# Modified-Wald tests with Weak Identification

**Bertille Antoine**

Université de Montréal

UNC-CH, CIREQ

bertille.antoine@umontreal.ca

**Eric Renault**

University of North Carolina, CH

CIRANO, CIREQ

renault@email.unc.edu

# Framework:

We consider the parameter of interest $\theta \in \Theta \subset \mathbb{R}^p$. $\hat{\theta}_n$ (calculated given the observed data $x = (x_1 \cdots x_n)^T$ on a random variable $X$) is an asymptotically normal estimator of $\theta$.

$$\sqrt{n}\left[\hat{\theta}_n - \theta\right] \xrightarrow{d} \mathcal{N}(0, \Sigma_\theta)$$

where the asymptotic variance of $\hat{\theta}_n$, $\Sigma_\theta$, is supposed to be known.

We are interested in constructing **confidence regions** for a parameter ${}_r\xi(\theta)$ where ${}_r\xi : \Theta \longrightarrow \mathbb{R}^r$ with $r \leq p$ such that $\partial_r\xi/\partial\theta^T$ is of full row rank.

These regions are defined as the values ${}_r\xi_0$ for which the null hypothesis $H_0({}_r\xi_0) : \ {}_r\xi(\theta) = \ {}_r\xi_0$ is not rejected.

Note that the following theory is not affected if $\Sigma_\theta$ is unknown as long as it can be consistently estimated.

# A common procedure, Wald statistic:

A common procedure is to use a Wald-type statistic. The Wald-type confidence region is then defined as:

$$CR_W(\alpha) = \left\{ \,_r\xi_0 \,/\, n \left[ _r\xi(\hat{\theta}_n) - \,_r\xi_0 \right]^T \Sigma^{-1}_{_r\xi}(\hat{\theta}_n) \left[ _r\xi(\hat{\theta}_n) - \,_r\xi_0 \right] \leq \chi^2_\alpha(r) \right\}$$

where $\Sigma^{-1}_{_r\xi}(\hat{\theta}_n)$ is the estimated asymptotic variance of $_r\xi(\hat{\theta}_n)$:

$$\Sigma_{_r\xi}(\hat{\theta}_n) = \frac{\partial_r\xi}{\partial\theta^T}(\hat{\theta}_n)\Sigma_\theta \frac{\partial_r\xi^T}{\partial\theta}(\hat{\theta}_n)$$

The Wald statistic is a natural one but it (often) yields to **ellipsoidal confidence regions** ie symmetric bounded regions.

Note that the original Wald statistic was defined for a parametric model $\mathcal{M} = \{f(x;\theta)|\theta \in \Theta\}$ and using the Maximum Likelihood estimator. The Wald-type statistics use any asymptotically normal estimator of $\theta$.

# Necessity for Unboundedness:

**Boundedness** has become a real issue since Dufour (1997) in the context of locally almost unidentified parameter (LAU) (ie parameter with identification issues coming for instance from transformation with discontinuities, or an unappropriate choice of instruments)

**Necessary Condition** (Dufour, 1997): *When an LAU parametric function has an unbounded range, under regularity conditions, any valid confidence set should have nonzero probability of being unbounded under any distribution compatible with the model.*

A "valid" confidence set $CR_{_r\xi}(X)$ for the transformation $_r\xi(\theta)$ with **level** $(1-\alpha)$ is such that:

$$\inf_{\theta\in\Theta} P_\theta\left( {}_r\xi(\theta) \in CR_{_r\xi}(X)\right) \geq 1-\alpha$$

Further it has **size** $(1-\alpha)$ (or coverage probability) when the above infimum equals $(1-\alpha)$.

Clearly if the problem is stated as before, Wald statistic does not (always) satisfy the above necessary condition.

# Why does Wald fail?

$$W_n({}_r\xi_0) = n\left[{}_r\xi(\hat{\theta}_n) - {}_r\xi_0\right]^T \Sigma^{-1}_{{}_r\xi}(\hat{\theta}_n)\left[{}_r\xi(\hat{\theta}_n) - {}_r\xi_0\right]$$

Intuitively, one could think that Wald-(type) statistics do not fully incorporate the information contained in $H_0$. Note that the matrix $\Sigma_{{}_r\xi}(\theta)$ is a function of $\theta$ rather than the parameter of interest ${}_r\xi(\theta)$. In particular, if $\Sigma_{{}_r\xi}(.)$ was a function of ${}_r\xi_0$ as well, it is not clear that the confidence region would remain bounded since the statistic would not be a quadric form.

This observation is related to the work of Critchley, Marriott and Salmon in **differential geometry**. They note that the Wald statistic is not a genuine geometrical object: it is neither a squared length of a vector in a tangent space, nor a squared distance between two points in a manifold since it uses a fixed metric, whereas the metric should in general vary with $\xi$. For an introduction to these concepts, see Critchley *et al.* and for a more complete treatment Amari.

# The Idea of *Completion*:

Since in general ${}_r\xi(\theta)$ does not constitute a complete reparametrization of the problem, one can not replace $\theta$ by some function of ${}_r\xi$. Hence, the first idea is to **"complete" the partial reparametrization**: we define the new (p,1)-vector of parameter as,

$$\xi(\theta) = \begin{pmatrix} {}_r\xi(\theta) \\ {}_{p-r}\xi(\theta) \end{pmatrix} \text{ with } Rank\left[\frac{\partial \xi}{\partial \theta^T}\right] = p$$

Basically if we were working in a parametric model, the initial model $\mathcal{M} = \{f(x;\theta)|\theta \in \Theta\}$ would be rewritten as $\mathcal{M} = \{f^*(x;\xi)|\xi \in \Theta\}$ via a legitimate change of parameter.

# Modified Wald Procedure (1):

**Comments:**

1) The existence of such a vector $\xi$ is always guaranteed at least in a neighborhood of $H_0$.

2) The asymptotically normal estimator $\xi$ is obviously obtained as $\hat{\xi}_n = \xi(\hat{\theta}_n)$.

3) Also we have,

$$\Sigma_\xi(\hat{\xi}_n) = \frac{\partial \xi(\hat{\theta}_n)}{\partial \theta^T} \Sigma_\theta \frac{\partial \xi^T(\hat{\theta}_n)}{\partial \theta} = \begin{pmatrix} \Sigma_{_r\xi}(\hat{\theta}_n) & ** \\ ** & \Sigma_{_{p-r}\xi}(\hat{\theta}_n) \end{pmatrix}$$

$\longrightarrow$ $\Sigma_{_r\xi}(\theta_n)$ can be expressed as a function of $\xi$ only, $\Sigma_{_r\xi}(\xi)$.

$\longrightarrow$ a consistent estimator of the variance under $H_0$ is enough, so we use

$$\Sigma_\xi(\hat{\xi}_{n,0}) \text{ where } \hat{\xi}^T_{n,0} = \left[{}_r\xi^T_0 \;\; {}_{p-r}\hat{\xi}^T_n\right]$$

Note the slight abuse of notation: unless ${}_{p-r}\xi$ is chosen such that, at $\hat{\theta}_n$, the $r$-constant lines are orthogonal to the *(p-r)*-constant lines with respect to the metric $\Sigma_\theta(\hat{\theta}_n)$, there is no such a direct relationship between $\left[\Sigma_\xi^{-1}(\hat{\xi}_n)\right]_{r,r}$ and $\Sigma^{-1}_{_r\xi}(\hat{\theta}_n)$. Such a choice is always available!

# Modified Wald Procedure (2):

Hence we want to base inference on $\Gamma(\xi)$ for a *well-chosen* transformation $\Gamma$ such that

$$\Gamma : \Theta \to \mathbb{R}^r, \qquad Rank\left[\partial\Gamma/\partial\xi^T\right] = r \qquad \text{and}$$

$$\sqrt{n}\left[\Gamma(\hat{\xi}_n) - \Gamma(\hat{\xi}_{n,0})\right] \sim \frac{\partial\Gamma(\hat{\xi}_{n,0})}{\partial_r\xi^T}\sqrt{n}\left[{}_r\hat{\xi}_n - {}_r\xi_0\right] \text{ where } \hat{\xi}_{n,0} = \begin{pmatrix} {}_r\xi_0 \\ {}_{p-r}\hat{\xi}_n \end{pmatrix}$$

The **Modified-Wald statistic** is defined as:

$$\begin{aligned} MW_n \quad = \quad & n\left[\Gamma(\hat{\xi}_n) - \Gamma(\hat{\xi}_{n,0})\right]^T \left(\frac{\partial\Gamma(\hat{\xi}_{n,0})}{\partial_r\xi^T}\Sigma_{r\xi}(\hat{\xi}_{n,0})\frac{\partial\Gamma^T(\hat{\xi}_{n,0})}{\partial_r\xi}\right)^{-1} \\ & \times\left[\Gamma(\hat{\xi}_n) - \Gamma(\hat{\xi}_{n,0})\right] \end{aligned}$$

# Modified Wald Procedure (3):

**THEOREM 1** *Under* $H_0({}_r\xi_0): \; {}_r\xi(\theta) = {}_r\xi_0$*, we have:* $MW_n \xrightarrow{d} \chi^2(r)$ *and the associated confidence region with level* $\alpha$ *is defined as:*

$$CR_{{}_r\xi}(\alpha) = \left\{ {}_r\xi_0 \;/\; MW_n \leq \chi^2_\alpha(r) \right\}$$

We define the power function associated with the above test as:

$$\pi_n({}_r\xi) = P_{{}_r\xi}(MW_n \notin CR_{{}_r\xi}(\alpha))$$

**THEOREM 2** *The sequence of tests with power function* $\pi_n({}_r\xi)$ *is asymptotically consistent at level* $\alpha$ *at the alternative* ${}_r\xi$*.*

# Modified Wald Procedure (4):

**Comments:**

1) We keep the appeal of a classic Wald statistic because constrained estimation is not needed.

2) We have a general procedure that works without the need for a parametric model.

3) The choice of the function $\Gamma$ might be questioned. Simply using $\Gamma(\xi) = {}_r\xi$ would lead to the following statistic:

$$n \left[ {}_r\hat{\xi}_n - {}_r\hat{\xi}_0 \right]^T \Sigma_{r\xi}^{-1}(\hat{\xi}_{n,0}) \left[ {}_r\hat{\xi}_n - {}_r\hat{\xi}_0 \right]$$

which we may hope will circumvent the boundedness problem of the Wald procedure. However we see that $\Sigma_{r\xi}(\hat{\xi}_{n,0})$ depends on ${}_r\xi_0$ as well as on ${}_{p-r}\hat{\xi}_n$. We pretend that in some cases the extra flexibility coming from the function $\Gamma(.)$ may be used to incorporate special features of the problem of interest and for instance "kill" the dependence on ${}_{p-r}\hat{\xi}_n$. However a general theory appears unachievable.

# Example 1: Ratio of parameters

$$\begin{array}{ccccc} \psi(\beta_1, \beta_2) & = & \beta_1^{-1} & \times & \beta_2 \\ (r, 1) & & (r, r) & & (r, 1) \end{array}$$

Assumption: $\sqrt{n}\left[vec(\hat{\beta}) - vec(\beta)\right] \xrightarrow{d} \mathcal{N}(0, \Sigma_\beta)$

| Initial Parameter $\theta$ | Transformation $_r\xi(\theta)$ | Completion $\xi(\theta)$ |
|---|---|---|
| $vec(\beta)$ | $\psi(\beta) = \beta_1^{-1}\beta_2$ | $\xi = vec[\psi(\beta)\ \beta_1]$ |

The modified variance is:

$$\Sigma_\psi(\psi_0, \hat{\beta}_1) = \hat{\beta}_1^{-1}\left(\left[-\psi_0^T\ 1\right] \otimes I_r\right)\Sigma_\beta\left(\left[-\psi_0^T\ 1\right] \otimes I_r\right)'(\hat{\beta}_1^{-1})^T$$

and the confidence region

$$CR_\psi(\alpha) = \left\{\psi_0 \,/\, n\left[\psi_0 - \hat{\psi}_n\right]^T \Sigma_\psi^{-1}(\psi_0, \hat{\beta}_1)\left[\psi_0 - \hat{\psi}_n\right] \leq \chi^2_\alpha(r)\right\}$$

■ In the **univariate case** the CR is written as:

$$CR_\psi(\alpha) = \left\{\psi_0 \,/\, \psi_0^2 \left[\hat{\beta}_1^2 - \frac{\chi^2_\alpha(1)}{n}\sigma_{11}\right] - 2\psi_0 B + C \le 0\right\}$$ for some real constants $B, C$.

which is simply a quadratic region.

**Comments:**

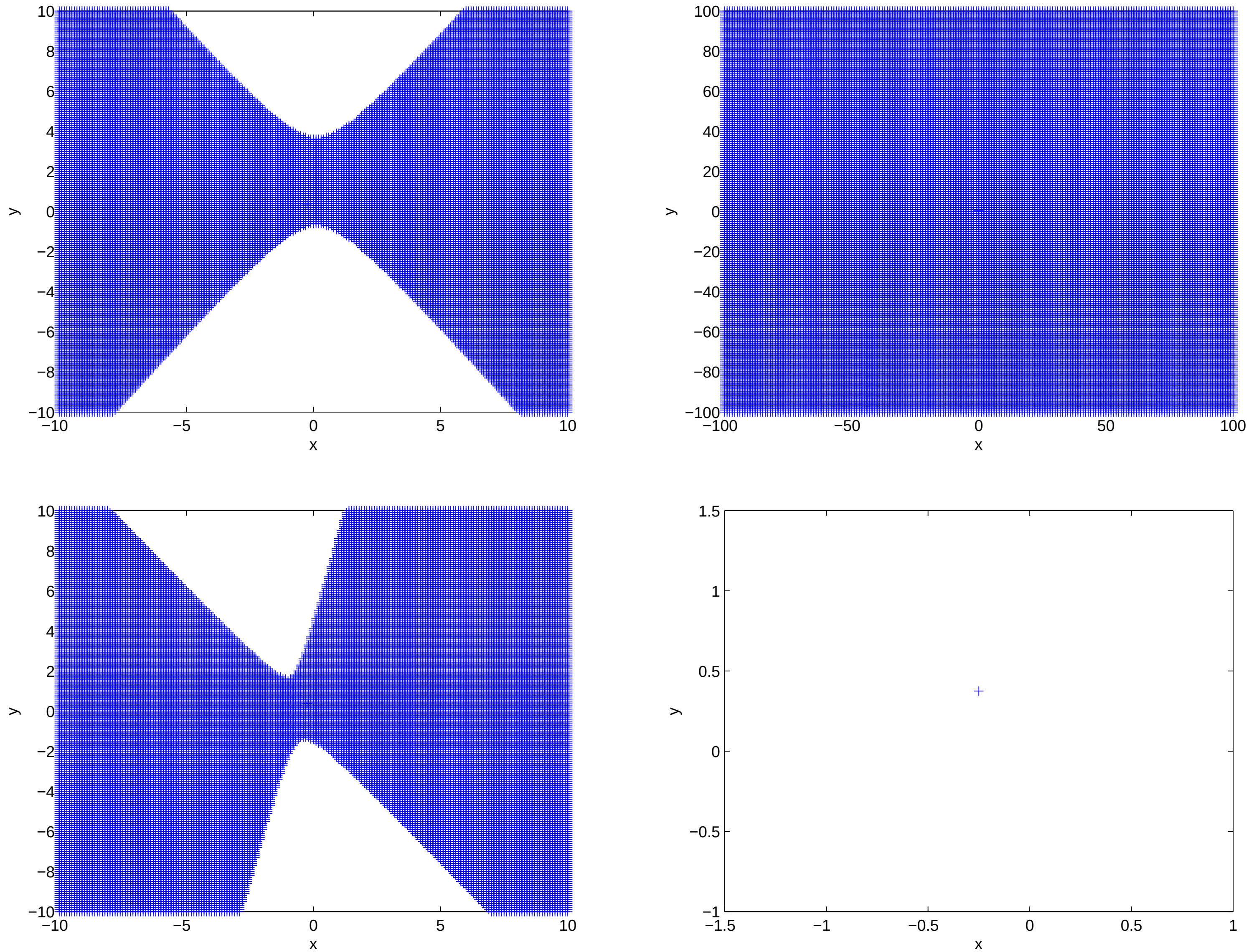
1) We can determine the probability of an unbounded CR:

$$P\left(CR_\psi(\alpha)\; unbounded\right) = P\left(n\frac{\hat{\beta}_1^2}{\sigma_{11}} < \chi^2_\alpha(1)\right) \xrightarrow{n} 1-\alpha \text{ when } \beta_1 \text{ arbitrary close to 0}$$

2) $CR_\psi$ corresponds to the confidence region derived by Dufour 1997 in the context of a parametric regression model with uncorrelated normal errors.

3) $CR_\psi$ can also be reinterpreted as the confidence region that would have been obtained with a classic Wald test performed on a linear equivalent reformulation of the null hypothesis: $H_0 : \beta_2 - \beta_1\psi_0 = 0$. Recall however that Wald test is (in general) not invariant to reformulations of the null hypothesis.

■ **Some possible Shapes** of the confidence region (shaded area) when $r = 2$



# Example 2: Weak instruments

We consider the following just-identified structural model:

$$\begin{cases} \underset{(r,1)}{y} = \underset{(n,r)}{Y} \underset{(r,1)}{\psi} + \underset{(n,1)}{u} \\ \underset{(n,r)}{Y} = \underset{(n,r)}{X} \underset{(r,r)}{\Pi} + \underset{(n,r)}{V} \end{cases}$$

We want to provide inference on the structural parameter $\psi$.

$\psi$ is identifiable if and only if $Rank\left[\Pi\right] = r$.

A case of interest appears when $\Pi$ might be a weak set of instruments in the sense of Staiger and Stock (1997):

$$\Pi = \frac{C}{\sqrt{n}} \text{ where C is a fixed deterministic matrix of rank r.}$$

The well-known 2SLS estimator of $\psi$ (in the justidentified case) is: $\hat{\psi} = (X'Y)^{-1}X'y$

**LEMMA 1** *Under $H_0 : \psi = \psi_0$ and some regularity conditions, we have:*

*i) If the instruments are valid:* $\sqrt{n}\left(\hat{\psi} - \psi_0\right) \xrightarrow{d} \mathcal{N}(0, \Sigma)$

*ii) If the instruments are weakly identified in the sense of Staiger and Stock (1997),*

$$vec\left[\frac{X'Y}{\sqrt{n}}\ \frac{X'y}{\sqrt{n}}\right] \xrightarrow{d} \mathcal{N}\left(vec\left[QC\ QC\psi_0\right], \Sigma_\beta(\psi_0)\right)$$

*iii) If the instruments are invalid, we get the result ii) by replacing $C$ by the null matrix.*

$\longrightarrow$ Hence in the 3 above cases (valid, weak and invalid instruments), the assumption of an asymptotically normal estimator is fulfilled.

Note that:

- for case i) valid instruments, inference is provided on $\psi$ taken as the original parameter.

- for cases ii) weak and iii) invalid instruments, inference is provided on $\psi$ taken as a transformation of $[X'Y/n \; X'y/n]$.

We can now state the main result:

**THEOREM 3** *In the three cases of interest i) valid, ii) weak and iii) invalid instruments, under* $H_0 : \psi = \psi_0$ *and some regularity conditions, we have:*

$$MW_n(\psi_0) = n \left[\hat{\psi} - \psi_0\right]^T \Sigma_\psi^{-1}(\psi_0) \left[\hat{\psi} - \psi_0\right] \xrightarrow{d} \chi^2(r)$$

*with* $\Sigma_\psi(\psi_0) = (y - Y\psi_0)^T (y - Y\psi_0) \times [Y' P_X Y]^{-1}$

The **Confidence Region** is then:

$$CR_\psi(\alpha) = \left\{ \psi_0 \; / \; \left[\hat{\psi}_n - \psi_0\right]^T \Sigma_\psi^{-1}(\psi_0, \hat{\beta}_1) \left[\hat{\psi}_n - \psi_0\right] \leq \chi^2_\alpha(r) \right\}$$

which can be rewritten as:

$$CR_\psi(\alpha) = \left\{ \psi_0 \; / \; \psi_0^T Y^T A Y \psi_0 - 2\psi_0^T A y + y^T A y \leq 0 \right\}$$

with $A = P_X - \chi^2_\alpha(r) I_n / n$

- In the **univariate case** (r=1), we can show that:

i) with strong identification: $P(unbounded\, CR_\psi(\alpha)) \xrightarrow{n} 0$

ii) with weak identification: $P(unbounded\, CR_\psi(\alpha)) \xrightarrow{n} 1 - \tilde{\alpha}$ with $\tilde{\alpha} > \alpha$

iii) with unidentification: $P(unbounded\, CR_\psi(\alpha)) \xrightarrow{n} 1 - \alpha$

- Comparison with the **literature**

Other papers obtain a similar structure of the confidence region:

$$CR_{\psi}(\alpha, A) = \left\{\psi_0 \,/\, \psi_0^T Y^T A Y \psi_0 - 2\psi_0^T A y + y^T A y \leq 0\right\}$$

where the matrix $A$ depends on the method used:

| Method | Matrix A |
|---|---|
| Modified Wald | $P_X - \chi^2_{\alpha}(r) I_n / n$ |
| Anderson-Rubin | $P_X - M_X \chi^2_{\alpha}(r)/(n-r)$ |
| Wang and Zivot (GMM0) | $P_X Y (Y^T P_X Y)^{-1} Y^T P_X - I_n \chi^2_{\alpha}(r)/n$ |
| Wang and Zivot (LR-LIML) | $I_n - k(LIML) \exp[\chi^2_{\alpha}(r)/n] M_X$ |

■ Inference about a **subvector of structural parameter**:

- We now want to produce inference on a subvector of $\psi$: *wlog* we suppose that inference is provided about the first $m$ components of $\psi$ denoted as ${}_m\psi$.

- The classical hypothesis in the literature is to suppose that the remaining components (ie the ones not involved in $H_0$ and denoted as ${}_{r-m}\psi$) are strongly identified.

- We can show that all the previous results remain valid when ${}_{r-m}\psi$ are simply replaced by their respective consistent estimators.

$\longrightarrow$ No conservative projection methods are needed with modified Wald statistic.

# Incoming Research:

★ **Power under local alternatives**

★ **Choice of** $_{p-r}\xi$

★ **General theory for the shape of Confidence Regions**

★ **Choice of the function** $\Gamma(.)$

★ **More examples**: structural change and near unit root