Finite Sample System Identification

Exact, Distribution-Free Confidence Regions

Balázs Csanád Csáji

Institute for Computer Science and Control (SZTAKI)
Hungarian Academy of Sciences (MTA)

joint work with **Marco Campi** and **Erik Weyer**

Mathematical Modelling Seminar

October 13, 2015

Table of contents

Dynamical Systems

Mathematical models of dynamical systems are widely used in science and engineering, e.g., for prediction and control purposes.

Dynamical System (Markov)

$$x_t \triangleq f(x_{t-1}, u_t, w_t)$$

where

- t — time (discrete)
- x_t — output (state)
- u_t — input (external)
- w_t — noise (innovation)
- f — transition function

Point Estimation

Consider the parametric identification problem of the system

$$x_t \triangleq f_{\theta^*}(x_{t-1}, u_t, w_t),$$

parametrized with $\theta^* \in \Theta \subseteq \mathbb{R}^d$

Given: a finite sample, $\mathcal{Z}$, of outputs $\{x_t\}$ and inputs $\{u_t\}$

Point Estimate (Parametric)

$$\widehat{\theta}_{\mathcal{Z}} \triangleq \arg\min_{\theta \in \Theta} \mathcal{V}(\theta \mid \mathcal{Z})$$

where $\mathcal{V}$ is a criterion function.

Confidence Regions

In practice often some quality tag is needed to judge the estimate.

Safety, stability, or quality requirements? $\Rightarrow$ confidence regions

Confidence Region (Level μ)

$$\mathbb{P}\big(\theta^* \in \widehat{\Theta}_{\mathcal{Z},\mu}\big) \geq \mu$$

for some $\mu \in (0,1)$, where θ^* is the "true" parameter, $\widehat{\Theta}_{\mathcal{Z},\mu} \subseteq \Theta$.

Typically the level sets of the (scaled) limiting distribution is used.

Issues: only approximately correct for finite samples,
requires the existence of a (known) limiting distribution.

Main Objectives

- We aim at building confidence regions for dynamical systems.
- With non-asymptotic guarantees ("finite sample" method).
- Which are distribution-free: namely, do not make strong statistical assumption on the innovations of the process.
- They should be built around specific point estimates.
- The Sign-Perturbed Sums (SPS) method is presented.
- Its main assumption is that the noise terms are symmetric.
- Under which it can even provide exact confidence sets.
- Main models: linear regression, general linear dynamical systems, volatility models (not covered by this talk).

Linear Regression

Consider a standard linear regression problem:

Linear Regression

$$x_t \triangleq \varphi_t^{\mathrm{T}} \theta^* + w_t$$

where

x_t — output (for time $t = 1, \ldots, n$)

φ_t — regressor (deterministic, d dimensional)

w_t — noise (zero mean, uncorrelated)

σ^2 — variance of the noise (homoscedastic)

θ^* — true parameter (deterministic, d dimensional)

$\Phi_n = [\varphi_1, \ldots, \varphi_n]^{\mathrm{T}}$ — skinny and full rank

Least Squares

Given: a sample, $\mathcal{Z}$, of size n of outputs $\{x_t\}$ and regressors $\{\varphi_t\}$

A classical approach is the least squares criterion, namely

$$\mathcal{V}(\theta \mid \mathcal{Z}) \triangleq \sum_{t=1}^{n} (x_t - \varphi_t^{\mathrm{T}}\theta)^2.$$

The least squares estimate (LSE) can be found by solving

Normal Equation

$$\nabla_\theta \mathcal{V}(\hat{\theta}_n \mid \mathcal{Z}) = \sum_{t=1}^{n} \varphi_t (x_t - \varphi_t^{\mathrm{T}}\hat{\theta}_n) = 0$$

Asymptotic Normality

LSE can be explicitly formulated as

$$\hat{\theta}_n = \left(\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \right)^{-1} \left(\sum_{t=1}^{n} \varphi_t x_t \right).$$

LSE is asymptotically normal

Limiting Distribution

$$\sqrt{n}\,(\hat{\theta}_n - \theta^*) \xrightarrow{d} \mathcal{N}(0, \sigma^2 R^{-1}) \quad \text{as} \quad n \to \infty$$

where R is the limit of $R_n = \frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}$ as $n \to \infty$ (if exists).

Confidence Ellipsoids

The standard confidence region construction is then

Confidence Ellipsoid

$$\widetilde{\Theta}_{n,\mu} \triangleq \left\{ \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq \frac{\mu\, \hat{\sigma}_n^2}{n} \right\}$$

where $\mathbb{P}(\theta^* \in \widetilde{\Theta}_{n,\mu}) \approx F_{\chi^2(d)}(\mu)$, where $F_{\chi^2(d)}$ is the CDF of $\chi^2(d)$,

$$\hat{\sigma}_n^2 \triangleq \frac{1}{n-d} \sum_{t=1}^{n} (x_t - \varphi_t^{\mathrm{T}} \hat{\theta}_n)^2,$$

is an estimate of σ^2 based on the sample.

Reference and Sign-Perturbed Sums

Let us introduce a reference sum and $m-1$ sign-perturbed sums.

Reference Sum

$$S_0(\theta) \triangleq R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t (x_t - \varphi_t^{\mathrm{T}} \theta)$$

Sign-Perturbed Sums

$$S_i(\theta) \triangleq R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t \, \alpha_{i,t} (x_t - \varphi_t^{\mathrm{T}} \theta)$$

for $i = 1, \ldots, m-1$, where $\alpha_{i,t}$ $(t = 1, \ldots, n)$ are i.i.d. random signs, that is $\alpha_{i,t} = \pm 1$ with probability $1/2$ each (Rademacher).

Intuitive Idea: Distributional Invariance

Assume $\{w_t\}$ are independent and each w_t is symmetric about zero.

Observe that, if $\theta = \theta^*$, we have $(i = 1, \ldots, m-1)$

Distributional Invariance

$$S_0(\theta^*) = R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t w_t$$

$$S_i(\theta^*) = R_n^{-\frac{1}{2}} \sum_{t=1}^{n} \varphi_t \alpha_{i,t} w_t$$

Consider the ordering $\|S_{(0)}(\theta^*)\|^2 \prec \cdots \prec \|S_{(m-1)}(\theta^*)\|^2$

Note: relation "$\prec$" is the canonical "$<$" with random tie-breaking

$\boxed{\text{All orderings are equally probable!}}$ (they are conditionally i.i.d.)

Intuitive Idea: Reference Dominance

What if $\theta \neq \theta^*$?

In fact, the reference paraboloid $\|S_0(\theta)\|^2$ increases faster than $\{\|S_i(\theta)\|^2\}$, thus will eventually dominate the ordering.

Intuitively, for "large enough" $\|\tilde{\theta}\|$, where $\tilde{\theta} \triangleq \theta^* - \theta$

Eventual Dominance of the Reference Paraboloid

$$\left\| \sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \varphi_t w_t \right\|_{R_n^{-1}}^2 > \left\| \sum_{t=1}^{n} \pm \varphi_t \varphi_t^{\mathrm{T}} \tilde{\theta} + \sum_{t=1}^{n} \pm \varphi_t w_t \right\|_{R_n^{-1}}^2$$

with "high probability" (for simplicity $\pm$ is used instead of $\{\alpha_{i,t}\}$).

Non-Asymptotic Confidence Regions

The rank of $\|S_0(\theta)\|^2$ in the ordering of $\{\|S_i(\theta)\|^2\}$ w.r.t. $\prec$ is

$$\mathcal{R}(\theta) = 1 + \sum_{i=1}^{m-1} \mathbb{I}(\|S_i(\theta)\|^2 \prec \|S_0(\theta)\|^2),$$

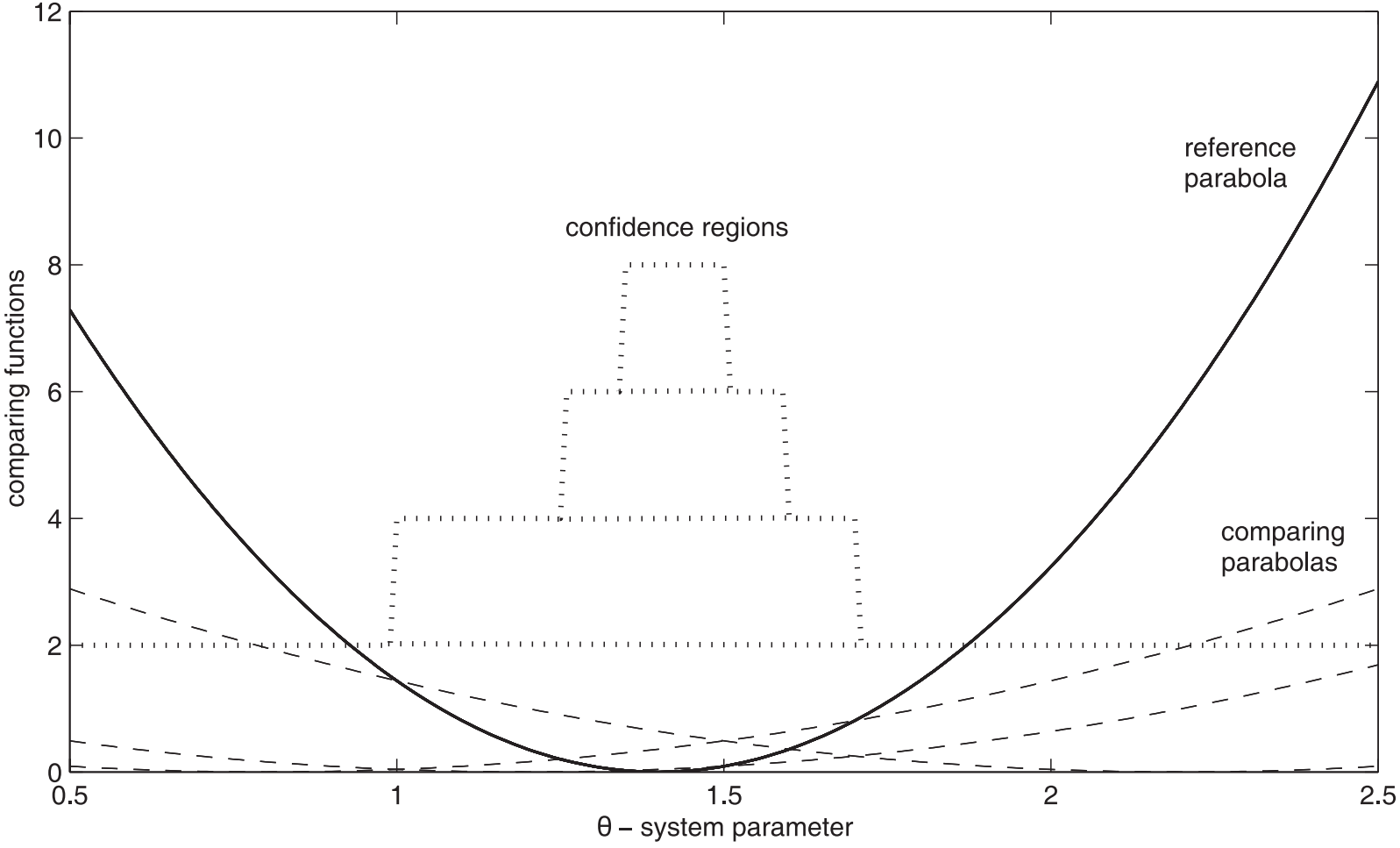
where $\mathbb{I}(\cdot)$ is an indicator function.

Sign-Perturbed Sums (SPS) Confidence Regions

$$\widehat{\Theta}_n \triangleq \left\{ \theta \in \mathbb{R}^d \, : \, \mathcal{R}(\,\theta\,) \leq m - q \right\}$$

where $m > q > 0$ are user-chosen integers (design parameters).

Simple Illustration, Constant in Noise, $n = 3$, $m = 4$



Exact Confidence

(A1) $\{w_t\}$ is a sequence of independent random variables.
Each w_t has a symmetric probability distribution about zero.

(A2) The outer product of regressors is invertible, $\det(R_n) \neq 0$.

Exact Confidence of SPS

$$\mathbb{P}\big(\theta^* \in \widehat{\Theta}_n\big) = 1 - \frac{q}{m}$$

for finite samples. Parameters m and q are under our control.

Note that $\|S_0(\hat{\theta}_n)\|^2 = 0$, thus $\hat{\theta}_n \in \widehat{\Theta}_n$, assuming it is non-empty.

Star Convexity

Set $\mathcal{X} \subseteq \mathbb{R}^d$ is star convex if there is a star center $c \in \mathbb{R}^d$ with

$$\forall x \in \mathcal{X}, \forall \beta \in [0,1] \,:\, \beta\, x + (1-\beta)\, c \in \mathcal{X}.$$

Star Convexity of SPS

$\widehat{\Theta}_n$ is star convex with the LSE, $\hat{\theta}_n$, as a star center

Hint $\widehat{\Theta}_n$ is the union and intersection of ellipsoids containing LSE.

Strong Consistency

(A1) independence, symmetricity: $\{w_t\}$ are independent, symmetric

(A2) invertibility: $R_n \triangleq \frac{1}{n}\sum_{t=1}^{n} \varphi_t \varphi_t^{\mathrm{T}}$ is invertible

(A3) regressor growth rate: $\sum_{t=1}^{\infty} \|\varphi_t\|^4 / t^2 < \infty$

(A4) noise moment growth rate: $\sum_{t=1}^{\infty} \left(\mathbb{E}[w_t^2]\right)^2 / t^2 < \infty$

(A5) Cesàro summability: $\lim_{n\to\infty} R_n = R$, which is positive definite

Strong Consistency of SPS

$$\forall\, \varepsilon > 0 : \exists\, (\text{a.s.})\, N : \forall\, n > N : \widehat{\Theta}_n \subseteq B_\varepsilon(\theta^*)$$

where $B_\varepsilon(\theta^*) \triangleq \{\, \theta \in \mathbb{R}^d : \|\theta - \theta^*\| \leq \varepsilon \,\}$ is a norm ball.

Ellipsoidal Outer Approximation

The reference paraboloid can be rewritten as

$$\|S_0(\theta)\|^2 = (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n).$$

From which an alternative description of the confidence region is

$$\widehat{\Theta}_n \subseteq \left\{ \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq r(\theta) \right\},$$

where $r(\theta)$ is the qth largest value of $\{\|S_i(\theta)\|^2\}_{i \neq 0}$.

Ellipsoidal Outer Approximation

$$\widehat{\Theta}_n \subseteq \left\{ \theta \in \mathbb{R}^d \,:\, (\theta - \hat{\theta}_n)^{\mathrm{T}} R_n (\theta - \hat{\theta}_n) \leq r^* \right\}$$

The question is of course how to find such an r^* efficiently.

Quadratically Constrained Quadratic Program

$\max\{\|S_i(\theta)\|^2 : \|S_0(\theta)\|^2 \leq \|S_i(\theta)\|^2\}$ can be obtained by

$$\begin{aligned} \text{maximize} \quad & \|z\|^2 \\ \text{subject to} \quad & z^{\mathrm{T}} A_i z + 2 z^{\mathrm{T}} b_i + c_i \leq 0 \end{aligned}$$

$$\begin{aligned} A_i &\triangleq I - R_n^{-\frac{1}{2}} Q_i R_n^{-1} Q_i R_n^{-\frac{1}{2}\mathrm{T}}, \\ b_i &\triangleq R_n^{-\frac{1}{2}} Q_i R_n^{-1} (\psi_i - Q_i \hat{\theta}_n), \\ c_i &\triangleq -\psi_i^{\mathrm{T}} R_n^{-1} \psi_i + 2\hat{\theta}_n^{\mathrm{T}} Q_i R_n^{-1} \psi_i - \hat{\theta}_n^{\mathrm{T}} Q_i R_n^{-1} Q_i \hat{\theta}_n. \end{aligned}$$

$$Q_i \triangleq \sum_{t=1}^{n} \alpha_{i,t} \varphi_t \varphi_t^{\mathrm{T}}, \qquad \psi_i \triangleq \sum_{t=1}^{n} \alpha_{i,t} \varphi_t x_t.$$

Semi-Definite Program

Problem: the previous QCQP is not convex.

Fortunately, strong duality holds and its dual can be written as:

Dual Problem

$$\begin{aligned} \text{minimize} \quad & \gamma \\ \text{subject to} \quad & \lambda \geq 0 \\ & \begin{bmatrix} -I + \lambda A_i & \lambda b_i \\ \lambda b_i^{\mathrm{T}} & \lambda c_i + \gamma \end{bmatrix} \succeq 0 \end{aligned}$$

where "$\succeq 0$" denotes that a matrix is positive semidefinite.

Radius r^* can then be found by solving $m-1$ such convex problems, obtaining $\{\gamma_i^*\}$, and defining r^* the qth largest one.

Simulation Experiment

Finite Impulse Response (FIR) System (2nd order)

$$x_t = 0.7\, u_{t-1} + 0.3\, u_{t-2} + w_t$$

where $\{w_t\}$ are i.i.d. zero mean Laplacian, with variance 0.1.

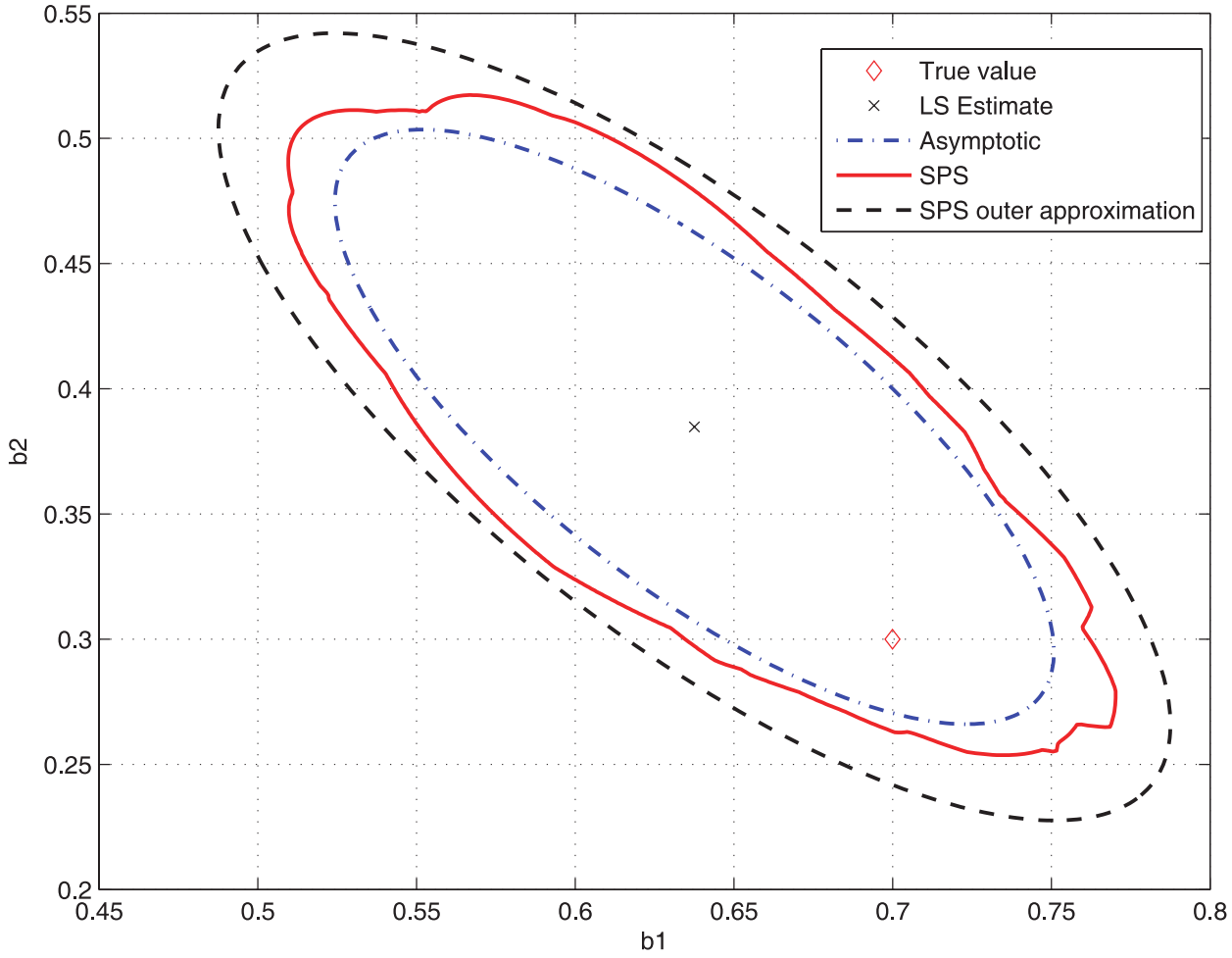
The input signal $\{u_t\}$ is given by the autoregression

$$u_t = 0.75\, u_{t-1} + v_t,$$

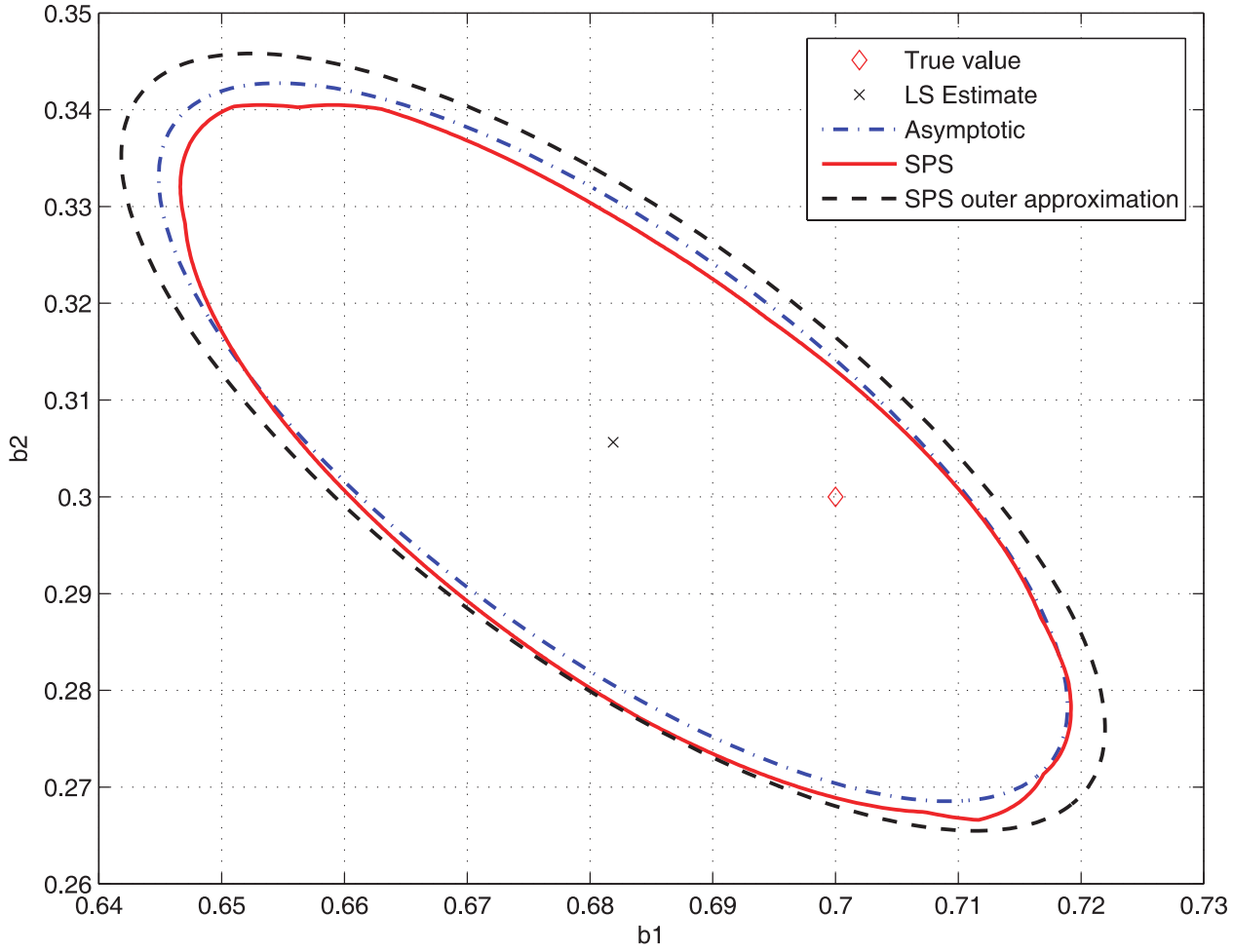
where $\{v_t\}$ is a sequence of i.i.d. standard normal variables.

Confidence regions (level 95 %) of SPS, its outer approximation and the standard asymptotic ellipsoids are compared.

95% Confidence Regions, $n = 25$, $m = 100$, $q = 5$



95% Confidence Regions, $n = 400$, $m = 100$, $q = 5$



Linear Dynamical Systems

Consider systems written using (rational) transfer functions:

General Linear Systems

$$x_t \triangleq G(z^{-1};\theta^*)\, u_t + H(z^{-1};\theta^*)\, w_t$$

(A1) The true system is in the model class, the orders are known.

(A2) The transfer function $H(z^{-1};\theta^*)$ has a (stable) inverse, as well as $G(0;\theta^*) = 0$ and $H(0;\theta^*) = 1$.

(A3) Noises $\{w_t\}$ are independent and symmetrically distributed.

(A4) Inputs $\{u_t\}$ are observed and independent of $\{w_t\}$.

(A5) Initialization: for all $t \leq 0$, we have $x_t = w_t = u_t = 0$.

Prediction Error Estimate

Prediction Error or Residual (for parameter θ)

$$\widehat{\varepsilon}_t(\theta) \triangleq H^{-1}(z^{-1};\theta)\left(x_t - G(z^{-1};\theta)\, u_t\right)$$

Note that $\widehat{\varepsilon}_t(\theta^*) = w_t$, hence, it is accurate for $\theta = \theta^*$.

Prediction Error Estimate (for model class Θ)

$$\hat{\theta}_{\mathrm{PEM}} \triangleq \arg\min_{\theta\in\Theta} \sum_{t=1}^{n} \widehat{\varepsilon}_t^2(\theta)$$

In general, there is no closed-form solution for PEM.

Prediction Error Equation

The PEM estimate can be found, e.g., by using the equation

PEM Equation

$$\nabla_\theta \mathcal{V}(\hat{\theta}_{\mathrm{PEM}} \mid \mathcal{Z}) = \sum_{t=1}^{n} \psi_t(\hat{\theta}_{\mathrm{PEM}})\, \widehat{\varepsilon}_t(\hat{\theta}_{\mathrm{PEM}}) = 0$$

where $\psi_t(\theta)$ is the negative gradient of the prediction error,

$$\psi_t(\theta) \triangleq -\nabla_\theta \widehat{\varepsilon}_t(\theta).$$

These gradients can be directly calculated in terms of the defining polynomials of the rational transfer functions G and H.

Perturbed Samples

Perturbed Output Trajectories

$$\bar{x}_t(\theta, \alpha_i) \triangleq G(z^{-1}; \theta)\, u_t + H(z^{-1}; \theta)\, (\alpha_{i,t}\, \widehat{\varepsilon}_t(\theta))$$

where $\{\alpha_{i,t}\}$ are random signs, as previously.

Recall that $\psi_t(\theta)$ is a linear filtered version of $\{x_t\}$ and $\{u_t\}$,

$$\psi_t(\theta) = W_0(z^{-1}; \theta)\, x_t + W_1(z^{-1}; \theta)\, u_t,$$

where W_0 and W_1 are vector-valued, and $\psi_t(\theta) \in \mathbb{R}^d$.

Perturbed (Negative) Gradients

$$\bar{\psi}_t(\theta, \alpha_i) \triangleq W_0(z^{-1}; \theta)\, \bar{x}_t(\theta, \alpha_i) + W_1(z^{-1}; \theta)\, u_t$$

Sign-Perturbed Sums for General Linear Systems

Reference and Sign-Perturbed Sums

$$S_0(\theta) \triangleq \Psi_n^{-\frac{1}{2}}(\theta) \sum_{t=1}^{n} \psi_t(\theta)\, \widehat{\varepsilon}_t(\theta)$$

$$S_i(\theta) \triangleq \bar{\Psi}_n^{-\frac{1}{2}}(\theta, \alpha_i) \sum_{t=1}^{n} \bar{\psi}_t(\theta, \alpha_i)\, \alpha_{i,t}\, \widehat{\varepsilon}_t(\theta)$$

Reference and Sign-Perturbed Covariances

$$\Psi_n(\theta) \triangleq \frac{1}{n} \sum_{t=1}^{n} \psi_t(\theta) \psi_t^{\mathrm{T}}(\theta)$$

$$\bar{\Psi}_n(\theta, \alpha_i) \triangleq \frac{1}{n} \sum_{t=1}^{n} \bar{\psi}_t(\theta, \alpha_i) \bar{\psi}_t^{\mathrm{T}}(\theta, \alpha_i)$$

Non-Asymptotic Confidence Regions

$\mathcal{R}(\theta)$ is again the rank of $\|S_0(\theta)\|^2$ among $\{\|S_i(\theta)\|^2\}$ w.r.t. $\prec$

SPS Confidence Regions for General Linear Systems

$$\widehat{\Theta}_n \triangleq \left\{ \theta \in \mathbb{R}^d \,:\, \mathcal{R}(\theta) \leq m - q \right\}$$

where $m > q > 0$ are user-chosen (integer) parameters.

We have $S_0(\hat{\theta}_{\mathrm{PEM}}) = 0$, thus, $\hat{\theta}_{\mathrm{PEM}} \in \widehat{\Theta}_n$, if it is non-empty.

Exact Confidence of SPS for General Linear Systems

$$\mathbb{P}\left(\theta^* \in \widehat{\Theta}_n\right) = 1 - \frac{q}{m}$$

Simulation Experiment

Autoregressive Moving Average: ARMA(1,1)

$$x_t + a^* x_{t-1} = w_t + c^* w_{t-1}$$

where $\theta^* = (a^*, c^*)$ and $\{w_t\}$ are i.i.d. standard normal.

The inverse filter of $C(z^{-1};\theta)\, w_t = w_t + c\, w_{t-1}$ is

$$C^{-1}(z^{-1};\theta) = \sum_{k=0}^{\infty} (-1)^k\, c^k\, z^{-k}$$

Can be used to define the prediction errors for $\theta = (a, c)$

$$\widehat{\varepsilon}_t(\theta) = C^{-1}(z^{-1};\theta)\,(x_t + a\, x_{t-1}).$$

Simulation Experiment

Perturbed Output Trajectories

$$\bar{x}_t(\theta, \alpha_i) = -a\,\bar{x}_{t-1}(\theta, \alpha_i) + \alpha_{i,t}\,\widehat{\varepsilon}_t(\theta) + c\,\alpha_{i,t-1}\widehat{\varepsilon}_{t-1}(\theta)$$

for $1 \leq i \leq m$ and $1 \leq t \leq n$, where $\{\alpha_{i,t}\}$ are random signs.

Perturbed (Negative) Gradients

$$\bar{\psi}_t(\theta, \alpha_i) = \begin{bmatrix} -C^{-1}(z^{-1};\theta)\,\bar{x}_t(\theta, \alpha_i) \\ C^{-1}(z^{-1};\theta)\,\alpha_{i,t}\,\widehat{\varepsilon}_t(\theta) \end{bmatrix}$$

which can be used to define the sign-perturbed sums.

99% Confidence Regions, $n = 500$, $m = 100$, $q = 1$

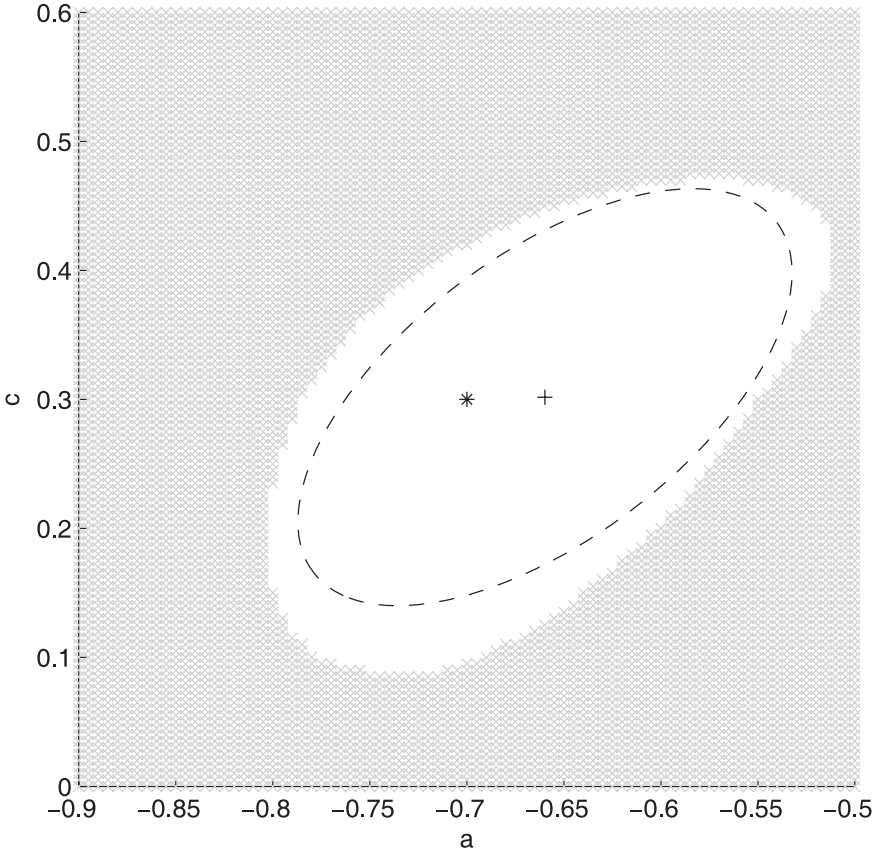
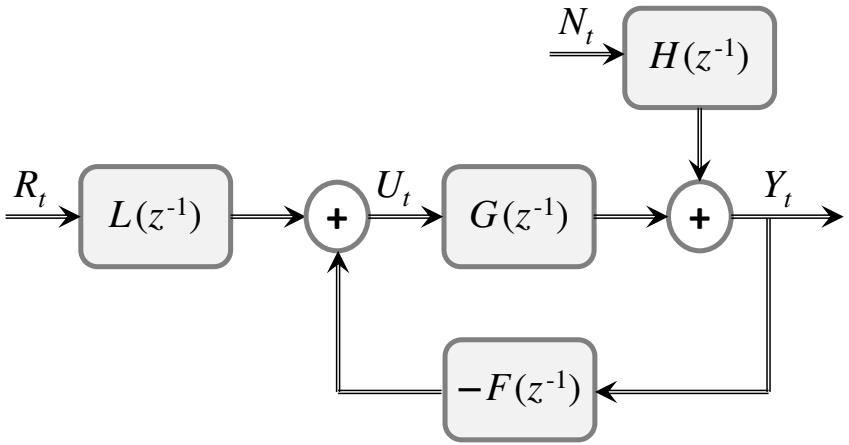


Figure: "$\times$": SPS (compl.); "$*$": θ^*; "$+$": PEM; "--": asymp. ellipsoid

Extensions

- SPS can be modified for instrumental variable estimates, which simplifies to construction, but needs instrumentals.
- SPS can be extended to closed-loop systems, to the direct, indirect and joint input-output approach of PEM.



- The core ideas of SPS can be transferred to volatility models.

Summary

- A new finite sample estimation method (SPS) was presented.
- It builds confidence regions around the least squares estimate.
- Only mild statistical assumptions are needed, e.g., symmetry.
- Not needed: stationarity, moments, particular distributions.
- For (rational) probabilities, exact confidence sets can be built.
- SPS is strongly consistent, i.e., the confidence regions almost surely shrink around the true parameter (for lin.reg.).
- SPS is star convex with the LSE as a center (for lin.reg.).
- Efficient ellipsoidal outer approximation exists (for lin.reg.).
- The algorithm was also presented for general linear systems.
- Its fundamental ideas work for other kinds of systems, as well.

Thank you for your attention.