Instrumental variables system identification with L^p consistency

Simon Kuang SLKU@UCDAVIS.EDU

Xinfan Lin LXFLIN@UCDAVIS.EDU

University of California, Davis

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

Instrumental variables (IV) eliminate the bias that afflicts least-squares identification of dynamical systems through noisy data, yet traditionally relies on external instruments that are seldom available for nonlinear time series data. We propose an IV estimator that synthesizes instruments from the data. We establish finite-sample L^p consistency for *all* $p \geq 1$ in both discrete- and continuous-time models, recovering a nonparametric $\sqrt{n}$-convergence rate. On a forced Lorenz system our estimator reduces parameter bias by 200x (continuous-time) and 500x (discrete-time) relative to least squares and reduces RMSE by up to tenfold. Because the method only assumes that the model is linear in the unknown parameters, it is broadly applicable to modern sparsity-promoting dynamics learning models.

Keywords: system identification,finite-sample,instrumental variables

1. Introduction

Filtering, smoothing, prediction, and control benefit from accurate models of a plant's dynamics. Our present data-rich world facilitates and demands dynamics models that can learn efficiently from long time series. Just as in large-scale machine learning, recurrent neural network architectures are being succeeded by autoregressive architectures such as transformers (Vaswani et al., 2023), it is now preferable to specify an autoregressive model that maps past outputs to future outputs, rather than a state-space model that advances an unobserved state. We further make the simplification (Brunton et al., 2016; Mezić, 2021) of assuming a parametric form in which the prediction is nonlinear in the inputs (past outputs, exogenous inputs, etc.) but linear in the parameter vector. Thus the parameter vector may be estimated (following a feature selection process) by processing the data with a linear filter, applying a nonlinearity, and then running least squares estimation. This is common practice in engineering (Kutz et al., 2016; Haller, 2025).

But least-squares estimation is problematic when viewed as point estimation of the parameter vector. Assuming that the data is contaminated by realistic levels of noise (such as from quantization (Maity and Goswami, 2025)), least-squares regressions of (noisy) future outputs on (noisy) past outputs are biased (Kutz et al., 2016, Chapter 8) (Yi et al., 2021; Söderström, 2018). For this reason, an autoregressive linear system identification method introduced in the 1970s has since been supplanted by instrumental variables (IV) and other bias-avoiding methods (Garnier and Wang, 2008, Chapter 1) (González, 2022).

The current wave of autoregressive system identification methods recognizes that output measurement noise degrades parameter estimation with a bias that persists even at the low noise levels achievable by judicious data prefiltering (Brunton et al., 2016; Wentz and Doostan, 2023; Hsin et al., 2024). Therefore, it is not enough to filter away the noise and then pretend it is not there. Two bias

mitigation methods are bias correction, which depends strongly on accurate estimation of the noise variance, and instrumental variables, which does not (Garnier and Wang, 2008).

We use instrumental variables to counteract the bias due to noise in the right-hand side of the model equation. In IV problems in econometrics and engineering, an instrumental variable is provided along with the data or generated using model predictions (Davidson and MacKinnon, 2004; González, 2022), and one usually settles for asymptotic normality (Pan et al., 2020) because the IV estimator has no finite expectation (Davidson and MacKinnon, 2004, §8.4). Zeiringer et al. (2026) use the model predictions to synthesize instruments for a linear-in-the-parameters nonlinear system, but it is hard to get statistical guarantees on the resulting estimator. We show that under reasonable sampling assumptions, the instruments can be synthesized from the same data as the regressors. Whereas the vanilla IV estimator's heavy tails result from inverting a matrix whose probability mass may concentrate near singularity, our estimator imposes minor regularizations, leading to a finite expectation and consistency in L^p for all $p \geq 1$.

Contributions

We define an instrumental variables estimator by applying local polynomial regression (De Brabanter et al., 2013) in a new way at the data filtering step. We analyze the benefits of singular value truncation, a form of ridge regularization, and instrumental variable truncation, which is a way to convert a subexponential tail to a subgaussian tail. The final consistency result, stated in Theorem 10, requires delicate tail bounds originally developed for a biased estimator in Kuang and Lin (2024).

Outline

We state the problem (discrete and continuous cases) in §3. We construct the estimator in §4 and state the theoretical principles behind its design. We state the main theoretical result in Theorem 10. We apply it to the Lorenz system (discrete and continuous cases) in §5.

2. Notation and conventions

We write $x \vee y$ for $\max(x, y)$ and $x \wedge y$ for $\min(x, y)$. We write $C^p(A, B)$ for the space of p-times continuously differentiable functions between Euclidean spaces A and B.

Unless otherwise specified, the notation $\|A\|$ refers to the operator 2-norm of A or the 2-norm of a vector. A subscript denotes a stochastic L^q norm: $\|x\|_q = \left(\mathbb{E}\,\|x\|^q\right)^{1/q}$.

The subgaussian norm of a random vector or matrix X is

$$\|X\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\exp\left(\frac{\|X\|^2}{t^2}\right) \leq 2\right\}.$$

The centered subgaussian norm of a random vector or matrix X is $\|X - \mathbb{E}\,X\|_{\psi_2}$.

The variable C denotes a constant that may depend on the dimension of the problem (d_y and d_ϕ), and its value may change from line to line. It never depends on the key statistical variables n, N, or h.[1]

1. Eliding dimensionality constants into C effectively commits our analysis to the "classical" low-dimensional regime, in which the design and parameter matrices are generic, full rank, etc. We reserve for future work the high-dimensional regime where intrinsic dimension may be far less than the ambient dimension.

3. Problem statement

We adopt the Output Error model of a deterministic system with a single unknown θ_0 (Ljung, 2010, p. 31). The data $\{z_i\}_{i=1}^n$ comprises measurements of the signal y at a sampling period $h > 0$, plus noise $\{\epsilon_i\}_{i=1}^n$:

$$z_i = y(ih) + \epsilon_i, \quad i \in [1 \dots n]. \tag{1}$$

The time horizon is $T = nh$. The signal $y \in C^p([0,T], \mathbb{R}^{\mathsf{d}_y})$ is taken as fixed (only the noise is random), and satisfies

$$\mathcal{H}y(t) = \theta_0^\intercal \phi(t, \mathcal{G}y(t)), \tag{2}$$

where n is the number of observations, h is the sampling period, $p > 0$ is the smoothness of y, $\phi \in C^1([0,T] \times \mathbb{R}^{\mathsf{d}_\mathcal{G}}, \mathbb{R}^{\mathsf{d}_\phi})$ is a static nonlinearity, $\theta_0 \in \mathbb{R}^{\mathsf{d}_\phi \times \mathsf{d}_\mathcal{H}}$, and $\mathcal{H} : C^p([0,T], \mathbb{R}^{\mathsf{d}_y}) \to C([0,T], \mathbb{R}^{\mathsf{d}_\mathcal{H}})$ and $\mathcal{G} : C^p([0,T], \mathbb{R}^{\mathsf{d}_y}) \to C([0,T], \mathbb{R}^{\mathsf{d}_\mathcal{G}})$ are *known* linear operators with compatible dimensions. Our paper concerns the cases: (a) *discrete-time*, where $\mathcal{H}$ is the left shift $\mathcal{H}x(t) = x(t+\tau)$ for some $\tau > 0$ and $\mathcal{G}$ is the identity; (b) first-order *continuous-time*, where $\mathcal{H} = \partial_t$ and $\mathcal{G}$ is the identity; and (c) continuous-time autoregression, where $\mathcal{H}$ is a high-order time derivative and $\mathcal{G}$ contains lower-order derivatives.

Example 1 *For example, a reduced-order model of vortex shedding behind a cylinder (Haller, 2025, Eq. 6.1),*

$$\begin{aligned}\dot\rho &= 0.0584\rho - 0.479\rho^3 + 1.27\rho^5 + 6.80\rho^7 - 58.9\rho^9 + 108\rho^{11}\\ \dot\gamma &= 0.553 + 0.441\rho^2 - 3.38\rho^4 + 55.5\rho^6 - 321\rho^8 + 626\rho^{10}\end{aligned}$$

can be expressed using $\mathcal{H} = \partial_t$, $\mathcal{G} = \mathrm{id}$, $y = (\rho, \gamma)$, and $\phi(y) = (1, \rho, \rho^2, \dots, \rho^{11})$.

Problem 1 *Given data $\{z_i\}_{i\in[1\dots n]}$, find a plug-in estimator $\hat\theta$ satisfying*

$$\left\| \hat\theta - \theta_0 \right\|_q \le \mathsf{error}(n, h, q)$$

where $\mathsf{error}(n,h,q)$ is a computable function satisfying $\limsup_{h\to 0} \limsup_{n\to\infty} \mathsf{error}(n,h,q) = 0$ for all q.

Assumption 1 *The nonlinearity ϕ is Lipschitz in its second argument, uniformly in its first:*

$$\sup_{t\in[0,T]} \sup_{x\in\mathbb{R}^{\mathsf{d}_\mathcal{G}}} \left\| \nabla_x \phi(t,x) \right\| < \infty.$$

Assumption 2 *For some $p \ge 2$, the signal y satisfies*

$$\sup_{t\in[0,T]} \left\| \frac{\partial^p y}{\partial t^p}(t) \right\| < \infty.$$

Assumption 3 *The noise $\{\epsilon_i\}_{i\in[1\dots n]}$ has zero mean, is independent across i, and is subgaussian:*

$$\sup_{i\in[1\dots n]} \|\epsilon_i\|_{\psi_2} \le K < \infty.$$

4. Design of the estimator

We use two different kinds of regularization:

Definition 1 (Regularization operators) *Express $A \in \mathbb{R}^{n\times n}$ as $A = \sum_{i=1}^n \sigma_i u_i v_i^\intercal$ with $\sigma_i \geq 0$ and $\{u_i\}$, $\{v_i\}$ orthonormal. The singular-value clipping operator is $[A]_{\vee\lambda} = \sum_{i=1}^n \max(\lambda, \sigma_i) u_i v_i^\intercal$. For $\mu > 0$, $\rho_\mu : \mathbb{R}^n \to \mathbb{R}^n$ is $\rho_\mu(x) = x/(1 + \|x\|/\mu)$.*

The estimator has two hyperparameters $\lambda, \mu > 0$. We select regression times $\{t_j\}_{j=1}^{n'}$ at which to impose (2), then approximate the *continuous functions* $\mathcal{H}y : \mathbb{R} \to \mathbb{R}^{\mathsf{d}_\mathcal{H}}$ and $\mathcal{G}y : \mathbb{R} \to \mathbb{R}^{\mathsf{d}_\mathcal{G}}$ at these times using the filtered estimates $\hat{\mathcal{H}}$, $\hat{\mathcal{G}}$, and $\tilde{\mathcal{G}}$ (described in Section 4). Define the matrices $Y \in \mathbb{R}^{n'\times \mathsf{d}_\mathcal{H}}$, $X \in \mathbb{R}^{n'\times \mathsf{d}_\phi}$, and $Z \in \mathbb{R}^{n'\times \mathsf{d}_\phi}$ by

$$Y_j = \hat{\mathcal{H}}y(t_j), \qquad X_j = \phi(t_j, \hat{\mathcal{G}}y(t_j)), \qquad Z_j = \rho_\mu \circ \phi(t_j, \tilde{\mathcal{G}}y(t_j)), \tag{3}$$

where Y_j, X_j, and Z_j denote the j-th rows of the respective matrices. Note that $\tilde{\mathcal{G}}y(t_j)$ is stochastically independent of both $\hat{\mathcal{G}}y(t_j)$ and $\hat{\mathcal{H}}y(t_j)$, and each row of Z is bounded by μ in absolute value. The estimator is given by

$$\hat{\theta} = \left([Z^\intercal X]_{\vee\lambda}\right)^{-1} Z^\intercal Y, \tag{4}$$

and satisfies:

Corollary 2 *Let X, Y, and Z come from the estimator $\hat{\theta}$ defined in §4. Then $\hat{\theta}$ satisfies the first two hypotheses of Theorem 10. If N has the ideal scaling with respect to h, $\sigma^2 \sim n$, $\sup_i |y_i^\star| = \Theta(1)$, $\sup_i |x_i^\star| = O(1)$, and for all C,*

$$\exp\left(-Cn^3 h^{2p/(2p+1)}\right) \ll \lambda \ll n,$$

then for any $q \geq 1$,

$$\left\|\hat{\theta} - \theta_0\right\|_q \lesssim h^{(p-d)/(2p+1)} + \sqrt{\frac{1}{nh^{2p/(2p+1)}}}$$

where $d = 1$ for the first-order continuous-time estimator and $d = 0$ for the discrete-time estimator.

Remark 3 (Sensitivity to tuning parameters) *The range of favorable $\lambda = \lambda(n, h)$ as $n \to \infty$, $h \to 0$ is very wide; and at the $\lambda \sim n$ extreme, the IV estimator behaves like least squares. The role of μ is more subtle. On one hand, μ multiplies the entire RHS of the error bound (hidden behind "$\lesssim$"), so it would seem that smaller is better. On the other hand, μ also factors into the persistence of excitation condition (c), and an excessively small μ might make the condition fail. Fortunately (in practice), the persistence of excitation condition is checkable in that $\mathbb{E}\, Z^\intercal X \approx Z^\intercal X$, which can be computed from the data.*

The proof of this result relies on Theorem 10, which has the form

$$\text{error} \lesssim (\text{sensitivity to } \lambda)\,\lambda + (\text{sensitivity to data})\,(\text{bias} + \text{noise})\,.$$

In order to motivate and satisfy the hypotheses of Theorem 10, we now narrate the differences between our estimator and the least-squares estimator $\hat{\theta}_{\text{LS}} = (X^\intercal X)^{-1} X^\intercal Y$, which is obtained by replacing Z with X, setting $\lambda = 0$, and setting $\mu = \infty$.

Why Z? As the upcoming example shows, the noise-noise interaction in $X^\intercal X$ (and, to a lesser extent, in $X^\intercal Y$) is a source of bias. Replacing $X^\intercal X$ with $Z^\intercal X$ and $X^\intercal Y$ with $Z^\intercal Y$ eliminates this source of bias; the columns of Z are called instrumental variables (Davidson and MacKinnon, 2004, Chapter 8). We depict a one-dimensional miniature of the quadratic terms in $\hat{\theta}$ and $\hat{\theta}_{\mathrm{LS}}$: η_1 illustrates what is happening inside $\hat{\theta}_{\mathrm{LS}}$, and η_2 illustrates what is happening inside $\hat{\theta}$.

Example 2 *Suppose we are trying to estimate $\eta = \frac{1}{n}\sum_{i=1}^n \mu_i^2 \in \mathbb{R}$ from the datasets $\{X_{\mathsf{A},i}\}_{i=1}^n$ and $\{X_{\mathsf{B},i}\}_{i=1}^n$ where for $\beta \in \{\mathsf{A},\mathsf{B}\}$ we have $X_{\beta,i} = \mu_i + \epsilon_{\beta,i}$. The noise $\epsilon_{\beta,i}$ has distribution $\mathcal{N}(0,\sigma^2)$, and $\epsilon_{\beta,i}$ is independent of $\epsilon_{\beta',i'}$ if $\beta \neq \beta'$ or $i \neq i'$. Consider the two estimators*

$$\hat{\eta}_1 = \frac{1}{2n}\left(\sum_{i=1}^n X_{\mathsf{A},i}^2 + \sum_{i=1}^n X_{\mathsf{B},i}^2\right) \qquad \text{and} \qquad \hat{\eta}_2 = \frac{1}{n}\sum_{i=1}^n X_{\mathsf{A},i}X_{\mathsf{B},i}$$

with means

$$\mathbb{E}\,\hat{\eta}_1 = \eta + \sigma^2 \qquad \text{and} \qquad \mathbb{E}\,\hat{\eta}_2 = \eta$$

and variances

$$\operatorname{var}\hat{\eta}_1 = \operatorname{var}\hat{\eta}_2 = \frac{2\sigma^2\eta + \sigma^4}{n}.$$

While these two estimators have identical variances, $\hat{\eta}_2$ is unbiased and has a strictly smaller MSE than $\hat{\eta}_1$.

This example shows that the independence structure of $Z^\intercal X$ is key. We achieve it in our problem via a novel sample-split design for time series with a latent continuous-time structure.

The filters $\hat{\mathcal{H}}$, $\hat{\mathcal{G}}$, and $\tilde{\mathcal{G}}$. To approximate $\mathcal{H}$ and $\mathcal{G}$ from noisy discrete data, we use local polynomial regression (Fan and Gijbels, 2003; De Brabanter et al., 2013). Each filter is specified by a *differentiation stencil* $\mathbf{D}_{d,h}^{k,i_0,N} \in \mathbb{R}^N$, which is a vector of coefficients that linearly combines N consecutive measurements to approximate the d-th derivative at location $i_0 h$.

Definition 4 (Differentiation stencil) *Fix a window size $N \in \mathbb{N}$, step size $h > 0$, derivative order $d \in \mathbb{N}$, and location $i_0 \in \mathbb{R}$. The differentiation stencil $\mathbf{D}_{d,h}^{k,i_0,N} \in \mathbb{R}^N$ satisfies*

$$\sum_{k=1}^N \mathbf{D}_{d,h}^{k,i_0,N} f(kh) = \frac{\mathrm{d}^d f}{\mathrm{d}x^d}(i_0 h) \quad \text{for all polynomials } f \text{ of degree at most } p-1, \tag{5}$$

where $p \geq d+1$ is a design parameter. Among all stencils satisfying (5), we select the one with minimum Frobenius norm (see Appendix D for construction).

Lemma 5 (Local polynomial filtering) *Let $f \in C^p([0,T],\mathbb{R})$ be a univariate function with $R_p := \sup_{t\in[0,T]} |f^{(p)}(t)| < \infty$, and let $\{w_k\}_{k=1}^N$ be independent mean-zero noise with $\max_k \|w_k\|_{\psi_2} \leq \nu$. Define the filtered estimate*

$$\widehat{f^{(d)}}(i_0 h) := \sum_{k=1}^N \mathbf{D}_{d,h}^{k,i_0,N}\left(f(kh) + w_k\right).$$

Then the bias and fluctuation satisfy

$$\left|\mathbb{E}\,\widehat{f^{(d)}}(i_0h) - f^{(d)}(i_0h)\right| \le C(p,i_0)R_p(Nh)^{p-d},$$
$$\left\|\widehat{f^{(d)}}(i_0h) - \mathbb{E}\,\widehat{f^{(d)}}(i_0h)\right\|_{\psi_2} \le C(p,i_0)\nu N^{-d-\frac{1}{2}}h^{-d}.$$

Proof This relies on Assumptions 2 and 3. The construction and analysis are detailed in Appendix D. The stencil coefficients are obtained by solving a minimum-norm problem subject to the exactness constraint (5) (Proposition 14). The bias bound follows from Taylor expansion (Proposition 16). The fluctuation bound uses subgaussian concentration (Proposition 17). ■

We first "unzip" the time series into even and odd subsequences, then apply the local polynomial filter to each subsequence separately:

Definition 6 (Sample-split filters) *Let $\{z_i\}_{i=1}^n$ denote the measurements, and let $z^{even} = \{z_{2j}\}_{j=1}^{\lfloor n/2\rfloor}$ and $z^{odd} = \{z_{2j-1}\}_{j=1}^{\lceil n/2\rceil}$ denote the even- and odd-indexed subsequences. For a given operator $\mathcal{T}$ with parameters (i_0, d), define:*

$$\hat{\mathcal{T}}y(t_j) := \sum_{k=1}^{N} \mathbf{D}_{d,2h}^{k,i_0-1/4,N} z_k^{even}, \qquad \tilde{\mathcal{T}}y(t_j) := \sum_{k=1}^{N} \mathbf{D}_{d,2h}^{k,i_0+1/4,N} z_k^{odd},$$

where the stencils are evaluated at step size $2h$ (twice the original sampling period) and shifted locations $i_0 \pm 1/4$.

Both $\hat{\mathcal{T}}$ and $\tilde{\mathcal{T}}$ approximate the same continuous function at the same off-grid location (corresponding to "$z_{i+0.5}$"), but use disjoint subsets of the data; see Figure 2 in Appendix A for a visualization that explains the quarter-step offset.

Because the noise terms $\{\epsilon_i\}$ are independent across indices, we have:

Lemma 7 *The filters $\hat{\mathcal{H}}$, $\hat{\mathcal{G}}$, and $\tilde{\mathcal{G}}$ obey the same bounds as in Lemma 5; moreover, $\hat{\mathcal{H}}$ and $\hat{\mathcal{G}}$ are independent of $\tilde{\mathcal{G}}$.*

Why λ? The vanilla IV estimator $(Z^\intercal X)^{-1}Z^\intercal Y$ is well-studied for its unbiasedness and asymptotic normality and efficiency under standard identifying assumptions (González et al., 2024). However, it has a heavy-tailed distribution and does not have any finite moments of any order (Davidson and MacKinnon, 2004, §8.4). We seek to capture a finite-sample analog of the traditional asymptotic consistency result. By clipping the lower spectrum of $Z^\intercal X$, we ensure that for all $q \ge 1$, $\left\|\hat{\theta}\right\|_q < \infty$. It is furthermore helpful that singular value clipping is a bounded perturbation of the unclipped matrix:

Lemma 8 *Let A be any matrix. Then $\left\|A - [A]_{\vee\lambda}\right\| \le \lambda$.*

Proof Because A and $[A]_{\vee\lambda}$ have the same singular vectors, $\left\|A - [A]_{\vee\lambda}\right\|$ has singular values at most λ. ■

Why μ? We have made $\left\|\hat{\theta}\right\|_q$ finite by choosing $\lambda > 0$. We must now go further and ensure that $\left\|\hat{\theta} - \theta_0\right\|_q$ is small; in order to prove the IV estimator's consistency, we need a concentration inequality that bounds the deviations of $Z^\intercal X$ from a nominal value. If we set $\mu = \infty$, then both Z and X would have subgaussian rows, and one could apply a Hanson-Wright inequality to get subexponential concentration of $Z^\intercal X$ (Ziemann et al., 2023). However, it is easier to work with subgaussian concentration than subexponential concentration, especially when working in probabilistic L^p spaces. Hence, by choosing $\mu < \infty$, the rows of Z become bounded by construction, and we can get subgaussian concentration of $Z^\intercal X$, and the following moment inequality for its regularized inverse:

Proposition 9 (Three-way integral) *Let $0 < a < b < \infty$ be constants. Let $W > 0$ be a real-valued random variable satisfying $\mathbb{P}(W \geq t) \leq \exp(-t^2/K^2)$ for some $K > 0$. Then for all $r \geq 1$, X defined by*

$$X = \frac{1}{a + 0 \vee (b - W)}$$

satisfies

$$\|X\|_r \leq \gamma(r; a, b, K) = \gamma_{\mathit{head}}(r; a, b) + \gamma_{\mathit{body}}(r; a, b, K) + \gamma_{\mathit{tail}}(r; a, b, K),$$

where

$$\gamma_{\mathit{head}}(r; a, b) = \frac{2}{b}$$
$$\gamma_{\mathit{body}}(r; a, b, K) = C\frac{r^{1/r}K^{1/r}}{b^{2(1+1/r)}} \exp\left(CK^2\frac{(r+1)^2\log^2(a/b)}{r\,(b-a)^2}\right),$$

and

$$\gamma_{\mathit{tail}}(r; a, b, K) = \frac{\exp(-Cb^2/rK^2)}{a}$$

for some absolute constant $C > 0$.

Proof The full proof is given in Appendix B. The idea is to truncate X into three regions based on the support of W: a "head" where W is small, a "tail" where W is very large, and (our innovation) a "body" where W is large but not too large. In this region, the value of X transitions from $1/a$ to $1/b$. We integrate over this region using a "layer cake" Tonelli rearrangement to get a Gaussian integral. ■

Finally, we state a technical result which bounds the L^p risk of $\hat{\theta} = [Z^\intercal X]_{\vee\lambda}^{-1} Z^\intercal Y$ under abstract assumptions on Z, X, and Y.

Theorem 10 *Let $\lambda > 0$, $q \geq 2$, and $\epsilon > 0$. Suppose that unobserved $Y^\star \in \mathbb{R}^{n\times \mathsf{d}_y}$, $X^\star \in \mathbb{R}^{n\times \mathsf{d}_x}$, and $\theta^\star \in \mathbb{R}^{\mathsf{d}_x\times \mathsf{d}_y}$ satisfy*

$$Y^\star = X^\star\theta^\star. \tag{6}$$

Suppose we have data (z_i, x_i, y_i) for $i \in [1 \ldots n]$, such that

a. The data bounds hold:

$$\sup_{i\in[1\dots n]} \left\|\mathbb{E}\, z_i y_i^\intercal - z_i (y_i^\star)^\intercal\right\| = \bar{\nu}_{zy} < \infty, \qquad \sup_{i\in[1\dots n]} \left\|\mathbb{E}\, z_i y_i^\intercal - z_i y_i^\intercal\right\|_{\psi_2} = \tilde{\nu}_{zy} < \infty,$$
$$\sup_{i\in[1\dots n]} \left\|\mathbb{E}\, z_i x_i^\intercal - z_i (x_i^\star)^\intercal\right\| = \bar{\nu}_{zx} < \infty, \qquad \sup_{i\in[1\dots n]} \left\|\mathbb{E}\, z_i x_i^\intercal - z_i x_i^\intercal\right\|_{\psi_2} = \tilde{\nu}_{zx} < \infty.$$

b. For $i, i' \in [1 \dots n]$, if $|i - i'| \geq N$, then (x_i, y_i, z_i) is independent of $(x_{i'}, y_{i'}, z_{i'})$.

c. The data satisfies the persistence of excitation condition:

$$\sigma_{min}\left(\mathbb{E}\, Z^\intercal X\right) \geq \sigma^2 > 0.$$

Then the estimator defined by

$$\hat{\theta} = \left([Z^\intercal X]_{\vee\lambda}\right)^{-1} (Z^\intercal Y)$$

satisfies

$$\frac{\left\|\hat{\theta} - \theta^\star\right\|_q}{\|\theta^\star\|} \leq \gamma(q; \lambda, \sigma^2 - \lambda, \sqrt{nN}\tilde{\nu}_{zx})\lambda + C\sqrt{q(1+1/\epsilon)}\gamma(q(1+1/\epsilon); \lambda, \sigma^2 - \lambda, \sqrt{nN}\tilde{\nu}_{zx}) \cdot \left\{ n\left(\bar{\nu}_{zx} + \left\|\theta^\star\right\|^{-1} \bar{\nu}_{zy}\right) + \sqrt{nN}\left(\tilde{\nu}_{zx} + \left\|\theta^\star\right\|^{-1} \tilde{\nu}_{zy}\right) \right\}.$$

where γ is the function appearing in Proposition 9.

5. Numerical examples

Our data comes from the Lorenz system with sinusoidal forcing:

$$\dot{x}^1 = \sigma(x^2 - x^1), \quad \dot{x}^2 = x^1(\rho - x^3) - x^2, \quad \dot{x}^3 = u(t) + x^1 x^2 - \beta x^3, \quad u(t) = \sin(2\pi f t)$$

which we write as

$$\dot{\xi} = \theta_0^\intercal \phi(t, \xi) \tag{7}$$

where

$$\xi = (x^1, x^2, x^3) \qquad \text{and} \qquad \phi(t, x^1, x^2, x^3) = (\sin(2\pi f t), x^1, x^2, x^3, x^1 x^2, x^1 x^3).$$

Since our contribution focuses on the regression step, rather than (regularized or sequential) sparse model selection, we regard the entire matrix $\theta_0 \in \mathbb{R}^{6\times 3}$ as unknown. See Appendix F for details on the data, estimator, and reporting.

5.1. Continuous-time

Taking $\mathcal{H} = \frac{\partial}{\partial t}$ and $\mathcal{G} = 1$ recovers the true data-generating process (7). We compare a sample-split IV estimator of our design to a least-squares estimator, taking θ_0 as ground truth.

Our estimator achieves a $\sim$200x reduction in bias and $\sim$2x reduction in L^2 risk versus least squares (Table 1); its sampling distribution appears nearly unbiased, while least squares is biased toward zero with less dispersion (Figure 3, Appendix F).

Estimator	abs. bias (%)	std (%)	rmse (%)
Instrumental Variables (ours)	0.017(8)	0.800(7)	0.800(7)
Least Squares	2.382(3)	0.517(5)	2.437(3)

Table 1: Comparison of bias, standard deviation, and root mean square error of our estimator and a least squares estimator for the parameter of the continuous-time Lorenz system. Monte Carlo standard errors in parentheses.

Estimator	abs. bias (%)	std (%)	rmse (%)
Instrumental Variables (ours)	0.003 18(148)	0.144 31(126)	0.144 34(126)
Least Squares	1.511 12(59)	0.076 90(73)	1.513 07(59)

Table 2: Comparison of bias, standard deviation, and root mean square error (normalized by the pseudo-true value) of our estimator and a least squares estimator for the parameter of the discrete-time Lorenz system. Monte Carlo standard errors in parentheses.

5.2. Discrete-time

Taking $\mathcal{H}$ to be a left shift by h and $\mathcal{G} = 1$ produces a first-order discretization of (7). We compare a sample-split IV estimator of our design to a least-squares estimator. There is no ground truth in this case, so we compare against a pseudo-true value obtained by evaluating the least-squares estimator on noise-free data.

Our estimator achieves a $\sim$500x reduction in bias and $\sim$10x reduction in L^2 risk versus least squares (Table 2); the sampling distribution is again virtually unbiased, while least squares is biased toward zero (Figure 4, Appendix F).

5.3. Van der Pol oscillator

Unlike the Lorenz examples, the Van der Pol oscillator exercises the case $\mathcal{G} \neq 1$. The scalar observation $x(t)$ satisfies the second-order ODE

$$\ddot{x} = \mu(1 - x^2)\dot{x} - x,$$

which we write as

$$\mathcal{H}\, x(t) = \theta_0^{\mathsf{T}} \phi(\mathcal{G}\, x(t)) \tag{8}$$

where $\mathcal{H} = \partial_t^2$, $\mathcal{G} = \left(\partial_t^0, \partial_t^1\right)$,

$$\phi(x, \dot{x}) = (x,\ \dot{x},\ x^2\dot{x}), \qquad \theta_0 = (-1,\ \mu,\ -\mu)^{\mathsf{T}}.$$

Both $\mathcal{H}$ and $\mathcal{G}$ require derivative estimation from the noisy scalar data $\{z_i\}$. We compare a sample-split IV estimator (with $\mu = 2$) to a least-squares estimator, taking θ_0 as the ground truth. See Appendix G for details on the data, estimator, and reporting.

Our estimator achieves a substantial reduction in bias and L^2 risk versus least squares (Table 3, Appendix G); its sampling distribution appears nearly unbiased, while least squares is biased toward zero (Figure 1).

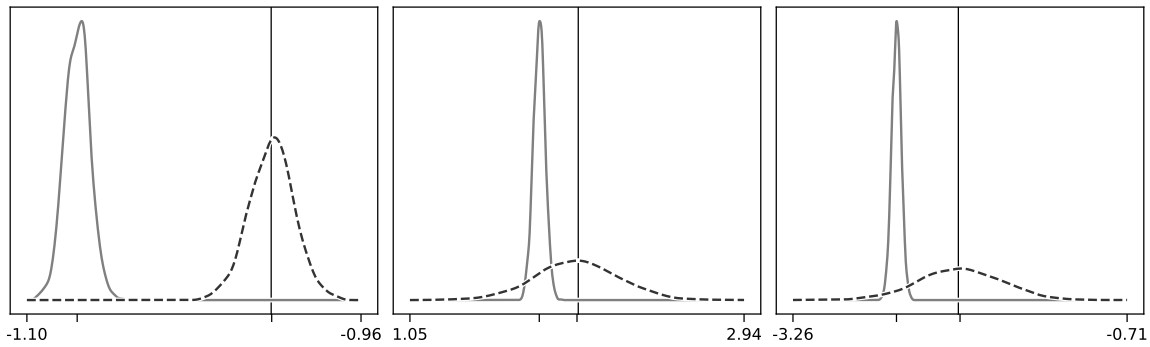

Figure 1: Elementwise marginal kernel density estimates of the sampling distributions of our estimator (dashed) and a baseline estimator (solid) for the Van der Pol parameter. Vertical line indicates ground truth; ticks indicate mean of sampling distribution.

6. Novelty

We develop a form of instrumental variables estimation for system identification problems where there are no obvious instruments. The instruments are based on a novel application of local polynomial regression to smooth or differentiate a function onto an off-grid point.

We then show that this estimator is consistent in L^p, whereas the vanilla IV estimator does not even have an expectation.

7. Significance

Many problems in driven engineering could benefit from the bias reduction of instrumental variables estimation. The only shortfall is that (especially in nonlinear models) there are no instruments. Our filtering constructions are applicable to any method that models time evolution linearly in the parameters (Kutz et al., 2016; Brunton et al., 2016; Mezić, 2021; Haller, 2025).

The L^p consistency of our estimator appeals to a trend towards non-asymptotic analyses of point estimators traditionally understood via asymptotic normality, such as system identification with noiseless data and random designs (Ziemann et al., 2023; Bakshi et al., 2023). Broadly speaking, one hopes to attain the classical $\sqrt{n}$-asymptotic rate of convergence, but with non-asymptotic constants. Whereas a large body of work focuses on bounding the quantiles of the estimation error, our work achieves this goal (with a nonparametric power-of-h caveat) by bounding the L^p risk for a fixed-design, random noise setting. A technique that may be of independent interest is the layer cake technique for bounding regularized matrix inverses (Proposition 9), and the μ-truncation for thinning the tails of a design matrix from subexponential to subgaussian without incurring any bias (an elementary workaround that could obviate more complex Hanson-Wright inequalities).

Acknowledgments

This work was supported by the National Science Foundation CAREER Program (Grant No. 2046292).

References

Ainesh Bakshi, Allen Liu, Ankur Moitra, and Morris Yau. A New Approach to Learning Linear Dynamical Systems. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 335–348, Orlando FL USA, June 2023. ACM. doi:10.1145/3564246.3585247. URL https://dl.acm.org/doi/10.1145/3564246.3585247.

Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, April 2016. doi:10.1073/pnas.1517384113. URL https://www.pnas.org/doi/10.1073/pnas.1517384113. Publisher: Proceedings of the National Academy of Sciences.

Russell Davidson and James G. MacKinnon. *Econometric theory and methods*. Oxford Univ. Press, New York, NY, 2004. ISBN 978-0-19-512372-2.

Kris De Brabanter, Jos De Brabanter, Bart De Moor, and Irène Gijbels. Derivative estimation with local polynomial fitting. *J. Mach. Learn. Res.*, 14(1):281–301, January 2013. ISSN 1532-4435.

Jianqing Fan and I. Gijbels. *Local polynomial modelling and its applications*. CRC Press, Boca Raton, 2003. ISBN 978-0-203-74872-5. OCLC: 1034989441.

Hugues Garnier and Liuping Wang, editors. *Identification of continuous-time models from sampled data*. Advances in industrial control. Springer, London, 2008. ISBN 978-1-84800-160-2 978-1-84800-161-9. OCLC: ocn183149174.

Rodrigo A. González. *Continuous-time System Identification: Refined Instrumental Variables and Sampling Assumptions*. Kungliga Tekniska högskolan, Stockholm, Sweden, 2022. ISBN 9789180401869.

Rodrigo A. González, Siqi Pan, Cristian R. Rojas, and James S. Welsh. Consistency analysis of refined instrumental variable methods for continuous-time system identification in closed-loop. *Automatica*, 166:111697, August 2024. ISSN 00051098. doi:10.1016/j.automatica.2024.111697. URL https://linkinghub.elsevier.com/retrieve/pii/S0005109824001912.

George Haller. *Modeling Nonlinear Dynamics from Equations and Data — with Applications to Solids, Fluids, and Controls*. Society for Industrial and Applied Mathematics, Philadelphia, PA, January 2025. ISBN 978-1-61197-834-6 978-1-61197-835-3. doi:10.1137/1.9781611978353. URL https://epubs.siam.org/doi/book/10.1137/1.9781611978353.

Junette Hsin, Shubhankar Agarwal, Adam Thorpe, Luis Sentis, and David Fridovich-Keil. Symbolic Regression on Sparse and Noisy Data with Gaussian Processes, October 2024. URL http://arxiv.org/abs/2309.11076. arXiv:2309.11076 [cs].

Simon Kuang and Xinfan Lin. Estimation Sample Complexity of a Class of Nonlinear Continuous-time Systems. In *Modeling, Estimation, and Control Conference*, volume 58, pages 786–791, Chicago, IL, 2024. doi:10.1016/j.ifacol.2025.01.069. URL https://linkinghub.elsevier.com/retrieve/pii/S2405896325000692.

J. Nathan Kutz, Steven L. Brunton, Bingni W. Brunton, and Joshua L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, November 2016. ISBN 978-1-61197-449-2 978-1-61197-450-8. doi:10.1137/1.9781611974508. URL http://epubs.siam.org/doi/book/10.1137/1.9781611974508.

Lennart Ljung. Perspectives on system identification. *Annual Reviews in Control*, 34(1):1–12, April 2010. ISSN 1367-5788. doi:10.1016/j.arcontrol.2009.12.001. URL https://www.sciencedirect.com/science/article/pii/S1367578810000027.

Dipankar Maity and Debdipta Goswami. On the Effect of Quantization on Extended Dynamic Mode Decomposition. In *2025 American Control Conference (ACC)*, pages 3176–3182, Denver, CO, USA, July 2025. IEEE. ISBN 9798331569372. doi:10.23919/ACC63710.2025.11107527. URL https://ieeexplore.ieee.org/document/11107527/.

Igor Mezić. Koopman Operator, Geometry, and Learning of Dynamical Systems. *Notices of the American Mathematical Society*, 68(07):1, August 2021. ISSN 0002-9920, 1088-9477. doi:10.1090/noti2306. URL https://www.ams.org/notices/202107/rnoti-p1087.pdf.

Siqi Pan, James S. Welsh, Rodrigo A. González, and Cristian R. Rojas. Efficiency analysis of the Simplified Refined Instrumental Variable method for Continuous-time systems. *Automatica*, 121:109196, November 2020. ISSN 00051098. doi:10.1016/j.automatica.2020.109196. URL https://linkinghub.elsevier.com/retrieve/pii/S0005109820303940.

Torsten Söderström. *Errors-in-Variables Methods in System Identification*. Communications and Control Engineering. Springer International Publishing, Cham, 2018. ISBN 978-3-319-75000-2 978-3-319-75001-9. doi:10.1007/978-3-319-75001-9. URL http://link.springer.com/10.1007/978-3-319-75001-9.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. URL http://arxiv.org/abs/1706.03762. arXiv:1706.03762 [cs].

Roman Vershynin. High-Dimensional Probability: An Introduction with Applications in Data Science, September 2018. URL https://www.cambridge.org/core/books/highdimensional-probability/797C466DA29743D2C8213493BD2D2102. ISBN: 9781108231596 9781108415194 Publisher: Cambridge University Press.

Jacqueline Wentz and Alireza Doostan. Derivative-based SINDy (DSINDy): Addressing the challenge of discovering governing equations from noisy data. *Computer Methods in Applied Mechanics and Engineering*, 413:116096, August 2023. ISSN 00457825. doi:10.1016/j.cma.2023.116096. URL https://linkinghub.elsevier.com/retrieve/pii/S0045782523002207.

Grace Y. Yi, Aurore Delaigle, and Paul Gustafson. *Handbook of Measurement Error Models*. Chapman and Hall/CRC, Boca Raton, 1 edition, September 2021. ISBN 978-1-315-10127-9. doi:10.1201/9781315101279. URL https://www.taylorfrancis.com/books/9781315101279.

Thomas Zeiringer, Hugues Garnier, and Martin Horn. Instrumental variable framework for nonlinear continuous-time system identification. *Journal of the Franklin Institute*, 363(7):108619, May 2026. ISSN 00160032. doi:10.1016/j.jfranklin.2026.108619. URL https://linkinghub.elsevier.com/retrieve/pii/S001600322600219X.

Ingvar Ziemann, Anastasios Tsiamis, Bruce Lee, Yassir Jedra, Nikolai Matni, and George J. Pappas. A Tutorial on the Non-Asymptotic Theory of System Identification, September 2023. URL http://arxiv.org/abs/2309.03873. arXiv:2309.03873 [cs, eess, stat].

Contents

Appendix A. Sample split illustration

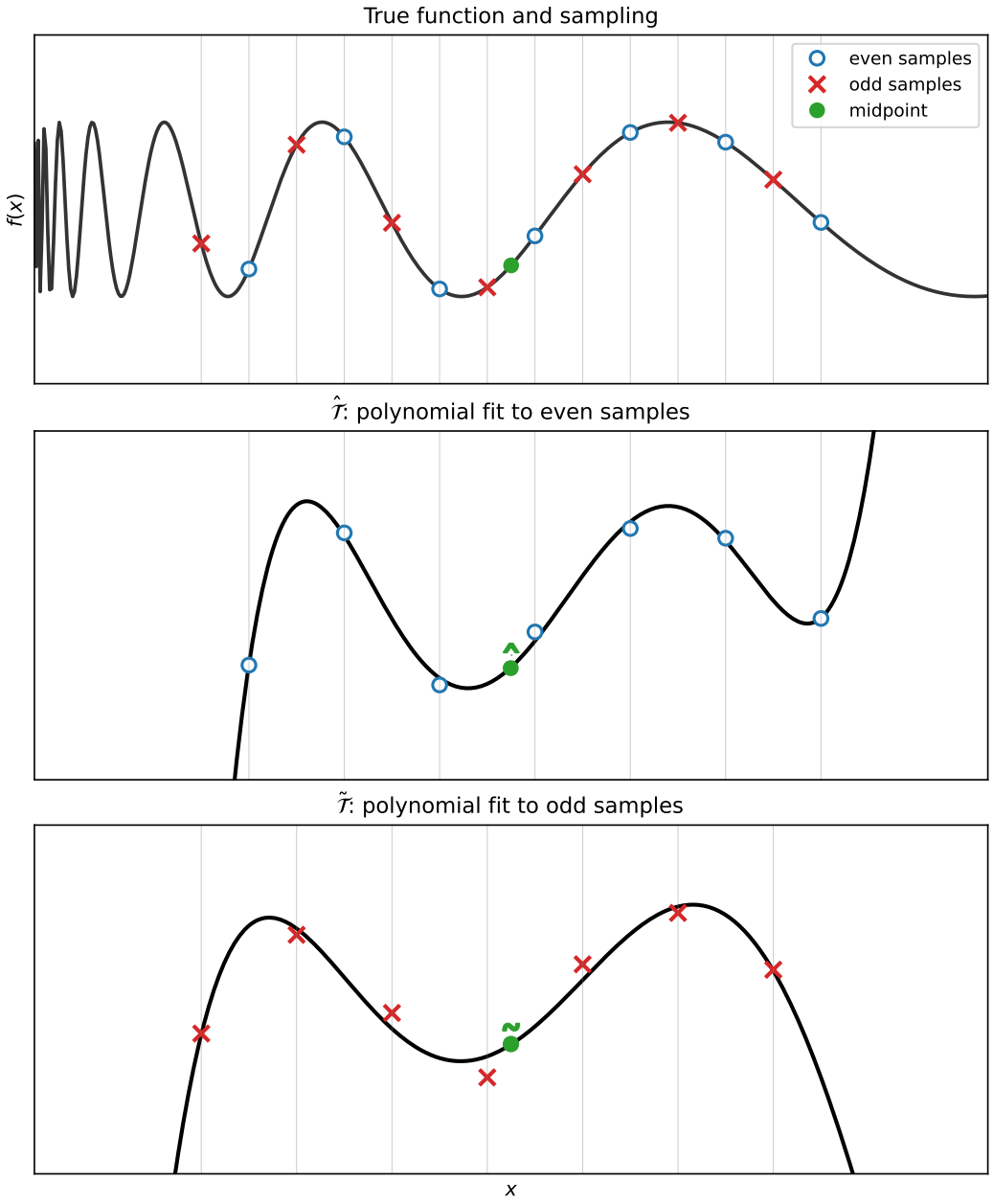

Figure 2: The top panel shows the true function with even (blue circles) and odd (red crosses) samples. The middle panel shows $\hat{\mathcal{T}}$ constructed from even samples, and the bottom panel shows $\tilde{\mathcal{T}}$ constructed from odd samples. Both filters interpolate to the same point "•" using disjoint data subsets, resulting in estimates $\hat{\bullet}$ and $\tilde{\bullet}$ respectively. Note that from the points of views of $\hat{\mathcal{T}}$ and $\tilde{\mathcal{T}}$, the evaluation point is one-quarter step off center.

Appendix B. Proof of Proposition 9

Proof Observe that $a + 0 \vee (b - W)$ enjoys the following upper bound according to three cases of W:

$$a + 0 \vee (b - W) \geq \begin{cases} b, & W \leq a \\ a + b - W, & a \leq W < b \\ a, & b < W \end{cases} \tag{9}$$

These three cases yield an additive decomposition of X as

$$X \leq X_1 + X_2 + X_3, \quad \text{where} \tag{10}$$

$$X_1 = \frac{\mathbf{1}_{W \geq a}}{b}, \qquad X_2 = \frac{\mathbf{1}_{a \leq W < b}}{a + b - W}, \qquad \text{and} \qquad X_3 = \frac{\mathbf{1}_{W \geq b}}{a}. \tag{11}$$

By the triangle inequality of L^r norms, we have

$$\|X\|_r \leq \|X_1\|_r + \|X_2\|_r + \|X_3\|_r \tag{12}$$

By taking expectations, we immediately obtain the bounds

$$\|X_1\|_r \leq \frac{1}{b} \quad \text{and} \tag{13}$$

$$\|X_3\|_r \leq \frac{\mathbb{P}\left(W \geq b\right)^{1/r}}{a} \leq \frac{\exp(-b^2/rK^2)}{a}. \tag{14}$$

To bound $\mathbb{E}\left\|X_2\right\|^r$, we use the Fundamental Theorem of Calculus:

$$X_2^r = X_{2,1}^r + X_{2,2}^r, \tag{15}$$

where

$$X_{2,1}^r = \left.\frac{1}{(a + b - s)^r}\right|_{s=a} = \frac{1}{b^r} \tag{16}$$

results in

$$\left\|X_{2,1}\right\|_r \leq \frac{1}{b} \tag{17}$$

and

$$X_{2,2}^r = \mathbf{1}_{a \leq W < b} \left.\frac{1}{(a + b - s)^r}\right|_{s=a}^{s=W} \tag{18}$$

$$= \mathbf{1}_{a \leq W < b} \int_a^W \frac{\mathrm{d}}{\mathrm{d}s}\left[\frac{1}{(a + b - s)^r}\right] \mathrm{d}s \tag{19}$$

Taking expectations,

$$\mathbb{E}\, X_{2,2}^r = \mathbb{E}\, \mathbf{1}_{a\le W<b} \int_a^W \frac{r}{(a+b-s)^{r+1}}\,\mathrm{d}s \tag{20}$$

$$= \mathbb{E} \int_a^b \frac{r\mathbf{1}_{W\ge s}}{(a+b-s)^{r+1}}\,\mathrm{d}s \tag{21}$$

$$= \int_a^b \frac{r\,\mathbb{P}\left(W\ge s\right)}{(a+b-s)^{r+1}}\,\mathrm{d}s \qquad \text{(Tonelli)}$$

$$= r\int_a^b \underbrace{\frac{1}{(a+b-s)^{r+1}}}_{=:g(s)} e^{-s^2/K^2}\,\mathrm{d}s \qquad \text{(Lemma 12)}$$

$$= r\int_a^b e^{-s^2/K^2+\log g(s)}\,\mathrm{d}s \tag{22}$$

$$\le r\int_a^b \exp\left(-\frac{s^2}{K^2} + \log(b^{-2(r+1)}) + \frac{\log\left[(a/b)^{-(r+1)}\right]}{b-a}s\right)\mathrm{d}s$$

(by convexity of $\log g(s)$ on $[a,b]$)

$$\le \frac{r}{b^{2(r+1)}}\int_{-\infty}^{\infty} \exp\left(-\frac{s^2}{K^2} + \frac{\log\left[(a/b)^{-(r+1)}\right]}{b-a}s\right)\mathrm{d}s \tag{23}$$

$$= C\frac{rK}{b^{2(r+1)}}\exp\left(CK^2\frac{(r+1)^2\log^2(a/b)}{(b-a)^2}\right) \tag{24}$$

by the Gaussian integral identity $\int_{-\infty}^{\infty} e^{-(ax^2+bx)}\,\mathrm{d}x = \sqrt{\frac{\pi}{a}}e^{\frac{b^2}{4a}}$. Raising both sides to the power $1/r$,

$$\left\|X_{2,2}\right\|_r \le C\frac{r^{1/r}K^{1/r}}{b^{2(1+1/r)}}\exp\left(CK^2\frac{(r+1)^2\log^2(a/b)}{r\,(b-a)^2}\right) \tag{25}$$

by a Gaussian integral. We conclude a bound on $\|X_{2,2}\|_r$ by raising both sides to the power $\frac{1}{r}$.

We re-associate the summands in (12),

$$\|X\|_r \le \underbrace{\|X_1\|_r + \left\|X_{2,1}\right\|_r}_{:=\gamma_{\text{head}}} + \underbrace{\left\|X_{2,2}\right\|_r}_{\gamma_{\text{body}}} + \underbrace{\|X_3\|_r}_{\gamma_{\text{tail}}}. \tag{26}$$

and conclude by inserting (13) for $\|X_1\|_r$, (16) for $\left\|X_{2,1}\right\|_r$, (25) for $\left\|X_{2,2}\right\|_r$, and (14) for $\|X_3\|_r$. ■

Appendix C. Proof of Theorem 10

Manipulating (6) yields the identity

$$\begin{aligned} Z^\intercal Y &= Z^\intercal Y^\star + \left[Z^\intercal Y - Z^\intercal Y^\star\right] && (27)\\ &= (Z^\star)^\intercal X^\star \theta^\star + \left[Z^\intercal Y - Z^\intercal Y^\star\right] && (28)\\ &= \left[Z^\intercal X\right]_{\vee\lambda} \theta^\star + \left[Z^\intercal X - \left[Z^\intercal X\right]_{\vee\lambda}\right] \theta^\star \\ &\quad + \left[(Z^\star)^\intercal X^\star - Z^\intercal X\right] \theta^\star + \left[Z^\intercal Y - Z^\intercal Y^\star\right] && (29) \end{aligned}$$

Multiplying both sides by the inverse of $S = \left[Z^\intercal X\right]_{\vee\lambda}$ and inserting the definition of $\hat{\theta}$, we have the decomposition

$$\hat{\theta} - \theta^\star = S^{-1}\left\{ \left[Z^\intercal X - \left[Z^\intercal X\right]_{\vee\lambda}\right] \theta^\star + \left[Z^\intercal X^\star - Z^\intercal X\right] \theta^\star + \left[Z^\intercal Y - Z^\intercal Y^\star\right] \right\}, \tag{30}$$

which may be combined with the Hölder conjugacy

$$\frac{1}{q} = \frac{1}{q(1+\epsilon^{-1})} + \frac{1}{q(1+\epsilon)}, \tag{31}$$

to yield

$$\begin{aligned} \frac{\left\|\hat{\theta} - \theta^\star\right\|_q}{\|\theta^\star\|} &\leq \underbrace{\left\|S^{-1}\right\|_q}_{\text{Lemma 13}} \underbrace{\left\|Z^\intercal X - \left[Z^\intercal X\right]_{\vee\lambda}\right\|_\infty}_{\text{Lemma 8}} \\ &\quad + \underbrace{\left\|S^{-1}\right\|_q}_{\text{Lemma 13}} \left\{ \underbrace{\left\|Z^\intercal X^\star - Z^\intercal X\right\|_{q(1+\epsilon)}}_{\text{Lemma 12 and Fact 1}} + \left\|\theta^\star\right\|^{-1} \underbrace{\left\|Z^\intercal Y - Z^\intercal Y^\star\right\|_{q(1+\epsilon)}}_{\text{Lemma 12 and Fact 1}} \right\} \end{aligned} \tag{32}$$

To finish the proof, we use Lemma 13 to bound the prefactor as a function of σ^2, λ, and $q(1+1/\epsilon)$:

$$\begin{aligned} \gamma(s; \lambda, \sigma^2 - \lambda, L_{ZX}) = \frac{2}{\sigma^2 - \lambda} + \frac{s^{1/s} C L_{ZX}^{1/s}}{\left(\sigma^2 - \lambda\right)^{2(1+1/s)}} \exp\left(C L_{ZX}^2 \frac{(s+1)^2 \log^2(\lambda/(\sigma^2 - \lambda))}{s\left(\sigma^2 - 2\lambda\right)^2} \right) \\ + \frac{\exp(-C(\sigma^2 - \lambda)^2/s L_{ZX}^2)}{\lambda}. \end{aligned} \tag{33}$$

We invoke Lemma 8 to bound $Z^\intercal X - \left[Z^\intercal X\right]_{\vee\lambda}$ almost surely. We invoke Lemma 12 to bound the subgaussian norms of $(Z^\star)^\intercal X^\star - Z^\intercal X$ and $Z^\intercal Y - (Z^\star)^\intercal Y^\star$, and then use Fact 1 to convert these into L^p norms. The result is

$$\begin{aligned} \frac{\left\|\hat{\theta} - \theta^\star\right\|_q}{\|\theta^\star\|} &\leq \gamma(q; \lambda, \sigma^2 - \lambda, L_{ZX})\lambda + C\sqrt{q(1+1/\epsilon)}\gamma(q(1+1/\epsilon); \lambda, \sigma^2 - \lambda, L_{ZX}) \\ &\quad \cdot \left\{ n\left(\bar{\nu}_{zx} + \left\|\theta^\star\right\|^{-1} \bar{\nu}_{zy}\right) + \sqrt{nN}\left(\tilde{\nu}_{zx} + \left\|\theta^\star\right\|^{-1} \tilde{\nu}_{zy}\right) \right\} \end{aligned} \tag{34}$$

C.1. Bounds on N-dependent sums

Before entering the proof of Theorem 10, we first recall some basic facts about (non-isotropic) vector- and matrix-valued subgaussian random variables; these are trivially adapted from the facts found in (Vershynin, 2018, Chapter 2).

Fact 1 *Let X be a vector- or matrix-valued random variables with $\|X\|_{\psi_2} = K$. Then*

i. The moments of X satisfy

$$\|X\|_p \leq CK\sqrt{p}.$$

ii. The tails of X satisfy

$$\mathbb{P}\left(\|X\| \geq t\right) \leq \exp\left(-\frac{t^2}{CK^2}\right).$$

Fact 2 ((Vershynin, 2018, Proposition 2.6.1)) *Let $\{X_i\}_{i=1}^n$ be a sequence of vector- or matrix-valued random variables. Then*

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2}^2 \leq C\sum_{i=1}^n \|X_i\|_{\psi_2}^2.$$

Proposition 11 (Local dependence in L^q) *Let $q \geq 2$. Suppose that $\{X_i\}_{i=1}^n$ are random variables satisfying*

$$\begin{aligned} \mathbb{E}\, X_i &= 0, && \forall i \in [n] \\ \max_{i\in[n]} \|X_i\|_{\psi_2} &= \nu, && \forall i \in [n] \\ X_i &\perp X_j && \forall i, j \in [n] \text{ with } |i-j| \geq N \end{aligned}$$

Then

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2} \leq C\sqrt{nN}\nu.$$

Proof For $k \in [N]$, define the index sets $I_k = \left\{i \in [n] : i \cong k \pmod N\right\}$.

$$\sum_{i=1}^n X_i = \sum_{k=1}^N \sum_{i\in I_k} X_i \tag{35}$$

$$\left\|\sum_{i=1}^n X_i\right\|_{\psi_2} \leq \sum_{k=1}^N \left\|\sum_{i\in I_k} X_i\right\|_{\psi_2} \qquad \text{(triangle inequality)}$$

$$\leq C\sum_{k=1}^N \sqrt{|I_k|}\nu \qquad \text{(Fact 2)}$$

$$\leq C\sqrt{nN}\nu \tag{36}$$

■

C.2. Term-by-term bounds

Lemma 12 *The random matrices $Z^\intercal X$ and $Z^\intercal Y$ satisfy*

$$\begin{aligned}\left\|Z^\intercal X - \mathbb{E}\, Z^\intercal X\right\|_{\psi_2} &\le C\sqrt{nN}\tilde{\nu}_{zx} \\ \left\|Z^\intercal X - Z^\intercal X^\star\right\|_{\psi_2} &\le C\left(n\bar{\nu}_{zx} + \sqrt{nN}\tilde{\nu}_{zx}\right) \\ \left\|Z^\intercal Y - Z^\intercal Y^\star\right\|_{\psi_2} &\le C\left(n\bar{\nu}_{zy} + \sqrt{nN}\tilde{\nu}_{zy}\right)\end{aligned}$$

Proof This follows from combining assumption a with Proposition 11. ■

Lemma 13 (Bounds on S^{-1}) *For all $r \ge 1$, the matrix S^{-1} satisfies*

$$\left\|S^{-1}\right\|_r \le \gamma(r; \lambda, \sigma^2 - \lambda, \sqrt{nN}\tilde{\nu}_{zx})$$

where γ is the function defined in Proposition 9.

Proof By the SVD,

$$\left\|S^{-1}\right\| = \frac{1}{\sigma_{\min}(S)}, \tag{37}$$

so our next task is to bound $\sigma_{\min}(S)$ from below. We have

$$\begin{aligned}\sigma_{\min}(S) &= \max\left(\lambda, \sigma_{\min}(Z^\intercal X)\right) && \text{(by } S = [Z^\intercal X]_{\vee\lambda}\text{)} \\ &= \lambda + \left(\sigma_{\min}(Z^\intercal X) - \lambda\right)_+ && \text{(by the identity } \max(a,b) = a + (b-a)_+\text{)} \\ &\ge \lambda + \left(\sigma_{\min}\left(\mathbb{E}\, Z^\intercal X\right) - \lambda - \sigma_{\max}\left(Z^\intercal X - \mathbb{E}\, Z^\intercal X\right)\right)_+ && \text{(Weyl's inequality)} \\ &\ge \lambda + \left(\sigma^2 - \lambda - D\right)_+, \quad D = \sigma_{\max}\left(Z^\intercal X - \mathbb{E}\, Z^\intercal X\right) \\ & && \text{(persistence of excitation hypothesis)}\end{aligned} \tag{38}$$

By Lem 12, $Z^\intercal X - \mathbb{E}\, Z^\intercal X$ is subgaussian with constant $C\sqrt{nN}\tilde{\nu}_{zx}$. By Fact 1, for all $t \ge 0$

$$\mathbb{P}\left(\left\|Z^\intercal X - \mathbb{E}\, Z^\intercal X\right\| \ge t\right) \le \exp\left(-\frac{t^2}{\left(C\sqrt{nN}\tilde{\nu}_{zx}\right)^2}\right).$$

Now (37) becomes

$$\left\|\hat{S}^{-1}\right\| \le \frac{1}{\lambda + (\sigma^2 - \lambda - D)_+}, \tag{39}$$

which is amenable to Lemma 9 with constants $a = \lambda$, $b = \sigma^2 - \lambda$, and $K = L_{ZX} = \sqrt{nN}\tilde{\nu}_{zx}$. ■

Appendix D. Construction of local polynomial filters

Proposition 14 *For any window size $N > 0$, step size $h > 0$ and location $i_0 \in \mathbb{R}$, there exist coefficients $\mathbf{D}_{d,h}^{k,i_0,N}$, $d \in [0 \dots m]$, $k \in [1 \dots N]$, such that for all polynomials f of degree at most $p-1 < N$,*

$$\frac{\mathrm{d}^d}{\mathrm{d}x^d} f(i_0 h) = \sum_{k=1}^{N} \mathbf{D}_{d,h}^{k,i_0,N} f(kh).$$

Considered as a matrix in (d,k),

$$\left\| \mathbf{D}_{\cdot,h}^{\cdot,i_0,N} \right\| \leq C(p,i_0) N^{-m-\frac{1}{2}} h^{-m}.$$

Proof We prescribe $\mathbf{D}$ as a solution to the following convex program:

$$\begin{aligned} \min_{\mathbf{D} \in \mathbb{R}^{m \times N}} \quad & \|\mathbf{D}\|_{\mathrm{F}} \\ \text{subject to} \quad & \mathbf{D}A = B \end{aligned} \tag{40}$$

where $A \in \mathbb{R}^{N \times p}$ and $B \in \mathbb{R}^{(m+1) \times p}$ are given by

$$A_{ij} = \left. \frac{(x - i_0 h)^j}{N^j} \right|_{x=ih} = \frac{(i-i_0)^j h^j}{N^j} \tag{41a}$$

$$B_j^d = \left. \frac{\frac{\mathrm{d}^d}{\mathrm{d}x^d} (x - i_0 h)^j}{N^j} \right|_{x=i_0 h} = \delta_{dj} \frac{d!}{N^d h^d} \tag{41b}$$

$$i \in [1 \dots N] \tag{41c}$$

$$j \in [0 \dots p-1] \tag{41d}$$

$$d \in [0 \dots m] \tag{41e}$$

Explicit solution Write the Frobenius inner product as $\langle X, Y \rangle_{\mathrm{F}} = \operatorname{tr}(X^\intercal Y)$. Let $\Lambda \in \mathbb{R}^{(m+1) \times p}$ be a Lagrange multiplier, and form the Lagrangian $\frac{1}{2} \langle D, D \rangle_{\mathrm{F}} - \langle \Lambda, DA - B \rangle_{\mathrm{F}}$. First-order optimality yields $D = \Lambda A^\intercal$. Right-multiplying by A, we get $B = \Lambda (A^\intercal A)$ which can be solved for Λ. The result is the min-norm solution $D = B(A^\intercal A)^{-1} A^\intercal$.

To bound D, use

$$\|D\| \leq \|B\| \left\| (A^\intercal A)^{-1} A^\intercal \right\| \tag{42}$$

$$\leq \frac{\|B\|}{\sigma_{\min}(A)} \tag{43}$$

Estimates To estimate $\sigma_{\min}(A) = \lambda_{\min}(A^\intercal A)^{1/2}$, notice that

$$(A^\intercal A)_{jk} = \sum_{i=1}^{N} \left(\frac{i - i_0}{N} \right)^{j+k} \tag{44}$$

is a right Riemann sum. Evaluating the integral (with an error estimate),

$$\left(\tilde{A}^\intercal \tilde{A}\right)_{jk} = \frac{N}{j+k+1} + S_{jk} \tag{45}$$

$$\left|S_{jk}\right| \leq \frac{2p}{N}. \tag{46}$$

As a consequence of this rescaling, we have the estimate

$$\left\| \left(\tilde{A}^\intercal \tilde{A}\right)^{-1} \right\| \leq C(p) N^{-1}. \tag{47}$$

■

Remark 15 (Numerics of $\mathbf{D}$) *Note that $A^\intercal A$ is a notoriously ill-conditioned Hilbert matrix (46). For numerical stability, we solve for $\mathbf{D}$ by rewriting the conditions (41) in a basis of Legendre polynomials.*

Proposition 16 (Bias) *For some i_0, h, and N, let $\mathbf{D}$ be result of Proposition 14. Let $[x_0, x_1]$ be an interval and $f \in C^p([x_0, x_1], \mathbb{R})$ with $R_p := \sup_{t\in[x_0,x_1]} |f^{(p)}(t)|$. Assume that $d < p$. Then*

$$\left| \frac{\mathrm{d}^d f}{\mathrm{d}x^d}(x + i_0 h) - \sum_{k=1}^{N} \mathbf{D}_{d,h}^{k,i_0,N} f(x + kh) \right| \leq C(m,p) R_p (Nh)^{p-m}.$$

Proof Expanding around $x + i_0 h$,

$$f(x+kh) = \sum_{\nu=0}^{p-1} \frac{f^{(\nu)}(x+i_0 h)}{\nu!} ((k-i_0)h)^\nu + R(k), \tag{48}$$

$$|R(k)| \leq C(m,p)\, R_p\, (Nh)^p. \tag{49}$$

Contracting the dth row of $\mathbf{D}$ with $\{f(x+kh)\}_{k=1}^N$,

$$\sum_{k=1}^{N} \mathbf{D}_d^k f(x+kh) = \sum_{\nu=0}^{p-1} \frac{f^{(\nu)}(x+i_0 h)}{\nu!} \sum_{k=1}^{N} \mathbf{D}_d^k ((k-i_0)h)^\nu + \sum_{k=1}^{N} \mathbf{D}_d^k R(k), \tag{50}$$

$$= f^{(d)}(x+i_0 h) + \sum_{k=1}^{N} \mathbf{D}_d^k R(k), \tag{51}$$

where the last equality uses the constraints (41). Therefore,

$$\left| f^{(d)}(x+i_0 h) - \sum_{k=1}^{N} \mathbf{D}_d^k f(x+kh) \right| \leq \left\|\mathbf{D}_{\cdot,d}\right\|_2 \left\|R(\cdot)\right\|_2 \tag{52}$$

By Cauchy–Schwarz, $\left|\sum_{k=1}^N \mathbf{D}_d^k R(k)\right| = |\langle \mathbf{D}_{\cdot,d}, R\rangle| \leq \|\mathbf{D}_{\cdot,d}\|_2 \|R\|_2$. Moreover, $\|R\|_2 \leq \sqrt{N} \sup_k |R(k)|$ and $\|\mathbf{D}_{\cdot,d}\|_2 \leq \|\mathbf{D}\|$, yielding:

$$\leq \|\mathbf{D}\| \sqrt{N} \sup_k |R(k)|. \tag{53}$$

By Proposition 14, $\|\mathbf{D}\| \leq C(p, i_0) N^{-m-\frac{1}{2}} h^{-m}$. Combining with (49) yields the claim. ■

Proposition 17 (Fluctuation) *For some i_0, h, and N, let $\mathbf{D}$ be the result of Proposition 14. Let $\{w_k\}_{k=1}^N$ be independent, mean-zero, and subgaussian with $\nu := \max_{k\in[1\dots N]} \|w_k\|_{\psi_2} < \infty$. Then for any $d < p$,*

$$\left\| \sum_{k=1}^{N} \mathbf{D}_{d,h}^{k,i_0,N} w_k \right\|_{\psi_2} \leq C(p, i_0)\, \nu\, N^{-m-\frac{1}{2}} h^{-m}.$$

Proof Write $a_k = \mathbf{D}_d^k$. Since $\mathbb{E}\, w_k = 0$, the sum equals $\sum_{k=1}^N a_k w_k$. By Fact 2,

$$\left\| \sum_{k=1}^{N} a_k w_k \right\|_{\psi_2}^2 \leq C \sum_{k=1}^{N} |a_k|^2 \|w_k\|_{\psi_2}^2 \leq C\nu^2 \sum_{k=1}^{N} a_k^2 = C\nu^2 \|\mathbf{D}_{\cdot,d}\|_2^2.$$

Finally, $\|\mathbf{D}_{\cdot,d}\|_2 \leq \|\mathbf{D}\|$ and Proposition 14 gives $\|\mathbf{D}\| \leq C(p, i_0) N^{-m-1/2} h^{-m}$, which completes the proof. ∎

Appendix E. Proof of Corollary 2

The latent model equation $Y^\star = X^\star \theta^\star$ holds if the hatted operators $\hat{\mathcal{H}}$, $\hat{\mathcal{G}}$ are replaced by the true operators. and $\theta^\star = \theta_0$. The n of the theorem should be replaced with the number of filter windows $\sim n$. The N of the theorem applies to the N of the estimator. Now we turn to verifying the data bounds. For $\bar{\nu}_{zy}$,

$$\begin{aligned}
\left\| \mathbb{E}\, z_i y_i^\intercal - z_i (y_i^\star)^\intercal \right\| &= \left\| \mathbb{E}\, z_i\, \mathbb{E}\left(y_i^\intercal - (y_i^\star)^\intercal \right) \right\| && \text{(independence of } z_i \text{ and } y_i\text{)} \\
&\leq \|\mathbb{E}\, z_i\| \left\| \mathbb{E}\left(y_i^\intercal - (y_i^\star)^\intercal \right) \right\| \\
&\leq \mu \left\| \mathbb{E}\left(y_i^\intercal - (y_i^\star)^\intercal \right) \right\| && \text{(boundedness of } z_i \text{ §4)} \\
&\leq C\mu (Nh)^{p-d}. && \text{(bias of Lemmas 5, 7)}
\end{aligned}$$

where $d = 1$ in continuous-time and $d = 0$ in discrete-time. For $\bar{\nu}_{zx}$, the same reasoning yields

$$\begin{aligned}
\left\| \mathbb{E}\, z_i x_i^\intercal - z_i (x_i^\star)^\intercal \right\| &= \left\| \mathbb{E}\, z_i\, \mathbb{E}\left(x_i^\intercal - (x_i^\star)^\intercal \right) \right\| && \text{(independence of } z_i \text{ and } x_i\text{)} \\
&\leq \|\mathbb{E}\, z_i\| \left\| \mathbb{E}\left(x_i^\intercal - (x_i^\star)^\intercal \right) \right\| \\
&\leq \mu \left\| \mathbb{E}\left(x_i^\intercal - (x_i^\star)^\intercal \right) \right\| && \text{(boundedness of } z_i \text{ §4)} \\
&\leq C\mu \left[(Nh)^p + N^{-1/2} \right]. \\
&\text{(bias and subgaussian norm of Lemmas 5, 7, propagated through Lipschitz } \phi\text{, Assumption 1)}
\end{aligned}$$

For $\tilde{\nu}_{zy}$,

$$\begin{aligned}
\left\| \mathbb{E}\, z_i y_i^\intercal - z_i y_i^\intercal \right\|_{\psi_2} &\leq \left\| z_i y_i^\intercal \right\|_{\psi_2} && \text{(centering)} \\
&\leq \|z_i\|_{\psi_2} \left\| y_i^\intercal \right\|_{\psi_2} = \|z_i\|_{\psi_2} \left\| y_i^\star + \left(\mathbb{E}\, y_i - y_i^\star \right) + \left(\mathbb{E}\, y_i - \mathbb{E}\, y_i \right) \right\|_{\psi_2} \\
&\leq C\mu \left(|y_i^\star| + (Nh)^{p-d} + N^{-d-1/2} h^{-d} \right). \\
&\text{(subgaussian norms Lemmas 5, 7)}
\end{aligned}$$

where $d = 1$ in continuous-time and $d = 0$ in discrete-time. The same bound applies to $\tilde{\nu}_{zx}$ after taking $d = 0$. Balancing the two terms,

$$(Nh)^{p-d} = N^{-d-1/2}h^{-d} \implies N = h^{-2p/(2p+1)}.$$

Thus the ideal scaling is

$$\begin{aligned}
\bar{\nu}_{zy} &\lesssim \mu h^{(p-d)/(2p+1)} \\
\bar{\nu}_{zx} &\lesssim \mu h^{p/(2p+1)} \\
\tilde{\nu}_{zy} &\lesssim \mu \left(\sup_i |y_i^\star| + h^{(p-d)/(2p+1)} \right) \\
\tilde{\nu}_{zx} &\lesssim \mu \left(\sup_i |x_i^\star| + h^{p/(2p+1)} \right)
\end{aligned}$$

The function $\gamma = \gamma_{\text{head}} + \gamma_{\text{body}} + \gamma_{\text{tail}}$ becomes:

$$\gamma_{\text{head}} = \frac{2}{\sigma^2 - \lambda} \lesssim \frac{1}{n - \lambda}$$

and

$$\begin{aligned}
\gamma_{\text{tail}}(r; a, b, K) &= \frac{1}{\lambda} \exp\left[-\frac{C(\sigma^2)^2}{r\left(\sqrt{nh^{-2p/(2p+1)}}\mu \sup_i |y_i^\star|\right)^2} \right] \\
&\lesssim \frac{1}{\lambda} \exp\left(-Cn^3 h^{2p/(2p+1)} \right),
\end{aligned}$$

with γ_{body} having a subdominant contribution. Thus the condition for γ_{head} to dominate is that for all C,

$$\exp\left(-Cn^3 h^{2p/(2p+1)} \right) \ll \lambda \ll n.$$

This also renders the first term in the conclusion of Theorem 10 to be of subdominant order, resulting in the scaling in n and h,

$$\begin{aligned}
\left\| \hat{\theta} - \theta^\star \right\|_q &\lesssim \left(\bar{\nu}_{zx} + \bar{\nu}_{zy} \right) + \sqrt{\frac{N}{n}} \left(\tilde{\nu}_{zx} + \tilde{\nu}_{zy} \right) \\
&\lesssim h^{(p-d)/(2p+1)} + \sqrt{\frac{1}{nh^{2p/(2p+1)}}}.
\end{aligned}$$

Appendix F. Supplement to §5

F.1. Data-generating process

We are given n measurements of ξ at $\{ih\}_{i=1}^n$, with i.i.d. Gaussian noise of mean zero and variance η:

$$z_i = \xi(ih) + \mathcal{N}(0, \eta)$$

The number of measurements is $n = 100000$; the sampling period is $h = 0.001$.

The true parameter is given by

$$\theta_0 = \begin{pmatrix} 0 & 0 & 1 \\ -10 & 28 & 0 \\ 10 & -1 & 0 \\ 0 & 0 & -8/3 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix},$$

and the initial condition is $\xi(0) = (-8, 8, 27)$.

In the input signal, the excitation frequency is $f = 1$.

F.2. Continuous-time estimator

We approximate these operators using local polynomial regression with a filter window size $N = 100$ and accuracy order $p = 75$. The measurement noise variance is $\eta = 0.1$. The estimator hyperparameters are $\lambda = 10$ and $\mu = 200$.

F.3. Discrete-time estimator

We approximate these operators using local polynomial regression with a filter window size $N = 100$ and accuracy order $p = 75$. The measurement noise variance is $\eta = 1$. The estimator hyperparameters are $\lambda = 10$ and $\mu = 200$.

F.4. Reporting

All values are normalized by the Frobenius norm of the (pseudo-) true parameter. The bias is computed as the Frobenius distances between the mean of the estimator and the (pseudo-) true parameter. The standard deviation is computed as the quadratic mean of the Frobenius distance between the estimator and its mean. The root mean square error is computed as the quadratic mean of the Frobenius distance between the estimator and the (pseudo-) true parameter.

We run 2000 Monte Carlo trials for each estimator and compute standard errors by bootstrapping.

F.5. Kernel density plots

Appendix G. Supplement to §5.3

G.1. Data-generating process

We observe n scalar measurements of $x(t)$ at $\{ih\}_{i=1}^n$, with i.i.d. Gaussian noise of mean zero and variance η:

$$z_i = x(ih) + \mathcal{N}(0, \eta)$$

The number of measurements is $n = 100000$, the sampling period is $h = 0.001$. The Van der Pol parameter is $\mu = 2$ and the initial condition is $(x(0), \dot{x}(0)) = (2, 0)$.

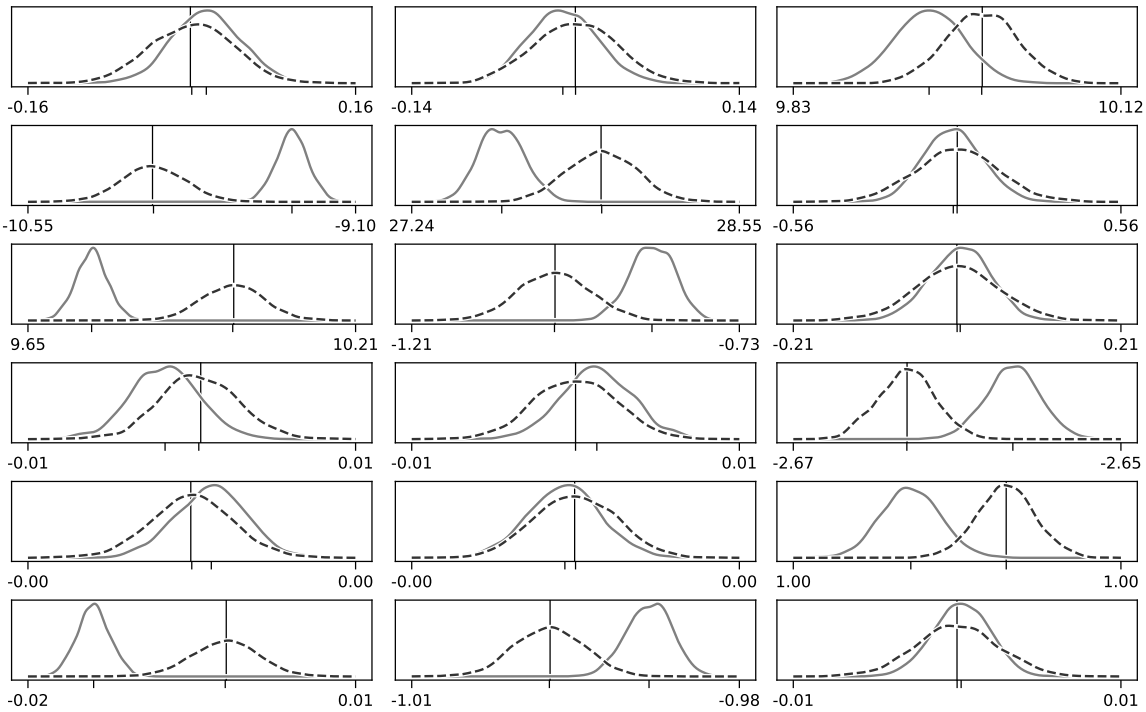

Figure 3: Continuous-time Lorenz: elementwise marginal kernel density estimates of the sampling distributions of our estimator (dashed) and a baseline estimator (solid). Vertical line indicates ground truth; ticks indicate mean of sampling distribution.

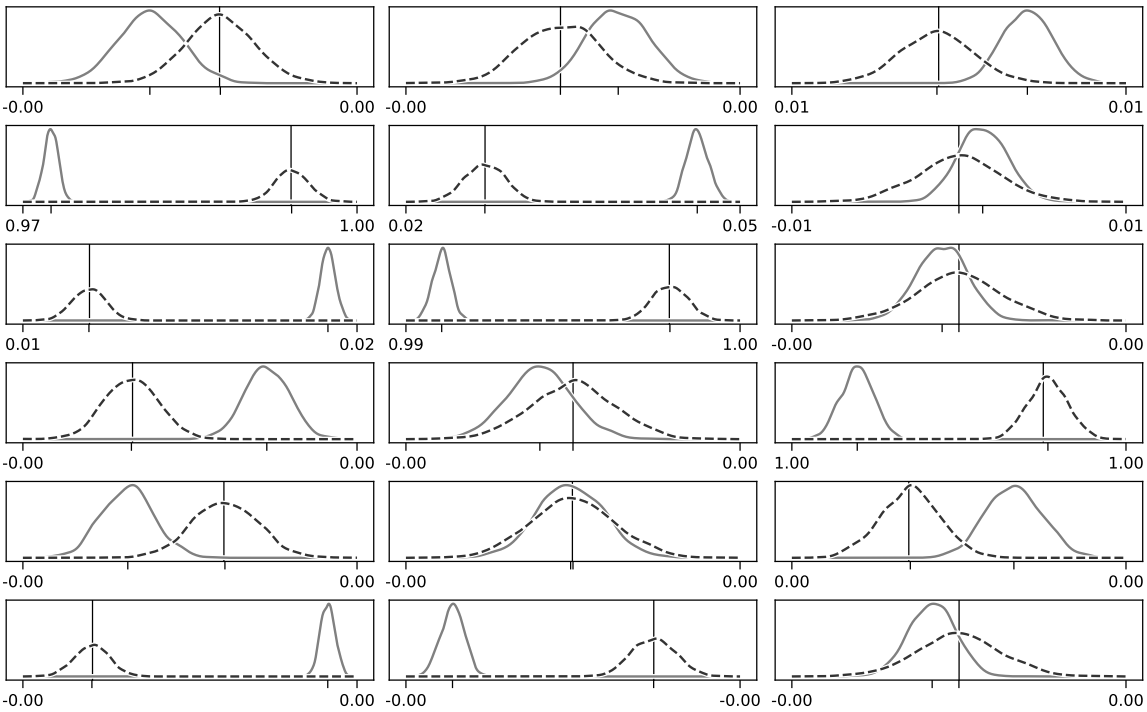

Figure 4: Discrete-time Lorenz: Elementwise marginal kernel density estimates of the sampling distributions of our estimator (dashed) and a baseline estimator (solid). Vertical line indicates pseudo-true value; ticks indicate mean of sampling distribution.

Estimator	abs. bias (%)	std (%)	rmse (%)
Instrumental Variables (ours)	$0.489(258)$	$13.205(211)$	$13.214(211)$
Least Squares	$17.533(18)$	$1.622(22)$	$17.608(18)$

Table 3: Comparison of bias, standard deviation, and root mean square error of our estimator and a least squares estimator for the parameter of the Van der Pol oscillator. Monte Carlo standard errors in parentheses.

G.2. Continuous-time estimator

We approximate $\mathcal{H}$ and $\mathcal{G}$ using local polynomial regression with a filter window size $N = 100$ and accuracy order $p = 20$. The stencil outputs derivatives of order $d = 0, 1, 2$: rows $d = 0, 1$ implement $\hat{\mathcal{G}}$ (or $\tilde{\mathcal{G}}$) and row $d = 2$ implements $\hat{\mathcal{H}}$. The measurement noise variance is $\eta = 10^{-4}$. The estimator hyperparameters are $\lambda = 1$ and $\mu = 200$.

G.3. Reporting

All values are normalized by the Euclidean norm of the true parameter θ_0. The bias is computed as the Euclidean distance between the mean of the estimator and θ_0. The standard deviation is computed as the quadratic mean of the Euclidean distance between the estimator and its mean. The root mean square error is computed as the quadratic mean of the Euclidean distance between the estimator and θ_0.

We run 2000 Monte Carlo trials for each estimator and compute standard errors by bootstrapping.