Revealing Unobservables by Deep Learning

Yingyao Hu, Yang Liu, Jiaxiong Yao

JHU, IMF, IMF

University of Toronto
Oct. 6, 2023

Latent variables in microeconomic models

empirical models	unobservables	observables
measurement error	true earnings	self-reported earnings
consumption function	permanent income	observed income
production function	productivity	output, input
wage function	ability	test scores
learning model	belief	choices, proxy
auction model	unobserved heterogeneity	bids
contract model	effort, type	outcome, state var.
...	...	...

Goal of this paper

- suppose the observables satisfy independence conditional on the latent variable.
- can we back out the latent variable such that the conditional independence holds?
- in other words, can we extract the common element from observables.
- use deep neural network to impute the (pseudo) true values

Related literature

- factor model

$$X = \Lambda F + u$$

  factors in F can be "estimated"

- generated regressors, e.g., control function

$$\begin{aligned} Y &= m(X) + e \\ X &= h(Z, U), \quad Z \perp (U, e) \\ U &= F_{X|Z}(X|Z = z) \end{aligned}$$

- imputation in missing data models (and treatment effect models)

- machine learning methods for latent variables
  – Variational autoencoders
  – Generative adversarial networks
  – This paper uses a semi-nonparametric approach with deep neural network

A general framework

- observed & unobserved variables

X	measurement	observables
X^*	latent true variable	unobservables

- economic models described by distribution function f_{X^*}

$$f_X(x) = \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*) f_{X^*}(x^*) dx^*$$

f_{X^*} : latent distribution
f_X : observed distribution
$f_{X|X^*}$: relationship between observables & unobservables

The discrete case: Hu (2008)

- key identification conditions:
  1) (X, Y, Z) are independent conditional on X^*
  2) Matrices $M_{X|X^*}$ and $M_{X|Z}$ are invertible.
  3) for all $\overline{x}^* \neq \widetilde{x}^*$ in $\mathcal{X}^*$, $E_{Y|X^*}\left[w(Y)|\overline{x}^*\right] \neq E_{Y|X^*}\left[w(Y)|\widetilde{x}^*\right]$ for some $w(.)$.
  4) $Mode\left[f_{X|X^*}(\cdot|x^*)\right] = x^*$ for all $x^* \in \mathcal{X}^*$.
- then

$$f_{X,Y,Z} \text{ uniquely determines } f_{X,Y,Z,X^*}$$

  with

$$f_{X,Y,Z,X^*} = f_{X|X^*} f_{Y|X^*} f_{Z|X^*} f_{X^*}$$

- a global nonparametric point identification

Identification in the continuous case

- define a set of bounded and integrable functions containing f_{X^*}

$$\mathcal{L}^1_{bnd}\left(\mathcal{X}^*\right) = \left\{ h : \int_{\mathcal{X}^*} \left|h(x^*)\right| dx^* < \infty \text{ and } \sup_{x^* \in \mathcal{X}^*} \left|h(x^*)\right| < \infty \right\}$$

- define a linear operator

$$\begin{aligned} L_{X|X^*} &: \mathcal{L}^1_{bnd}\left(\mathcal{X}^*\right) \to \mathcal{L}^1_{bnd}\left(\mathcal{X}\right) \\ \left(L_{X|X^*} h\right)(x) &= \int_{\mathcal{X}^*} f_{X|X^*}(x|x^*) h(x^*) dx^* \end{aligned}$$

- operator equation

$$f_X = L_{X|X^*} f_{X^*}$$

- identification requires injectivity of $L_{X|X^*}$, i.e.,

$$L_{X|X^*} h = 0 \text{ implies } h = 0 \text{ for any } h \in \mathcal{L}^1_{bnd}\left(\mathcal{X}^*\right)$$

The Hu and Schennach (2008) Theorem

- key identification conditions:
  1) (X, Y, Z) are independent conditional on X^*. All densities are bounded
  2) the operators $L_{X|X^*}$ and $L_{Z|X}$ are injective.
  3) for all $\overline{x}^* \neq \widetilde{x}^*$ in $\mathcal{X}^*$, the set $\left\{y : f_{Y|X^*}\left(y|\overline{x}^*\right) \neq f_{Y|X^*}\left(y|\widetilde{x}^*\right)\right\}$ has positive probability.
  4) there exists a known functional M such that $M\left[f_{X|X^*}\left(\cdot|x^*\right)\right] = x^*$ for all $x^* \in \mathcal{X}^*$.
- then
$$f_{X,Y,Z} \text{ uniquely determines } f_{X,Y,Z,X^*}$$
  with
$$f_{X,Y,Z,X^*} = f_{X|X^*} f_{Y|X^*} f_{Z|X^*} f_{X^*}$$
- a global nonparametric point identification

A specification based on convolution

- a 3-measurement model

$$
\begin{aligned}
x_1 &= g_1(x^*) + \epsilon_1 \\
x_2 &= g_2(x^*) + \epsilon_2 \\
x_3 &= g_3(x^*) + \epsilon_3
\end{aligned}
$$

- normalization: $g_3(x^*) = x^*$
- advantage of this specification: testability of completeness

Testability of completeness in the convolution case

- a 3-measurement model

$$
\begin{aligned}
x_1 &= g_1(x^*) + \epsilon_1 \\
x_2 &= g_2(x^*) + \epsilon_2 \\
x_3 &= g_3(x^*) + \epsilon_3
\end{aligned}
$$

- $\phi_{x_1}(t) \neq 0$ implies that $\phi_{\epsilon_1}(t) \neq 0$
- Under this convolution specification and monotonicity of $g_1(\cdot)$, one can test $\phi_{x_1}(t) \neq 0$ using e.ch.f.
- under $H_0 : \phi_{x_1}(t)$ has zeros on the real line, existing algorithm can find the first zero. (Hu and Shiu, 2021)

An existing estimator: A sieve semiparametric MLE

- Based on :

$$f_{y,x|z}(y,x|z) = \int f_{y|x^*}(y|x^*) f_{x|x^*}(x|x^*) f_{x^*|z}(x^*|z) dx^*$$

- Approximate ∞-dimensional parameters, e.g., $f_{x|x^*}$, by truncated series

$$\widehat{f}_1(x|x^*) = \sum_{i=0}^{i_n} \sum_{j=0}^{j_n} \widehat{\gamma}_{ij} p_i(x) p_j(x^*),$$

– where $p_k(\cdot)$ are a sequence of known univariate basis functions.

- Sieve Semiparametric MLE

$$\begin{aligned} \widehat{\alpha} &= \left(\widehat{\beta}, \widehat{\eta}, \widehat{f}_1, \widehat{f}_2\right) \\ &= \underset{(\beta,\eta,f_1,f_2)\in\mathcal{A}_n}{\arg\max} \frac{1}{n} \sum_{i=1}^{n} \ln \int f_{y|x^*}(y_i|x^*;\beta,\eta) f_1(x_i|x^*) f_2(x^*|z_i) dx^* \end{aligned}$$

$$\begin{cases} \beta : & \text{parameter vector of interest} \\ \eta, f_1, f_2 : & \infty\text{-dimensional nuisance parameters} \\ \mathcal{A}_n : & \text{space of series approximations} \end{cases}$$

Uncorrelated deviations and identification in observation

- Question: Are the true values in each observation identified?
- Let X_i^* be a draw of X^* and we define *an uncorrelated deviation* from that draw as

$$X_i^* + \delta_i \quad \textit{with} \quad E(X_i^* \delta_i) = E(\delta_i) = 0 \tag{1}$$

  where (X_i^*, δ_i) is a i.i.d. random draw from the joint distribution of (X^*, δ).

Corollary

Suppose that the assumptions in Theorem HS2008 hold. Given an observed sample $\{X_i^1, X_i^2, ..., X_i^k\}$, which is a subset of the infeasible full sample $\{X_i^1, X_i^2, ..., X_i^k, X_i^*\}$, no uncorrelated deviation from latent draws X_i^*, defined in Equation (1), is observationally equivalent to X_i^*.

-

Uncorrelated deviations and identification in observation

- The identification result in 1 can be extended to the case where δ_i is dependent of the observables $(X^1, X^2, ..., X^k)$ because the conditional distribution $f(X^*|X^1, X^2, ..., X^k)$ is identified by the HS2008 Theorem.
- We define *a conditionally uncorrelated deviation* from X_i^* as

$$X_i^* + \delta_i \quad with \quad E(X_i^* \delta_i \,|\, X_i^1, X_i^2, ..., X_i^k) = E(\delta_i \,|\, X_i^1, X_i^2, ..., X_i^k) = 0 \tag{2}$$

where $(X_i^*, \delta_i, X_i^1, X_i^2, ..., X_i^k)$ is a i.i.d. random draw from their corresponding joint distribution.

Conditionally uncorrelated deviations

Corollary

Suppose that the assumptions in Theorem HS2008 hold. Given an observed sample $\{X_i^1, X_i^2, ..., X_i^k\}$*, which is a subset of the infeasible full sample* $\{X_i^1, X_i^2, ..., X_i^k, X_i^*\}$*, no conditionally uncorrelated deviation from latent draws* X_i^**, defined in Equation (2), is observationally equivalent to* X_i^**.*

- from identification in distribution to identification in observation

convergence argument

- suppose our estimator $\hat{X}_i^* = X_i^* + \delta_i$ with uncorrelated δ_i across observations, which may not be true.
- the identification argument makes sure the distribution of $X_i^* + \delta_i$ is consistent with that of X_i^*.
- our identification results suggest that our estimates $\hat{X}_i^*$ should have the same distribution (and variance) as X_i^*. Then the sample moments of $\hat{X}_i^*$ should converges to the true moments. In other words, we have

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{X}_i^*)^2 - \frac{1}{N}\sum_{i=1}^{N}(X_i^*)^2 = o_p(1) \tag{3}$$

convergence argument

Theorem

Suppose that the estimator $\hat{X}_i^* = X_i^* + \delta_i$ *for* $i = 1, 2, ..., N$ *satisfying*

$$\frac{1}{N}\sum_{i=1}^{N} X_i^* \delta_i = o_p(1). \tag{4}$$

Then, the consistency of the sample moment in Equation 3, implies that for any $\epsilon > 0$*, the sample proportion of large deviations goes to zero, i.e.,*

$$P_N\left(\left|\hat{X}_i^* - X_i^*\right| > \epsilon\right) := \frac{1}{N}\sum_{i=1}^{N} I(|\delta_i| > \epsilon) = o_p(1)$$

- That means for any $\epsilon > 0$, proportion of large mistakes goes to zero.
- no convergence argument for a fixed $X_{i_0}^*$
- Probably Approximately Correct (PAC) bounds?

$$lim_{N\to\infty} P\left(\left|\hat{X}_{i_0}^* - X_{i_0}^*\right| > \epsilon\right) < \eta$$

Point identification in observation

- $$\mathcal{P}_{X,X^*} = \{(x_i, x_i^*) : i = 1, 2, ..., \infty\}.$$
- Assumption: No two different subjects in the population are observationally identical, i.e., for any (x_i, x_i^*) and (x_j, x_j^*) in $\mathcal{P}_{X,X^*}$, $i \neq j$ implies $x_i \neq x_j$.

Theorem

Suppose that the assumptions in Hu and Schennach (2008) and Assumption above hold. Given an observed sample $\{X_i^1, X_i^2, ..., X_i^k\}$, which is a subset of the infeasible full sample $\{X_i^1, X_i^2, ..., X_i^k, X_i^*\}$, the realization of X_i^* in observation i is uniquely determined by the realization of $\{X_i^1, X_i^2, ..., X_i^k\}$ in the observation. In particular,

$$x_i^* = E[X^* | X_i^1 = x_i^1, X_i^2 = x_i^2, ..., X_i^k = x_i^k].$$

-

Estimation: Latent variable models with machine learning

- Variational autoencoders
- Generative adversarial networks
- This paper uses a semi-nonparametric approach with deep neural network

Variational Autoencoders

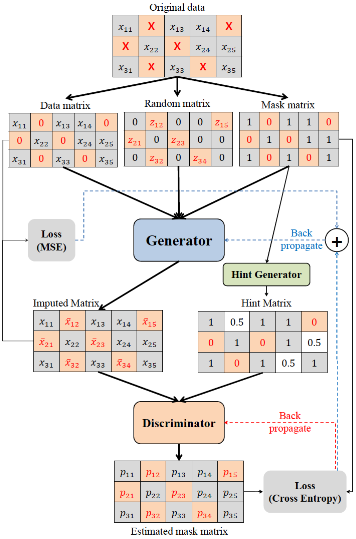
It uses a parametric specification to approximate the distribution

$$\begin{aligned}
\ln f_{X;\theta} &= \ln \int f_{X,X^*;\theta} dx^* \\
&= \ln \int f_{X,X^*;\theta} \frac{\hat{f}_{X^*;\lambda}}{\hat{f}_{X^*;\lambda}} dx^* \\
&\geq \int \hat{f}_{X^*;\lambda} ln \frac{f_{X,X^*;\theta}}{\hat{f}_{X^*;\lambda}} dx^* \\
&= E_{\hat{f}_{X^*;\lambda}} \left[ln \frac{f_{X,X^*;\theta}}{\hat{f}_{X^*;\lambda}} \right] \\
&= ELBO(X;\theta,\lambda)
\end{aligned}$$

The Evidence Lower Bound (ELBO) admits a tractable unbiased Monte Carlo estimator

$$\max_{\theta} \sum_{X} \max_{\lambda} E_{\hat{f}_{X^*;\lambda}} \left[ln \frac{f_{X,X^*;\theta}}{\hat{f}_{X^*;\lambda}} \right] \tag{5}$$

Generative Adversarial Imputation Nets (GAIN) (Yoon et al, 2018)



Our semi-nonparametric estimator

We use a deep neural network G to generate the unobservable satisfying the conditional independence. Let $\vec{V}$ stand for the vector of draws of variable V in the sample, i.e.,

$$\vec{X}^* = (X_1^*, X_2^*, ..., X_N^*)^T \tag{6}$$

$$\vec{X}^j = (X_1^j, X_2^j, ..., X_N^j)^T \tag{7}$$

We generate $\vec{X}^*$ as follows:

$$\vec{X}^* = G(\vec{X}^1, \vec{X}^2, ..., \vec{X}^k). \tag{8}$$

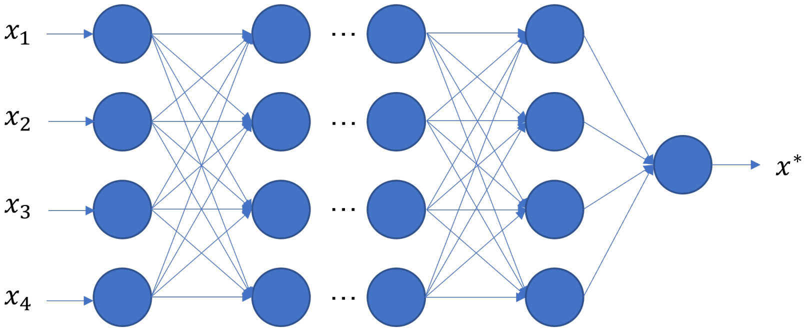


Figure: DNN model to generate $\vec{X}^*$

- $hidden_{l+1} = ReLU(W_l \times hidden_l + b_l)$
- Rectified Linear Units use activation function: $ReLU(z) = m\{0, z\}$

Latent variable models: Estimation

We train G to minimize the Kullback–Leibler divergence

$$\min_{G} D_{KL}\left(\hat{p} \,||\, \hat{p}_{ci}\right) \tag{9}$$

with

$$\hat{p} = \hat{f}_{X^1,X^2,\ldots,X^k,X^*}$$

and

$$\hat{p}_{ci} = \hat{f}_{X^1|X^*}\hat{f}_{X^2|X^*}\ldots\hat{f}_{X^k|X^*}\hat{f}_{X^*}$$

where $\hat{f}$ are empirical distribution functions based on sample $(\vec{X}^1, \vec{X}^2, \ldots, \vec{X}^k, \vec{X}^*)$.

Simulation

$$X_i^j = m^j(X_{i,true}^*) + \epsilon_i^j \tag{10}$$

for $j = 1, 2, ..., k$ and $i = 1, 2, ..., N$. WLOG, we normalize

$$m^1(x) = x$$

and

$$E[\epsilon^j | X_{true}^*] = 0.$$

baseline case

$$
\begin{aligned}
k &= 4 \\
m^1(x) &= x \\
m^2(x) &= \frac{1}{1+e^x} \\
m^3(x) &= x^2 \\
m^4(x) &= \ln(1+exp(x)) \\
\epsilon^1 &= N(0,1) \\
\epsilon^2 &= Beta(2,2) - \frac{1}{2} \\
\epsilon^3 &= Laplace(0,1) \\
\epsilon^4 &= Uniform(0,1) - \frac{1}{2} \\
X^* &= N(0,4)
\end{aligned}
$$

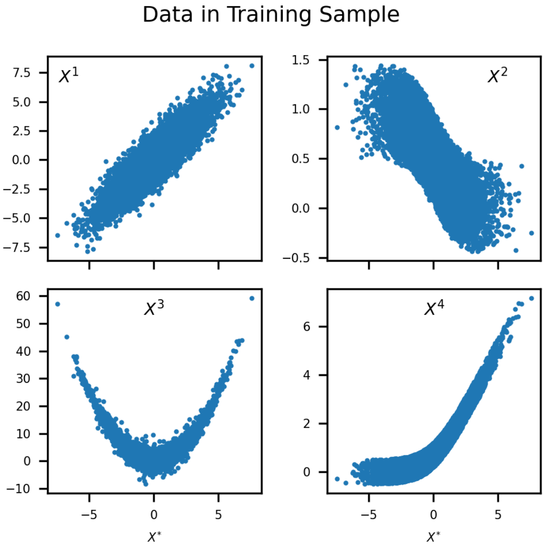


Figure: Baseline Training Sample

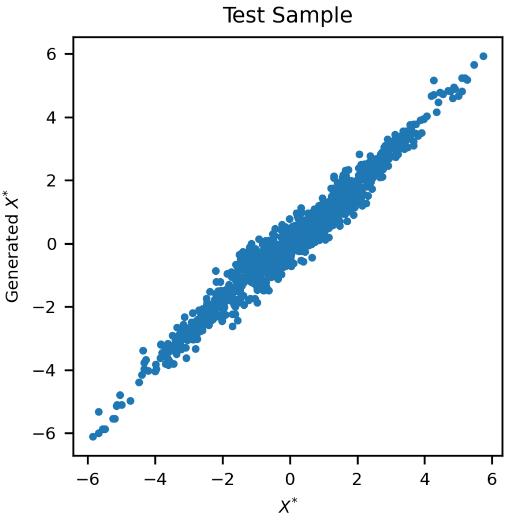


Figure: Results in Baseline Experiment

Case 2: Error terms correlate with X^*

$$
\begin{aligned}
k &= 4 \\
m^1(x) &= x \\
m^2(x) &= \frac{1}{1+e^x} \\
m^3(x) &= x^2 \\
m^4(x) &= \ln(1+exp(x)) \\
\epsilon^1 &= N(0, \frac{1}{4}x^2) \\
\epsilon^2 &= Beta(2,2) - \frac{1}{2} \\
\epsilon^3 &= Laplace(0, 0.5|x|) \\
\epsilon^4 &= Uniform(0, 0.5|x|) - \frac{1}{4}|x| \\
X^* &= N(0,4)
\end{aligned}
$$

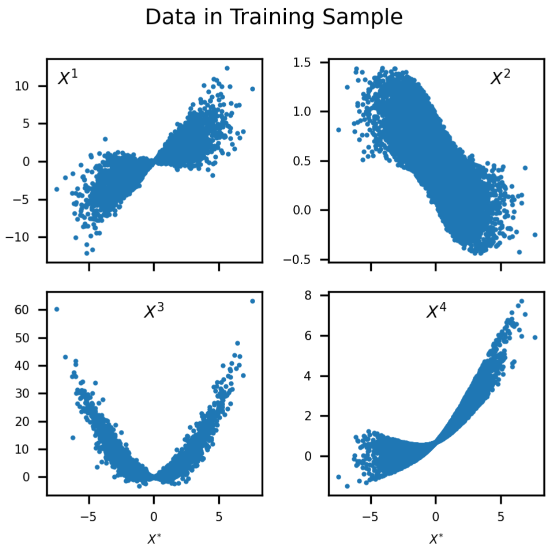


Figure: Linear Error Training Sample

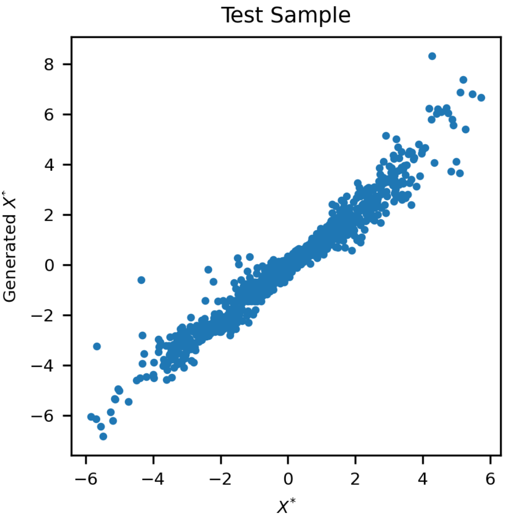


Figure: Results in Linear Error Experiment

Case 3: Larger error variances

$$
\begin{aligned}
k &= 4 \\
m^1(x) &= x \\
m^2(x) &= \frac{1}{1+e^x} \\
m^3(x) &= x^2 \\
m^4(x) &= \ln(1+exp(x)) \\
\epsilon^1 &= N(0,4) \\
\epsilon^2 &= Beta(2,4) - \frac{1}{3} \\
\epsilon^3 &= Laplace(0,2) \\
\epsilon^4 &= Uniform(0,2) - 1 \\
X^* &= N(0,4)
\end{aligned}
$$

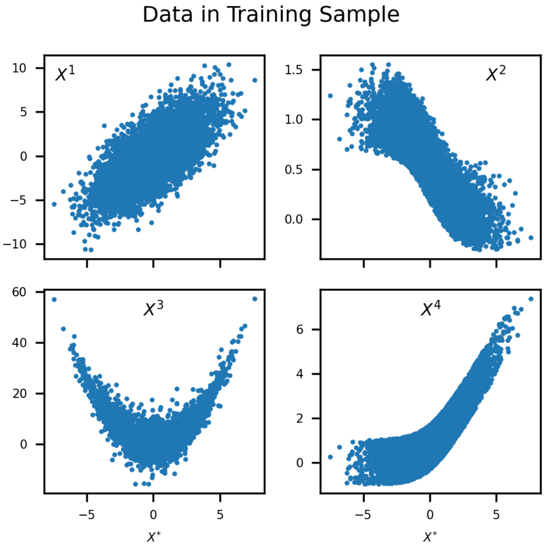


Figure: Double Error Training Sample

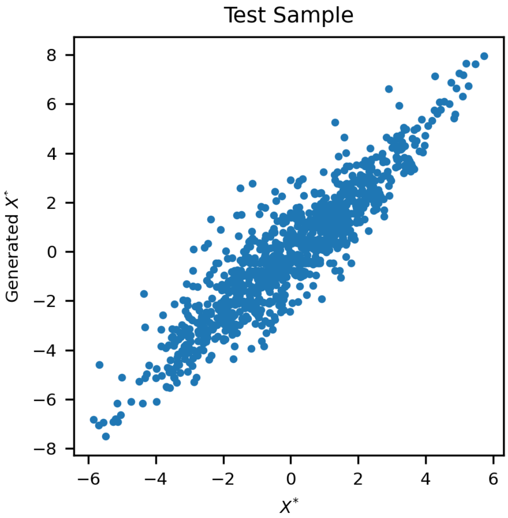


Figure: Results in Double Error Experiment

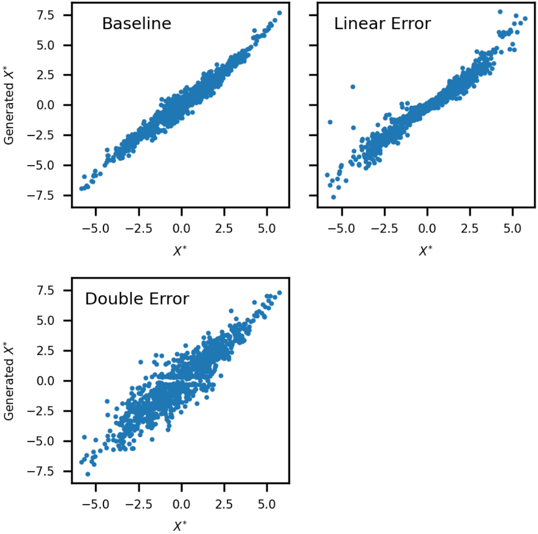


Figure: Results in the First Three Experiments

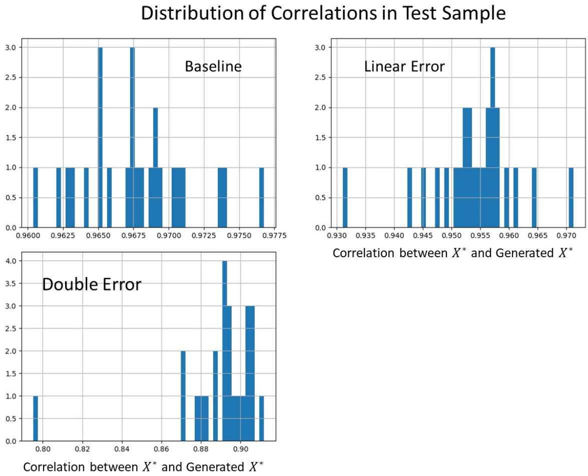


Figure: Results in the First Three Experiments

Case 4: Without normalization

$$
\begin{aligned}
k &= 4 \\
m^1(x) &= x^2 + x \\
m^2(x) &= \frac{1}{1+e^x} \\
m^3(x) &= x^2 \\
m^4(x) &= \ln(1 + exp(x)) \\
\epsilon^1 &= N(0,1) \\
\epsilon^2 &= Beta(2,2) - \frac{1}{2} \\
\epsilon^3 &= Laplace(0,1) \\
\epsilon^4 &= Uniform(0,1) - \frac{1}{2} \\
X^* &= N(0,4)
\end{aligned}
$$

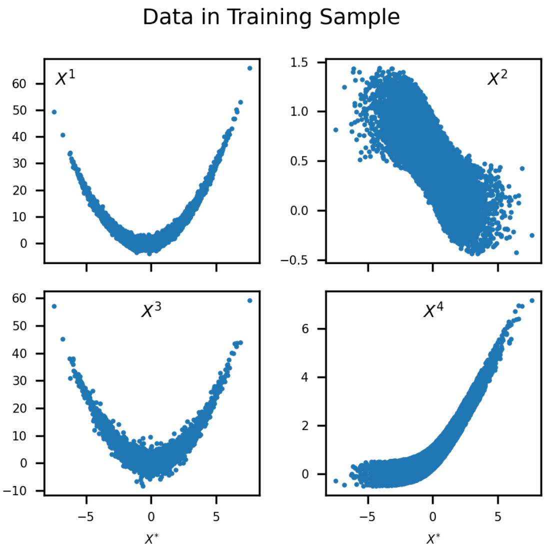


Figure: No Normalization Training Sample

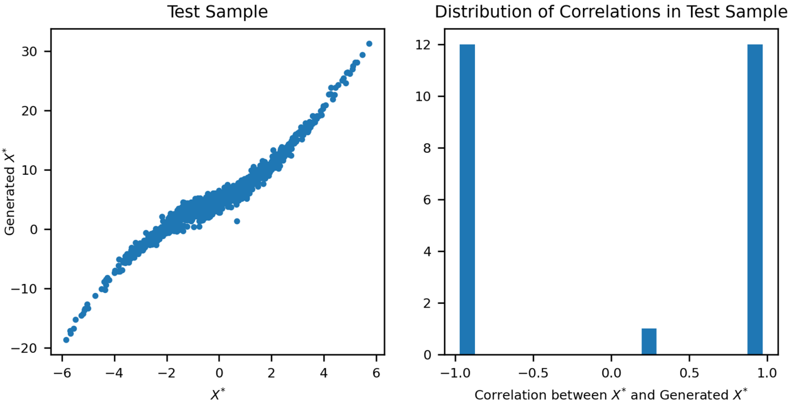
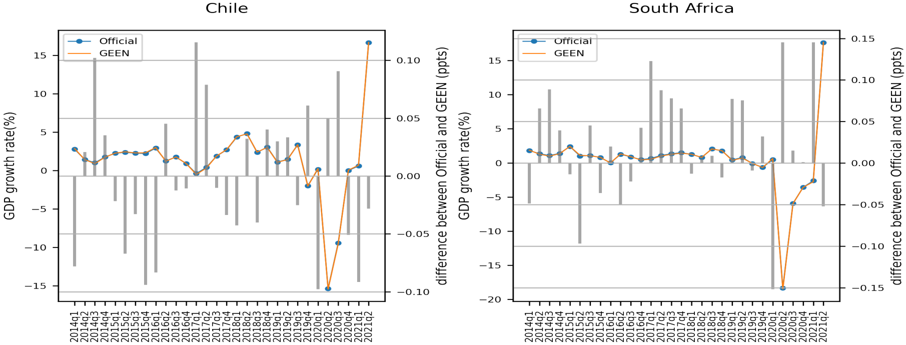


Figure: Results in No Normalization Experiment

Empirical application: Refining GDP growth

- Estimation of GDP growth with three measures
  - Official data
  - Nighttime light growth (Hu and Yao, 2022)
  - Google Search Volume growth (Woloszko, 2021)

Examples: official and refined GDP growth well aligned



- Both Chile and South Africa have differences within 0.15 percentage points despite volatile economic growth
- Suggests that GEEN could be useful in leveraging alternative data to understand economic activity of countries without timely official GDP data

Examples: official GDP data excessively smooth

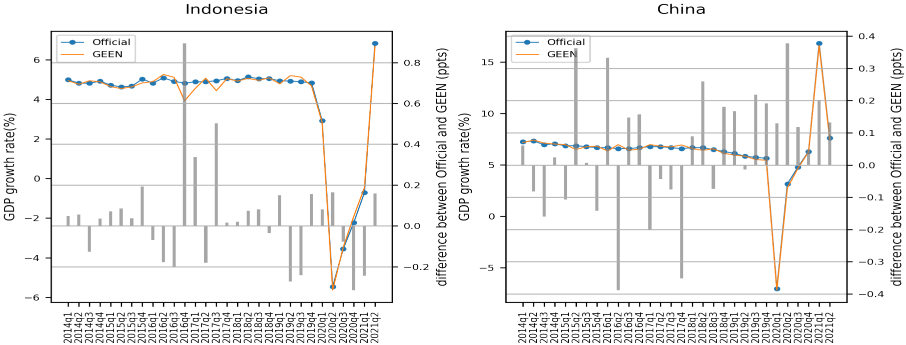


Figure: Country Examples of Official and GEEN-refined GDP Growth

- Excess smoothness masks underlying dynamics and volatility of economic activity
- Estimates of underlying economic growth could enrich policymakers' understanding of the state of macroeconomy, including output gap and inflationary pressures, and inform efficient policy making

Examples: official GDP data systematically different

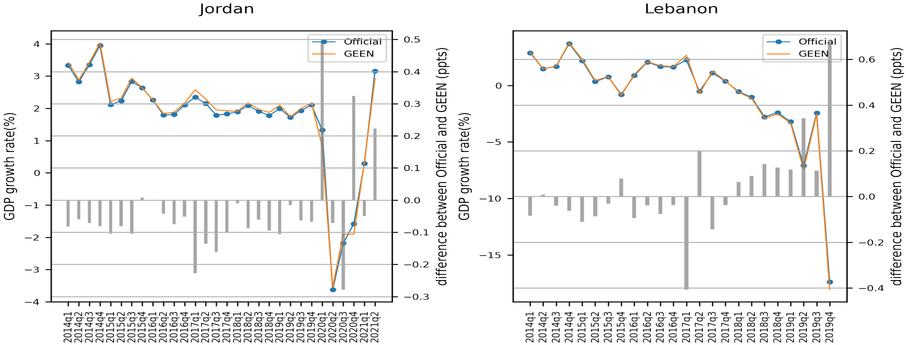


Figure: Country Examples of Official and GEEN-refined GDP Growth

- When Lebanon's economy shrank after 2017, official data systematically overstated the performance of the economy
- Jordan's official data systematically understated economic growth
- A plausible explanation is the existence of the informal sector is missing in official data

Conclusion

- This paper uses deep neural network to impute latent variable under conditional independence
- It provides a semi-nonparametric machine learning method
- It is useful to extract common information from observables at the observation level
- Empirical application: GDP growth refinement