An alternative paradigm for probabilistic model uncertainty bounding in prediction error identification

Paul M.J. Van den Hof

Sippe Douma, Arjan den Dekker, Xavier Bombois

Delft Center for Systems and Control
Delft University of Technology
The Netherlands

November 13, 2007

TU Delft
Delft University of Technology

Outline

TU Delft
Delft University of Technology

Introduction

Motivation

- General interest in bounding model uncertainty, motivated e.g. by "identification for control"
- Simple experiments that acquire sufficient process knowledge
- "Simple experiment" implies also "short"
- Almost all results in PE uncertainty bounding are asymptotic in N (at least), require $\mathcal{S} \in \mathcal{M}$, and the exact covariance matrix P_0 to be known
- Estimator properties versus single-experiment uncertainty bounding
- Finite time results, as well as handling of $\mathcal{S} \notin \mathcal{M}$ only available for FIR-like model structures
- Some alternative results available from Campi and Weyer (2002,2006).

Main question

Explore the possiblities for (parametric) model uncertainty bounding beyond the classical situation that relies heavily on asymptotics and consistent parameter estimates

PE uncertainty bounds and hypothesis testing

The set-up

Data-generating system:

$$y(t) = G_0(q)u(t) + H_0(q)e(t)$$

with e a Gaussian white noise (for the moment).
One-step ahead prediction error:

$$\varepsilon(t,\theta) = H^{-1}(q,\theta)\left[y(t) - G(q,\theta)u(t)\right]$$

and the parameter estimator:

$$\hat{\boldsymbol{\theta}}_N = \arg\min_{\theta} V_N(\theta, Z^N)$$

$$\text{with } V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t,\theta),$$

Note:
boldface $\hat{\boldsymbol{\theta}}_N$ denotes random variable, as opposed to single realization $\hat{\theta}_N$.

Reasoning to arrive at uncertainty bounds

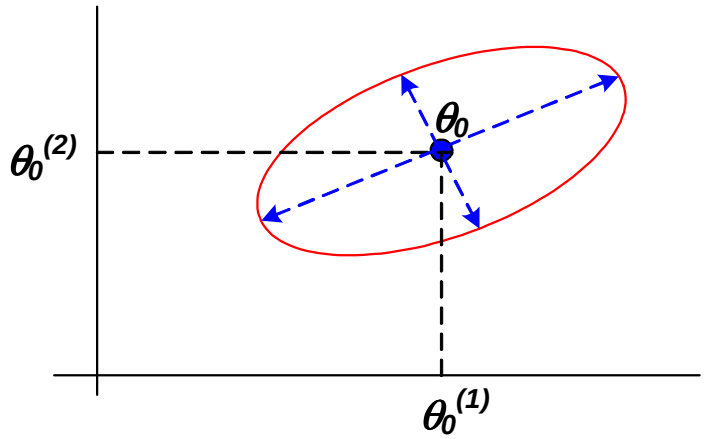
Start with asymptotic distribution:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_0) \underset{N\to\infty}{\to} \mathcal{N}(0, P_0)$$

and in quadratic form:

$$N(\hat{\boldsymbol{\theta}}_N - \theta_0)^T P_0^{-1}(\hat{\boldsymbol{\theta}}_N - \theta_0) \underset{N\to\infty}{\to} \chi^2_d$$

Then $\hat{\boldsymbol{\theta}}_N \in \mathcal{D}(\alpha, \theta_0)$ w.p. α



with

$$\mathcal{D}(\alpha, \theta_0) := \left\{\theta \mid N(\theta - \theta_0)^T P_0^{-1}(\theta - \theta_0) \le \chi^2_{d,\alpha}\right\}$$

However; we want to say something about θ_0!

Since for every realization $\hat{\theta}_N$:

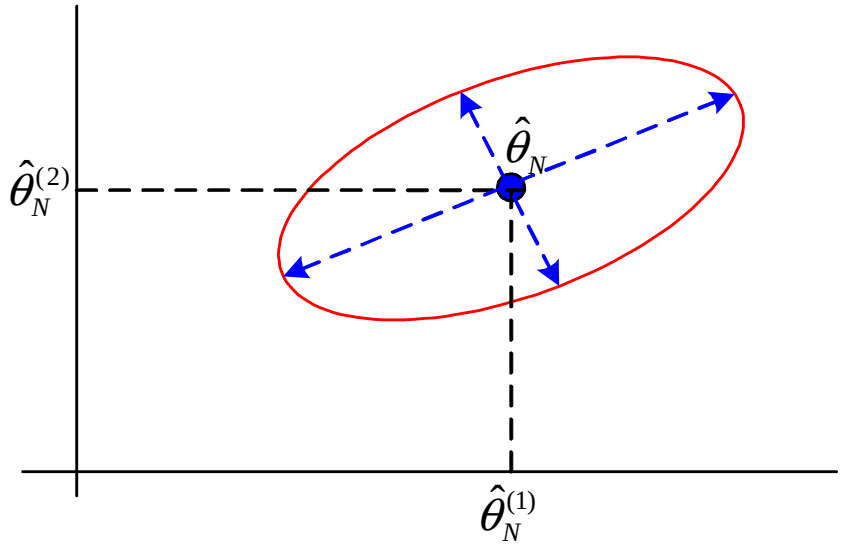
$$\hat{\theta}_N \in \mathcal{D}(\alpha, \theta_0) \Leftrightarrow \theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$$

it follows that

$$\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N) \text{ with probability } \alpha,$$

with

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \left\{ \theta \mid N(\hat{\theta}_N - \theta)^T P_0^{-1} (\hat{\theta}_N - \theta) \leq \chi^2_{d,\alpha} \right\}.$$



Note that this is an uncertainty set that varies with every realization $\hat{\theta}_N$ of $\hat{\boldsymbol{\theta}}_N$.

Interpretation in terms of hypothesis test

Test the null hypothesis H_0: $\theta = \theta_0$
against the alternative hypothesis H_1: $\theta \neq \theta_0$
on the basis of the test statistic:

$$N(\hat{\boldsymbol{\theta}}_N - \theta)^T P_0^{-1}(\hat{\boldsymbol{\theta}}_N - \theta)$$

that under H_0 is known to have a χ^2_d distribution.

Standard hypothesis test

For a given realization $\hat{\theta}_N$, select those values of θ that are within an α confidence bound of the test statistic.

This leads to exactly the same ellipsoid

Test statistic is directly based on the estimator pdf. Is this necessary?

Illustrative example

Consider the data generating system:

$$\mathbf{y} = \theta_0 \boldsymbol{x}_1 + \boldsymbol{x}_2$$

and one available measurement $\{y, x_1\}$ of $\mathbf{y}$ and $\boldsymbol{x}_1$.
Given: $\boldsymbol{x}_1, \boldsymbol{x}_2$ are random Gaussian numbers, with an unknown correlation, and with $\boldsymbol{x}_2 \in \mathcal{N}(0, 2)$.
We consider the following estimator of θ_0:

$$\hat{\boldsymbol{\theta}} = \frac{\mathbf{y}}{\boldsymbol{x}_1}.$$

Then

$$\hat{\boldsymbol{\theta}} = \frac{\mathbf{y}}{\boldsymbol{x}_1} = \theta_0 + \frac{\boldsymbol{x}_2}{\boldsymbol{x}_1}.$$

Analysis of pdf of estimator is cumbersome because of quotient of rv's

Alternative

Write:

$$\boldsymbol{x}_1(\hat{\boldsymbol{\theta}} - \theta_0) = \boldsymbol{x}_2 \in \mathcal{N}(0, 2)$$

so that

$$(\hat{\boldsymbol{\theta}} - \theta_0)\frac{\boldsymbol{x}_1^2}{2}(\hat{\boldsymbol{\theta}} - \theta_0) \in \chi_1^2.$$

Using the related test statistic

$$(\hat{\boldsymbol{\theta}} - \theta)\frac{\boldsymbol{x}_1^2}{2}(\hat{\boldsymbol{\theta}} - \theta)$$

then leads to the uncertainty bounding result:

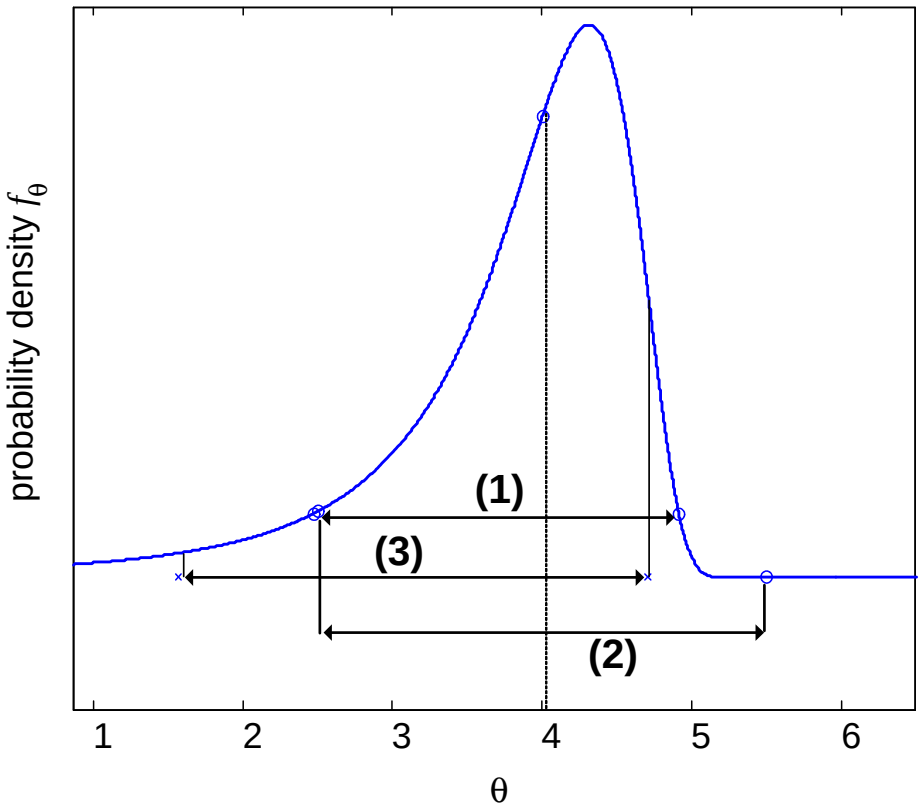
$$\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}) \text{ w.p. } \alpha$$

with

$$\mathcal{D}(\alpha, \hat{\theta}) = \left\{ \theta \,|\, \frac{x_1^2}{2}(\hat{\theta} - \theta)^2 \leq \chi_{1,\alpha}^2 \right\}.$$

Whereas analysis of the pdf of $\hat{\boldsymbol{\theta}}$ is hard, this result is exact and much more simple

Illustration for the particular correlation $\boldsymbol{x}_1 = 3 + \frac{0.5}{\boldsymbol{x}_2}$:



curve of pdf $f_{\hat{\boldsymbol{\theta}}}(\theta)$
(1): smallest 99% region
(2): symmetric 99% region around $\mathbb{E}\hat{\boldsymbol{\theta}}$
(3): realizations of $\hat{\theta}_N$ for which the new bound is correct

ARX models

Can this principle be applied to ARX models?:

$$G(q,\theta) = \frac{B(q^{-1},\theta)}{A(q^{-1},\theta)}, \qquad H(q,\theta) = \frac{1}{A(q^{-1},\theta)}$$

$$\hat{y}(t|t-1;\theta) = \varphi^T(t)\theta$$

with $\varphi^T(t) = [-y(t-1)\cdots -y(t-n_a)\; u(t)\cdots u(t-n_b+1)]$.
By denoting

$$\mathbf{\Phi} = \begin{pmatrix} \varphi^T(1) \\ \vdots \\ \varphi^T(N) \end{pmatrix} \text{ and } \mathbf{e} = [e(1)\cdots e(N)]^T$$

and using $\mathcal{S} \in \mathcal{M}$, it follows that

$$\hat{\theta}_N = \theta_0 + (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{e}.$$

Following classical PE analysis:

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \underset{N\to\infty}{\to} \mathcal{N}(0, P_0)$$

with

$$P_0 = (\lim_{N\to\infty} \mathbb{E}[\frac{1}{N}\mathbf{\Phi}^T\mathbf{\Phi}])^{-1} \cdot \sigma_e^2.$$

According to the hypothesis testing procedure discussed before:

Classical Result 1

On the basis of the test statistic

$$N(\hat{\boldsymbol{\theta}}_N - \theta)^T P_0^{-1}(\hat{\boldsymbol{\theta}}_N - \theta)$$

it follows that asymptotically in N, $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ w.p. α, with

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \{\theta \mid N(\hat{\theta}_N - \theta)^T P_0^{-1}(\hat{\theta}_N - \theta) \leq \chi^2_{d,\alpha}\}. \tag{1}$$

This result is built on the asymptotic normality of the term $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{e}$.

Note: P_0 will generally be unknown.

Applying the “trick” from the example:

$$\hat{\boldsymbol{\theta}}_N - \theta_0 = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{e}$$

can be written as:

$$(\boldsymbol{\Phi}^T \boldsymbol{\Phi})(\hat{\boldsymbol{\theta}}_N - \theta_0) = \boldsymbol{\Phi}^T \mathbf{e}$$

and with the Central Limit Theorem:

$$\frac{1}{\sqrt{N}} \boldsymbol{\Phi}^T \mathbf{e} \underset{N \to \infty}{\to} \mathcal{N}(0, Q_0) \quad \text{with } Q_0 = \lim_{N \to \infty} \mathbb{E} \frac{1}{N} \boldsymbol{\Phi}^T \boldsymbol{\Phi} \cdot \sigma_e^2.$$

We can now use the test statistic:

$$\frac{1}{N} (\hat{\boldsymbol{\theta}}_N - \theta)^T \boldsymbol{\Phi}^T \boldsymbol{\Phi} Q_0^{-1} \boldsymbol{\Phi}^T \boldsymbol{\Phi} (\hat{\boldsymbol{\theta}}_N - \theta)$$

known to have a χ^2_d distribution under $\theta = \theta_0$.

Alternative Result 2

On the basis of the test statistic

$$N(\hat{\theta}_N - \theta)^T \mathbf{\Phi}^T \mathbf{\Phi} Q_0^{-1} \mathbf{\Phi}^T \mathbf{\Phi} (\hat{\theta}_N - \theta)$$

it follows that asymptotically in N, $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ w.p. α, with

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \{\theta \mid N(\hat{\theta}_N - \theta)^T \Phi^T \Phi Q_0^{-1} \Phi^T \Phi (\hat{\theta}_N - \theta) \leq \chi^2_{d,\alpha}\}. \quad (2)$$

This result is built on the asymptotic normality of the term $\frac{1}{\sqrt{N}} \mathbf{\Phi}^T \mathbf{e}$.

Note: Q_0 will generally be unknown.

Can we take this one step further?

$$\mathbf{\Phi}^T \mathbf{\Phi}(\hat{\theta}_N - \theta_0) = \mathbf{\Phi}^T \mathbf{e}$$

Apply svd: $\mathbf{\Phi}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$
and write:

$$\mathbf{\Sigma}^{-1}\mathbf{U}^T \mathbf{\Phi}^T \mathbf{\Phi}(\hat{\theta}_N - \theta_0) = \mathbf{V}^T \mathbf{e}.$$

With the central limit theorem:

$$\mathbf{V}^T \mathbf{e} \underset{N\to\infty}{\to} \mathcal{N}(0, \sigma_e^2 I)$$

leading to the test statistic:

$$\frac{1}{\sigma_e^2}(\hat{\theta}_N - \theta)^T \mathbf{\Phi}^T \mathbf{\Phi}\mathbf{U}^T \mathbf{\Sigma}^{-2}\mathbf{U}^T \mathbf{\Phi}^T \mathbf{\Phi}(\hat{\theta}_N - \theta) =$$

$$= \frac{1}{\sigma_e^2}(\hat{\theta}_N - \theta)^T \mathbf{\Phi}^T \mathbf{\Phi}(\hat{\theta}_N - \theta)$$

known to be χ^2_d distributed under $\theta = \theta_0$.

Alternative Result 3

On the basis of the test statistic

$$\frac{1}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N - \theta)^T \boldsymbol{\Phi}^T \boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta)$$

it follows that asymptotically in N, $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ w.p. α, with

$$\mathcal{D}(\alpha, \hat{\theta}_N) := \{\theta \mid \frac{1}{\sigma_e^2}(\hat{\theta}_N - \theta)^T \Phi^T \Phi(\hat{\theta}_N - \theta) \leq \chi^2_{d,\alpha}\}. \tag{3}$$

This result is built on the asymptotic normality of the term $\mathbf{V}^T\mathbf{e}$.

Note: All terms in the uncertainty set are known, except σ_e^2.

Intermediate discussion

- If unknowns P_0 (Result 1)and Q_0 (Result 2) are replaced by sample-estimates, all 3 results become exactly the same!
- In Results 1 and 2, this replacement is a compromise, in Result 3 not.
- When σ_e^2 is to be estimated, the χ_d^2 distributions need to be replaced by $\mathcal{F}$ distributions.

Reflection on finite-time perspectives

The three results rely on asymptotic normality of:

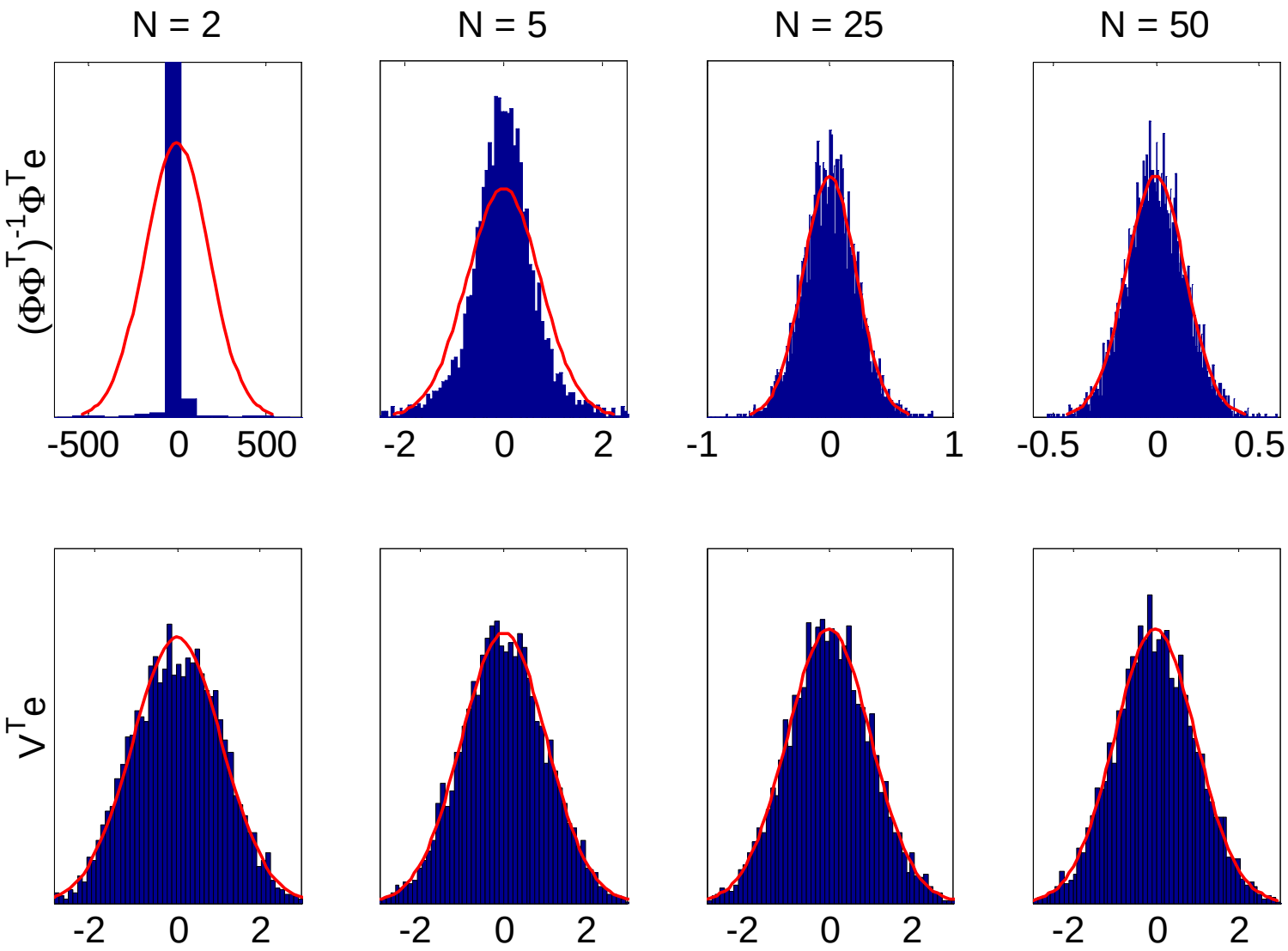
$$\begin{array}{ll} (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{e} & \text{Result 1} \\ \frac{1}{\sqrt{N}}\mathbf{\Phi}^T\mathbf{e} & \text{Result 2} \\ \mathbf{V}^T\mathbf{e} & \text{Result 3} \end{array}$$

Verification in simulation example of 1st order ARX model:

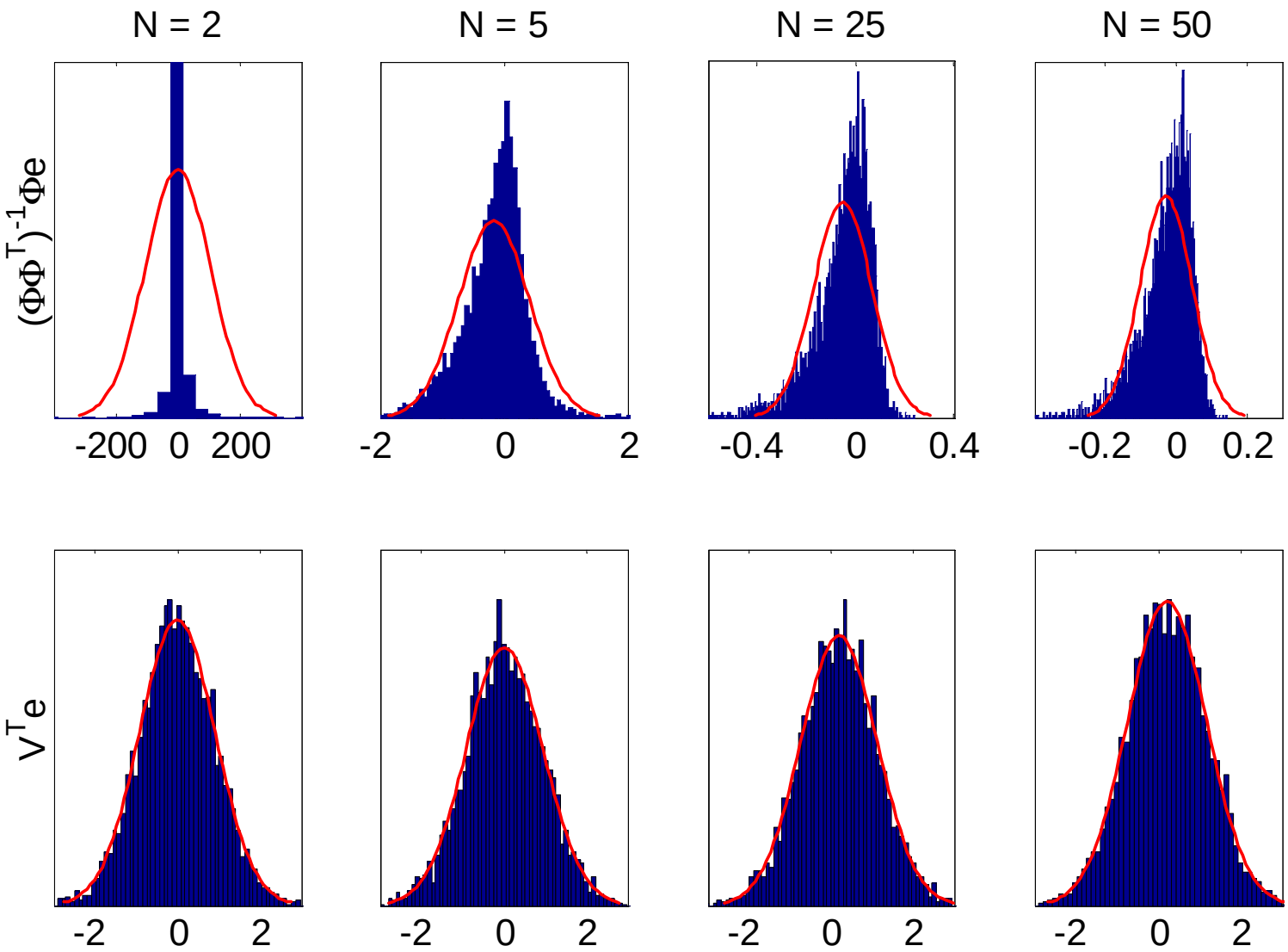
$$\varepsilon(t,\theta) = (1 + 0.5q^{-1})y(t) + 0.9u(t)$$

Input signal and noise signal are unit variance Gaussian white noise. Evaluate, for different values of N, 5000 Monte Carlo results of the relevant random variables above.

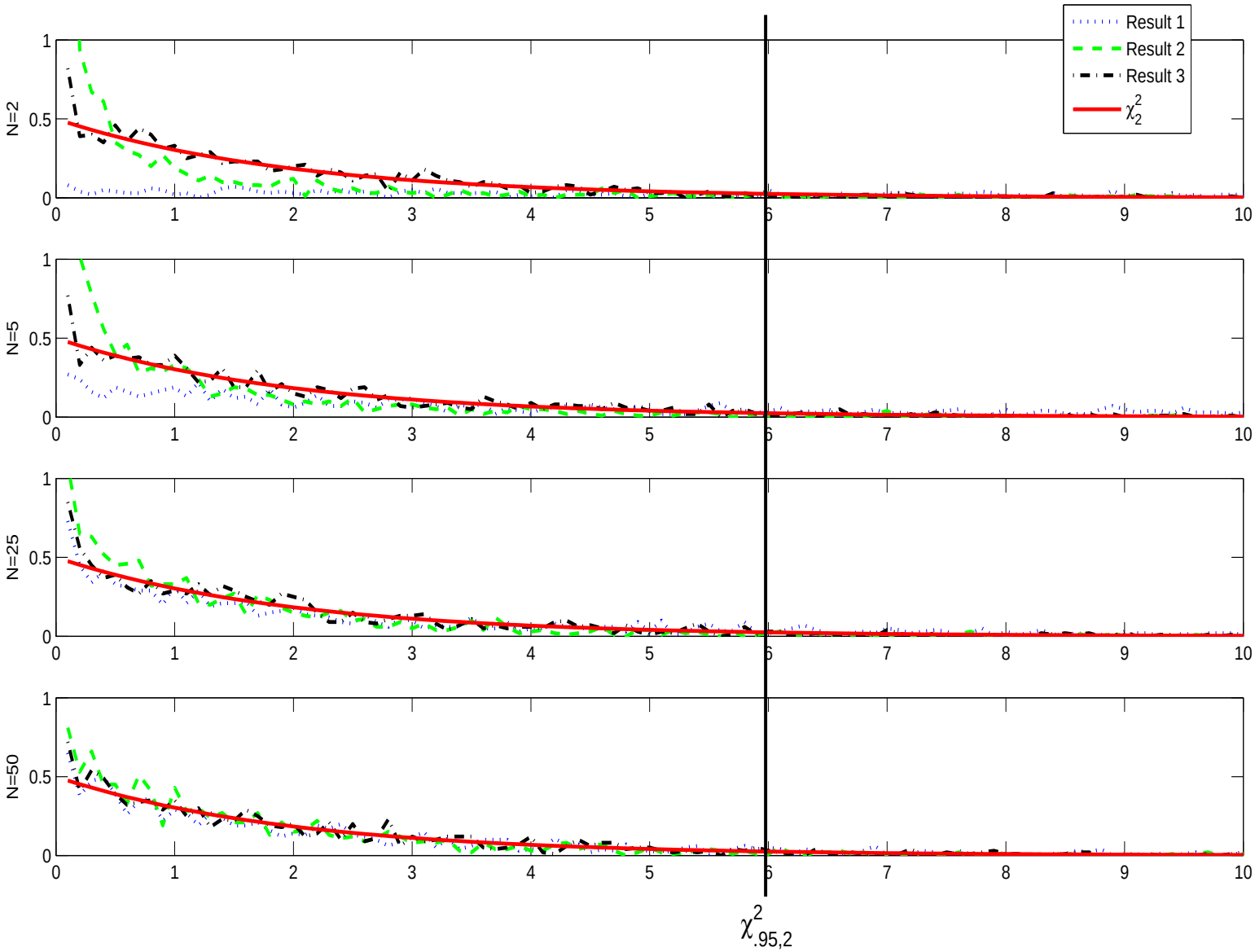
Numerator parameter



Denominator parameter



χ^2_2 test statistics:



Coverage rates for $\alpha = 0.95$:

	$N = 2$	$N = 5$	$N = 25$	$N = 50$
Result 1	0.1810	0.5530	0.8320	0.8760
Result 2	0.9690	0.9610	0.9550	0.9550
Result 3	0.9610	0.9500	0.9590	0.9550

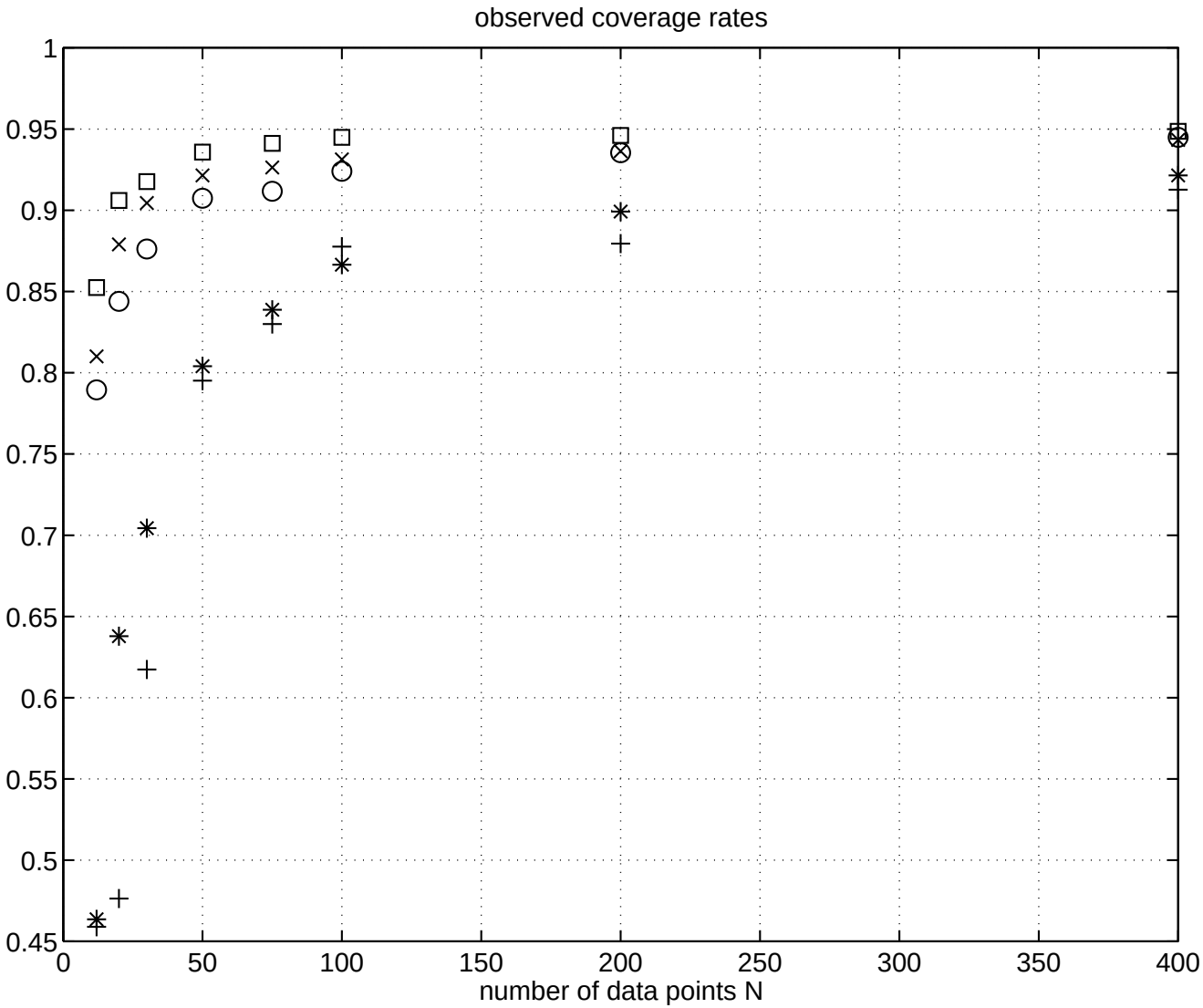
Further justification of this result

Lemma

If $\mathbf{V}$ and $\mathbf{e}$ are independent, $\mathbf{V}$ unitary and $e(t) \in \mathcal{N}(0, \sigma_e^2)$, then $\mathbf{V}^T \mathbf{e} \in \mathcal{N}(0, \sigma_e^2)$ for any value of N.

Note: Independence of $\mathbf{V}$ and $\mathbf{e}$ is formally not satisfied for ARX models (due to correlation between $\mathbf{\Phi}$ and $\mathbf{e}$).
However the deteriorating effect of this seems minimal.

Coverage rates for a second order ARX model ($\alpha = 0.95; M = 50,000$)



Result 1 (+); Result 3 (□)

Observation

Replacing (covariance) matrices by sample estimates in parameter uncertainty bounds is not at all a comprise, but rather well justified from a statistical and a practical point of view.

A likelihood perspective

When **e** is Gaussian (as assumed here) we can phrase the results in a likelihood framework, with the log-likelihood:

$$\log f_y(\theta; y^N) = -\frac{N}{2}\log(2\pi) - N\log\sigma_e - \frac{N}{2\sigma_e^2}V_N(\theta).$$

Define the generalized likelihood ratio $L_G(\theta)$ as (Kay, 1998):

$$L_G(\theta) = \frac{f_y(\theta; y^N)}{\sup_\theta f_y(\theta; y^N)} = \frac{f_y(\theta; y^N)}{f_y(\hat{\theta}_N; y^N)}$$

with $\hat{\theta}_N$ the maximum likelihood estimator.
Then (Kay, 1998):

$$-2\log L_G(\theta_0) \to \chi^2_d$$

As a result, $-2\log L_G(\theta)$ can be used as a test statistic for quantifying the model uncertainty.

Proposition

For the ARX models considered,

$$\begin{aligned} -2\log L_G(\theta_0) &= \frac{N}{\sigma_e^2}\left[V_N(\theta_0) - V_N(\hat{\boldsymbol{\theta}}_N)\right] \\ &= \frac{N}{\sigma_e^2}(\hat{\boldsymbol{\theta}}_N - \theta_0)^T \frac{1}{N}\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\hat{\boldsymbol{\theta}}_N - \theta_0) \end{aligned}$$

Observation

Uncertainty bounding Result 3 is equivalent to a generalized likelihood ratio test, and comes down to considering a level set of the identification cost function $V_N(\theta)$, i.e. $\theta_0 \in \mathcal{D}(\alpha, \hat{\theta}_N)$ w.p. α with

$$\mathcal{D}(\alpha, \hat{\theta}_N) = \{\theta \,|\, \frac{N}{\sigma_e^2}[V_N(\theta) - V_N(\hat{\theta}_N)] \leq \chi^2_{d,\alpha}\}.$$

Output error models

Can the same mechanisms be used for nonlinearly parametrized (Output error) models?

$$\varepsilon(t,\theta) = y(t) - \frac{B(q^{-1},\theta)}{F(q^{-1},\theta)}u(t)$$

Classical result:

$$\sqrt{N}(\hat{\boldsymbol{\theta}}_N - \theta_0) \to \mathcal{N}(0, P_{oe})$$

with

$$P_{oe} = \sigma_e^2[\lim_{N\to\infty} \frac{1}{N}\Psi(\theta_0)^T\Psi(\theta_0)]^{-1}$$

with $\Psi(\theta)$ a matrix with rows $\psi^T(t,\theta) = \frac{\partial}{\partial\theta}\hat{y}(t|t-1,\theta)$.

An ellipsoidal confidence bound for θ_0 is obtained on the basis of the test statistic

$$\frac{1}{N}(\hat{\boldsymbol{\theta}}_N - \theta)^T P_{oe}^{-1}(\hat{\boldsymbol{\theta}}_N - \theta)$$

Result relies on Taylor approximation: $(\hat{\boldsymbol{\theta}}_N - \theta_0) \approx -[\bar{V}_N''(\theta_0)]^{-1}[V_N'(\theta_0)]$

Alternatives

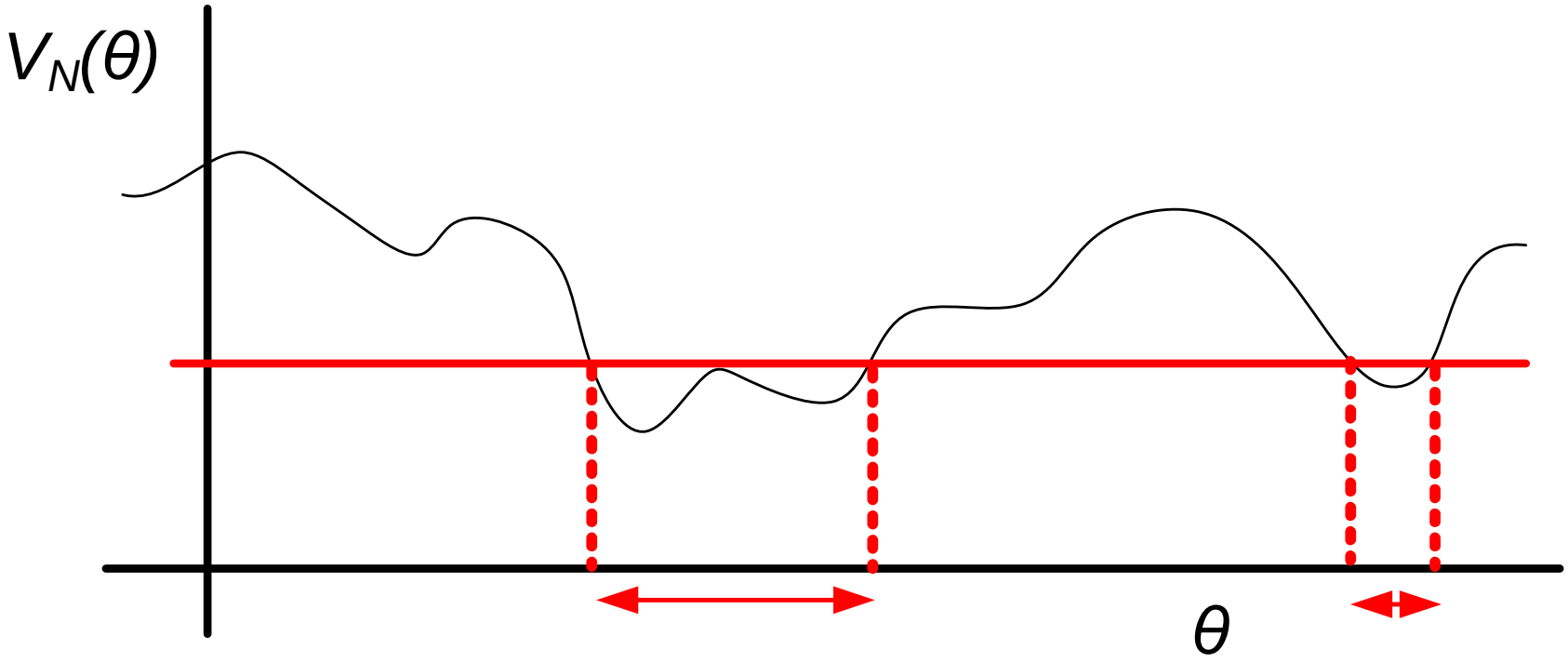
- Similar step as for ARX: $\boldsymbol{\Psi}^T(\hat{\boldsymbol{\theta}}_N)\boldsymbol{\Psi}(\hat{\boldsymbol{\theta}}_N)[\hat{\boldsymbol{\theta}}_N - \theta_0] = \boldsymbol{\Psi}^T(\hat{\boldsymbol{\theta}}_N)\mathbf{e}$
  (Taylor approxim)
  with $\boldsymbol{\Psi}$ containing predictor derivatives w.r.t. θ
- Using relations similar as in pseudo-linear regression (avoiding Taylor approxim):

$$\boldsymbol{\Psi}^T(\hat{\boldsymbol{\theta}}_N)\boldsymbol{\Phi}_{oe}(\hat{\boldsymbol{\theta}}_N)(\hat{\boldsymbol{\theta}}_N - \theta_0) = \boldsymbol{\Psi}^T(\hat{\boldsymbol{\theta}}_N)\mathbf{e}_F(\hat{\boldsymbol{\theta}}_N)$$

- Likelihood ratio approach:

$$-2\log L_G = \frac{N}{\sigma_e^2}[V_N(\theta) - V_N(\hat{\boldsymbol{\theta}}_N)] \to \chi_d^2$$

under $\theta = \theta_0$.
This leads to level sets of the identification cost function (no ellipsoid).



Uncertainty set on the basis of “level sets” can easily lead to disconnected regions,

but are very appealing from an engineering perspective.

Alternatives (contin.)

- The cost function derivative:

$$V_N'(\theta_0) = \boldsymbol{\Psi}^T(\theta_0)\mathbf{e} \in \mathcal{N}(0, J(\theta_0); \qquad J(\theta_0) = \frac{1}{\sigma_e^2}\boldsymbol{\Psi}^T(\theta_0)\boldsymbol{\Psi}(\theta_0))$$

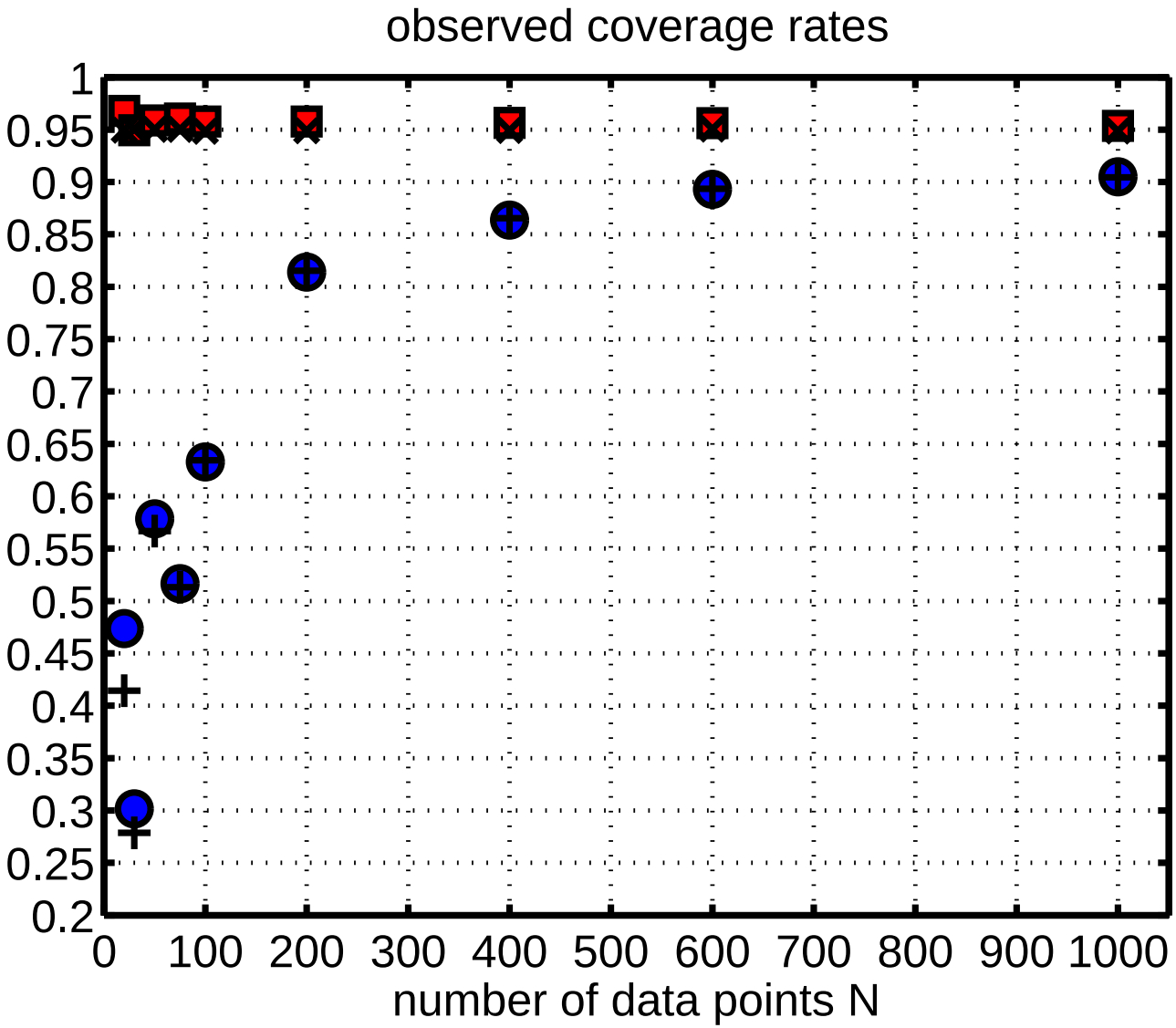
Fisher score or Rao test.
As in OE models $\boldsymbol{\Psi}(\theta_0)$ is not correlated with $\mathbf{e}$ the distribution holds for finite-time also (provided that $\mathbf{e}$ is Gaussian).

Related test statistic:

$$\frac{1}{\sigma_e^2}\varepsilon^T(\theta)\boldsymbol{\Psi}(\theta)[\boldsymbol{\Psi}^T(\theta)\boldsymbol{\Psi}(\theta)]^{-1}\boldsymbol{\Psi}^T(\theta)\varepsilon(\theta) \in \chi_d^2$$

This is no ellipsoid!

Coverage rates for a second order OE model ($\alpha = 0.95, M = 50,000$)



Classical (blue circles); LR (red □); Rao-test (black ×)

Extensions

Simple extension to IV

For IV estimators:

$$\hat{\theta}_N - \theta_0 = (Z^T \mathbf{\Phi})^{-1} Z^T \mathbf{e}$$

where the instrument matrix Z is uncorrelated with the noise $\mathbf{e}$. Choosing a test static based on

$$\frac{1}{\sqrt{N}}(Z^T \mathbf{\Phi})(\hat{\theta}_N - \theta_0) = \frac{1}{\sqrt{N}} Z^T \mathbf{e}$$

leads to an ellipsoidal uncertainty set that is correct for finite N if $\mathbf{e}$ is Gaussian distributed.

- Extension to BJ models and to handling unmodelled dynamics (bias term)
- Possibly useful for other identification methods (subspace ID), where it is hard to derive the estimator pdf

Conclusions

- We have explored a degree of freedom in formulating model uncertainty bounds
- Current PE theory and practice maybe somehow limited, and directed towards exact knowledge
- Bounds on the basis of data-based “covariance matrices” seem better than the theoretical ones
- Finite-time results are available for OE models (but not in ellipsoidal form)
- In nonlinearly parametrized model sets (OE), linearization seems to be a severe limitation

References

S.G. Douma and P.M.J. Van den Hof.
An alternative paradigm for probabilistic uncertainty bounding in prediction error identification. In *Proc. 44th IEEE Conf. Decision and Control and European Control Conf., CDC-ECC'05*, pages 4970–4975, Seville, Spain, December 2005.

S.G. Douma.
From Data to Performance - System Identification Uncertainty and Robust Control Design. PhD thesis, Delft Univ. Technology, Delft, The Netherlands, 2006.

S.G. Douma and P.M.J. Van den Hof.
Probabilistic model uncertainty bounding: an approach with finite-time perspectives. In *Prepr. 14th IFAC Symp. System Identification*, pages 1021–1026, Newcastle, NSW, Australia, March 2006.

S.G. Douma and P.M.J. Van den Hof.
Probabilistic uncertainty bounding in output error models with unmodelled dynamics. In *Proc. 2006 American Control Conf.*, pages 1677–1682.

A.J. den Dekker, X. Bombois, and P.M.J. Van den Hof.
Likelihood based uncertainty bounding in prediction error identification using ARX models: a simulation study. In *Proc. 2007 European Control Conf.*, 2879–2886.