic. The CEM can be approximated locally at θ_0 by a FEM of dimension d . In this approximation $\hat{\theta}$, or equivalently the score statistic, is a complete sufficient statistic.

Define orthogonal complements $C(\theta)_\perp, B(\theta)_\perp : k \times (k-d)$ chosen such that the determinant $|B(\theta) : C(\theta)_\perp| = |i(\theta)| / |\Sigma(\theta)|$ and $B(\theta)_\perp = \Sigma(\theta)C(\theta)_\perp$ which implies $B(\theta)'B(\theta)_\perp = C(\theta)'\Sigma(\theta)^{-1}\Sigma(\theta)C(\theta)_\perp = 0$.⁵

Definition 4 *The affine ancillary statistic is defined as:*

$$\mathbf{a} = A(\hat{\theta})' \left(T - \tau(\hat{\theta}) \right), \quad (15)$$

with

$$A(\hat{\theta}) = C(\hat{\theta})_\perp \left(C(\hat{\theta})'_\perp \Sigma(\hat{\theta}) C(\hat{\theta})_\perp \right)^{-1/2}. \quad (16)$$

Note that for the true value θ_0 , the statistic $\mathbf{a}_0 = (C(\theta_0)'_\perp \Sigma(\theta_0) C(\theta_0)_\perp)^{-1/2} C(\theta_0)'_\perp (T - \tau(\theta_0))$, has $E[\mathbf{a}_0] = 0$ and $Var(\mathbf{a}_0) = I_{k-d}$, and is uncorrelated with the score statistic $s(\theta_0; T) = B(\theta_0)'(T - \tau(\theta_0))$:

$$Cov(\mathbf{a}_0, s(\theta_0; T)) = (C(\theta_0)'_\perp \Sigma(\theta_0) C(\theta_0)_\perp)^{-1/2} C(\theta_0)'_\perp \Sigma(\theta_0) B(\theta_0) = 0 \quad (17)$$

This obviously does not prove ancillarity when θ_0 is replaced by the MLE. We can establish approximate ancillarity by relating \mathbf{a} to an LM test statistic which, under general regularity conditions, is asymptotically χ^2_{k-d} distributed independent of parameters.

Proposition 1 *The LM test statistic for testing the CEM against the embedding FEM equals*

$$LM = \mathbf{a}'\mathbf{a}. \quad (18)$$

It should be clear that other test statistics, such as the Wald or LR test would also qualify as approximate ancillaries. These tests could alternatively be used as specification tests, similar to Chesher and Smith (1997), but in our setting the CEM is never rejected and we condition on the value of the test statistic. Note that the augmented density of Chesher and Smith (1997, p.631) is like the FEM embedding here and can be used to show our results outside the CEM setting.

⁵The sign of each element in \mathbf{a} can be freely chosen since post-multiplying $C(\theta)_\perp$ by an orthogonal matrix U , $U'U = UU' = I_{k-d}$ still satisfies $|B(\theta) : C(\theta)_\perp U| = |i(\theta)| / |\Sigma(\theta)|$ without affecting the properties $E[\mathbf{a}_0] = 0$, $Var(\mathbf{a}_0) = I_{k-d}$ and $Cov(\mathbf{a}_0, s(\theta_0; T)) = 0$ when \mathbf{a} is based on $C(\theta)_\perp U$ because the space spanned by the columns of $C(\theta)_\perp$ is not affected. Hence, we can freely choose directions in the column space of $C(\theta)_\perp$ in which the distribution of $\hat{\theta}|\mathbf{a}$ is most- and least affected.

The final property of \mathbf{a} to be confirmed is that it is a maximal ancillary of dimension $(k - d)$ and jointly with $\hat{\theta}$ able to recover T . To show the invertibility combine the first order conditions with the definition of \mathbf{a} to obtain $\left[B(\hat{\theta}) : A(\hat{\theta}) \right]' (T - \tau(\hat{\theta})) = (0 : \mathbf{a})'$. The matrix $\left[B(\hat{\theta}) : A(\hat{\theta}) \right]$ is of full rank as shown in the proof of the following lemma.

Lemma 1 $[B(\theta) : A(\theta)]$ is invertible.

This leads to the following proposition.

Proposition 2 The inverse of the mapping $T \mapsto (\hat{\theta}, \mathbf{a})$ from the MLE and affine ancillary to the minimal sufficient statistic $T = t(\hat{\theta}, \mathbf{a})$ is:

$$t(\hat{\theta}, \mathbf{a}) = \tau(\hat{\theta}) + \begin{bmatrix} B(\hat{\theta})' \\ A(\hat{\theta})' \end{bmatrix}^{-1} \begin{pmatrix} 0 \\ \mathbf{a} \end{pmatrix} \quad (19)$$

This mapping can be extended to arbitrary values q and a for which $\tau(q)$, $A(q)$ and $B(q)$ are defined and is not limited to q being the MLE.

All values $t(q, a)$ will be in F_q but not necessarily in \tilde{F}_q or C_q^+ when, for instance in our nonlinear regression example, a is too negative. The inverse mapping is straightforward to evaluate and will play an important role in the application of the adjusted p^{**} formula below.

5 The Adjusted p^{**} Approximation

Based on the arguments that led to the definition of the set C_q^+ , we propose to adapt the p^* -formula using an indicator function to mark whether a particular combination of MLE and ancillary, $(\hat{\theta}, \mathbf{a}) = (q, a)$ can occur, or that the density should be set to zero instead. The second important element of the formula is the explicit calculation of the norming constant since for multimodal distributions it can vary substantially with a and θ . We will denote it $c^+(\theta, a)$ since it can be very different from $c(\theta, a)$ used in the original formula.

Proposition 3 The multimodality adapted p^{**} -formula is defined as:

$$p_{\hat{\theta}}^{**}(q | \theta; a) = c^+(\theta, a) |j(q, a)|^{1/2} \exp \{ \ell(\theta; t) - \ell(q; t) \} I_{C_q^+}, \quad (20)$$

with $I_{C_q^+} = 1$ if $t \in C_q^+$ and 0 otherwise and $c^+(\theta, a)$ determined such that the density integrates to 1.

Note that p^{**} can be expressed in terms of (q, a) only or in terms of t only, because of the one-to-one relation between t and (q, a) . As a function of (q, a) we could write for instance $\ell(\theta; q, a) = \ell(\theta; t(q, a))$ and $j(q, a) = -\partial^2 \ell(\theta; t(q, a)) / \partial \theta \partial \theta' |_{\theta=q}$.

The essential elements in the derivation of p^{**} given in the appendix are:

(i) the quality of the approximation of embedding FEM density of T (e.g. Barndorff-Nielsen and Cox, 1979),

(ii) the Jacobian of the transformation from T to $(\hat{\theta}, \mathbf{a})$ derived by Barndorff-Nielsen (1980) and recognizing its continuing validity as we do here for multimodal likelihoods in the neighborhood of the true (global) MLE,

(iii) identifying (q, a) combinations that are impossible and must have corresponding density 0, which is a principal contribution of this paper.

The key adjustment in p^{**} is setting the density to zero when t is not in C_q^+ . When $t \notin C_q^+$ it would never be mapped onto the value q for the MLE and could never lead to an outcome (q, a) for $(\hat{\theta}, \mathbf{a})$. This provides a simple way of calculating $I_{C_q^+}$ which we use in practice. We calculate $t = t(q, a)$ using the inversion formula (19) and calculate the associated ML estimate, $\hat{\theta}(t)$ say, based on this t . If $\hat{\theta}(t) = q$ then $t \in C_q^+$ and $I_{C_q^+} = 1$. If $\hat{\theta}(t) \neq q$ then $t(q, a)$ will not be mapped on (q, a) and $t \notin C_q^+$ and the density at (q, a) must be zero. Obviously this is true when $j(q, t(q, a))$ is not positive definite. This could be used as a first check and the density set to 0. There may be other points however, with positive definite $j(q, t(q, a))$ but $t \notin C_q^+$. Hence determining the eigenvalues of $j(q, t(q, a))$ is not sufficient.

We obtain the normalizing constant $c^+(\theta, a)$ numerically by integration of the un-normalized $|j(q, a)|^{1/2} \exp\{\ell(\theta; t) - \ell(q; t)\} I_{C_q^+}$ over all relevant values of q , keeping a fixed at the conditioning value. This is straightforward using $t(q, a)$ of Proposition 2 again and requires only a d -dimensional integration, rather than a k or n dimensional integration, irrespective of the dimension of a or t .

Nonlinear Regression

Figure 2 compares the empirical density of $\hat{\theta}$ given $\mathbf{a} = -0.5$ and $\theta_0 = 0$ in the nonlinear regression model with the basic p^* -approximation as originally formulated. The original p^* clearly has a hump around zero that is *inappropriate*. The value $\mathbf{a} = -0.5$ is such that no $\hat{\theta}$ near 0 could occur and any T value with this \mathbf{a} will lead to $\hat{\theta}$ bounded away from 0. The actual distribution is nil at zero. This is confirmed by the nonparametric kernel density estimate using simulated data shown as a dashed line.

Figure 3 shows the adapted p^{**} -approximations compared with the empirical distributions for different outcomes of the ancillary statistic. The true parameter value is $\theta_0 = 1$. We see that there are two modes around 1 and -1 respectively. The second mode is more important when the value of a is smaller. When a is more negative the realized t is closer to the set \widetilde{M} . For fixed q there will typically be a value for a such that $t(q, a) \in \widetilde{M}$ and the two modes will in this case be equal in height.

The p^{**} -formula approximates the true density extremely well, including its bimodal features, and equals zero where it should be. For $a = -0.75$ the p^{**} is 0 between -0.7 and $+0.7$. In the 100000 simulations there are no values for $\hat{\theta}$ between -0.7 and $+0.7$ when a is (close to) -0.75 . The nonparametric density estimate is slightly misleading in this case. Despite the absence of observations in this interval, the nonparametric estimate still gives positive density near ± 0.7 , but this is the well known boundary effect of kernel density estimators. Without

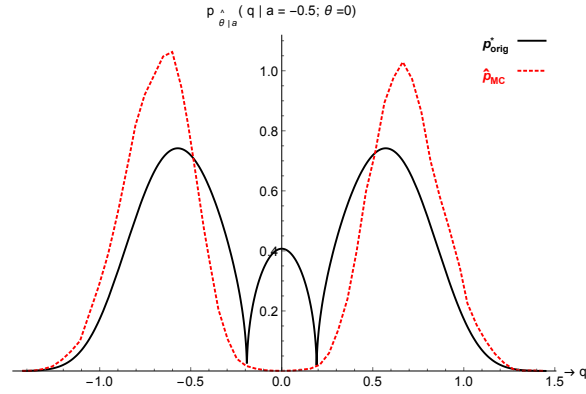


Figure 2: p^* -approximation and simulated conditional density of $\hat{\theta}$ given $\mathbf{a} = -0.5$. \hat{p}_{MC} is kernel density estimate based on 100000 Monte-Carlo replications. True $\theta = 0$, $n = 50$, $n_1 = 10$, $n_2 = 40$, $\sigma^2 = 10$.

p^{**} this could easily have gone unnoticed and we would not even have located the boundaries around which the kernel density is biased.

Figure 4 shows the values of $c^+(\theta, a)$ in the nonlinear regression model with $\theta \in (-3, 3)$ and $a \in (-1, 1)$. It illustrates the importance of the explicit calculation of $c^+(\theta, a)$. For positive values of a and θ away from 0, $c^+(\theta, a) \approx 0.4$ and virtually $(2\pi)^{-d/2}$ as predicted by Barndorff-Nielsen, but for negative values of a , and θ around 0 it can be much larger. In the graph the largest value of c^+ shown is 1.5, about 4 times larger than its asymptotic equivalent, but for (θ, a) values closer to $(0, -1)$ are even more extreme.

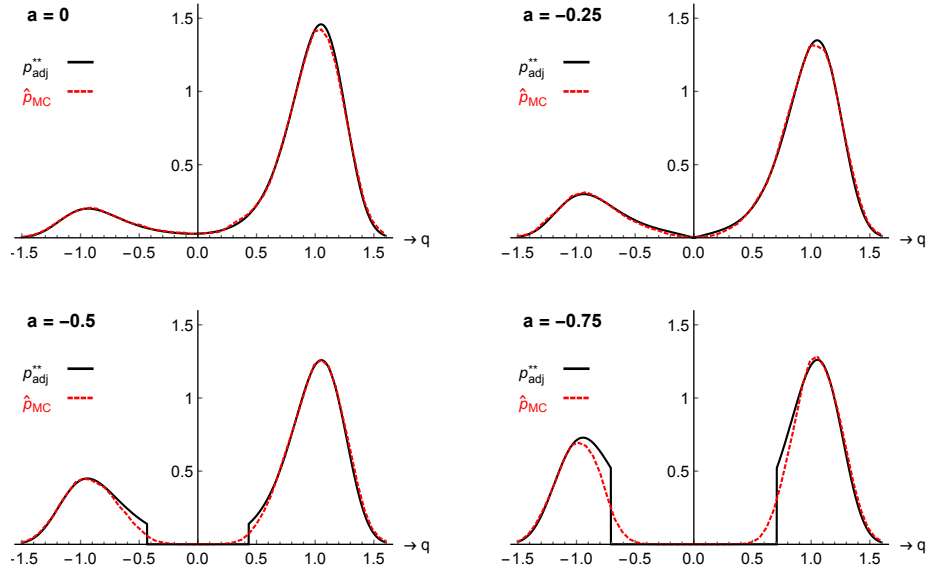


Figure 3: p^{**} -approximation in the nonlinear regression model for conditional density of $\hat{\theta}$ given a , ranging from 0 to -0.75 . Kernel density \hat{p}_{MC} misleading for $a = -0.5$ and -0.75 because of boundary effects. Large positive values of a are not illustrated since the second mode is negligible and p^{**} and \hat{p}_{MC} coincide with p^* . True $\theta = 1$, $n = 50$, $n_1 = 10$, $n_2 = 40$, $\sigma^2 = 10$.

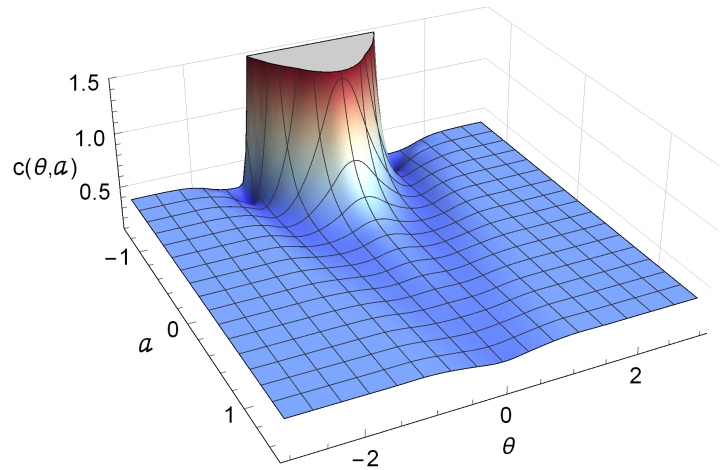


Figure 4: $c^+(\theta, a)$ in the nonlinear regression model showing large variation depending on the values of θ and a . Extreme values are attained for $\theta \approx 0$ and $a < -1$ but values larger than 1.5 are not shown. $n = 50$, $n_1 = 10$, $n_2 = 40$, $\sigma^2 = 10$.

Errors in Variables Model

If the explanatory variable in the single structural equations model (8) is unobservable we can substitute (9) in (8) and obtain a very limited information model that does not depend on x , but only on z , as proxy or mis-measured value of x , and $y_i|z_i \sim N(z_i\gamma\beta, \sigma^2(\beta))$ with $\sigma^2(\beta) = \sigma_u^2 + 2\beta\rho\sigma_u\sigma_v + \beta^2\sigma_v^2$. The log-likelihood for β based on y given z equals

$$\ell(\beta; y) = -\frac{1}{2} \frac{1}{\sigma^2(\beta)} \sum_{i=1}^n y_i^2 + \frac{\beta\gamma}{\sigma^2(\beta)} \sum_{i=1}^n y_i z_i - \frac{1}{2} \frac{\beta^2\gamma^2}{\sigma^2(\beta)} \sum_{i=1}^n z_i^2 - \frac{n}{2} \log(\sigma^2(\beta)). \quad (21)$$

This model is a CEM(2, 1) when assuming σ_u , σ_v , ρ , and γ known, since the log-likelihood can be written in canonical form $\eta(\beta)' t(y) - \kappa(\eta(\beta))$, so the generic θ is β here and we denote an outcome of the MLE by b (instead of q). Further details given in Appendix A5. ⁶

Figure 5 shows p^{**} , p^* , and p_{MC} the distribution obtained by simulation and using a non-parametric kernel density estimator for an outcome of the ancillary statistic of -0.75 .

The original p^* -formula is zero for $b = -0.82$ and $b = -0.68$ since the term $j(b, t(b, -0.75))$ is zero. For values of b in between -0.68 and -0.82 , the observed information is negative and taking those values for β would locally minimize the log-likelihood $\ell(\beta; t(b, -0.75))$. Taking the absolute value leads to the inappropriate hump in p^* . The adjusted p^{**} -formula sets the density to zero, not only in this interval, but for the larger interval $(-0.87, -0.62)$. The observed information evaluated at b in the intervals $(-0.87, -0.82)$ and $(-0.68, -0.62)$ is in fact positive. A value b for β in this interval would *locally* maximize the log-likelihood $\ell(\beta; t(b, -0.75))$, i.e. $\beta = b$ gives a local maximum, but another value for β will *globally* maximize the likelihood. This implies that the combinations $(b, -0.75)$ with b in these intervals would never be observed. The p^{**} density is set to zero, even though the observed information is positive definite. The simulations confirm that this is correct: there are no realizations for b in the whole of the interval $(-0.87, -0.62)$ when a is (near) -0.75 . Again the nonparametric kernel density is misleading as it assigns positive density to outcomes that never occurred in 100000 replications.

The observed information is not sufficient for determining if the density is zero. It identifies points $F_q \setminus \tilde{F}_q$ which have density zero because the observed information is not positive definite and q is a local minimum, but cannot identify $\tilde{F}_q \setminus C_q^+$ when q is not the global maximum. Coherency of q and $\hat{\theta}$ needs to be checked by confirming that the MLE $\hat{\theta}(t(q, a))$ equals q . This also provides straightforward implementation in more complicated models. Invert (q, a) to obtain t and use the global optimizer required to find the estimator and if it equals q , then t is in C_q^+ .

The model is closely related to the SSE model and we can consider a weak instrument type situation where z is a proxy that is only weakly correlated with x by letting γ go to zero as the sample size increases. In Figure 6 the sample size is $n = 1000$ and $\gamma = 0.1$. We see that

⁶Again it might be possible to start with a more general model and reduce the dimension by concentrating the likelihood for instance, but the resulting profile likelihood is not a proper likelihood based directly on a density and the implications will be left for future investigations.

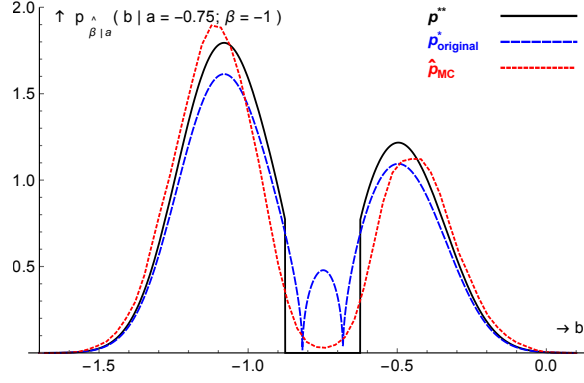


Figure 5: Conditional distributions in the Errors in Variables model: p^{**} , p^* and nonparametric kernel density estimate \hat{p}_{MC} based on 100000 Monte Carlo simulations but misleading because of boundary effects: no draws in the interval $(-0.87, -0.62)$. $a = -0.75$, true $\beta = -1$, $n = 25$, $\sigma_u^2 = \sigma_v^2 = 1$, $\rho = 0.75$, $\gamma^2 \sum_{i=1}^n z_i^2 = 4.5$.

bimodality persists, but the density around the two modes is much more concentrated. The adjusted p^{**} still differs from the original p^* -formula which displays a small hump and is positive in the region where it should be zero, but less so than for $n = 25$. The (marginal) probability of observing an \mathbf{a} which is smaller than -0.4 , is more than 10%, so not too extreme or impossible.

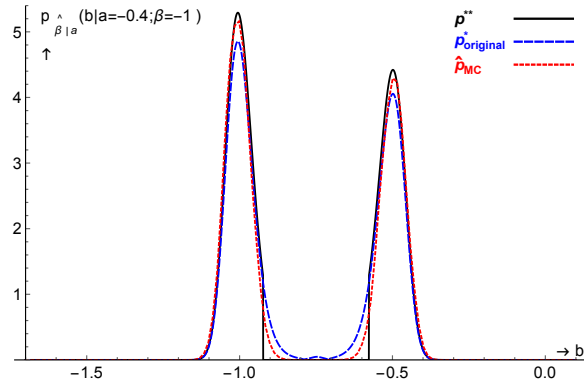


Figure 6: Conditional densities in the Errors in Variable model with large $n = 1000$ and $\gamma = 0.1$. z a weak instrument for unobservable x . Density is still bimodal and p^* is (slightly) positive around $b = -0.75$. $a = -0.4$, true $\beta = -1$, $\sigma_u^2 = \sigma_v^2 = 1$, $\rho = 0.75$, $\gamma^2 \sum_{i=1}^n z_i^2 = 2.36$.

6 Confidence Regions based on p^{**}

Bimodality of the density of the MLE has consequences for inference. In particular for confidence regions, it can result in confidence sets that are disjoint, rather than a single connected interval. Standard confidence intervals based on the usual normal approximation of the MLE can lead to coverage rates that are very different from their presumed nominal level. Figure 7 below shows this shortcoming in the nonlinear regression model for a 95% confidence interval when the true parameter value is $\theta_0 = 0$, using the standard first order method with boundaries $\hat{\theta} \pm 1.96 \cdot \widehat{se}(\hat{\theta})$. The standard error $\widehat{se}(\hat{\theta})$ is based on the square root of either the expected information $i(\hat{\theta})$, or the observed information $j(\hat{\theta}; t)$ as advocated by Efron and Hinkley (1978).

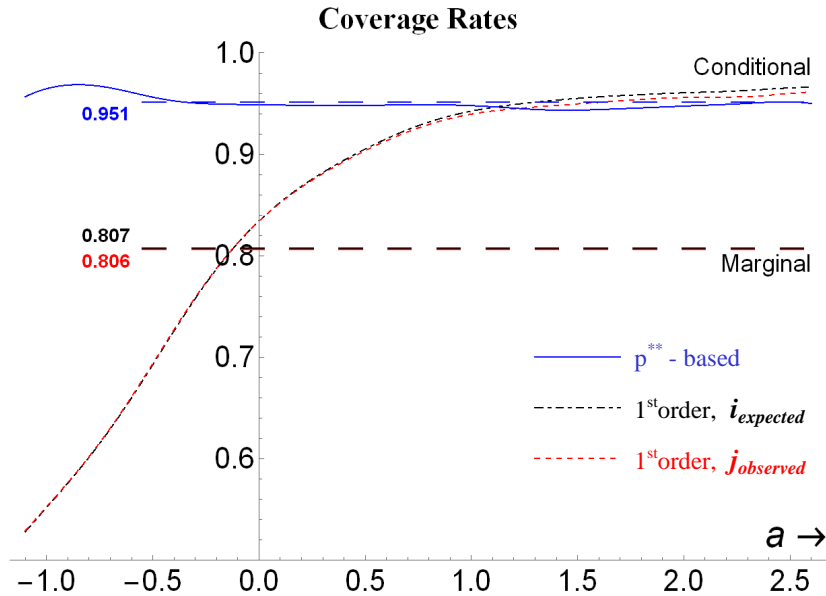


Figure 7: Conditional coverage rates as function of a , the observed value of the ancillary statistic, for standard methods $\hat{\theta} \pm 1.96 \widehat{se}(\hat{\theta})$ and standard errors se based on the square root of the (estimated) expected information $i(\hat{\theta})$ and expected information $j(\hat{\theta}; a)$. Confidence regions based on p^{**} are explained below. Straight horizontal lines are the marginal coverage rates: 95.1% for p^{**} based and 80.7% and 80.6% for standard first order method based on expected- and observed information respectively, $n = 50$, $n_1 = 10$, $n_2 = 40$, $\sigma^2 = 10$. Results based on 100000 Monte-Carlo replications.

The graph shows that the coverage rate drops dramatically for negative realizations of the affine ancillary. The results based on the expected information are shown, but also those based on the observed information. Efron and Hinkley (1978) showed that the observed information provides a more accurate measure of the conditional variance and the affine ancillary is

equivalent in this case to the ancillary statistic they introduced.

For a bimodal distribution the variance in one point is not very useful since it can only give a local approximation at one of the modes. The graph illustrates this. If the ancillary a is more negative then the bimodality is more severe. Confidence intervals based on the observed information cannot resolve this issue and are very similar to those based on the expected information. Overall coverage of 80%, instead of 95%, is also a consequence of the curvature of the model, causing the marginal variance to be larger than the inverse of the expected information. The coverage rate can be adjusted of course by adjusting the critical values via a (parametric) Bootstrap. This, however, will not change the fact that the coverage level drops dramatically for small values of a .

Figure 7 illustrates three problems with standard first order methods. First, the fact that the coverage rate depends heavily on the ancillary, is ignored. Second, bimodality is not taken into account. Third, even marginal coverage rates are well below the nominal level. The figure also shows that p^{**} -based confidence regions deliver correct coverage rate conditionally, as well as unconditionally. Construction of these confidence regions is explained next.

The p^{**} -formula gives very accurate approximations to the exact density. This can be used to construct predictive regions for $\hat{\theta}$ given θ . These predictive- or acceptance regions can be inverted to obtain confidence regions for θ on the basis of an observed value for $\hat{\theta}$ using standard arguments on the duality between tests and confidence regions (e.g. Fraser, 1976, p.580).

Define for each θ the set $\mathbf{A}(\theta)$ that contains the estimator $\hat{\theta}$ with probability $(1 - \alpha)$ as:

$$\mathbf{A}(\theta) = \{q : p_{\hat{\theta}}(q|\theta) > \delta\}, \quad (22)$$

and δ such that $P_{\theta}[\hat{\theta} \in \mathbf{A}(\theta)] = 1 - \alpha$, i.e. the probability of the event when the parameter value is θ . If q_0 is an observed value of $\hat{\theta}$, define the set $\mathbf{B}(q_0)$ that collects all parameter values that are acceptable for the outcome q_0 according to the same criterion:

$$\mathbf{B}(q_0) = \{\theta : p_{\hat{\theta}}(q_0|\theta) > \delta\}. \quad (23)$$

We have the logical equivalence $\theta \in \mathbf{B}(q_0) \leftrightarrow q_0 \in \mathbf{A}(\theta)$ and the confidence region $\mathbf{B}(q_0)$ based on the outcome q_0 is a proper $(1 - \alpha)$ level confidence set as a result.

Our approach is to replace the unknown exact density by the approximate conditional density $p_{\hat{\theta}}^{**}(q_0|a, \theta)$ and choosing c such that the conditional coverage level, given a , is $(1 - \alpha)$.

The sets $\mathbf{A}(\theta)$ and $\mathbf{B}(q_0)$ are motivated by Corollary 1 in Appendix A.6, which shows that $\mathbf{A}(\theta)$ is an optimal $(1 - \alpha)$ -level prediction region that is smallest in the terms of Lebesgue measure. When the density of the MLE is multimodal, it may happen that the sets consist of disjoint intervals when $p_{\hat{\theta}}(q|\theta)$ is larger than c for θ around two distinct modes, but smaller than c somewhere in between these two modes. This occurs for the two leading examples, depending on the value of a . The prediction and confidence regions depend heavily on the value of the ancillary statistic.

Non-linear Regression

Figure 8 shows confidence regions for θ based on p^{**} as a function of the estimated $\hat{\theta} = q$, when

the actual value of the observed ancillary is $a = -0.5$. The figure shows that the confidence regions for θ will indeed consist of two disjoint intervals when certain values of $\hat{\theta}$ are realized. As an example, the result for outcome $\hat{\theta} = -1.2$ and $a = -0.5$ is shown. The confidence region is the union of the two disjoint sets $(-1.62, -0.72)$ and $(0.92, 1.43)$. These disjoint intervals are quite far apart because the density of the MLE has a mode around the true (hypothesized) value θ and a second around $-\theta$ in this model.

Another notable observation is that for $\mathbf{a} = -0.5$ and q close to 0, no confidence region is, or has to be, defined. The reason is that we would never observe an MLE value q close to 0 when $\mathbf{a} = -0.5$. The value $(q, a) = (0, -0.5)$ corresponds to a value of the sufficient statistic $t(0, -0.5) = (0, 0.25)'$ by the inversion formula (19) of Proposition 2. The point $(0, 0.25)'$ is in \tilde{M} and leads to $\hat{\theta} = \pm 0.35$ and $a = -0.43$. The observed information is negative and the likelihood actually has a local minimum for $\theta = 0$.

The intervals vary significantly with a as is evident when comparing Figure 8 for the value $a = -0.5$ with Figure 11 in Appendix A.7 for $a = 0$. For large positive a there are no disjoint regions. The intervals should of course change substantially with a because Figure 7 illustrated that standard methods with intervals that do not vary with a , can show a dramatic drop in coverage rate as a decreases.

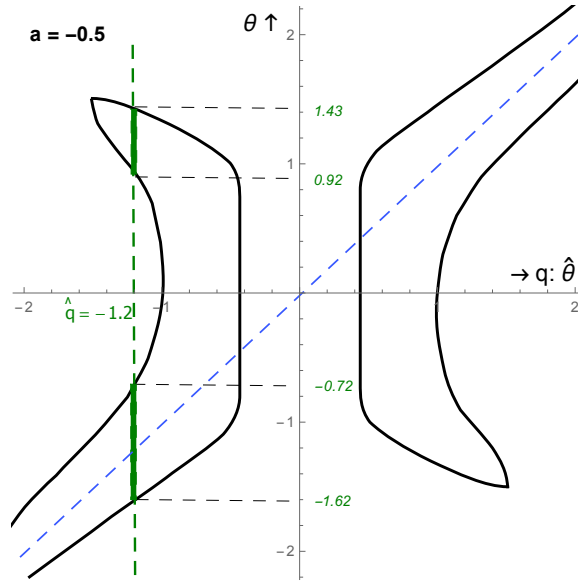


Figure 8: Confidence regions $B(q)$ as a function of q (observed $\hat{\theta}$) for the nonlinear regression example. Illustrated is an observed value $\hat{\theta} = -1.2$ which leads to the confidence region $(-1.62, -0.72) \cup (0.92, 1.43)$ for this value of $a = -0.5$. $n = 50$, $n_1 = 10$, $n_2 = 40$, $\sigma^2 = 10$.

If conditional coverage rates are correct for each given a , then they are also marginally correct. Confidence regions based on p^{**} have, conditional on a , very accurate coverage rates close to their nominal levels. This implies that also marginally, after averaging out a , the coverage rate is close to the nominal level. The reliability of the p^{**} based confidence regions was illustrated in Figure 7 with a conditional and overall coverage rate close to the 0.95 horizontal

line.

Errors in Variables Model

The p^{**} based confidence regions for β in the Errors in Variables model when $\mathbf{a} = -0.4$ are given in Figure 12 in Appendix A.7 and are very similar to Figure 8. For estimates b around 0 and around -1.4 the confidence regions are disjoint. With $a = -0.4$ the MLE will never be close to -0.7 .

The weak proxy type setting with $n = 1000$ and $\gamma = 0.1$ is of particular interest and different. We still have disjoint confidence regions, but both parts are much narrower. Figure 9 shows the regions as a function of the observed value b for $\hat{\beta}$ and $\gamma^2 \sum_{i=1}^n z_i^2 = 2.36$.

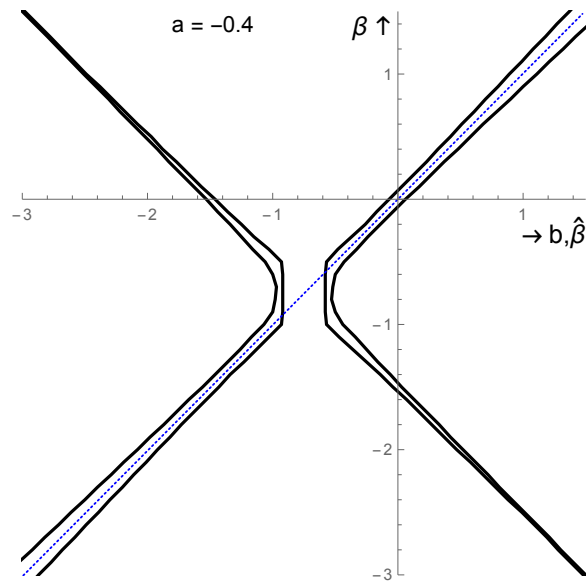


Figure 9: Confidence regions as a function of $\hat{\beta} = b$ in the Errors in Variables model when z is a proxy weakly correlated with unobservable x . $n = 1000$, $\gamma = 0.1$, $\sigma_u^2 = \sigma_v^2 = 1$, $\rho = 0.75$, $\gamma^2 \sum_{i=1}^n z_i^2 = 2.36$.

7 Conclusion

In this paper we have proposed a new p^{**} -formula that is a multimodality adjusted version of Barndorff-Nielsen's well known p^* -formula. It allows for the possible non-uniqueness of the solution to the first order conditions. We identified combinations of outcomes for $\hat{\theta}$ and \mathbf{a} that are theoretically impossible. For those cases p^{**} equals zero and is exact. It is therefore strictly superior to any other approximation that assigns positive density to such outcomes, such as a straight application of the original p^* , or even the non-parametric kernel density estimate we used in the examples.

Using the observed information is not sufficient to identify all the points where the density should be zero, but could be used as a first practical step. It is easy to calculate and to set the density equal to zero when one of the eigenvalues is not positive. (Only checking the determinant is not sufficient: an even number of negative eigenvalues would still result in $\det(j(q, t)) > 0$).

The adjusted formula is shown to be very accurate and capable of capturing the bimodality of the MLE's distribution in the non-linear regression and errors in variable models. The bimodality depends heavily on the ancillary statistic. Previous asymptotic results have been concerned with how the *variance* of the MLE depends on ancillary statistics. This paper has focused on bimodality instead and has shown that it depends heavily on the value of a . By itself a has approximately no information on the parameter, but it contains valuable information on the shape of the distribution. We should therefore condition on the observed value of a calculated from the actual sample. The non-linear regression and errors in variables model have been chosen as leading examples because they have interesting global curvature properties and a changing local curvature and allowed for the illustration of weak instrument (proxy) type effects.

Both examples have only one parameter and one ancillary statistic and the first order conditions for the MLE are in both cases cubic in the parameter. This facilitated the exposition of the main points and the explicit calculation of the confidence regions. The results are not restricted to CEM(2,1) families however, and hold for arbitrary k and d . The definition of \mathbf{a} , the partitioning of the sample space, definitions of the sets F_q, \tilde{F}_q, C_q^+ are still applicable when d and/or $(k - d)$ are larger than one and the p^{**} -formula can still be applied. Practically, having d larger requires integration over a d -dimensional surface, but is conceptually not difficult and only has to be carried out for one outcome of \mathbf{a} . If $(k - d)$ is larger than one then this higher dimension of \mathbf{a} is practically not difficult but opens the possibility of different directions in the orthogonal space of the expectation manifold and choosing direction in which the density changes most.

We have shown how the adjusted p^{**} -formula can be successfully used to obtain accurate conditional confidence intervals when the distribution of the MLE is bimodal. The resulting confidence intervals can be disjoint, depending on the value of the observed ancillary and the estimated parameter value. The resulting p^{**} based confidence regions are very accurate, both conditionally and unconditionally.

A Appendix

A.1 Proof of Lemma 1

Since the covariance matrix is a symmetric positive definite matrix we have an eigendecomposition $\Sigma(\theta)^{-1} = U'\Lambda^{-1}U$ with $U'U = UU' = I_k$ and $\Lambda^{-1} = \text{diag}(1/\lambda_1, \dots, 1/\lambda_k)$ and $\{\lambda_i\}_{i=1}^k$ the strictly positive eigenvalues of $\Sigma(\theta)$. Pre-multiplication by the full rank matrix ΛU leaves the rank of $[B(\theta) : A(\theta)]$ unchanged and post-multiplication of $A(\theta)$ by $(C(\theta)'_{\perp}\Sigma(\theta)C(\theta)_{\perp})^{-1/2}$ does not alter the column space, nor the rank of $[B(\theta) : A(\theta)]$. Hence:

$$\begin{aligned} \text{rank}([B(\theta) : A(\theta)]) &= \text{rank}([B(\theta) : C_{\perp}(\theta)]) \\ &= \text{rank}\left(\left[B(\theta) : \Sigma(\theta)^{-1} B_{\perp}(\theta)\right]\right) \\ &= \text{rank}([\Lambda U B(\theta) : U B(\theta)_{\perp}]) \\ &= \text{rank}\left([\Lambda \tilde{B} : \tilde{B}_{\perp}]\right). \end{aligned}$$

For the rank to be smaller than k , there must exist a vector $v = (v'_1 : v'_2)' \neq 0$ with $v_1 : d \times 1$ and $v_2 : (k-d) \times 1$ such that $0 = [\Lambda \tilde{B} : \tilde{B}_{\perp}]v = \Lambda \tilde{B}v_1 + \tilde{B}_{\perp}v_2$. The $k \times 1$ vectors $\tilde{b} \equiv \tilde{B}v_1$ and $\tilde{b}_{\perp} \equiv \tilde{B}_{\perp}v_2$ are in the column spaces of \tilde{B} and \tilde{B}_{\perp} respectively and hence orthogonal. The rank is therefore reduced only if there exist \tilde{b} and \tilde{b}_{\perp} such that $\tilde{b}_{\perp} = \Lambda \tilde{b} = (\lambda_1 \tilde{b}_1, \dots, \lambda_k \tilde{b}_k)'$. Orthogonality implies that \tilde{b}_i and $(\tilde{b}_{\perp})_i$ must have opposite signs for at least one i , or have a zero when the other one is non-zero for at least one i , since otherwise $\tilde{b}'\tilde{b}_{\perp}$ is a sum of only positive terms and $\tilde{b}'\tilde{b}_{\perp} > 0$. In these cases with $\text{sign}(\tilde{b}_{\perp i}) = -\text{sign}(\tilde{b}_i)$, or $\tilde{b}_{\perp i} = 0 \neq \tilde{b}_i$, or $\tilde{b}_i = 0 \neq \tilde{b}_{\perp i}$, there does not exist a $\lambda_i > 0$ such that $\tilde{b}_{\perp i} = \lambda_i \tilde{b}_i$. Hence no $v \neq 0$ exists such that $[\Lambda \tilde{B} : \tilde{B}_{\perp}]v = 0$ and $[B(\theta) : A(\theta)]$ has rank k and is invertible. This holds for any value of $\theta \in \Theta$ for which $B(\theta)$ and $\Sigma(\theta)$ are regular with full column rank d and k respectively, and in particular for $\hat{\theta}$. Hence the inverse of $[B(\hat{\theta}) : A(\hat{\theta})]$ exists. \square

A.2 Proof of Proposition 1

The score function of the embedding FEM is:

$$s(\eta; T) = T - \frac{\partial \kappa(\eta)}{\partial \eta} = T - \tau(\eta), \quad (24)$$

which can be evaluated at the restricted CEM as $\eta(\hat{\theta})$. Using the covariance matrix of T at the restricted estimate $\eta(\hat{\theta})$ we have:

$$LM = [T - \tau(\hat{\theta})]'\Sigma(\hat{\theta})^{-1}[T - \tau(\hat{\theta})], \quad (25)$$

which asymptotically will have χ_{k-d}^2 distribution under the implicit regularity conditions. Since $[B(\hat{\theta}) : A(\hat{\theta})]$ is invertible by Lemma 1 and because $B(\hat{\theta})'(T - \tau(\hat{\theta})) = 0$ we may write:

$$\begin{aligned} (T - \tau(\hat{\theta})) &= \begin{bmatrix} B(\hat{\theta})' \\ A(\hat{\theta})' \end{bmatrix}^{-1} \begin{bmatrix} B(\hat{\theta})' \\ A(\hat{\theta})' \end{bmatrix} (T - \tau(\hat{\theta})) \\ &= \begin{bmatrix} B(\hat{\theta})' \\ A(\hat{\theta})' \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ A(\hat{\theta})'(T - \tau(\hat{\theta})) \end{bmatrix}. \end{aligned} \quad (26)$$

Hence:

$$LM = \begin{bmatrix} 0 \\ A(\hat{\theta})'(T - \tau(\hat{\theta})) \end{bmatrix}' \left[\begin{bmatrix} B(\hat{\theta})' \\ A(\hat{\theta})' \end{bmatrix} \Sigma(\hat{\theta}) \begin{bmatrix} B(\hat{\theta}) \\ A(\hat{\theta}) \end{bmatrix} \right]^{-1} \begin{bmatrix} 0 \\ A(\hat{\theta})'(T - \tau(\hat{\theta})) \end{bmatrix}.$$

We have $A(\hat{\theta})'\Sigma(\hat{\theta})B(\hat{\theta}) = 0$ since $A(\hat{\theta})' = \{C_{\perp}(\hat{\theta})'\Sigma(\hat{\theta})C_{\perp}(\hat{\theta})\}^{-1}C_{\perp}(\hat{\theta})'$ and $\Sigma(\hat{\theta})B(\hat{\theta}) = C(\hat{\theta})$ we obtain:

$$\begin{aligned} LM &= \begin{bmatrix} 0 \\ A(\hat{\theta})'(T - \tau(\hat{\theta})) \end{bmatrix}' \begin{bmatrix} B(\hat{\theta})'\Sigma(\hat{\theta})B(\hat{\theta}) & 0 \\ 0 & A(\hat{\theta})'\Sigma(\hat{\theta})A(\hat{\theta}) \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ A(\hat{\theta})'(T - \tau(\hat{\theta})) \end{bmatrix} \\ &= (T - \tau(\hat{\theta}))' A(\hat{\theta}) \{A(\hat{\theta})'\Sigma(\hat{\theta})A(\hat{\theta})\}^{-1} A(\hat{\theta})'(T - \tau(\hat{\theta})) \\ &= (T - \tau(\hat{\theta}))' C_{\perp}(\hat{\theta}) \{C_{\perp}(\hat{\theta})'\Sigma(\hat{\theta})C_{\perp}(\hat{\theta})\}^{-1} C_{\perp}(\hat{\theta})'(T - \tau(\hat{\theta})) \\ &= \mathbf{a}'\mathbf{a}. \quad \square \end{aligned}$$

A.3 Derivation of p^{**}

The derivation of the p^* -formula as given by Barndorff-Nielsen (1980) depends on having an accurate approximation for the density of T in the FEM and determining the Jacobian of the transformation from T to $(\hat{\theta}, a)$. The derivation of the Jacobian here is essentially the same and the saddlepoint approximation for FEMs was derived in Barndorff-Nielsen and Cox (1979, p.299):

$$p_T^*(t; \eta) = \frac{1}{(2\pi)^{k/2}} |\Sigma(\hat{\eta})|^{1/2} \exp \{(\eta - \hat{\eta})'t - (\kappa(\eta) - \kappa(\hat{\eta}))\}, \quad (27)$$

which has proved to be highly accurate, not only asymptotically in an *i.i.d.* context when the relative error is of order $O(n^{-1})$, but also in small samples with dependent data.

The derivation of the Jacobian is based on the relation:

$$\begin{bmatrix} B(\hat{\theta})' \\ C(\hat{\theta})'_{\perp} \end{bmatrix} (T - \tau(\hat{\theta})) = \begin{pmatrix} 0 \\ \mathbf{c} \end{pmatrix}. \quad (28)$$

The first d zeros correspond to the first order conditions for the MLE $\hat{\theta}$, which are commonly used for in the derivation of the asymptotic distribution of the MLE in an expansion of the type $0 = s(\hat{\theta}) = s(\theta_0) + \frac{\partial s(\theta_0)}{\partial \theta'}(\hat{\theta} - \theta_0) + remainder$. The term \mathbf{c} stems from the second term in the

definition of the ancillary. Denote particular values of $\hat{\theta}$, \mathbf{c} , \mathbf{a} , and T by q , c , a and t respectively. On taking differentials we obtain:

$$\begin{bmatrix} B(q)' \\ C(q)'_{\perp} \end{bmatrix} dt + \sum_{l=1}^k (t - \tau(q))_l \begin{bmatrix} F(q)_l \\ G(q)_l \end{bmatrix} dq - \begin{bmatrix} B(q)' \\ C(q)'_{\perp} \end{bmatrix} \Sigma(q) B(q) dq = I_{k-d} dc, \quad (29)$$

with $G(q)_l = \left[\frac{\partial}{\partial \theta} \{ (C(\theta)'_{\perp})_l \} \right]_{\theta=q}$. Using $C(q)'_{\perp} \Sigma(q) B(q) = C(q)'_{\perp} C(q) = 0$ we have:

$$dt = \begin{bmatrix} B(q)' \\ C(q)'_{\perp} \end{bmatrix}^{-1} \begin{bmatrix} j(q; t) & 0 \\ -\sum_{l=1}^k (t - \tau(q))_l G(q)_l & I_{k-d} \end{bmatrix} \begin{pmatrix} dq \\ dc \end{pmatrix}, \quad (30)$$

and recalling that by construction $|B(\theta) : C(\theta)_{\perp}| = |i(\theta)| / |\Sigma(\theta)|$ the Jacobian of the transformation from t to (q, a) is $|i(q)|^{-1} |\Sigma(q)| |j(q; t)|$. Now $dc = \left| (C(q)'_{\perp} \Sigma(q) C(q)_{\perp})^{1/2} \right| da = |i(q)|^{1/2} |\Sigma(q)|^{-1/2} da$, since the Jacobian matrix of this transformation does not depend on a (or c). The Jacobian of the overall transformation from T to $(\hat{\theta}, \mathbf{a})$ equals $|i(q)|^{-1/2} |\Sigma(q)|^{1/2} |j(q; t)|$. Given the exponential distribution of T , the density of $(\hat{\theta}, a)$ becomes:

$$p_{\hat{\theta}, a}(q, a; \theta) = |i(q)|^{-1/2} |\Sigma(q)|^{1/2} |j(q; t(q; a))| \exp \{ \eta(\theta)' t(q, a) - \kappa(\theta) \} \tilde{h}(t(q, a)). \quad (31)$$

Now letting $r = \ell(\hat{\eta}; t) - \ell(q; t) = (\hat{\eta} - \eta(q))' t - \{ \kappa(\hat{\eta}) - \kappa(q) \}$, with $\hat{\eta}$ the MLE of the embedding FEM,

$$p_{\hat{\theta}, a}(q, a; \theta) = |j(q; t(q; a))|^{1/2} \exp \{ (\eta(\theta) - \eta(q))' t(q, a) - \{ \kappa(\theta) - \kappa(q) \} \} \cdot \left\{ \frac{|j(q; t(q; a))|}{|i(q)|} \frac{|\Sigma(q)|}{|\Sigma(\hat{\eta})|} e^{-2r} \right\}^{1/2} |\Sigma(\hat{\eta})| \exp \{ \hat{\eta}' t - \kappa(\hat{\eta}) \} \tilde{h}(t). \quad (32)$$

In an *i.i.d.* setting all ratios in the second line are shown by Barndorff-Nielsen (1980) to have an expansion of the form $\frac{|j(q; t(q; a))|}{|i(q)|} = 1 + n^{-1/2} c_1(\theta, a) + O_p(n^{-1})$, $\frac{|\Sigma(q)|}{|\Sigma(\hat{\eta})|} = 1 + n^{-1/2} c_2(\theta, a) + O_p(n^{-1})$ and $e^{-2r} = c_3(\theta, a) \{ 1 + n^{-1/2} c_4(\theta, a) + O_p(n^{-1}) \}$ when a is the affine ancillary. The remaining term is of order $O(n^{-1})$ by the saddlepoint expansion for the FEM. The leading term of this result is the p^* -formula.

When the distribution of $\hat{\theta}$ is bimodal, there is no guarantee that $\hat{\theta}$ is in a neighborhood of the true θ_0 . The Jacobian matrix is still valid however, because it concerns the mapping between T and $(\hat{\theta}, \mathbf{a})$, irrespective of the distance from $\hat{\theta}$ to the true θ_0 . This mapping is locally a diffeomorphism as long as T is bounded away from sets where the mapping between T and $(\hat{\theta}, \mathbf{a})$ is not uniquely defined: i.e. \tilde{M} or where $|j(q; t)| = 0$. To make the point more explicit consider a realization t which is in $F_{q^{(1)}}$ and $\tilde{F}_{q^{(2)}}$. Although $q^{(1)}$ satisfies the first order conditions $s(q^{(1)}; t) = 0$ this point $q^{(1)}$ and associated ancillary $a^{(1)}$ are irrelevant, because t is mapped uniquely on to $q^{(2)}$ and the associated $a^{(2)}$. Hence the relevant neighborhood is around $(q^{(2)}, a^{(2)})$ and the Jacobian is evaluated in this point.

In Figure 1 for the nonlinear regression example, b^- maps uniquely to the MLE value 1.5 and b^+ maps uniquely to -1.54 and their associated values for a . The Jacobian is evaluated in the uniquely defined point and the transformation theorem is valid in this neighborhood where the rank is constant.

When t is not in C_q^+ , the MLE as a function of t will not map t onto (q, a) . For a given value of a this t will be mapped onto a different value $\hat{\theta}(t) \neq q$ for the MLE. This implies that the combination of (q, a) cannot occur and the density must be 0.

The derivation of the Jacobian is independent of the sample size n . Asymptotically, if T is defined as a sample average, T will converge by a law of large numbers to $\tau(\eta_0) = \tau(\theta_0)$ since the true model is the CEM with parameter value θ_0 and expansions in terms of orders of n make sense. In small samples (or when information accrues slower than linearly in n) this is no longer true and the quality of the adjusted p^{**} -formula depends on how well the second line can be approximated by a function of a alone. This may vary by case and sample size. Our examples show that this approximation can be very good in the small samples or in a weak instrument (proxy) type setting. Asymptotically we can fall back on the original result by Barndorff-Nielsen (1980).

A.4 Further details of the Nonlinear Regression Model

The model is embedded in a bivariate normal density:

$$pdf_{\hat{\beta}}(b|\beta, \Omega) \sim |2\pi|^{-n/2} |\Omega|^{-1/2} \exp \left\{ -\frac{1}{2} (b - \beta)' \Omega^{-1} (b - \beta) \right\},$$

with $\Omega = \sigma^2 (X'X)^{-1}$ and σ^2 assumed known. This is a FEM with canonical loglikelihood parameterization:

$$\ell(\beta|b) = b'\eta - \kappa(\eta) + \text{constant},$$

with $\eta = \Omega^{-1}\beta = \sigma^{-2}X'X\beta$, $\kappa(\eta) = \frac{1}{2}\eta'\Omega\eta$. The constant includes terms involving the known σ^2 , observable b and $X'X$, but not β . Our nonlinear regression model is a CEM(2,1) because the two dimensional η is a smooth function of the single parameter θ . Orthogonality of the columns in X implies that we can write $\Omega^{-1} = \text{diag}\{r_1, r_2\}$ and in the example with complementary dummies, $r_1 = n_1/\sigma^2$ and $r_2 = n_2/\sigma^2$. The loglikelihood of the CEM is:

$$\begin{aligned} \ell(\theta) &= r_1 b_1 \theta + r_2 b_2 \theta^2 - \frac{1}{2} r_1 \theta^2 - \frac{1}{2} r_2 \theta^4, \\ \eta(\theta) &= \begin{pmatrix} r_1 \theta \\ r_2 \theta^2 \end{pmatrix}, \quad B(\theta) = \frac{\partial \eta(\theta)}{\partial \theta'} = \begin{pmatrix} r_1 \\ 2r_2 \theta \end{pmatrix}, \quad B_{\perp}(\theta) = \begin{pmatrix} 2r_2 \theta \\ -r_1 \end{pmatrix}, \\ \tau(\theta) &= E[\hat{\beta}] = \begin{pmatrix} \theta \\ \theta^2 \end{pmatrix}, \quad C(\theta) = \frac{\partial \tau(\theta)}{\partial \theta'} = \begin{pmatrix} 1 \\ 2\theta \end{pmatrix}, \quad C_{\perp}(\theta) = \frac{1}{r_1 r_2} \begin{pmatrix} -2\theta \\ 1 \end{pmatrix}, \\ \Sigma(\theta) &= \text{diag}\{r_1, r_2\}. \end{aligned}$$

The constant term $(r_1 r_2)^{-1}$ in C_{\perp} drops out in the construction of $A(\theta)$ since $(k-d) = 1$, but is chosen in line with the definition which specifies that $|B(\theta) : C_{\perp}(\theta)| = |i(\theta)| / |\Sigma(\theta)| = \frac{r_1 + 4r_2 \theta}{r_1 r_2}$. The determinant $|\Sigma(\theta)| = r_1 r_2$ and:

$$i(\theta) = \text{var}(s(\theta)) = r_1 + 4r_2 \theta^2 = B(\theta)' \Sigma(\theta) B(\theta),$$

with $s(\theta)$ the score satisfying:

$$s(\theta) = \frac{\partial \ell(\theta)}{\partial \theta} = r_1 b_1 + 2r_2 b_2 \theta - r_1 \theta - 2r_2 \theta^3 = B(\theta)'(b - \tau(\theta)).$$

The observed information equals:

$$\begin{aligned} j(\theta; b) &= -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'} = -2r_2 b_2 + r_1 + 6r_2 \theta^2 \\ &= r_1 + 4r_2 \theta^2 - 2r_2 (b_2 - \theta^2) = i(\theta) - \sum_{j=1}^2 \frac{\partial^2 \eta_j(\theta)}{\partial \theta \partial \theta'} (b - \tau(\theta))_j, \end{aligned}$$

since $\frac{\partial^2 \eta_1(\theta)}{\partial \theta \partial \theta'} = 0$ and $\frac{\partial^2 \eta_2(\theta)}{\partial \theta \partial \theta'} = 2r_2$. From this equation it is immediate that $E[j(\theta; b)] = i(\theta)$.

The first order conditions $s(\hat{\theta}) = 0$ for the MLE can be written as:

$$\hat{\theta}^3 + d_1 \hat{\theta} + d_0 b_1 = 0,$$

with $d_0 = -\frac{r_1}{2r_2}$, $d_1 = -\left(b_2 + \frac{r_1}{2r_2}\right)$. The type of roots of this (suppressed) cubic equation in $\hat{\theta}$ depend on the discriminant:

$$\Delta = -4d_1^3 - 27d_0^2 b_1^2.$$

If $\Delta > 0$ then there are three distinct real roots. If $\Delta = 0$ has three real roots of which at least two are equal, and if $\Delta < 0$, then there is one real root and two non-real complementary roots. Any one of these three cases can occur since Δ depends on the outcomes b_1 and b_2 which are normally distributed and independent. This cubic can be exploited to derive algebraically which values of $\hat{\theta}$ can occur for given value of a defined next.

$$\begin{aligned} A(\hat{\theta}) &= C(\hat{\theta})_{\perp} \left(C(\hat{\theta})'_{\perp} \Sigma(\hat{\theta}) C(\hat{\theta})_{\perp} \right)^{-1/2} \\ &= \frac{1}{r_1 r_2} \begin{pmatrix} -2\hat{\theta} \\ 1 \end{pmatrix} \left\{ \frac{1}{r_1 r_2} \begin{pmatrix} -2\hat{\theta} \\ 1 \end{pmatrix}' \begin{pmatrix} 1/r_1 & 0 \\ 0 & 1/r_2 \end{pmatrix} \frac{1}{r_1 r_2} \begin{pmatrix} -2\hat{\theta} \\ 1 \end{pmatrix} \right\}^{-1/2} \\ &= \begin{pmatrix} -2\hat{\theta} \\ 1 \end{pmatrix} \frac{1}{\sqrt{4\hat{\theta}^2/r_1 + 1/r_2}} \quad \text{or} \quad \frac{\sqrt{r_1 r_2}}{\sqrt{4r_2 \hat{\theta}^2 + r_1}} \begin{pmatrix} -2\hat{\theta} \\ 1 \end{pmatrix}. \end{aligned}$$

Hence the affine ancillary statistic equals:

$$\begin{aligned} \mathbf{a} &= \frac{1}{\sqrt{4\hat{\theta}^2/r_1 + 1/r_2}} \begin{pmatrix} -2\hat{\theta} \\ 1 \end{pmatrix}' \begin{pmatrix} b_1 - \hat{\theta} \\ b_2 - \hat{\theta}^2 \end{pmatrix} \\ &= \frac{-2\hat{\theta} b_1 + \hat{\theta}^2 + b_2}{\sqrt{4\hat{\theta}^2/r_1 + 1/r_2}}. \end{aligned} \tag{33}$$

Using Proposition 2 we can determine the outcome of the minimal sufficient statistic b from

(q, a) .

$$\begin{aligned} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} &= \tau(q) + \begin{bmatrix} B(q)' \\ A(q)' \end{bmatrix}^{-1} \begin{pmatrix} 0 \\ a \end{pmatrix} \\ &= \begin{pmatrix} q \\ q^2 \end{pmatrix} + \frac{1}{r_1 + 4r_2q^2} \begin{pmatrix} 1 & 2r_2q\sqrt{4q^2/r_1 + 1/r_2} \\ -2q & r_1\sqrt{4q^2/r_1 + 1/r_2} \end{pmatrix} \begin{pmatrix} 0 \\ a \end{pmatrix} \\ &= \begin{pmatrix} q \\ q^2 \end{pmatrix} + \mathbf{a} \frac{1}{\sqrt{4q^2/r_1 + 1/r_2}} \begin{pmatrix} 2q/r_1 \\ 1/r_2 \end{pmatrix}. \end{aligned}$$

Note that by this result if $a = -\frac{1}{2}r_1\sqrt{4\frac{q^2}{r_1} + \frac{1}{r_2}}$ then $b_1 = 0$ and $b_2 = q^2 - \frac{1}{2}\frac{r_1}{r_2}$ and the observed information $j(q; b) = 4r_2q^2 + 2r_1 > 0$, but the likelihood is symmetric in q : $\ell(q) = -\frac{1}{2}q^2(2r_1 - q^2r_2)$ and q and $-q$ have the same value of the likelihood and $b = \left(0, q^2 - \frac{1}{2}\frac{r_1}{r_2}\right)' \in \widetilde{M}$, unless the likelihood is globally maximized for $\theta = 0$. This occurs if $j(0; (0, b_2)') = r_1 - 2r_2b_2 > 0$ so for $b_2 < \frac{r_1}{2r_2} = 1/8$ in the example, $j(0; (0, 1/8)') = 0$.

A.5 Further details of the Errors in Variables Model

In this example we continue with β as model parameter in the CEM, rather than θ , and still $\sigma^2(\beta) = \sigma_u^2 + 2\beta\rho\sigma_u\sigma_v + \beta^2\sigma_v^2$ and we denote $\dot{\sigma}^2(\beta) = \frac{\partial\sigma^2(\beta)}{\partial\beta} = 2\rho\sigma_u\sigma_v + 2\beta\sigma_v^2$.

From the loglikelihood (21) it follows:

$$\begin{aligned} \eta(\beta) &= \frac{1}{\sigma^2(\beta)} \begin{pmatrix} -1/2 \\ \beta\gamma \end{pmatrix}; \quad t = \begin{pmatrix} \sum_{i=1}^n y_i^2 \\ \sum_{i=1}^n y_i z_i \end{pmatrix}; \\ \kappa(\beta) &= \frac{1}{2} \frac{\beta^2\gamma^2}{\sigma^2(\beta)} \sum_{i=1}^n z_i^2 + \frac{n}{2} \log(\sigma^2(\beta)). \end{aligned}$$

The embedding FEM has cumulant function, writing $s_{zz} = \sum_{i=1}^n z_i^2$:

$$\kappa(\eta) = -\frac{1}{4}s_{zz}\frac{\eta_2^2}{\eta_1} - \frac{n}{2}\log|-2\eta_1|$$

and it is easily checked that $\kappa(\beta) = \kappa(\eta(\beta))$. The expectation of T as a function of β is quadratic in β :

$$\tau(\beta) = E_\beta[T] = \begin{pmatrix} \beta^2\gamma^2s_{zz} + n\sigma^2(\beta) \\ \beta\gamma s_{zz} \end{pmatrix},$$

and the variance of the minimal sufficient statistic equals:

$$\Sigma(\beta) = Var[T] = \sigma^2(\beta) s_{zz} \begin{pmatrix} 4\beta^2\gamma^2 + 2\frac{n}{s_{zz}}\sigma^2(\beta) & 2\beta\gamma \\ 2\beta\gamma & 1 \end{pmatrix},$$

with $|\Sigma(\beta)| = 2ns_{zz}\sigma^6(\beta)$.

The score can be written as:

$$s(\beta) = \sum_{j=0}^3 \beta^j d_j,$$

with coefficients:

$$\begin{aligned} d_0 &= \{2t_1\rho\sigma_u\sigma_v + 2t_2\gamma\sigma_u^2 - 2n\rho\sigma_u^3\sigma_v\} / (2\sigma^2(\beta)), \\ d_1 &= 2t_1\sigma_v^2 - 2\sigma_u^2 \cdot (s_{zz}\gamma^2 + n\sigma_v^2\{1 + 2\rho^2\}) / (2\sigma^2(\beta)), \\ d_2 &= -2t_2\gamma\sigma_v^2 - 2\rho\sigma_u\sigma_v (s_{zz}\gamma^2 + 3n\sigma_v^2) / (2\sigma^2(\beta)), \\ d_3 &= -2n\sigma_v^4 / (2\sigma^2(\beta)). \end{aligned}$$

After multiplying $s(\hat{\beta}) = 0$ by the positive $\sigma^2(\beta)$ the first order condition is a cubic in $\hat{\beta}$.

The expected Fisher information equals

$$i(\beta) = \frac{1}{\sigma^2(\beta)} \left(\gamma^2 s_{zz} - 2n \frac{(\beta\sigma_v + \rho\sigma_u)\sigma_v^2}{\sigma^2(\beta)} \right).$$

The observed information is more involved, but can of course be written as $i(\beta)$ plus a term linear in $(t - \tau(\beta))$ as in (7) $j(\beta; t) = i(\beta) - \sum_{l=1}^k (t_l - \tau_l(\beta)) \frac{\partial^2 \eta(\beta)}{\partial \beta \partial \beta^l}$.

The gradients to the canonical manifold and expectation manifold and there orthogonal complements are

$$\begin{aligned} B(\beta) &= \frac{\partial \eta(\beta)}{\partial \beta} = \frac{\dot{\sigma}^2(\beta)}{\sigma^4(\beta)} \begin{pmatrix} 1/2 - 1/2\sigma^2(\beta)/\dot{\sigma}^2(\beta) \\ -\gamma\beta + \gamma\sigma^2(\beta)/\dot{\sigma}^2(\beta) \end{pmatrix}; & B_{\perp}(\beta) &= \begin{pmatrix} 2\gamma(\beta - \sigma^2(\beta)/\dot{\sigma}^2(\beta)) \\ 1 - \sigma^2(\beta)/\dot{\sigma}^2(\beta) \end{pmatrix}; \\ C(\beta) &= \frac{\partial \tau(\beta)}{\partial \beta} = \begin{pmatrix} 2\beta\gamma^2 s_{zz} + n\dot{\sigma}^2(\beta) \\ \gamma s_{zz} \end{pmatrix}; & C_{\perp}(\beta) &= \lambda \begin{pmatrix} -\gamma s_{zz} \\ 2\beta\gamma^2 s_{zz} + n \cdot \dot{\sigma}^2(\beta) \end{pmatrix}; \end{aligned}$$

with λ such that $|B(\beta) : C_{\perp}(\beta)| = |i(\theta)| / |\Sigma(\beta)|$. This scalar is slightly complicated and cancels in the expression for $A(\beta)$ so there is little use for an explicit expression here. Since:

$$C_{\perp}(\beta)' \Sigma(\beta) C_{\perp}(\beta) = \lambda^2 n s_{zz} \sigma^2(\beta) (2 s_{zz} \beta \gamma^2 \sigma^2(\beta) + n \dot{\sigma}^2(\beta)),$$

we obtain:

$$A(\beta) = \frac{1}{\sqrt{n s_{zz} \sigma^2(\beta) (2 s_{zz} \beta \gamma^2 \sigma^2(\beta) + n \dot{\sigma}^2(\beta))}} \begin{pmatrix} -\gamma s_{zz} \\ 2\beta\gamma^2 s_{zz} + n \cdot \dot{\sigma}^2(\beta) \end{pmatrix},$$

and

$$\mathbf{a} = \frac{-\gamma s_{zz} \left(t_1 - \left(\hat{\beta}^2 \gamma^2 s_{zz} + n \sigma^2(\hat{\beta}) \right) \right) + \left(2\hat{\beta} \gamma^2 s_{zz} + n \dot{\sigma}^2(\hat{\beta}) \right) \left(t_2 - \hat{\beta} \gamma s_{zz} \right)}{\sqrt{n s_{zz} \sigma^2(\hat{\beta}) \left(2 s_{zz} \hat{\beta} \gamma^2 \sigma^2(\hat{\beta}) + n \dot{\sigma}^2(\hat{\beta}) \right)}}. \quad (34)$$

Although this formula is analytically more involved, it can easily be stored and derived using symbolic algebra packages and numerically evaluated. All calculations of the derivatives, the ancillary statistic, and the inverse mapping are straightforward and we will not present their analytic formulas here.

A.6 Smallest Prediction Regions

Analogues to a standard proof of the Neyman-Pearson Lemma we have the following proposition for the smallest prediction region, see for instance Juola, (1993).

Theorem 1 *Suppose $p_T(t)$ is a continuous density for $T \in \mathbb{T}$, $g(t)$ a continuous and strictly positive function, and \mathbf{A} a measurable subset of \mathbb{T} and let $P[\mathbf{A}] = \int_{\mathbf{A}} p_T(t)dt$ then:*

$$\min_{\mathbf{A}} \int_{\mathbf{A}} g(t)dt, \tag{35}$$

$$\text{s.t. } P[\mathbf{A}] = 1 - \alpha, \tag{36}$$

with $0 < \alpha < 1$, is solved by:

$$\mathbf{A} = \{t \in \mathbb{T} : p_T(t)/g(t) > \delta\}, \tag{37}$$

with $\delta > 0$ chosen to satisfy the condition $P[\mathbf{A}] = 1 - \alpha$.

Proof. Let \mathbf{R} be any measurable subset of \mathbb{T} that satisfies (36) and denote the complements of \mathbf{A} and \mathbf{R} by $\bar{\mathbf{A}}$ and $\bar{\mathbf{R}}$ respectively, then:

$$\begin{aligned} 0 &= \int_{\mathbf{R}} p_T(t)dt - \int_{\mathbf{A}} p_T(t)dt = \int_{\bar{\mathbf{A}} \cap \mathbf{R}} p_T(t)dt - \int_{\mathbf{A} \cap \bar{\mathbf{R}}} p_T(t)dt \\ &\leq \int_{\bar{\mathbf{A}} \cap \mathbf{R}} cg(t)dt - \int_{\mathbf{A} \cap \bar{\mathbf{R}}} cg(t)dt = c \left(\int_{\mathbf{R}} g(t)dt - \int_{\mathbf{A}} g(t)dt \right), \end{aligned}$$

and since $\delta > 0$ it follows that $\int_{\mathbf{R}} g(t)dt \geq \int_{\mathbf{A}} g(t)dt$ which means that \mathbf{A} is better in the sense of (35), or at least as good, as any other set. \square

Corollary 1 *If $g(t) = 1$ and the density is $p_T(t; \theta)$ then the (Lebesgue) smallest prediction region is:*

$$\mathbf{A}(\theta) = \{t \in \mathbb{T} : p_T(t; \theta) > \delta\}. \tag{38}$$

A.7 Additional Figures

Approximate Ancillarity in Nonlinear Regression model.

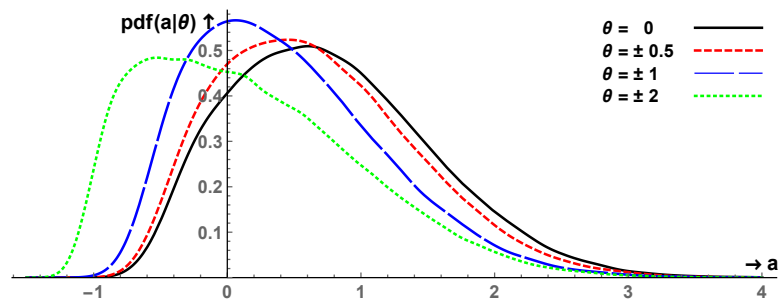


Figure 10: Density of the approximate ancillary in the Nonlinear Regression for different values of θ . If \mathbf{a} was truly ancillary, this density would not depend on θ and all four would be identical. $n = 50$, $n_1 = 10$, $n_2 = 40$, $\sigma^2 = 10$.

Confidence Regions in the Nonlinear Regression Model when $a = 0$.

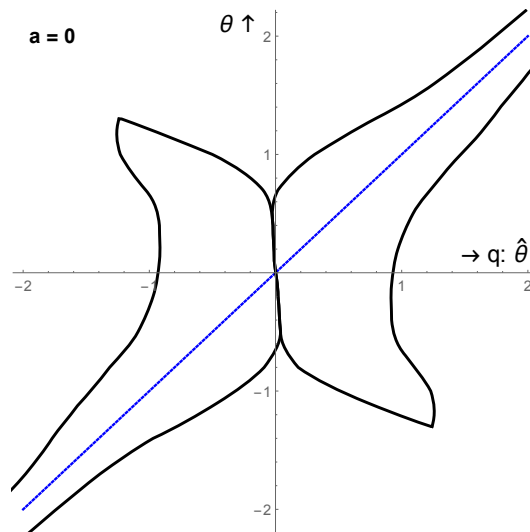


Figure 11: Confidence regions $B(q)$ as a function of q (observed $\hat{\theta}$) in the nonlinear regression model when $a = 0$. $n = 50$, $n_1 = 10$, $n_2 = 40$, $\sigma^2 = 10$.

Confidence Regions for the Errors in Variables model.

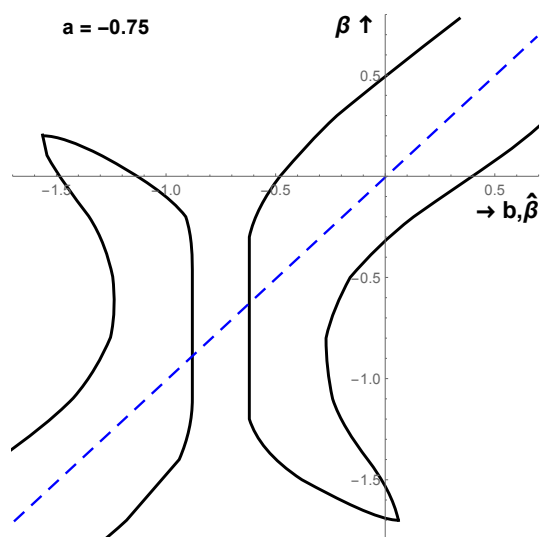


Figure 12: Confidence regions $B(q)$ as a function of q (observed $\hat{\theta}$) in the Errors in Variable model when $a = -0.75$. $n = 50$, $n_1 = 10$, $n_2 = 40$, $\sigma^2 = 10$.

References

- Amari, S. I. (1982). Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, 357-385.
- Amari,S.-I. (1985) *Differential Geometrical Methods in Statistics*. Springer Lecture Notes in Statistics, 28. Springer Verlag.
- Ariza, C.,& Van Garderen, K. J. (2010). Conditional bimodality in a structural equations model. University of Amsterdam Econometrics Discussion Paper: 2009/12 (rev 2010).
- Barndorff-Nielsen, O.E.(1980). Conditionality resolutions. *Biometrika* 67, 293-310.
- Barndorff-Nielsen, O.E.(1983) On a formula for the distribution of the maximum likelihood estimator. *Biometrika* 70, 343-365.
- Barndorff-Nielsen, O.E. & D.R. Cox.(1979) Edgeworth and saddlepoint approximations with statistical applications (with discussion). *Journal of the Royal Statistical Society B* 41: 279-312.
- Barndorff-Nielsen, O.E. & D.R. Cox. (1989) *Asymptotic Techniques for Use in Statistics*. Chapman and Hall.
- Barndorff-Nielsen, O.E. & D.R. Cox. (1994) *Inference and Asymptotics*. Chapman and Hall.
- Bergstrom, A.R. (1962). The exact sampling distributions of least squares and maximum likelihood estimators of the marginal propensity to consume. *Econometrica* 30, 480-490.
- Basu, D. (1955). On Statistics Independent of a Complete Sufficient Statistic. *Sankhyā* 15: 377-380.
- Cox, D. R. (1980). Local ancillarity. *Biometrika* 67, 273-8.
- Chesher, A., & Smith, R. J. (1997). Likelihood ratio specification tests. *Econometrica* 65, 627-646.
- Daniels, H.E. (1954) Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* 25, 631-650.
- Daniels, H.E. (1980) Exact saddlepoint approximations. *Biometrika* 67, 59-63.
- De Nadai, M., & Lewbel, A. (2016). Nonparametric errors in variables models with measurement errors on both sides of the equation. *Journal of Econometrics* 191, 19-32.
- Durbin, J. (1954). Errors in variables. *Revue de l'institut International de Statistique*, 23-32.
- Durbin, J. (1980). Approximations for densities of sufficient estimators. *Biometrika* 67, 311-33.

- Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *The Annals of Statistics*, 1189-1242.
- Efron, B. (1978). The geometry of exponential families. *The Annals of Statistics*, 6, 362-376.
- Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, 65(3), 457-483.
- Fisher, R. A. (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society of London. A* 144, 285-307.
- Forchini, G. (2006), On the bimodality of the exact distribution of the TSLS estimator. *Econometric Theory* 22, 932-946.
- Fraser, D.A.S., (1968), *The Structure of Inference*. Wiley.
- Fraser, D.A.S., (1979), *Probability and Statistics*. Duxbury.
- Hillier, G.H. (2006), Yet more on the exact properties of IV estimators. *Econometric Theory* 22, 913-931.
- Hillier, G., & Armstrong, M. (1999). The density of the maximum likelihood estimator. *Econometrica* 67, 1459-1470.
- Hinkley, D. V. (1980). Likelihood as approximate pivotal distribution. *Biometrika* 67, 287-92.
- Holly, A., & Phillips, P.C.B. (1979). A Saddlepoint Approximation to the Distribution of the k-Class Estimator of a Coefficient in a Simultaneous System. *Econometrica*, 47(6), 1527-1547.
- Juola, R.C. (1993). More on Shortest Confidence Intervals. *The American Statistician*, 47, 117-119.
- Maddala, G.S. & Jeong, J. (1992) On the exact small sample distribution of the instrumental variable estimator. *Econometrica* 60, 181-183.
- Mavroeidis, S. & Van Garderen, K.J.(2006), Conditional Inference in the Cointegrated Vector Autoregressive Model, mimeograph Oxford & Amsterdam.
- Nelson, C.R. & Startz, R. (1990) Some further results on the small sample properties of the instrumental variable estimator. *Econometrica* 58, 967-976.
- Phillips, P.C.B. (2006), A remark on bimodality and weak instrumentation in structural equation estimation. *Econometric Theory* 22, 947-960.
- Spady, R. H. (1991). Saddlepoint approximations for regression models. *Biometrika*, 78, 879-889.
- Stock, J. H., Wright, J. H., & Yogo, M. (2012). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*.

Van Garderen, K. J. (1997). Curved exponential models in econometrics. *Econometric Theory* 13, 771-790.

Woglom, G. (2001) More results on the exact small sample properties of the instrumental variable estimator. *Econometrica* 69, 1381–1389.