BGM-IV: AN AI-POWERED BAYESIAN GENERATIVE MODELING APPROACH FOR INSTRUMENTAL VARIABLE

Anonymous authors
Paper under double-blind review

ABSTRACT

Instrumental-variable (IV) regression enables causal estimation under endogeneity, but modern IV problems often involve nonlinear structural effects and high-dimensional covariates. Existing methods typically operate in observed or generic learned feature spaces, and they often yield point estimates without uncertainty quantification. We introduce BGM-IV, a Bayesian generative modeling approach that performs nonlinear IV regression through posterior inference in a causally structured latent space. BGM-IV separates covariate variation by the role in the treatment and outcome mechanism, and accounts for endogeneity through an IV-integrated pseudo-likelihood that averages over instrument-induced treatment variation. The resulting model provides both structural-function estimates and predictive intervals for outcomes under intervention. Across various benchmark datasets, BGM-IV outperforms existing nonlinear IV methods overall, with significant gains in high-dimensional settings, while achieving near-nominal predictive coverage. These results highlight structured latent generative modeling as a flexible approach to uncertainty-aware IV inference with rich covariates. The code of BGM-IV is available at `https://github.com/anonymous`.

1 INTRODUCTION

Estimating causal effects from observational data is challenging when treatment is endogenous. Unobserved confounders jointly affect treatment and outcome and a predictive model may fit the observed outcome well but fail to recover the effect of an intervention (Imbens & Rubin, 2015; Hernan & Robins, 2020; Hernán et al., 2019). Instrumental variables (IVs) address this problem by exploiting exogenous variation that shifts treatment while affecting the outcome only through treatment (Angrist et al., 1996; Angrist & Pischke, 2009; Imbens & Angrist, 1994). IV estimation is therefore inherently indirect, as the structural effect must be recovered from instrument-induced treatment variation.

Modern IV problems are further complicated by nonlinear structural effects and high-dimensional covariates, such as images, electronic health records, or molecular measurements. Rich covariates may contain useful information about confounding, but treating them directly is statistically and computationally difficult (Chernozhukov et al., 2022). Simply plugging predictive models into an IV estimator is often not enough as poor handling of high-dimensional covariates can lead to biased causal estimates (D'Amour et al., 2021; Chernozhukov et al., 2018). Addressing such settings requires both flexible nonlinear causal effect modeling and a representation that separates causally relevant covariate information from nuisance variation (Johansson et al., 2016; Louizos et al., 2017; Shalit et al., 2017; Schölkopf et al., 2021).

Existing work has largely expanded the scope of IV estimation beyond classical linear models (Wu et al., 2025). Two-stage least squares (2SLS) remains widely used for its simplicity and interpretability under linear specifications (Angrist et al., 1996; Stock & Trebbi, 2003). Nonparametric IV methods relax these restrictions through series, sieve, and regularized inverse formulations (Newey & Powell, 2003; Blundell et al., 2007; Darolles et al., 2011; Chen & Pouzo, 2012). Recent machine learning IV methods further increase flexibility: KIV generalizes 2SLS to reproducing kernel Hilbert spaces (Singh et al., 2019), DeepIV integrates the outcome mean over an estimated conditional treatment distribution (Hartford et al., 2017), DFIV learns deep treatment and instrument features within 2SLS by alternating between the two stages (Xu et al., 2021), DeepGMM and related estimators

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

extend the generalized method of moments to conditional moment restrictions through minimax optimization (Hansen, 1982; Bennett et al., 2019; Dikkala et al., 2020), DML-IV applies double machine learning to mitigate regularization bias of the two-stage estimators (Shao et al., 2024), and FIL combines kernelized IV regression with adaptive instrument learning (Wang et al., 2022).

Despite this progress, nonlinear IV estimation remains challenging with high-dimensional covariates, where causally relevant information may be dominated by nuisance variation. Existing approaches typically operate on observed covariates or generic learned features, without explicitly capturing the roles of the covariates in the treatment and outcome mechanisms. In addition, most methods emphasize point estimation, while uncertainty quantification for outcomes under intervention remains limited. These gaps motivate an AI-powered generative framework that separates shared, treatment-specific, outcome-specific, and nuisance variation while supporting uncertainty-aware IV inference.

In this work, we propose BGM-IV, a Bayesian generative modeling approach for nonlinear IV regression. The central idea is to perform IV regression through posterior inference in a causally structured latent space, rather than directly in observed or generic learned feature space. By partitioning the structured latent space into independent subspaces that play different roles in the causal modeling, BGM-IV provides a principled way for estimating causal effects with IVs with uncertainty quantification. Our main contributions are summarized as follows.

- BGM-IV bridges the flexibility of deep generative models with principled posterior inference for nonlinear IV regression. The structured latent space partitions high-dimensional covariate variation based on the causal roles, and a novel IV-integrated pseudo-likelihood accounts for endogeneity by integrating the outcome density over the IV-induced treatment distribution.
- BGM-IV provides uncertainty quantification together with the point estimate. Posterior inference over the covariate-conditioned latent state yields structural-function estimates and predictive intervals for outcomes under intervention.
- Across diverse benchmarks, BGM-IV achieves the best performance compared to strong nonlinear IV baselines in nearly all evaluated settings, with the largest gains under high-dimensional covariates, while its predictive intervals attain near-nominal coverage.

2 Problem Setup

We observe $i.i.d.$ samples $\{(X^{(i)}, Y^{(i)}, V^{(i)}, W^{(i)})\}_{i=1}^{n}$ from an observational distribution, where $X \in \mathcal{X}$ is the treatment (or exposure), $Y \in \mathbb{R}$ is the outcome, $V \in \mathcal{V}$ denotes the high-dimensional covariates, and $W \in \mathcal{W}$ is an instrumental variable. Our target is the structural function

$$g_0(x, \mathbf{v}) = \mathbb{E}[Y \mid \mathrm{do}(X = x),\, V = \mathbf{v}]\,,$$

using Pearl's intervention notation (Pearl, 2009). Under the structural formulation, we write

$$Y = g_0(X, V) + \epsilon.$$

The challenge is that treatment is endogenous: in general, $\mathbb{E}[\epsilon \mid X, V] \neq 0$, so the observational regression function $\mathbb{E}[Y \mid X = x, V = \mathbf{v}]$ generally differs from the causal target $g_0(x, \mathbf{v})$.

Conditional on V, the instrument satisfies three standard conditions. Relevance requires the conditional distribution $P(X \mid V, W)$ to vary with W (Bound et al., 1995). Exclusion requires the instrument to affect the outcome only through its effect on treatment. Exogeneity requires the instrument to be mean-independent of the structural error conditional on the observed covariates (i.e., conditional mean exogeneity), $\mathbb{E}[\epsilon \mid V, W] = 0$.[1] Under these conditions,

$$\mathbb{E}[Y \mid V, W] = \mathbb{E}[g_0(X, V) \mid V, W] = \int g_0(x, V)\, dP(x \mid V, W).$$

This moment equation shows that the instrument does not identify the structural function $g_0(x, \mathbf{v})$ directly. Instead, it identifies how g_0 averages over the treatment distribution induced by the instrument, conditional on the covariates. In nonparametric IV, recovering g_0 requires conditions such as completeness of the conditional expectation operator (Newey & Powell, 2003; Hall & Horowitz, 2005; Carrasco et al., 2007; D'Haultfoeuille, 2011).

[1] This is a sufficient conditional mean exogeneity restriction and does not require $W \perp \epsilon \mid V$.

The remaining challenge is how to model the causal target $g_0(x, \mathbf{v})$ and the observable moment equation induced by the IV assumptions, particularly when the covariates V are high-dimensional. The next section addresses this by introducing the latent generative structure used in BGM-IV.

3 METHOD

3.1 CAUSALLY STRUCTURED BAYESIAN GENERATIVE MODEL

Inspired by recent advances in structural latent space modeling (Liu & Wong, 2026a;b), we further extend it to nonlinear IV regression. Specifically, we partition the latent space $Z = (Z_0, Z_1, Z_2, Z_3)$ to account for distinct causal roles. The shared block Z_0 enters both the treatment and outcome mechanisms and captures the observed confounding structure. Z_1 enters only the outcome mechanism, Z_2 enters only the treatment mechanism, and Z_3 is a nuisance block that explains the remaining covariate variation.

In the inference stage, the structured latent state is inferred from the observed covariates V alone. BGM-IV addresses the unobserved confounding between X and Y through the instrument and the IV-integrated objective. This partition provides a causally structured representation of how high-dimensional covariates impact treatment and outcome.

The generative components of BGM-IV (Figure 1) are specified as

$$Z \sim p_Z, \qquad V \sim p_\theta(\mathbf{v} \mid \mathbf{z}), \qquad X \sim p_\phi(x \mid w, \mathbf{z}_0, \mathbf{z}_2), \qquad Y \sim p_\omega(y \mid x, \mathbf{z}_0, \mathbf{z}_1). \tag{1}$$

Here p_Z is the latent prior, which is set to be $\mathcal{N}(0, I)$ by default. p_θ, p_ϕ, and p_ω are neural generative models for the covariates, treatment, and structural outcome, respectively. The instrumental variable W enters only through the treatment mechanism, reflecting the exclusion restriction. Because X is endogenous, p_ω is not fitted as the observational conditional distribution of Y given X. Instead, its structural mean is learned through the IV-integrated pseudo-likelihood introduced in Section 3.2.

By default, the three generative models p_θ, p_ϕ, and p_ω are specified as multivariate normal distributions for continuous variables and Bernoulli or Categorical distributions for discrete variables. In a typical continuous setting, we parameterize the models as

$$\begin{cases} V \mid Z \sim \mathcal{N}\big(\boldsymbol{\mu}_\theta(\mathbf{z}),\, \boldsymbol{\Sigma}_\theta(\mathbf{z})\big), \\ X \mid W, Z_0, Z_2 \sim \mathcal{N}\big(\mu_\phi(w, \mathbf{z}_0, \mathbf{z}_2),\, \sigma^2_\phi(w, \mathbf{z}_0, \mathbf{z}_2)\big), \\ Y \mid X, Z_0, Z_1 \sim \mathcal{N}\big(\mu_\omega(x, \mathbf{z}_0, \mathbf{z}_1),\, \sigma^2_\omega(x, \mathbf{z}_0, \mathbf{z}_1)\big). \end{cases} \tag{2}$$

Here $(\boldsymbol{\mu}_\theta,\ \boldsymbol{\Sigma}_\theta)$, $(\mu_\phi,\ \sigma^2_\phi)$ and $(\mu_\omega,\ \sigma^2_\omega)$ denote learnable functions, represented by three neural networks (G, H, and F) each with two output heads. For further simplification, we use a diagonal covariance structure for the covariate model $\boldsymbol{\Sigma}_\theta(\mathbf{z}) = \mathrm{diag}\big(\sigma^2_{\theta,1}(\mathbf{z}), \ldots, \sigma^2_{\theta,p}(\mathbf{z})\big)$. For binary treatments, the treatment model is replaced by a Bernoulli distribution $X \mid W, Z_0, Z_2 \sim \mathrm{Bernoulli}\big(\pi_\phi(w, \mathbf{z}_0, \mathbf{z}_2)\big)$ where $\pi_\phi(w, \mathbf{z}_0, \mathbf{z}_2) = \mathrm{logit}^{-1}(H(w, \mathbf{z}_0, \mathbf{z}_2; \phi))$ represents the propensity score. Analogous Bernoulli or categorical parameterizations can be used for discrete outcomes or covariates.

3.2 IV-INTEGRATED PSEUDO-LIKELIHOOD

With endogenous treatment, directly training the outcome generator through $p_\omega(y \mid x, \mathbf{z}_0, \mathbf{z}_1)$ using the observed treatment can bias estimation of the structural function g_0 as the observed treatment reflects both causal variation and unobserved confounding. BGM-IV therefore uses an IV-integrated working likelihood for outcome learning. Specifically, it integrates the outcome density over the treatment distribution induced by the instrument, which is denoted as:

$$p_{\mathrm{IV}}(y \mid w, \mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2) = \int p_\omega(y \mid x, \mathbf{z}_0, \mathbf{z}_1)\, p_\phi(x \mid w, \mathbf{z}_0, \mathbf{z}_2)\, dx. \tag{3}$$

This is our derived IV-integrated pseudo-likelihood term, which is also related to the integrated treatment-distribution used by DeepIV (Hartford et al., 2017), but the objectives differ in both

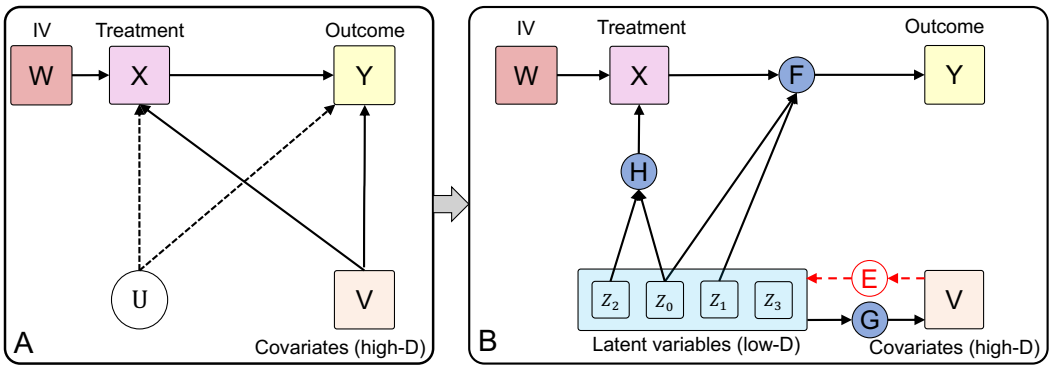


Figure 1: Illustration of the BGM-IV framework. (A) Causal graph for the IV setting. U is the unobserved confounder. (B) The causal graph in BGM-IV model. Rectangles denote variables and circles denote learned neural networks. G, H, and F denote the covariate, treatment, and outcome generative models, respectively. The encoder E is used in the EGM initialization and is not updated afterwards.

form and role. DeepIV minimizes a squared loss between the observed outcome and the response averaged over the estimated first-stage treatment distribution. BGM-IV instead integrates the outcome density first and uses the resulting mixture density as a component of latent quasi-posterior inference. Consequently, the IV-integrated term informs both the structural outcome generator and the update of the latent states during training.

Let μ_ω denote the mean of p_ω, the conditional mean implied by Eq. (3) is

$$\mathbb{E}_{\omega,\phi}[Y \mid W = w, Z = \mathbf{z}] = \int \mu_\omega(x, \mathbf{z}_0, \mathbf{z}_1) p_\phi(x \mid w, \mathbf{z}_0, \mathbf{z}_2)\, dx, \tag{4}$$

which has the same conditional-expectation operator form as the IV moment in Section 2.

3.3 Identification and Population Target

Appendix A analyzes the population structural-mean target of the conditional moment equation (4) for a covariate-measurable latent representation $A = (Z_0, Z_1, Z_2) = \varphi(V)$, with $A_Y = (Z_0, Z_1)$ and $A_X = (Z_0, Z_2)$, such that $g_0(x, V) = m_0(x, A_Y)$.

Mean identification. Under latent conditional mean exogeneity, $\mathbb{E}[\epsilon \mid W, A] = 0$, the structural mean satisfies the latent nonparametric IV moment equation with the same conditional-expectation operator form as Eq. (4). If the corresponding true conditional expectation operator is complete on the considered function class, the structural mean m_0 is the unique solution of this moment equation on the identified treatment support (Proposition A.1).

Consistency of structural prediction. Proposition A.3 bounds the IV-moment risk of the selected covariate-only MAP predictor. Under final-candidate approximation, first-stage, empirical, and Monte Carlo error control, and controlled ill-posedness on the candidate class (Xu et al., 2021; Dikkala et al., 2020), Corollary A.1 translates this risk bound into $L^2(P_{X,V})$ consistency of the selected structural-mean predictor.

3.4 Tempered Latent Quasi-Posterior

We combine the IV-integrated term with the latent prior, the covariate likelihood, and the first-stage treatment likelihood to define the quasi-posterior as

$$q_{\mathrm{IV}}(\mathbf{z} \mid x, y, \mathbf{v}, w) \propto p(\mathbf{z})\, p_\theta(\mathbf{v} \mid \mathbf{z})\, p_\phi(x \mid w, \mathbf{z}_0, \mathbf{z}_2)\, p_{\mathrm{IV}}(y \mid w, \mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2)^\gamma, \qquad \gamma \in [0, 1]. \tag{5}$$

γ is a tempering parameter used to control how strongly the IV-integrated outcome term influences latent-state inference (Bissiri et al., 2016). We set $\gamma = 0.01$ by default and report a sensitivity analysis over γ in Appendix C.

During training, Eq. (5) is used to update subject-specific latent variables using Y, raising the concern that they could absorb the structural error and reintroduce endogeneity. The key distinction is that

these outcome-informed latent states are not reused for structural prediction. In the inference stage, BGM-IV re-infers the latent state from V alone through the covariate-only posterior in Section 3.6. Thus, for a new subject, the latent state entering the final structural predictor preserves conditional mean exogeneity (Lemma A.1). Appendix B empirically checks that the resulting evaluation states show no detectable association with the structural error or the unobserved confounder.

3.5 Stochastic Optimization Training

We first develop an IV-aware initialization strategy based on EGM (Liu et al., 2024), which helps initialize both the model parameters and latent states (see details in Appendix D). Since the joint posterior distribution of latent variables and model parameters is not tractable under BGM-IV, we designed an iterative stochastic optimization procedure that alternates between updating subject-specific latent variables and updating the generator parameters. At each iteration, we draw a mini-batch $\mathcal{B}$. Each iteration consists of four steps:

(a) Given the current latent states $\{\mathbf{z}^{(i)}|i = 1, ..., n\}$, the covariate generative model is updated using

$$\mathcal{L}_V(\theta) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log p_\theta(\mathbf{v}^{(i)} \mid \mathbf{z}^{(i)}). \tag{6}$$

(b) Given the current latent states $\{\mathbf{z}^{(i)}|i = 1, ..., n\}$, the treatment generative model is updated using

$$\mathcal{L}_X(\phi) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log p_\phi(x^{(i)} \mid w^{(i)}, \mathbf{z}_0^{(i)}, \mathbf{z}_2^{(i)}). \tag{7}$$

(c) Given the current latent states and IV, the outcome generative model is updated using the IV-integrated pseudo-likelihood term,

$$\mathcal{L}_Y(\omega) = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log p_{\mathrm{IV}}(y^{(i)} \mid w^{(i)}, \mathbf{z}_0^{(i)}, \mathbf{z}_1^{(i)}, \mathbf{z}_2^{(i)}). \tag{8}$$

The pseudo-likelihood term is calculated as:

$$\log p_{\mathrm{IV}}(y^{(i)} \mid w^{(i)}, \mathbf{z}_0^{(i)}, \mathbf{z}_1^{(i)}, \mathbf{z}_2^{(i)}) \approx \log\left\{\frac{1}{M}\sum_{m=1}^{M} p_\omega(y^{(i)} \mid x^{(i,m)}, \mathbf{z}_0^{(i)}, \mathbf{z}_1^{(i)})\right\}, \tag{9}$$

where $\{x^{(i,m)}|m = 1, ..., M\}$ are Monte Carlo samples, M is the Monte Carlo sample size, and $x^{(i,m)}$ is sampled from treatment generative model $x^{(i,m)} \sim p_\phi(x \mid w^{(i)}, \mathbf{z}_0^{(i)}, \mathbf{z}_2^{(i)})$.

(d) Given the current parameters (θ, ϕ, ω) of generative models, subject-specific latent variables are updated by maximizing the corresponding IV quasi-posterior. For each $i \in \mathcal{B}$, we update $\mathbf{z}^{(i)}$ by gradient ascent on the logarithm of Eq. (5),

$$\log p(\mathbf{z}^{(i)}) + \log p_\theta(\mathbf{v}^{(i)} \mid \mathbf{z}^{(i)}) + \log p_\phi(x^{(i)} \mid w^{(i)}, \mathbf{z}_0^{(i)}, \mathbf{z}_2^{(i)}) + \gamma \log p_{\mathrm{IV}}(y^{(i)} \mid w^{(i)}, \mathbf{z}_0^{(i)}, \mathbf{z}_1^{(i)}, \mathbf{z}_2^{(i)}). \tag{10}$$

We draw $K = 10$ candidates from independent initializations and retain the one with the smallest IV-moment residual on the training sample (see details in Appendix F). The complete training procedure of BGM-IV is given in Algorithm 3 in Appendix D.

3.6 Structural Prediction and Uncertainty Quantification

After model fitting, we can estimate the target structural function $g_0(x, \mathbf{v})$. Given any intervention value x and the covariates $\mathbf{v}$ in the inference stage, BGM-IV first infers the latent variable under the *covariate-only* posterior using maximum a posteriori (MAP) estimation,

$$\hat{\mathbf{z}}(\mathbf{v}) = \arg\max_{\mathbf{z}} \left\{\log p(\mathbf{z}) + \log p_\theta(\mathbf{v} \mid \mathbf{z})\right\}. \tag{11}$$

The structural function is then estimated as

$$\hat{g}(x, \mathbf{v}) = \mu_\omega(x, \hat{\mathbf{z}}_0(\mathbf{v}), \hat{\mathbf{z}}_1(\mathbf{v})), \tag{12}$$

where μ_ω is the mean function of the outcome generator. This step does not condition on W, Y, or the observed endogenous treatment assignment. x is the intervention value provided to the fitted structural function.

The MAP readout is a point estimate. A key novelty of the generative model is that the same fitted model also quantifies the uncertainty of its predictions. To this end, BGM-IV samples the covariate-only posterior $p(\mathbf{z} \mid \mathbf{v}) \propto p(\mathbf{z})\, p_\theta(\mathbf{v} \mid \mathbf{z})$ by Hamiltonian Monte Carlo with several overdispersed chains. Each draw $\mathbf{z}^{(j)}$ yields a structural value $\mu_\omega(x, \mathbf{z}_0^{(j)}, \mathbf{z}_1^{(j)})$ and an outcome distribution $\mathcal{N}\big(\mu_\omega(x, \mathbf{z}_0^{(j)}, \mathbf{z}_1^{(j)}),\, \sigma_\omega^2(x, \mathbf{z}_0^{(j)}, \mathbf{z}_1^{(j)})\big)$. Averaging the latter over the J draws gives the model's posterior predictive distribution at $(x, \mathbf{v})$,

$$\hat{p}(y \mid x, \mathbf{v}) = \frac{1}{J} \sum_{j=1}^{J} \mathcal{N}\big(y;\, \mu_\omega(x, \mathbf{z}_0^{(j)}, \mathbf{z}_1^{(j)}),\, \sigma_\omega^2(x, \mathbf{z}_0^{(j)}, \mathbf{z}_1^{(j)})\big), \tag{13}$$

whose quantiles define predictive intervals. These predictive intervals propagate uncertainty in the latent state and incorporate the component variance σ_ω^2 of the fitted model. Throughout the paper, we report MAP point estimates and use the predictive distribution for interval estimates.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate BGM-IV on three benchmarks of increasing covariate complexity. The first is the demand-design benchmark derived from the airline-pricing simulation, which has low-dimensional covariates and has been widely used by existing IV methods (Hartford et al., 2017; Shao et al., 2024; Xu et al., 2021). The other two are high-dimensional benchmarks, in which the covariate variable is observed through a 1-D high-dimensional vector or through a 2-D image. The structural function is known in all three benchmarks, so that both the point estimates and the predictive intervals can be evaluated against the truth. The full details of the data-generating processes are given in Appendix G.

Baselines. We compare BGM-IV with five nonlinear IV methods, including KIV (Singh et al., 2019), DeepIV (Hartford et al., 2017), DFIV (Xu et al., 2021), DML-IV (Shao et al., 2024), and FIL (Wang et al., 2022), using their publicly released implementations and default hyperparameters. Every method except KIV, which has no random initialization, is trained from ten independent initializations per repetition, and one run is retained by that method's own training-side criterion to maximize the baseline performance. The baseline implementations and their selection criteria are described in Appendix F, and the implementation details of BGM-IV are given in Appendix E.

Evaluation metrics. For point estimation, we report the structural mean squared error (MSE) on the original outcome scale, computed on the evaluation grid of prior demand-design benchmarks. Each configuration is repeated 20 times with independent training samples shared by all methods to ensure fair comparison, and paired comparisons with the strongest baseline on the two high-dimensional benchmarks are reported in Appendix I. For uncertainty quantification, we report the coverage of new outcomes and the width of the predictive intervals of BGM-IV on the same grid, whose evaluation is detailed in Appendix F. All baseline implementations return point estimates only.

4.2 Demand Design Experiments

The demand-design benchmark models counterfactual demand estimation in airline pricing with low-dimensional covariates. The treatment P is the ticket price, the outcome Y is demand, and the observed covariates are time of year T and customer group S, which induce heterogeneous price sensitivity. Prices are endogenous because an unobserved demand shock affects both the price and the realized demand, with the strength of this confounding set by a correlation parameter ρ. The instrument is a cost shifter that changes the price but is excluded from the demand equation. The structural function is $f_0(P, T, S) = 100 + (10 + P)S\psi(T) - 2P$ with a nonlinear seasonal function ψ, and the full design is given in Appendix G. We vary the training sample size $n \in \{1{,}000, 5{,}000, 10{,}000\}$ and the confounding level $\rho \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

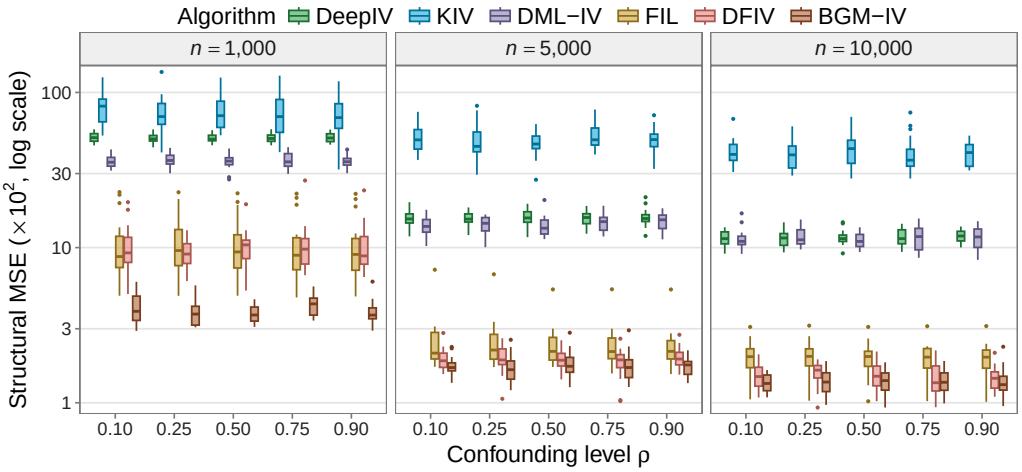


Figure 2: Structural MSE on the demand-design benchmark by sample size and confounding level.

Figure 2 summarizes the results and the complete values are in Table 8 of Appendix H. BGM-IV attains the lowest mean structural MSE in all 15 configurations. For example, when the sample size is $n = 1{,}000$, BGM-IV reduces the MSE by 57.9%-62.8% relative to the strongest baseline, while exhibiting 2.7-fold to 7.1-fold lower variability across repetitions. Taken together, these results show the superiority of BGM-IV in accurate and robust structural estimates on this widely used classical IV benchmark.

Section 3.4 argues that the latent states used for prediction are inferred from the covariates alone and therefore preserve conditional mean exogeneity. The unobserved confounder and the structural error are explicit in the demand design experiment, which allows this argument to be checked empirically. We run the check at the strongest confounding level $\rho = 0.9$, where the unobserved confounder accounts for 81% of the variance of the structural error and a latent state that absorbed this error would reintroduce the most endogeneity. The latent states show no detectable association with the structural error or with the unobserved confounder, on training and on held-out data. Appendix B provides details, including the negative and positive controls, and the complete results.

4.3 High-Dimensional Vector-Proxy Experiments

The vector-proxy benchmark is the high-dimensional extension of the demand design experiment. The structural equation, price equation, and instrument remain unchanged, but the customer group is no longer observed as a discrete label. Instead, each group $s \in \{1, \ldots, 7\}$ has a prototype $\boldsymbol{\mu}_s \in \mathbb{R}^{784}$ drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and we observe the noisy proxy $\mathbf{r} = \boldsymbol{\mu}_S + \sigma_{\text{rep}}\boldsymbol{\eta}$ with $\boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This construction follows the proxy-variable view that latent causal factors may be observed only through noisy measurements (Kuroki & Pearl, 2014; Miao et al., 2018), but here the proxy mechanism is fully controlled so the benchmark isolates representation learning rather than identification by proxy variables. To control how much of the proxy each method has to digest, every method receives the covariate $(T, \text{PCA}_k(\mathbf{r}))$, the first $k \in \{1, 2, 4, 6, 8, 16, 32, 64, 128\}$ principal components fitted on the training proxies, and keeps its demand-design architecture and hyperparameters. We fix $\rho = 0.5$ and vary $n \in \{5{,}000, 10{,}000, 20{,}000, 50{,}000\}$.

Figure 3 shows the results at $n = 5{,}000$. The complete values for all sample sizes are in Table 9 of Appendix H. Across all sample sizes, BGM-IV achieves the lowest mean MSE in 35 of the 36 configurations. With a single component with limited information to separate the seven groups, all methods perform poorly and BGM-IV still has the lowest mean MSE, and for $k \geq 2$ its improvement over the strongest baseline is statistically significant in 30 of the 32 configurations (Table 11 of Appendix I). For instance, when $k = 2$-8, it reduces structural MSE by 32%-58% relative to the strongest baseline. For $k > 8$, where the additional components primarily introduce nuisance variation in this design, BGM-IV retains the lowest mean error for every k when $n \geq 10{,}000$, while several competing methods deteriorate sharply as additional nuisance dimensions are included.

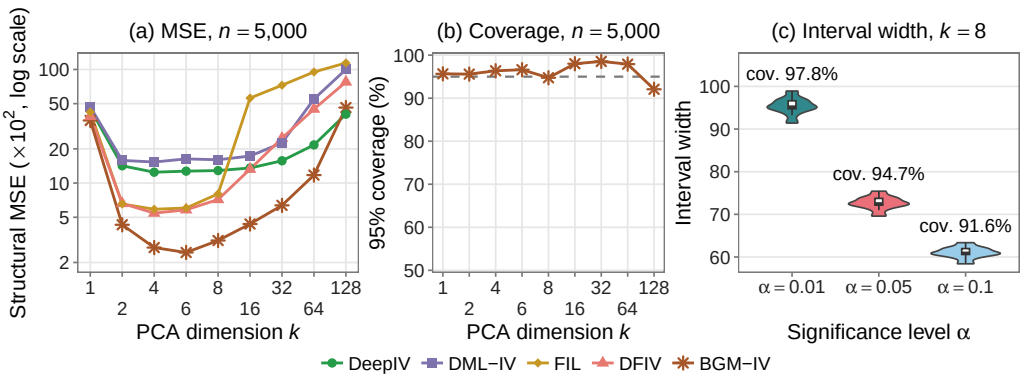


Figure 3: Vector-proxy benchmark at $n = 5{,}000$. (a) Structural MSE against the number k of principal components. (b) Coverage of the 95% predictive intervals of BGM-IV, the dashed line is the nominal level. (c) Widths of the 99%, 95%, and 90% predictive intervals at $k = 8$ over repetitions with the mean coverage rate.

Beyond point estimation, BGM-IV uniquely quantifies the uncertainty of its predictions. Panel (b) shows that the coverage of the 95% predictive intervals remains near the nominal level for every k, and panel (c) shows that at $k = 8$ the 90%, 95%, and 99% intervals all attain close to nominal coverage, with widths that increase with the nominal coverage level $1 - \alpha$.

4.4 High-Dimensional Image Experiments

The image-covariate benchmark follows the high-dimensional demand setting introduced by DeepIV (Hartford et al., 2017), which has since been adopted as a standard test of nonlinear IV methods with image covariates (Xu et al., 2021; Shao et al., 2024). The causal mechanism of the demand design is unchanged, but the customer group is no longer given as a label. It is observed only through a 28×28 handwritten-digit image from MNIST (LeCun et al., 1998) whose digit equals the group, so the observed covariates are T and a 784-dimensional image, and a method has to recover the group from raw pixels in order to estimate the structural function. We use the sample sizes and confounding levels of Section 4.2.

Table 1 reports the complete benchmark results. BGM-IV attains the lowest mean MSE in all 15 configurations, and each improvement over the strongest baseline is statistically significant (Table 12 of Appendix I). Its advantage grows with the sample size: relative to the strongest baseline, its MSE is about 1.3 times smaller at $n = 1{,}000$, 2.6 to 2.9 times smaller at $n = 5{,}000$, and more than three times smaller at $n = 10{,}000$. These results show the superiority of BGM-IV in accurate and robust structural estimates on this widely used high-dimensional IV benchmark.

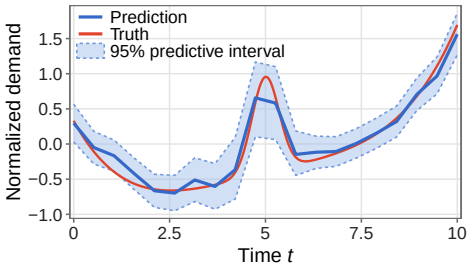


Figure 4: A slice of the demand curve estimated by BGM-IV on the image-covariate benchmark ($n = 50{,}000$, $\rho = 0.5$).

Figure 4 illustrates the unique uncertainty quantification of BGM-IV on this benchmark with a model trained on $n = 50{,}000$ samples at $\rho = 0.5$. It shows a slice of the estimated demand curve with its 95% predictive band. Over 20 repetitions of this setting, the 90%, 95%, and 99% predictive intervals cover 90.7%, 93.8%, and 96.8% of new outcomes.

In addition, building on the image-covariate benchmark, we consider the case in which the instrument is only weakly correlated with the treatment in Appendix H.4, where BGM-IV again attains the lowest MSE among all methods at every instrument strength and sample size considered.

4.5 Ablation Studies

We examine the three components of BGM-IV on the demand design at $n = 1{,}000$ with three settings: the standard design, a design in which the structural disturbance is amplified by a factor of 100, and a

Table 1: Structural MSE on the image-covariate benchmark (entries divided by 10^2), mean $\pm$ standard deviation over 20 repetitions. The lowest mean in each row is in bold.

n	ρ	DeepIV	KIV	DML-IV	FIL	DFIV	**BGM-IV**
1,000	0.1	153.21 $\pm$ 6.13	208.31 $\pm$ 2.45	133.32 $\pm$ 2.66	131.51 $\pm$ 13.87	102.65 $\pm$ 5.72	**80.04 $\pm$ 15.05**
	0.25	155.08 $\pm$ 5.31	207.54 $\pm$ 2.76	132.71 $\pm$ 3.91	133.82 $\pm$ 17.40	102.92 $\pm$ 5.71	**79.18 $\pm$ 11.15**
	0.5	155.46 $\pm$ 5.84	208.57 $\pm$ 2.62	133.54 $\pm$ 3.98	133.48 $\pm$ 16.94	103.20 $\pm$ 4.47	**75.85 $\pm$ 11.54**
	0.75	153.93 $\pm$ 5.25	209.06 $\pm$ 2.88	133.21 $\pm$ 3.93	133.16 $\pm$ 16.65	102.90 $\pm$ 5.85	**80.50 $\pm$ 8.48**
	0.9	154.11 $\pm$ 5.40	207.33 $\pm$ 2.32	132.23 $\pm$ 4.35	133.70 $\pm$ 16.89	101.58 $\pm$ 4.71	**78.36 $\pm$ 10.16**
5,000	0.1	85.81 $\pm$ 9.16	192.53 $\pm$ 2.21	75.51 $\pm$ 3.74	89.70 $\pm$ 12.85	69.11 $\pm$ 4.07	**26.50 $\pm$ 4.59**
	0.25	84.40 $\pm$ 7.25	192.19 $\pm$ 2.21	76.93 $\pm$ 3.00	89.75 $\pm$ 12.91	71.22 $\pm$ 1.93	**25.32 $\pm$ 2.62**
	0.5	85.04 $\pm$ 8.80	192.47 $\pm$ 2.19	75.13 $\pm$ 2.30	91.08 $\pm$ 13.52	71.10 $\pm$ 3.46	**25.54 $\pm$ 2.01**
	0.75	84.66 $\pm$ 7.48	192.05 $\pm$ 1.91	75.58 $\pm$ 2.85	91.20 $\pm$ 13.04	71.12 $\pm$ 4.00	**24.91 $\pm$ 2.13**
	0.9	87.67 $\pm$ 7.60	192.22 $\pm$ 1.78	76.15 $\pm$ 3.11	91.19 $\pm$ 12.98	69.33 $\pm$ 4.58	**25.94 $\pm$ 3.57**
10,000	0.1	80.57 $\pm$ 6.44	187.57 $\pm$ 1.80	63.76 $\pm$ 2.06	78.08 $\pm$ 19.88	55.38 $\pm$ 5.92	**17.12 $\pm$ 1.64**
	0.25	79.75 $\pm$ 7.40	187.63 $\pm$ 1.39	63.87 $\pm$ 2.40	73.48 $\pm$ 7.70	55.82 $\pm$ 6.06	**17.16 $\pm$ 1.24**
	0.5	78.48 $\pm$ 6.22	187.20 $\pm$ 1.73	63.75 $\pm$ 2.89	79.24 $\pm$ 19.30	55.73 $\pm$ 6.37	**17.14 $\pm$ 1.39**
	0.75	77.96 $\pm$ 5.53	187.31 $\pm$ 1.72	63.69 $\pm$ 2.27	79.91 $\pm$ 19.75	55.68 $\pm$ 5.70	**17.13 $\pm$ 1.25**
	0.9	80.88 $\pm$ 6.45	187.63 $\pm$ 2.00	63.68 $\pm$ 2.03	79.77 $\pm$ 19.98	54.69 $\pm$ 6.51	**17.14 $\pm$ 1.84**

Table 2: Ablation study on the demand design at $n = 1{,}000$ (structural MSE divided by 10^2, mean $\pm$ standard deviation over 20 repetitions). A check mark indicates that the component is kept: the EGM initialization, the IV-integrated pseudo-likelihood (IV-PL), and the latent partition.

Setting	EGM init.	IV-PL	Partition	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
Standard	✓	✓	✓	**4.15 $\pm$ 0.91**	**3.74 $\pm$ 0.47**	**3.87 $\pm$ 0.69**
	×	✓	✓	73.39 $\pm$ 8.63	70.56 $\pm$ 7.13	69.14 $\pm$ 7.29
Amplified confounding	✓	✓	✓	**35.21 $\pm$ 8.03**	**33.82 $\pm$ 9.01**	**31.65 $\pm$ 6.23**
	✓	×	✓	41.63 $\pm$ 9.75	67.09 $\pm$ 19.91	114.77 $\pm$ 36.21
Nuisance covariates	✓	✓	✓	**32.34 $\pm$ 9.98**	**32.86 $\pm$ 9.89**	**30.53 $\pm$ 8.69**
	✓	✓	×	45.16 $\pm$ 4.50	47.16 $\pm$ 4.93	43.73 $\pm$ 4.84

design with five additional nuisance covariates. Each variant changes one component and keeps every other setting unchanged, so that it is paired with the full model by repetition. Appendix K describes the settings and variants and compares BGM-IV with the strongest baselines in all three settings, and Table 2 reports the results.

IV-aware EGM initialization. Without the IV-aware EGM initialization, the structural MSE is 17.7 to 18.9 times larger than that of the full model, and it is larger in every one of the 60 paired repetitions. These results indicate that the initialization plays a critical role in optimization.

IV-integrated pseudo-likelihood. Under amplified confounding, replacing the IV-integrated pseudo-likelihood by the likelihood at the observed treatment increases the structural MSE by a factor of 1.18, 1.98, and 3.63 at $\rho = 0.1$, 0.5, and 0.9. The increase grows with the confounding level because the observed-treatment likelihood fits the outcome at the endogenous treatment, and the resulting bias grows with the correlation ρ between the confounder and the structural disturbance.

Structured latent space. With five nuisance covariates and a partition specified from the design, replacing the latent partition by an unstructured latent space increases the structural MSE by a factor of 1.40 to 1.44. A regression of the nuisance covariates on the trained latent states shows that they are captured by the nuisance block and not by the treatment and outcome generators (Appendix K).

5 Conclusion

We introduced BGM-IV, an AI-powered Bayesian generative framework for nonlinear IV regression with high-dimensional covariates. By combining a causally structured latent representation with an IV-integrated pseudo-likelihood, BGM-IV learns structural effects from instrument-induced treatment variation and provides both point estimates and uncertainty quantification. Across diverse benchmarks, it achieves lower structural MSE in nearly all evaluated settings, with the largest gains under high-dimensional covariates, while attaining near-nominal predictive coverage. Computational cost is one of the limitations (see Appendix J). Future work will extend BGM-IV to more complex treatment or instrument settings, such as longitudinal or dynamic regimes.

AI USE STATEMENT

In this work, we used generative AI tools to assist in implementing the scripts that run the released baseline implementations from multiple initializations and select one run, and the preprocessing pipeline of the vector-proxy benchmark, which fits the principal components on the training proxies and passes the same covariates to every method. We also used them to provide feedback on the design of the experiments and the evaluation protocols, including the ablation studies and the sensitivity analysis, to assist in summarizing some of the experimental results, and to check derivations and suggest edits to the presentation of the theoretical appendix. We have used AI tools to help organize the codebase of BGM-IV to enhance readability and prepare the installable software.

We have not used generative AI tools to implement the BGM-IV model, to generate synthetic data sets, to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to formulate mathematical claims, to provide critical ingredients for proofs, or to write proofs. Cleaning and reformatting data sets and qualitative or thematic data analysis are not applicable to this work. We have reviewed and verified all AI-assisted work, including the code, the summaries of the results, and every statement and proof step in the theoretical appendix. In particular, for every baseline, the authors examined the original paper and the released code to confirm that the criterion used to select among its runs is appropriate for that method and keeps the comparison fair. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide anonymized source code of BGM-IV, together with scripts that generate the benchmark data and run the reported experiments, in the supplementary material. The complete training, model-selection, and prediction procedure of BGM-IV is specified in Section 3 and Appendix D (Algorithms 1–4). All hyperparameters and preprocessing steps of BGM-IV are listed in Appendix E, and the baseline implementations, their configurations, and their training-side selection criteria are described in Appendix F. The data-generating processes, evaluation grids, and evaluation protocols of all benchmarks are given in Appendix G, results for every configuration with standard deviations over repetitions are reported in Appendix H, and the computational resources are documented in Appendix J. The assumptions and complete proofs of the theoretical results are given in Appendix A. Upon acceptance, the anonymous GitHub repository link will be replaced with a link to the public GitHub repository in the camera-ready version.

REFERENCES

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. *Advances in neural information processing systems*, 32, 2019.

P. G. Bissiri, C. C. Holmes, and S. G. Walker. A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130, 2016. doi: https://doi.org/10.1111/rssb.12158. URL `https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12158`.

Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric iv estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007.

John Bound, David A Jaeger, and Regina M Baker. Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association*, 90(430):443–450, 1995.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 02 2018. ISSN 1368-4221. doi: 10.1111/ectj.12097. URL https://doi.org/10.1111/ectj.12097.

Victor Chernozhukov, Whitney K Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Xavier D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.

Peter Hall and Joel L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904 – 2929, 2005. doi: 10.1214/009053605000000714. URL https://doi.org/10.1214/009053605000000714.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912775.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep iv: A flexible approach for counterfactual prediction. In *International conference on machine learning*, pp. 1414–1423. PMLR, 2017.

M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2020. ISBN 9781420076165. URL https://books.google.com/books?id=_KnHIAAACAAJ.

Miguel A. Hernán, John Hsu, and Brian Healy. A second chance to get causal inference right: A classification of data science tasks. *CHANCE*, 32(1):42–49, 2019. doi: 10.1080/09332480.2019.1579578. URL https://doi.org/10.1080/09332480.2019.1579578.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

Rob J Hyndman, George Athanasopoulos, et al. *Forecasting: principles and practice*. OTexts Melbourne, Australia, 2014.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/2951620.

Guido W. Imbens and Donald B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, pp. 423–437, 2014.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *EPIDEMIOLOGY*, 21(3):383–388, MAY 2010. ISSN 1044-3983. doi: 10.1097/EDE.0b013e3181d61eeb.

Qiao Liu and Wing Hung Wong. An ai-powered bayesian generative modeling approach for arbitrary conditional inference, 2026a. URL https://arxiv.org/abs/2601.05355.

Qiao Liu and Wing Hung Wong. An ai-powered bayesian generative modeling approach for causal inference in observational studies. *Journal of the American Statistical Association*, 121(554):976–987, 2026b. doi: 10.1080/01621459.2026.2654227. URL https://doi.org/10.1080/01621459.2026.2654227.

Qiao Liu, Zhongren Chen, and Wing Hung Wong. An encoding generative modeling approach to dimension reduction and covariate adjustment in causal inference with observational studies. *Proceedings of the National Academy of Sciences*, 121(23):e2322376121, 2024.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International conference on machine learning*, pp. 3076–3085. PMLR, 2017.

Daqian Shao, Ashkan Soleymani, Francesco Quinzan, and Marta Kwiatkowska. Learning decision policies with instrumental variables through double machine learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 44489–44514. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/shao24d.html.

Rahul Singh, Maneesh Sahani, and Arthur Gretton. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

James H Stock and Francesco Trebbi. Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194, 2003.

Ziyu Wang, Yuhao Zhou, and Jun Zhu. Fast instrument learning with faster rates. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Anpeng Wu, Kun Kuang, Ruoxuan Xiong, and Fei Wu. Instrumental variables in causal inference and machine learning: A survey. *ACM Computing Surveys*, 57(11):1–36, 2025.

Liyuan Xu, Yutian Chen, Siddarth Srinivasan, Nando de Freitas, Arnaud Doucet, and Arthur Gretton. Learning deep features in instrumental variable regression. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=sy4Kg_ZQmS7.

A Structural-Mean Identification and Conditional Consistency

This appendix studies the structural mean. We establish conditional mean exogeneity for covariate-only readouts and identify the structural function under representation sufficiency and completeness. We characterize the population target of the integrated EGM outcome component, then bound the final selected predictor's IV-moment risk and derive consistency under explicit approximation, error-control and inverse-stability conditions. A complementary observed-data comparison gives another sufficient condition for the selected predictor.

A.1 Covariate Readouts and Mean Identification

Write

$$A = (Z_0, Z_1, Z_2), \qquad A_Y = (Z_0, Z_1), \qquad A_X = (Z_0, Z_2).$$

For the fixed-representation population arguments, $A = \varphi(V)$ is a measurable map. Representation sufficiency is a separate condition: the structural target must admit $g_0(x, V) = m_0(x, A_Y)$ on the identified support. Measurability alone does not ensure this equality. Let λ_X denote Lebesgue measure for continuous treatment and counting measure for discrete treatment.

Assumption A.1 (Latent mean-IV model)**.** *The observations are iid and satisfy $Y = m_0(X, A_Y) + \epsilon$, with $Y, m_0(X, A_Y) \in L^2$ and $\mathbb{E}[\epsilon \mid V, W] = 0$. The representation $A = \varphi(V)$ is sufficient for the structural mean in the sense stated above. The conditional law of X given (W, A) has density $k_0(x \mid w, a_X)$, so that its dependence on a is through a_X. The causal interpretation of m_0 is inherited from the structural assumptions in Section 2 and the stated sufficiency condition.*

Assumption A.2 (Completeness on structural-mean differences)**.** *For P_A-almost every a, define*

$$(T_a h)(w) = \int h(x) k_0(x \mid w, a_X)\, d\lambda_X(x).$$

On the difference class $h(x) = m(x, a_Y) - m_0(x, a_Y)$ under consideration, $(T_a h)(w) = 0$ for $P_{W|A=a}$-almost every w implies $h(x) = 0$ for $P_{X|A=a}$-almost every x.

Lemma A.1 (Exogeneity of fixed and learned covariate readouts)**.** *For a fixed measurable map $A = \varphi(V)$, $\mathbb{E}[\epsilon \mid V, W] = 0$ implies $\mathbb{E}[\epsilon \mid W, A] = 0$. More generally, let $\mathcal{D}$ contain training observations, fitted preprocessing, model selection and training randomness. For a new observation independent of $\mathcal{D}$, let R be inference randomness independent of the joint tuple $(\mathcal{D}, V, W, \epsilon)$ and let $A = \varphi_{\mathcal{D}}(V, R)$. If $\mathbb{E}[\epsilon \mid V, W] = 0$, then*

$$\mathbb{E}[\epsilon \mid W, A, \mathcal{D}] = 0.$$

The same conclusion applies to any measurable sub-block of A.

Proof. For the fixed map, $\sigma(W, A) \subseteq \sigma(W, V)$, so iterated expectations give the claim. For the learned map, independence implies $\mathbb{E}[\epsilon \mid V, W, \mathcal{D}, R] = 0$. Since $\sigma(W, A, \mathcal{D}) \subseteq \sigma(V, W, \mathcal{D}, R)$, another conditional expectation proves the result. Sub-blocks generate smaller sigma-fields and inherit the same conclusion. □

Remark A.1 (Training states and evaluation readouts)**.** *The lemma permits training the network parameters with outcomes. It covers the numerical covariate-only MAP rule in Section 3.6, the encoder, and posterior sampling with independent inference randomness: each uses only the new unit's covariates and fitted quantities. For a reused training observation, the fitted parameters may depend on that observation's error, so the new-unit argument does not apply directly. Appendix B diagnoses associations in the fitted MAP readout on both training and held-out samples. The outcome-informed particles in Eq.* (10) *enter the training IV term and are re-inferred from covariates for final prediction. The lemma concerns conditional mean exogeneity of that prediction readout; information sufficiency and consistency of the paired outcome function are separate conditions.*

Proposition A.1 (Latent nonparametric IV moment)**.** *Under Assumption A.1,*

$$\mathbb{E}[Y \mid W = w, A = a] = \int m_0(x, a_Y) k_0(x \mid w, a_X)\, d\lambda_X(x). \tag{14}$$

Under Assumption A.2, m_0 is the unique solution in the stated class, up to equality for P_A-almost every a and $P_{X|A=a}$-almost every x.

Proof. Lemma A.1 gives $\mathbb{E}[\epsilon \mid W, A] = 0$. Taking the conditional expectation of $Y = m_0(X, A_Y)+\epsilon$ and using the true first-stage law proves Eq. (14). If another m satisfies the same equation, subtraction gives $T_a\{m(\cdot, a_Y) - m_0(\cdot, a_Y)\} = 0$. Completeness on these differences proves uniqueness on the stated support. □

A.2 Population Target of the Integrated EGM Outcome Component

The integrated squared loss targets the IV mean restriction under a fixed sufficient representation and a correct first stage.

Proposition A.2 (Population target of the integrated EGM outcome component)**.** *Fix $A = \varphi(V)$ and suppose $Y = m_0(X, A_Y) + \epsilon$ with $\mathbb{E}[\epsilon \mid V, W] = 0$. Assume $Y, m_0(X, A_Y) \in L^2$, all candidates satisfy $m(X, A_Y) \in L^2$, m_0 is in the candidate class, and the first-stage kernel in the objective equals the full conditional law $P(X \mid W, V)$ through A_X. For*

$$R(m) = \mathbb{E}\left[\left\{Y - \int m(x, A_Y)k_0(x \mid W, A_X)\, d\lambda_X(x)\right\}^2\right],$$

the minimum is $\mathbb{E}[\mathrm{Var}(Y \mid W, V)]$. Under completeness on candidate mean differences, every minimizer equals m_0 on the identified support.

Proof. Put $T_m(W, V) = \int m(x, A_Y)k_0(x \mid W, A_X)\, d\lambda_X(x)$. The conditional-mean restriction and correct first stage imply $\mathbb{E}[Y \mid W, V] = T_{m_0}(W, V)$. Conditional-mean orthogonality gives

$$R(m) = \mathbb{E}[\mathrm{Var}(Y \mid W, V)] + \|T_m - T_{m_0}\|^2_{L^2(P_{W,V})}.$$

The second term vanishes at m_0. If it vanishes at another m, condition the resulting zero conditional moment on (W, A) and apply completeness to $m - m_0$. □

Remark A.2 (Relation to EGM initialization)**.** *Proposition A.2 characterizes the integrated outcome loss for a fixed sufficient representation and a correct first-stage kernel. EGM uses this loss together with its reconstruction, adversarial, and first-stage objectives. The final covariate-only MAP predictor is analyzed in Proposition A.3 and Corollary A.1.*

A.3 Final Predictors and Conditional Consistency

The following results concern the final outcome function composed with the covariate-only readout. Population optimality at a fixed representation and exogeneity of the final readout address distinct parts of the argument; a learning condition must connect the fitted composite predictor to the identified moment target.

Remark A.3 (A quantitative training-to-readout condition)**.** *Let $g_E(x, v) = \mu_{\omega_E}(x, e_E(v)_Y)$ be an EGM composite predictor and let $g_F(x, v) = \mu_{\omega_F}(x, a_F(v)_Y)$ be the final readout. Define the true observed-covariate operator $(T_0 g)(w, v) = \mathbb{E}[g(X, V) \mid W = w, V = v]$. Conditional expectation is an L^2 contraction, so*

$$\|T_0(g_F - g_0)\|_2 \le \|T_0(g_E - g_0)\|_2 + \|g_F - g_E\|_2.$$

For outcome maps uniformly Lipschitz in their parameters and latent inputs on the relevant region, the last term is bounded by

$$L_\omega\|\omega_F - \omega_E\| + L_z\|a_F - e_E\|_{L^2(P_V)}.$$

This inequality isolates the transfer error; it does not assert that the reported fixed training schedule makes that error vanish. Exogeneity of $a_F(V)$ alone cannot control changes in its paired outcome map.

The final selection criterion. We now state a result for the prediction functions actually compared by the training-side selection rule, rather than for numerical latent coordinates. For candidate k, write

$$g_k(x, v) = \mu_{\widehat{\omega}_k}\big(x, \widehat{z}_{k,0}(v), \widehat{z}_{k,1}(v)\big), \qquad \widehat{K}_k(dx \mid w, v) = p_{\widehat{\phi}_k}\big(dx \mid w, \widehat{z}_{k,0}(v), \widehat{z}_{k,2}(v)\big).$$

The covariate-only inference rule is assumed to define the same measurable function $v \mapsto \widehat{z}_k(v)$ in scoring and deployment, with the same fitted preprocessing, initialization and numerical inference

settings. The results below concern this implemented prediction function. All training, preprocessing and model-selection inputs remain in the training sample. Assume $Y, g_0(X,V) \in L^2$ and define

$$(T_0 g)(w,v) = \int g(x,v) P_0(dx \mid w,v), \qquad \mathcal{B}(g) = \|T_0(g-g_0)\|^2_{L^2(P_{W,V})}.$$

The original mean-IV restriction gives the projection identity

$$R_0(g) := \mathbb{E}[(Y - T_0 g(W,V))^2] = C + \mathcal{B}(g), \qquad C = \mathbb{E}[\mathrm{Var}(Y \mid W,V)]. \tag{15}$$

Let the training sample consist of n iid observations, let P_n be its empirical measure, and set $\widehat{h}_k(w,v) = \int g_k(x,v)\widehat{K}_k(dx \mid w,v)$. Conditional on the fitted candidates and training data, draw treatment values independently across observations and Monte Carlo replicates, using randomness not used to fit the candidates. Denote the resulting implemented squared-residual score by S_k^{MC} and set

$$\nu_k = \frac{1}{nM} \sum_{i=1}^{n} \mathrm{Var}_{\widehat{K}_k(\cdot \mid W_i, V_i)}\{g_k(X, V_i)\}.$$

Assume these conditional second moments and all displayed risks are finite. The score then satisfies the exact identity

$$\mathbb{E}_{\mathrm{MC}}[S_k^{\mathrm{MC}} \mid \text{data and fitted candidates}] = P_n(Y - \widehat{h}_k)^2 + \nu_k. \tag{16}$$

For a deterministic class $\mathcal{G}_n \subset L^2(P_{X,V})$ containing all possible fitted candidates, define

$$\begin{aligned} u_n &= \sup_{g \in \mathcal{G}_n} \left| P_n(Y - T_0 g)^2 - R_0(g) \right|, \\ e_n &= \max_{k \le K} \left| P_n(Y - \widehat{h}_k)^2 - P_n(Y - T_0 g_k)^2 \right|, \\ t_n &= \max_{k \le K} \left| S_k^{\mathrm{MC}} - P_n(Y - \widehat{h}_k)^2 - \nu_k \right|. \end{aligned}$$

These quantities separate empirical generalization, first-stage error at the final readout, and Monte Carlo fluctuation.

Proposition A.3 (Oracle bound for final-readout selection)**.** *Under the preceding conditions, if* $\widehat{k} \in \arg\min_{k \le K} S_k^{\mathrm{MC}}$*, then, pathwise,*

$$\mathcal{B}(g_{\widehat{k}}) \le \min_{k \le K}\{\mathcal{B}(g_k) + \nu_k\} + 2u_n + 2e_n + 2t_n. \tag{17}$$

Proof. Equation (15) follows by expanding $Y - T_0 g = (Y - \mathbb{E}[Y \mid W,V]) + T_0(g_0 - g)$; the cross term has zero expectation. Equation (16) follows by taking the variance of the Monte Carlo mean. For every comparator k, the definitions of u_n, e_n, t_n and the selection rule give

$$\begin{aligned} R_0(g_{\widehat{k}}) &\le S_{\widehat{k}}^{\mathrm{MC}} - \nu_{\widehat{k}} + u_n + e_n + t_n \\ &\le S_k^{\mathrm{MC}} + u_n + e_n + t_n \\ &\le R_0(g_k) + \nu_k + 2u_n + 2e_n + 2t_n. \end{aligned}$$

The second line uses $\nu_{\widehat{k}} \ge 0$. Subtract C and minimize over k. □

Assumption A.3 (Final-candidate approximation)**.** *For nonnegative sequences* a_n *and* b_n^{MC}*, with probability tending to one there is a fitted candidate* $k_n^\dagger$ *such that*

$$\|T_0(g_{k_n^\dagger} - g_0)\|_{L^2(P_{W,V})} \le a_n, \qquad \nu_{k_n^\dagger} \le b_n^{\mathrm{MC}}.$$

To distinguish unique identification from stable recovery, define the restricted inverse modulus

$$\mathfrak{m}_n(r) = \sup\left\{ \|g - g_0\|_{L^2(P_{X,V})} : g \in \mathcal{G}_n \cup \{g_0\}, \quad \|T_0(g - g_0)\|_{L^2(P_{W,V})} \le r \right\}.$$

The modulus quantifies how IV-moment error translates into structural-function error on the candidate class, following ill-posedness measures used in NPIV analyses (Xu et al., 2021; Dikkala et al., 2020). For example, if a finite constant τ_n satisfies

$$\|g - g_0\|_{L^2(P_{X,V})} \le \tau_n \|T_0(g - g_0)\|_{L^2(P_{W,V})} \qquad \text{for all } g \in \mathcal{G}_n,$$

then $\mathfrak{m}_n(r) \le \tau_n r$. This gives a linear error bound of the form used by Xu et al. (2021); the modulus formulation also accommodates nonlinear error bounds.

Completeness alone does not make this modulus small uniformly over a growing class. For a fixed L^2-compact class containing g_0, continuity of T_0 and uniqueness of its moment solution imply $\mathfrak{m}(r) \to 0$ as $r \downarrow 0$: otherwise compactness would produce a distinct limit with the same conditional moments.

Corollary A.1 (Conditional consistency of the selected MAP predictor)**.** *Under Proposition A.3 and Assumption A.3, let*

$$\Delta_n = a_n^2 + b_n^{\mathrm{MC}} + 2u_n + 2e_n + 2t_n.$$

If $\mathfrak{m}_n(\sqrt{\Delta_n}) \to_p 0$*, then*

$$\|g_{\widehat{k}} - g_0\|_{L^2(P_{X,V})} \longrightarrow_p 0. \tag{18}$$

Proof. On the event in Assumption A.3, Eq. (17) gives $\mathcal{B}(g_{\widehat{k}}) \le \Delta_n$. The definition of $\mathfrak{m}_n$ bounds the structural error by $\mathfrak{m}_n(\sqrt{\Delta_n})$. The complement of that event has probability tending to zero. □

Remark A.4 (Conditions and scope)**.** *The uniform term u_n handles reuse of training observations; no exact same-sample latent-exogeneity claim is needed. A fixed uniformly bounded, sup-norm totally bounded class with $\mathbb{E}Y^2 < \infty$ satisfies the required uniform law, whereas growing neural classes require additional complexity control. For bounded candidate outputs $|g_k| \le B$, the Monte Carlo bias obeys $\nu_k \le B^2/M$. Conditional on a sample with $\max_i |Y_i| \le A$, bounded differences additionally give*

$$t_n \le B(A+B)\sqrt{\frac{8\log(2K/\delta)}{nM}}$$

with conditional probability at least $1-\delta$. Thus increasing n can reduce score fluctuation without removing the candidate-dependent bias at fixed M. The corollary retains that term explicitly.

Assumption A.3 describes the approximation quality of the fitted candidates under the final covariate-only MAP readout. Corollary A.1 combines this condition with first-stage, empirical, and Monte Carlo error control and the restricted inverse modulus to establish $L^2(P_{X,V})$ consistency of the selected predictor.

A.4 A Complementary Observed-Data Comparison

The candidate-approximation condition above can also be investigated through a reduced-form comparison after the original selection has been completed. This comparison is an optional diagnostic of the unchanged selected predictor, not a replacement selection rule or an experiment reported here. Let g be that predictor and let $\widehat{K}$ be its fitted treatment kernel at the covariate-only readout. Put

$$m_{\mathrm{RF}}(w,v) = \mathbb{E}[Y \mid W=w, V=v], \quad h_0 = T_0 g, \quad \widehat{h} = \int g(x,v)\widehat{K}(dx \mid w,v).$$

Let $\widehat{m}_{\mathrm{RF}}$ be a benchmark fitted using observed variables, with $a_n^{\mathrm{RF}} = \|\widehat{m}_{\mathrm{RF}} - m_{\mathrm{RF}}\|_2$. After fixing both the selected model and this benchmark, generate fresh draws X_{im} from $\widehat{K}(\cdot \mid W_i, V_i)$, conditionally independently over (i,m). For $M \ge 2$, write $G_{im} = g(X_{im}, V_i)$, $\overline{G}_i = M^{-1}\sum_m G_{im}$, and define

$$\mathcal{U}_n = \frac{1}{n}\sum_i \left\{ (Y_i - \overline{G}_i)^2 - \frac{s_{G,i}^2}{M} \right\}, \quad s_{G,i}^2 = \frac{1}{M-1}\sum_m (G_{im} - \overline{G}_i)^2,$$

$$\widehat{D}_n = \mathcal{U}_n - P_n(Y - \widehat{m}_{\mathrm{RF}})^2.$$

The diagnostic gap $\widehat{D}_n$ uses no structural truth. For deterministic classes $\mathcal{G}_n, \mathcal{H}_n$ containing all possible selected predictors and fitted benchmarks, respectively, define

$$u_n^G = \sup_{g' \in \mathcal{G}_n} |(P_n - P)(Y - T_0 g')^2|,$$

$$u_n^{\mathrm{RF}} = \sup_{m \in \mathcal{H}_n} |(P_n - P)(Y - m)^2|.$$

Put

$$e_n^{\mathrm{stage}} = |P_n(Y - \widehat{h})^2 - P_n(Y - h_0)^2|, \qquad \zeta_n = \mathcal{U}_n - P_n(Y - \widehat{h})^2.$$

Proposition A.4 (Reduced-form comparison bound)**.** *Under the mean-IV restriction, the stated uniform bounds, square-integrable risks and finite conditional second moments of the diagnostic draws,*

$$\left|\mathcal{B}(g) - \widehat{D}_n - (a_n^{\mathrm{RF}})^2\right| \le u_n^G + u_n^{\mathrm{RF}} + e_n^{\mathrm{stage}} + |\zeta_n|. \tag{19}$$

Consequently, with $[u]_+ = \max(u, 0)$*,*

$$\mathcal{B}(g) \le [\widehat{D}_n]_+ + (a_n^{\mathrm{RF}})^2 + u_n^G + u_n^{\mathrm{RF}} + e_n^{\mathrm{stage}} + |\zeta_n|.$$

Applying the restricted inverse modulus to the square root of this upper bound gives a sufficient condition for structural consistency of this already selected predictor, without an assumption about another good candidate.

Proof. Conditional-mean orthogonality gives

$$\mathcal{B}(g) = R_0(g) - \mathbb{E}[(Y - \widehat{m}_{\mathrm{RF}}(W, V))^2] + (a_n^{\mathrm{RF}})^2.$$

Conditional on the fitted quantities and data, the rowwise diagnostic equals

$$\frac{1}{M(M-1)} \sum_{m \neq l} (Y_i - G_{im})(Y_i - G_{il}),$$

whose expectation is $(Y_i - \widehat{h}(W_i, V_i))^2$ by independence. Insert the two empirical risks and this conditional expectation into the preceding population identity. The triangle inequality gives Eq. (19); replacing the observed gap by its positive part can only increase the upper bound. □

Fresh diagnostic draws after selection are essential for the displayed conditional expectation. If $v_i = \mathrm{Var}(G_{im} \mid \text{data and fitted quantities})$ and $a_i = Y_i - \widehat{h}(W_i, V_i)$, the conditional diagnostic variance is exactly

$$\mathrm{Var}(\mathcal{U}_n \mid \text{data and fitted quantities}) = \frac{1}{n^2} \sum_i \left\{ \frac{4a_i^2 v_i}{M} + \frac{2v_i^2}{M(M-1)} \right\}.$$

To obtain this formula, center each draw and expand the products; distinct centered pairs are uncorrelated, including pairs sharing one index. With suitable moment control, the diagnostic fluctuation vanishes even at fixed $M \ge 2$. This does not remove the variance term from the original selection score, which is unchanged.

The observable gap alone is not a complete confidence certificate: benchmark consistency, first-stage accuracy, uniform generalization and inverse stability still require justification. A poor benchmark can make the gap misleadingly small, which is why $(a_n^{\mathrm{RF}})^2$ is retained. Nor does the current algorithm automatically make the diagnostic gap tend to zero. Using the same observations for fitting and this diagnostic requires the uniform bounds above; fixing fitted parameters afterwards does not establish sample independence. This comparison is not used here to reselect candidates or tune the estimator.

B Empirical Check of Latent-State Exogeneity

Lemma A.1 establishes conditional mean exogeneity of the covariate-only readout for independent new units. This appendix checks the argument empirically on the BGM-IV models of the demand design benchmark of Section 4.2, whose unobserved confounder is written explicitly and can be replayed exactly.

Design. In the demand design the structural error is $\epsilon = \rho U + \sqrt{1-\rho^2}\,\xi$, where U is the unobserved demand shock that also enters the price equation and ξ is an independent standard normal disturbance. Because the price equation is $P = 25 + (C+3)\psi(T) + U$, the shock can be recovered exactly in the simulation as $U = P - 25 - (C+3)\psi(T)$ from the observed (P, C, T) and the known seasonal function ψ, so the only part of Y that cannot be inferred from pre-outcome observables is $\sqrt{1-\rho^2}\,\xi$. The disturbance ξ is independent of (X, V, W) and enters the data only through Y, an association between a latent coordinate and ξ beyond sampling variability would therefore indicate that the latent state has absorbed outcome information. We correlate the latent states with both ϵ and ξ, and separately with U to check whether the latent representation encodes the unobserved confounder itself. We report the strongest-confounding setting, $\rho = 0.9$, where U accounts for 81% of the variance of ϵ and deconfounding is hardest.

Protocol. The models are the twenty repetitions of Table 8 at $\rho = 0.9$ for training sample sizes $n = 5{,}000$ and $n = 10{,}000$, each selected among ten candidates by the criterion of Eq. (23). Every model was restored from its archived checkpoint, the restored models reproduce their reported structural MSE (largest absolute difference 1.8×10^{-5}) on training data that is bit-identical to the archived run. Latent states are inferred by the covariate-only MAP of Eq. (11), exactly as at evaluation, on two samples: the training sample of each model, and one held-out sample of $5{,}000$ units generated from an independent seed and shared by all models in both columns of Table 3. For each model we compute the Pearson correlation between each of the eight latent coordinates and each of ϵ and ξ, giving $20 \times 8 \times 2 = 320$ coordinate-level tests per column. A correlation is called *outside the null band* when $|r| > 1.96/\sqrt{m}$, with m the number of units in the sample on which it is computed (± 0.0277 on the held-out sample and on the $n = 5{,}000$ training sample, ± 0.0196 on the $n = 10{,}000$ training sample). If a latent coordinate were unrelated to the target, about 5% of these correlations would fall outside the band by chance, which is the nominal level of the test. Therefore, we compare the observed fraction outside the band (*Outside the null band* in Table 3) with this 5% reference, following the standard residual diagnostic for white noise (Hyndman et al., 2014). Multiplicity is handled by Holm's procedure (Holm, 1979) within families defined by one model and one target (ϵ or ξ) over its eight coordinates, i.e. 40 families per column. We count the tests that are outside the band and remain significant after the correction.

Controls. Three controls calibrate the tests. Negative controls replace the latent coordinate by the instrument W, the time covariate T, the customer group S, and an independent $\mathcal{N}(0, 1)$ variable, and apply the same correlation tests to each of them against ϵ and ξ on the training sample (160 tests per column). Their true correlation is zero, so their rejection rate shows the behavior of the procedure under the null (Lipsitch et al., 2010). Positive controls inject a known correlation with ξ into the first coordinate of $\mathbf{z}_0$ on the training sample, at $r = 0.03$ and 0.05 for $n = 5{,}000$ ($1.08\times$ and $1.80\times$ the training null band) and $r = 0.021$ and 0.035 for $n = 10{,}000$ ($1.07\times$ and $1.79\times$). An injection counts as detected when the F test of the $\mathbf{z}_0$ block against ξ has $p < 0.05$. Finally, $\mathrm{corr}(\epsilon, U)$ is the unobserved confounding present in the data and must be detected: its mean over the twenty models is 0.90 (range 0.896–0.906 at $n = 5{,}000$ and 0.896–0.904 at $n = 10{,}000$).

Results. Table 3 summarizes the outcome. On the held-out sample, 1.9% ($n = 5{,}000$) and 0.3% ($n = 10{,}000$) of the 320 coordinate-level correlations fall outside the null band, none survives the Holm correction, and the largest correlation is 1.26 times the band. On the training sample, where the network weights were fitted with the outcomes of the very same subjects, the rejection rates are 4.1% in both columns, below the nominal 5% and within the range of the negative controls (5.0% and 6.9%). Only one of the 320 tests at $n = 10{,}000$ remains significant after Holm correction, whereas under the null hypothesis at most two of the 40 families per column are expected to contain a rejection. The positive controls confirm that the tests have adequate statistical power at the relevant effect sizes, the injected correlations at about 1.8 times the band are detected in all twenty models, and $\mathrm{corr}(\epsilon, U)$ is detected at 32 to 46 times the band. Injections at about 1.1 times the band, which lie at the detection threshold of a 5%-level test, are detected in $3/20$ and $6/20$ models, as expected for a signal of that size. The check therefore resolves associations of about twice the null band and larger, which is why we describe the result as no detectable association. Finally, the latent states carry no detectable information about the unobserved confounder: the largest $|\mathrm{corr}(z, U)|$ is 0.036, and the linear R^2 of U on the $\mathbf{z}_0$ block is at most 0.002. The held-out results are consistent with Lemma A.1; the training-sample results provide a diagnostic of outcome-related associations in the re-inferred prediction states.

Table 3: Latent-state exogeneity check on the demand design at $\rho = 0.9$. Columns are the training sample sizes of the twenty BGM-IV models ($\gamma = 0.01$, ten candidates with training-set selection). Each percentage has denominator $320 = 20$ models $\times$ 8 latent coordinates $\times$ $\{\epsilon, \xi\}$, with latent states inferred by the covariate-only MAP. The held-out block uses one sample of 5,000 units shared by both columns; the null band is $\pm 1.96/\sqrt{m}$ for a sample of m units (± 0.0277 held-out and for the $n = 5{,}000$ training sample, ± 0.0196 for the $n = 10{,}000$ training sample), and "largest $|r|$ / null band" is relative to the band of the sample in question. Holm correction is applied within each model and target over its eight coordinates (40 families per column). Negative controls are W, T, S, and an independent $\mathcal{N}(0, 1)$ variable, each tested for correlation with ϵ and ξ (160 tests per column). Injected signals are correlations with ξ added to the first coordinate of $\mathbf{z}_0$ on the training sample ($r = 0.03/0.05$ at $n = 5{,}000$, $r = 0.021/0.035$ at $n = 10{,}000$), counted as detected when the $\mathbf{z}_0$-block F test has $p < 0.05$. $\mathrm{corr}(\epsilon, U)$ is the mean over the twenty models.

	$n = 5{,}000$	$n = 10{,}000$
Held-out sample (5,000 units, shared by both columns)		
Outside the null band (about 5% if unrelated)	1.9% (6/320)	0.3% (1/320)
Significant after within-family Holm correction	0/320	0/320
Largest $\|r\|$ / null band	1.25×	1.26×
Largest $\|\mathrm{corr}(z, U)\|$	0.035	0.028
Training sample		
Outside the null band (about 5% if unrelated)	4.1% (13/320)	4.1% (13/320)
Significant after within-family Holm correction	0/320	1/320
Largest $\|r\|$ / null band	1.26×	1.87×
Largest $\|\mathrm{corr}(z, U)\|$	0.036	0.027
Controls (training sample)		
Negative controls outside the null band	5.0% (8/160)	6.9% (11/160)
Injected signal at ≈1.1× null band detected	3/20	6/20
Injected signal at ≈1.8× null band detected	20/20	20/20
$\mathrm{corr}(\epsilon, U)$ (/ null band)	0.90 (32–33×)	0.90 (46×)

C Sensitivity Analysis

The exponent γ in Eq. (5) sets the weight of the IV-integrated pseudo-likelihood in the latent update of Eq. (10), it does not enter the objectives of the generators. All experiments in Section 4 use $\gamma = 0.01$. Figure 5 and Table 4 examine the sensitivity of BGM-IV to this choice on the demand-design benchmark over the full grid of sample sizes and confounding levels, for $\gamma \in \{0.01, 0.05, 0.1, 0.5, 1\}$, where $\gamma = 1$ uses the pseudo-likelihood without tempering. Only γ differs between the five settings, within each repetition they share the training data and the initialization seeds, and each is trained with the multi-start protocol of Appendix F, so the comparison is paired at the level of repetitions.

The three smallest values, $\gamma \in \{0.01, 0.05, 0.1\}$, give structural MSEs that differ by less than the repetition-to-repetition variability in every cell, with no consistent ordering across sample sizes or confounding levels. Larger weights degrade the estimates monotonically, and the degradation grows with the sample size, relative to $\gamma = 0.01$, the mean MSE at $\gamma = 1$ is higher by a factor of 1.18 at $n = 1{,}000$, 1.52 at $n = 5{,}000$, and 1.74 at $n = 10{,}000$, and $\gamma = 0.01$ has the lower MSE in 270 of the 300 paired repetitions (196 of 200 at $n \geq 5{,}000$). The comparison with the baselines of Table 8 follows the same pattern, $\gamma = 0.01$ attains the lowest mean MSE of all methods in every cell and $\gamma \in \{0.05, 0.1\}$ in 14 of 15, whereas $\gamma \in \{0.5, 1\}$ fall behind DFIV in all ten cells with $n \geq 5{,}000$. Giving the pseudo-likelihood term a weight close to that of a proper likelihood therefore harms the structural estimate, whereas a small positive weight leaves the results in a flat region that contains $\gamma = 0.01$. The selection criterion of Eq. (23), which uses only training data, gives the same ordering: in every cell, its value for the selected candidate, averaged over the twenty repetitions, is higher at $\gamma = 0.5$ and at $\gamma = 1$ than at each of $\gamma \in \{0.01, 0.05, 0.1\}$. The ordering among the three small values changes from cell to cell, so the criterion supports a small γ but does not select one value within that range. A small γ also limits the weight of the outcome in the latent update during training, consistent with the exogeneity check of Appendix B.

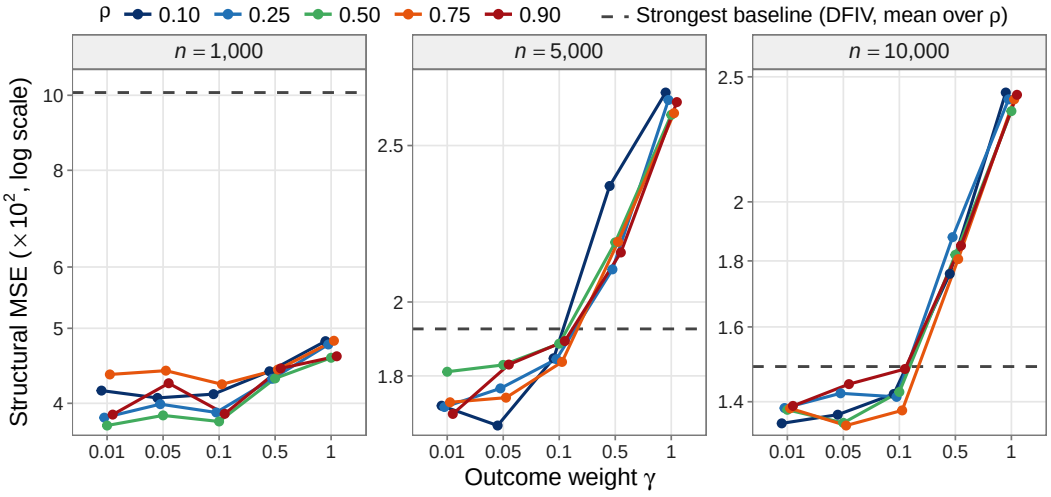


Figure 5: Sensitivity of BGM-IV to the tempering exponent γ on the demand-design benchmark. Panels correspond to training sample sizes $n = 1{,}000$, $5{,}000$, and $10{,}000$; each line is one confounding level ρ, and each point is the mean structural MSE over 20 repetitions (standard deviations are given in Table 4). The five values of γ are placed at equal horizontal spacing and the vertical axis is logarithmic, in units of 10^2. The dashed line marks the strongest baseline in each panel (DFIV), averaged over ρ.

Table 4: Sensitivity of BGM-IV to the tempering exponent γ on the demand-design benchmark. Entries are structural MSE divided by 10^2, mean $\pm$ sample standard deviation over 20 repetitions, grouped by training sample size n. Only γ differs between columns, within each repetition the five settings use the same training data and initialization seeds, and each is selected among ten candidates by the training-set criterion. The lowest mean in each row is in bold.

ρ	$\gamma = 0.01$	$\gamma = 0.05$	$\gamma = 0.1$	$\gamma = 0.5$	$\gamma = 1$
$n = 1{,}000$					
0.1	4.15 ± 0.91	$\mathbf{4.06 \pm 0.72}$	4.11 ± 0.80	4.40 ± 0.79	4.81 ± 0.74
0.25	$\mathbf{3.83 \pm 0.69}$	3.99 ± 0.79	3.89 ± 0.74	4.30 ± 0.84	4.76 ± 1.07
0.5	$\mathbf{3.74 \pm 0.47}$	3.86 ± 0.52	3.79 ± 0.52	4.31 ± 0.84	4.58 ± 0.83
0.75	4.36 ± 0.70	4.41 ± 1.06	$\mathbf{4.23 \pm 0.78}$	4.42 ± 0.74	4.82 ± 0.89
0.9	$\mathbf{3.87 \pm 0.69}$	4.25 ± 0.79	3.88 ± 0.55	4.44 ± 1.08	4.60 ± 1.13
$n = 5{,}000$					
0.1	1.72 ± 0.25	$\mathbf{1.68 \pm 0.19}$	1.85 ± 0.24	2.36 ± 0.55	2.70 ± 0.59
0.25	$\mathbf{1.72 \pm 0.37}$	1.77 ± 0.40	1.84 ± 0.43	2.10 ± 0.28	2.67 ± 0.64
0.5	$\mathbf{1.81 \pm 0.38}$	1.83 ± 0.36	1.88 ± 0.31	2.18 ± 0.32	2.61 ± 0.46
0.75	$\mathbf{1.73 \pm 0.39}$	1.74 ± 0.36	1.84 ± 0.38	2.18 ± 0.39	2.62 ± 0.47
0.9	$\mathbf{1.70 \pm 0.22}$	1.83 ± 0.30	1.89 ± 0.32	2.15 ± 0.38	2.66 ± 0.52
$n = 10{,}000$					
0.1	$\mathbf{1.35 \pm 0.18}$	1.37 ± 0.21	1.42 ± 0.23	1.76 ± 0.21	2.43 ± 0.42
0.25	$\mathbf{1.38 \pm 0.25}$	1.42 ± 0.27	1.41 ± 0.24	1.88 ± 0.37	2.40 ± 0.36
0.5	1.38 ± 0.26	$\mathbf{1.35 \pm 0.31}$	1.42 ± 0.25	1.82 ± 0.28	2.35 ± 0.41
0.75	1.38 ± 0.25	$\mathbf{1.34 \pm 0.21}$	1.38 ± 0.34	1.81 ± 0.25	2.40 ± 0.38
0.9	$\mathbf{1.39 \pm 0.31}$	1.44 ± 0.37	1.48 ± 0.34	1.85 ± 0.53	2.42 ± 0.64

D Detailed Training and Evaluation Algorithm

Algorithms 1 and 2 give the EGM initialization and the stochastic optimization training of Section 3.5 for one candidate; Algorithm 3 runs them for K candidates and applies the selection rule of

Eq. (23), Algorithm 4 gives the structural prediction and uncertainty quantification of Section 3.6. Hyperparameter values are listed in Appendix E.

IV-aware EGM initialization. The IV-aware EGM initialization provides the starting values of the generators and the latent states. It trains an encoder e_ψ from the covariates to the latent space jointly with the three generators (G, H, F) by minimizing

$$\mathcal{L}_{\mathrm{EGM}}(\psi, \theta, \phi, \omega) = \mathcal{L}_{\mathrm{adv}} + \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_x + \mathcal{L}_y, \tag{20}$$

where all terms are evaluated at the encoded state $\tilde{\mathbf{z}} = e_\psi(\mathbf{v})$. The adversarial term $\mathcal{L}_{\mathrm{adv}}$ matches the distribution of $e_\psi(V)$ to the prior p_Z through a discriminator trained with a gradient penalty, and the reconstruction term $\mathcal{L}_{\mathrm{rec}}$ penalizes the error of reconstructing $\mathbf{v}$ from $\tilde{\mathbf{z}}$ and of recovering a prior draw $\mathbf{z}$ through the generator–encoder cycle, as in EGM (Liu et al., 2024). The remaining two terms make the initialization IV-aware. $\mathcal{L}_x$ is the first-stage squared error $\{x - \mu_\phi(w, \tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_2)\}^2$, and $\mathcal{L}_y$ is the integrated outcome moment

$$\Big\{ y - \int \mu_\omega(x, \tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1)\, p_\phi(x \mid w, \tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_2)\, dx \Big\}^2, \tag{21}$$

which applies the integral of Eq. (4) to the outcome mean rather than to the outcome density. Under the conditions of Proposition A.2, the population minimizer of Eq. (21) is the structural mean, so the outcome component of the initialization targets the structural function rather than the regression of Y on the observed treatment.

For a scalar treatment with a Gaussian first stage, the integral in Eq. (21) is one-dimensional and its integrand is smooth, and we evaluate it by Gauss–Hermite quadrature with nodes t_q and normalized weights λ_q,

$$\int \mu_\omega(x, \tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1)\, p_\phi(x \mid w, \tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_2)\, dx \approx \sum_{q=1}^{Q} \lambda_q\, \mu_\omega\big(\mu_\phi(w, \tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_2) + \sqrt{2}\,\tilde{\sigma}\, t_q,\ \tilde{\mathbf{z}}_0, \tilde{\mathbf{z}}_1\big),$$

where $\tilde{\sigma}$ is a running estimate of the first-stage residual scale. The stochastic optimization training uses Monte Carlo instead, as in Eq. (9), because the outcome density can be sharply peaked in x when σ_ω is small. For a binary treatment the integral is an exact two-point sum over the treatment values weighted by p_ϕ. At the end of the initialization the latent states are set to $\mathbf{z}^{(i)} \leftarrow e_\psi(\mathbf{v}^{(i)})$, and the encoder is not updated afterwards. The Algorithm 1 gives the complete procedure.

Reproducibility of the multi-start procedure. All K candidates are trained on the same training sample with the same optimization schedule. Each candidate receives its own network-initialization seed and its own seed for the Monte Carlo draws of the selection criterion, whereas the mini-batch order of the EGM initialization and the random stream of the stochastic optimization training are shared across candidates, so that candidates differ only in their initialization. Candidates whose criterion is not finite are excluded before the minimum is taken, and ties are broken by the smallest candidate index. The K candidates are part of the estimator's computational budget and are not treated as independent repetitions; all repetitions reported in Section 4 use independent training samples.

Algorithm 1 IV-aware EGM initialization for one candidate

Require: Training data $\mathcal{D} = \{(x^{(i)}, y^{(i)}, \mathbf{v}^{(i)}, w^{(i)})\}_{i=1}^{n}$; prior $p(\mathbf{z})$; initialization seed s; EGM iterations T_{EGM} with n_d discriminator steps per generator step; Gauss–Hermite nodes and weights $\{(t_q, \lambda_q)\}_{q=1}^{Q}$; step size α_{EGM}.

Ensure: Initialized generators $(\theta^{(0)}, \phi^{(0)}, \omega^{(0)})$ and latent states $\{\mathbf{z}^{(i,0)}\}_{i=1}^{n}$.

Initialize the encoder e_ψ, the discriminator D_κ, and the generators (θ, ϕ, ω) from seed s.
for $t = 1, \ldots, T_{\mathrm{EGM}}$ **do**
  for n_d steps **do**
    Sample a mini-batch $\mathcal{B} \subset \{1, \ldots, n\}$ and prior draws $\mathbf{z}^{(i)} \sim p(\mathbf{z})$, $i \in \mathcal{B}$, and set $\tilde{\mathbf{z}}^{(i)} \leftarrow e_\psi(\mathbf{v}^{(i)})$.
    Update κ by gradient ascent on $|\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \{D_\kappa(\mathbf{z}^{(i)}) - D_\kappa(\tilde{\mathbf{z}}^{(i)})\}$ with a gradient penalty (WGAN-GP).
  end for
  Sample a new mini-batch $\mathcal{B} \subset \{1, \ldots, n\}$ and new prior draws $\mathbf{z}^{(i)} \sim p(\mathbf{z})$, $i \in \mathcal{B}$.
  Set $\tilde{\mathbf{z}}^{(i)} \leftarrow e_\psi(\mathbf{v}^{(i)})$ and partition $\tilde{\mathbf{z}}^{(i)} = (\tilde{\mathbf{z}}_0^{(i)}, \tilde{\mathbf{z}}_1^{(i)}, \tilde{\mathbf{z}}_2^{(i)}, \tilde{\mathbf{z}}_3^{(i)})$ for $i \in \mathcal{B}$.
  Set $\mathcal{L}_{\mathrm{adv}} \leftarrow -|\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} D_\kappa(\tilde{\mathbf{z}}^{(i)})$.
  Set $\mathcal{L}_{\mathrm{rec}} \leftarrow |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \big[\|\mathbf{v}^{(i)} - \boldsymbol{\mu}_\theta(\tilde{\mathbf{z}}^{(i)})\|^2 + \|\mathbf{z}^{(i)} - e_\psi(\boldsymbol{\mu}_\theta(\mathbf{z}^{(i)}))\|^2\big]$.
  Set $\mathcal{L}_x \leftarrow |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \{x^{(i)} - \mu_\phi(w^{(i)}, \tilde{\mathbf{z}}_0^{(i)}, \tilde{\mathbf{z}}_2^{(i)})\}^2$.
  for each $i \in \mathcal{B}$ **do**
    Compute the quadrature estimate of the integrated outcome mean $\int \mu_\omega(x, \tilde{\mathbf{z}}_0^{(i)}, \tilde{\mathbf{z}}_1^{(i)})\, p_\phi(x \mid w^{(i)}, \tilde{\mathbf{z}}_0^{(i)}, \tilde{\mathbf{z}}_2^{(i)})\, dx$:
    $\bar{\mu}^{(i)} \leftarrow \sum_{q=1}^{Q} \lambda_q\, \mu_\omega\big(\mu_\phi(w^{(i)}, \tilde{\mathbf{z}}_0^{(i)}, \tilde{\mathbf{z}}_2^{(i)}) + \sqrt{2}\, \tilde{\sigma}\, t_q,\ \tilde{\mathbf{z}}_0^{(i)}, \tilde{\mathbf{z}}_1^{(i)}\big)$, where $\tilde{\sigma}$ is the running first-stage residual scale.
  end for
  Set $\mathcal{L}_y \leftarrow |\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \{y^{(i)} - \bar{\mu}^{(i)}\}^2$.
  Update $(\psi, \theta, \phi, \omega)$ by one gradient step of size α_{EGM} on $\mathcal{L}_{\mathrm{adv}} + \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_x + \mathcal{L}_y$.
end for
Set $\mathbf{z}^{(i,0)} \leftarrow e_\psi(\mathbf{v}^{(i)})$ for $i = 1, \ldots, n$ and $(\theta^{(0)}, \phi^{(0)}, \omega^{(0)}) \leftarrow (\theta, \phi, \omega)$.
Return $(\theta^{(0)}, \phi^{(0)}, \omega^{(0)})$ and $\{\mathbf{z}^{(i,0)}\}_{i=1}^{n}$.

Algorithm 2 Stochastic optimization training for one candidate

Require: Training data $\mathcal{D}$; prior $p(\mathbf{z})$; initialized generators $(\theta^{(0)}, \phi^{(0)}, \omega^{(0)})$ and latent states $\{\mathbf{z}^{(i,0)}\}_{i=1}^{n}$ from Algorithm 1; BGM epochs T_{BGM}; Monte Carlo size M; step sizes $\alpha_\theta, \alpha_\phi, \alpha_\omega, \beta_z$; tempering exponent γ.
Ensure: Trained generators (θ, ϕ, ω).

Set $\tau \leftarrow 0$.
for epoch $= 1, \ldots, T_{\mathrm{BGM}}$ **do**
  for each mini-batch $\mathcal{B} \subset \{1, \ldots, n\}$ **do**
    Partition $\mathbf{z}^{(i,\tau)} = (\mathbf{z}_0^{(i,\tau)}, \mathbf{z}_1^{(i,\tau)}, \mathbf{z}_2^{(i,\tau)}, \mathbf{z}_3^{(i,\tau)})$ for $i \in \mathcal{B}$.
    Set $\mathcal{L}_V(\theta) \leftarrow -|\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \log p_\theta(\mathbf{v}^{(i)} \mid \mathbf{z}^{(i,\tau)})$.
    Update $\theta^{(\tau+1)} \leftarrow \theta^{(\tau)} - \alpha_\theta \nabla_\theta \mathcal{L}_V(\theta^{(\tau)})$.
    Set $\mathcal{L}_X(\phi) \leftarrow -|\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \log p_\phi(x^{(i)} \mid w^{(i)}, \mathbf{z}_0^{(i,\tau)}, \mathbf{z}_2^{(i,\tau)})$.
    Update $\phi^{(\tau+1)} \leftarrow \phi^{(\tau)} - \alpha_\phi \nabla_\phi \mathcal{L}_X(\phi^{(\tau)})$.
    for each $i \in \mathcal{B}$ **do**
      if X is continuous **then**
        Draw $x^{(i,m)} \sim p_{\phi^{(\tau+1)}}(x \mid w^{(i)}, \mathbf{z}_0^{(i,\tau)}, \mathbf{z}_2^{(i,\tau)})$, $m = 1, \ldots, M$.
        Evaluate $\log p^{(i,m)} \leftarrow \log p_{\omega^{(\tau)}}(y^{(i)} \mid x^{(i,m)}, \mathbf{z}_0^{(i,\tau)}, \mathbf{z}_1^{(i,\tau)})$.
        Set $c^{(i)} \leftarrow \max_m \log p^{(i,m)}$ and $\log \widehat{L}_{\mathrm{IV}}^{(i)} \leftarrow c^{(i)} + \log\{\sum_{m=1}^{M} \exp(\log p^{(i,m)} - c^{(i)})\} - \log M$, which evaluates Eq. (9) using the standard log-sum-exp stabilization.
      else
        Compute $\log \widehat{L}_{\mathrm{IV}}^{(i)}$ by evaluating Eq. (3) as an exact two-point mixture over $x \in \{0, 1\}$.
        The mixture uses weights from $p_{\phi^{(\tau+1)}}(x \mid w^{(i)}, \mathbf{z}_0^{(i,\tau)}, \mathbf{z}_2^{(i,\tau)})$ and component densities $p_{\omega^{(\tau)}}(y^{(i)} \mid x, \mathbf{z}_0^{(i,\tau)}, \mathbf{z}_1^{(i,\tau)})$.
      end if
    end for
    Set $\mathcal{L}_Y(\omega) \leftarrow -|\mathcal{B}|^{-1} \sum_{i \in \mathcal{B}} \log \widehat{L}_{\mathrm{IV}}^{(i)}$.
    Update $\omega^{(\tau+1)} \leftarrow \omega^{(\tau)} - \alpha_\omega \nabla_\omega \mathcal{L}_Y(\omega^{(\tau)})$.
    for each $i \in \mathcal{B}$ **do**
      Recompute $\log \widehat{L}_{\mathrm{IV}}^{(i)}$ as above with $\omega^{(\tau+1)}$ in place of $\omega^{(\tau)}$, using M fresh draws $x^{(i,m)} \sim p_{\phi^{(\tau+1)}}(x \mid w^{(i)}, \mathbf{z}_0^{(i,\tau)}, \mathbf{z}_2^{(i,\tau)})$ when X is continuous.
      Update the latent state by ascent on the objective in Eq. (10):
      $\mathbf{z}^{(i,\tau+1)} \leftarrow \mathbf{z}^{(i,\tau)} + \beta_z \nabla_{\mathbf{z}^{(i,\tau)}} \{ \log p(\mathbf{z}^{(i,\tau)}) + \log p_{\theta^{(\tau+1)}}(\mathbf{v}^{(i)} \mid \mathbf{z}^{(i,\tau)}) + \log p_{\phi^{(\tau+1)}}(x^{(i)} \mid w^{(i)}, \mathbf{z}_0^{(i,\tau)}, \mathbf{z}_2^{(i,\tau)}) + \gamma \log \widehat{L}_{\mathrm{IV}}^{(i)} \}$.
    end for
    Set $\mathbf{z}^{(i,\tau+1)} \leftarrow \mathbf{z}^{(i,\tau)}$ for $i \notin \mathcal{B}$, and $\tau \leftarrow \tau + 1$.
  end for
end for
Return $(\theta, \phi, \omega) \leftarrow (\theta^{(\tau)}, \phi^{(\tau)}, \omega^{(\tau)})$.

Algorithm 3 Multi-start training and model selection

Require: Training data $\mathcal{D}$; prior $p(\mathbf{z})$; number of candidates K; initialization seeds $s_1, \dots, s_K$; Monte Carlo size M; the inputs of Algorithms 1 and 2.

Ensure: Selected generators $(\hat{\theta}, \hat{\phi}, \hat{\omega})$.

for candidate $k = 1, \dots, K$ **do**
  Run Algorithm 1 with seed s_k to obtain $(\theta^{(0)}, \phi^{(0)}, \omega^{(0)})$ and $\{\mathbf{z}^{(i,0)}\}_{i=1}^{n}$.
  Run Algorithm 2 from this initialization to obtain $(\theta_k, \phi_k, \omega_k)$.
  for $i = 1, \dots, n$ **do**
    Compute $\hat{\mathbf{z}}^{(i)} = \arg\max_{\mathbf{z}} \{\log p(\mathbf{z}) + \log p_{\theta_k}(\mathbf{v}^{(i)} \mid \mathbf{z})\}$ (Eq. (11)).
    Draw $x^{(i,m)} \sim p_{\phi_k}(x \mid w^{(i)}, \hat{\mathbf{z}}_0^{(i)}, \hat{\mathbf{z}}_2^{(i)})$, $m = 1, \dots, M$, and set $\bar{\mu}_k^{(i)} \leftarrow M^{-1} \sum_{m=1}^{M} \mu_{\omega_k}(x^{(i,m)}, \hat{\mathbf{z}}_0^{(i)}, \hat{\mathbf{z}}_1^{(i)})$.
  end for
  Set $S_k \leftarrow n^{-1} \sum_{i=1}^{n} \{y^{(i)} - \bar{\mu}_k^{(i)}\}^2$ (Eq. (23)).
end for
Discard candidates with non-finite S_k; set $\hat{k} \leftarrow \arg\min_k S_k$, ties broken by the smallest index.
Return $(\hat{\theta}, \hat{\phi}, \hat{\omega}) \leftarrow (\theta_{\hat{k}}, \phi_{\hat{k}}, \omega_{\hat{k}})$.

Algorithm 4 Structural prediction and uncertainty quantification

Require: Selected generators $(\hat{\theta}, \hat{\phi}, \hat{\omega})$; evaluation pairs $\{(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)})\}_{r=1}^{R}$; benchmark structural values $g_0(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)})$; number of chains C, warm-up length T_{warm}, and retained draws per chain J_c; interval level $1 - \alpha$.

Ensure: Structural predictions $\hat{g}(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)})$, structural MSE, and $100(1-\alpha)\%$ predictive intervals.

MAP readout.
for test pair $(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)})$ **do**
  Compute $\hat{\mathbf{z}}(\mathbf{v}_{\text{eval}}^{(r)}) = \arg\max_{\mathbf{z}} \{\log p(\mathbf{z}) + \log p_{\hat{\theta}}(\mathbf{v}_{\text{eval}}^{(r)} \mid \mathbf{z})\}$.
  Set $\hat{g}(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)}) \leftarrow \mu_{\hat{\omega}}(x_{\text{eval}}^{(r)}, \hat{\mathbf{z}}_0(\mathbf{v}_{\text{eval}}^{(r)}), \hat{\mathbf{z}}_1(\mathbf{v}_{\text{eval}}^{(r)}))$.
end for
Set $\text{MSE}_{\text{struct}} \leftarrow R^{-1} \sum_{r=1}^{R} \{\hat{g}(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)}) - g_0(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)})\}^2$.
Posterior sampling.
for each distinct covariate value $\mathbf{v}$ among $\{\mathbf{v}_{\text{eval}}^{(r)}\}_{r=1}^{R}$ **do**
  Run a short pilot of Hamiltonian Monte Carlo on $\log p(\mathbf{z}) + \log p_{\hat{\theta}}(\mathbf{v} \mid \mathbf{z})$ to estimate a diagonal mass matrix.
  Run C chains from overdispersed starting points with this mass matrix, each with T_{warm} warm-up iterations, during which the step size is adapted by dual averaging, followed by J_c retained draws; pool the $J = C\,J_c$ draws $\{\mathbf{z}^{(j)}(\mathbf{v})\}_{j=1}^{J}$.
end for
Intervals.
for test pair $(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)})$ **do**
  For $j = 1, \dots, J$, with $\mathbf{z}^{(j)} = \mathbf{z}^{(j)}(\mathbf{v}_{\text{eval}}^{(r)})$: set $m_j \leftarrow \mu_{\hat{\omega}}(x_{\text{eval}}^{(r)}, \mathbf{z}_0^{(j)}, \mathbf{z}_1^{(j)})$ and $s_j^2 \leftarrow \sigma_{\hat{\omega}}^2(x_{\text{eval}}^{(r)}, \mathbf{z}_0^{(j)}, \mathbf{z}_1^{(j)})$.
  Predictive interval: let $\hat{F}(y) = J^{-1} \sum_{j=1}^{J} \Phi\big((y - m_j)/s_j\big)$ be the CDF of the mixture in Eq. (13); set $[\ell_\alpha, u_\alpha] \leftarrow [\hat{F}^{-1}(\alpha/2), \hat{F}^{-1}(1 - \alpha/2)]$ by bisection.
end for
Return $\{\hat{g}(x_{\text{eval}}^{(r)}, \mathbf{v}_{\text{eval}}^{(r)})\}_{r=1}^{R}$, $\text{MSE}_{\text{struct}}$, and the intervals.

E BGM-IV Implementation Details and Hyperparameters

All BGM-IV results use the settings below, fixed within each benchmark and shared across sample sizes, confounding levels, and repetitions. Table 5 lists the settings common to the three benchmarks and Table 6 the benchmark-specific ones.

Table 5: BGM-IV optimization and inference settings shared by all benchmarks.

Component	Setting
Optimizer	Adam, $\beta_1 = 0.9$, $\beta_2 = 0.99$; no weight decay
EGM initialization	50,000 iterations, learning rate 2×10^{-4}, 5 discriminator steps per generator step, gradient penalty 10, latent reconstruction on
EGM outcome term	8-node Gauss–Hermite quadrature; first-stage scale is a moving average (0.99) of the squared first-stage residual, capped at 1; variance-head penalty 0.001
Stochastic optimization training	200 epochs, learning rate 10^{-4} for θ, ϕ, ω and for the latent states
Tempering exponent	$\gamma = 0.01$
IV Monte Carlo	$M = 1{,}000$ treatment draws, in training and at evaluation
Multi-start selection	$K = 10$ candidates, selection by S_k in Eq. (23)
Test-time MAP	1,000 Adam steps, learning rate 10^{-4}
Networks	e_ψ and p_θ: five hidden layers of 64 units; p_ϕ and p_ω: hidden layers $[64, 32, 8]$ with mean and variance heads; discriminator: $[64, 32, 8]$ with batch normalization and tanh; LeakyReLU (slope 0.2) elsewhere
HMC	Pilot: 4 chains, 400 warm-up and 240 retained steps, 5 leapfrog steps in warm-up and 3, 5, or 7 in the retained steps, used only to set a diagonal mass matrix. Production: 4 overdispersed chains, dual-averaging step size (target acceptance 0.9); initial step sizes, production leapfrog steps, and chain lengths in Table 6

Table 6: Benchmark-specific BGM-IV settings.

	Demand	Vector proxy	MNIST
Latent partition $[\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3]$	$[3, 2, 1, 2]$	$[3, 2, 1, 2]$	$[2, 1, 1, 2]$
Batch size	64	64	32
Model covariate $\mathbf{v}$	(T, S)	$(T, \mathrm{PCA}_k(\mathbf{r}))$	(T, image)
$\dim(\mathbf{v})$	2	$1 + k$	785
Covariate likelihood	Gaussian	Gaussian	Gaussian / Bernoulli
Scale of T	learned	learned	fixed, 0.1
Covariate networks	MLP	MLP	CNN
Scaling of x, y	training mean/SD	training mean/SD	DFIV constants
HMC initial step size	0.05	0.05	0.02
HMC leapfrog steps, warm-up	5	5	31
HMC leapfrog steps, retained	3, 5, or 7	3, 5, or 7	7, 15, or 31
HMC warm-up / draws per chain	2,000 / 5,000	2,000 / 5,000	1,000 / 1,000

Architecture. The demand and vector-proxy benchmarks use the fully connected model of Table 5; in the vector-proxy benchmark the k PCA scores enter it directly as covariates. In the image benchmark, the image encoder has two 3×3 convolution layers with 64 filters, max pooling, dropout 0.1, dense layers of widths 128 and 64, and a final MLP with hidden units $[128, 64]$; the decoder maps the latent state to a $7 \times 7 \times 128$ tensor, applies two transposed-convolution upsampling blocks and a final convolutional block, and outputs pixel logits, with a Bernoulli likelihood on intensities in $[0, 1]$ and a separate Gaussian head for T. The treatment and outcome networks are the same in all benchmarks.

Preprocessing. In the demand benchmark, x, y, w, and the covariates are standardized with training-set means and standard deviations (standard deviations below 10^{-6} are replaced by 1). The vector-proxy benchmark standardizes x, y, w, and T in the same way; the PCA is fitted on the training proxies only, and the k scores are rescaled by one common factor so that their average variance equals that of the raw proxy, without per-component whitening, and then share one standardization scale. In the image benchmark we use the fixed scaling $x \mapsto (x - 17.779)/3.7$ and $y \mapsto (y + 292.1)/158.0$ of the DFIV code base, keep T on its raw scale, and divide pixel intensities by 255. Structural MSE is always reported on the original outcome scale.

F Baselines and Evaluation Details

KIV, DeepIV, and DFIV use the implementations in the DFIV code base (Xu et al., 2021)[2] (PyTorch (Paszke et al., 2019)), and DML-IV[3] and FIL[4] use the implementations released by their authors, each with the hyperparameters of the configuration its implementation provides for the demand design. Within a repetition all methods receive the same training sample: the baselines draw it from the demand-design generator of the DFIV code base with the repetition index as random seed, and the BGM-IV generator reproduces that random stream, so the training arrays are identical across methods.

For each repetition, BGM-IV, DeepIV, DFIV, DML-IV, and FIL are each trained $K = 10$ times on this sample. For DeepIV, DFIV, DML-IV, and FIL, the ten runs differ only in the seed of the model-side randomness: network initialization, mini-batch order, dropout, Monte Carlo draws, and the data partitions that are part of the estimator (the cross-fitting folds of DML-IV and the stage split of FIL; the stage split of DFIV is the fixed split of its implementation). The ten BGM-IV candidates differ only in their initialization, as described in Appendix D. After training, each run is scored by the method's own criterion defined below, computed on the training sample with all networks in evaluation mode. The run with the smallest criterion is retained; runs whose criterion is not finite are discarded first, and ties are broken by the smallest run index. In every experiment, all ten runs of every method in every cell had a finite criterion, so no run was discarded. The structural MSE of each run on the evaluation grid is recorded but never used in the selection. KIV has no random initialization: its two regularization parameters are chosen by its own stage-wise validation on the two halves of its training split over a grid of 50 values, and it is fit once per repetition, it was run on the demand design and on the image-covariate benchmark.

All selection criteria are training-sample versions of the IV moment restriction of Section 2. For a candidate structural function g, a first-stage law $\widehat{F}(x \mid w, \mathbf{v})$, and a set $\mathcal{I}$ of training units, write

$$\widehat{R}(g, \widehat{F}; \mathcal{I}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \Big\{ y^{(i)} - \int g(x, \mathbf{v}^{(i)})\, d\widehat{F}(x \mid w^{(i)}, \mathbf{v}^{(i)}) \Big\}^2, \tag{22}$$

the squared residual of the moment equation $\mathbb{E}[Y \mid W, V] = \int g(x, V)\, dP(x \mid W, V)$. The second stage of every baseline minimizes a version of Eq. (22) with its own estimate of the inner conditional expectation, and its selection criterion is that same quantity recomputed after training. BGM-IV is trained with the negative log working likelihood of Section 3.5, and its selection criterion is the IV-moment residual of Eq. (23), evaluated at the covariate-only latent states. The criteria differ in how the inner integral is evaluated, in the subsample $\mathcal{I}$, and in whether $y^{(i)}$ or a cross-fitted regression of it is the target. Each criterion is computed in its implementation's own outcome units ($\tilde{y}$ denotes a rescaled outcome) and is compared only across runs of the same method on the same training sample.

BGM-IV. For candidate k with parameters $(\theta_k, \phi_k, \omega_k)$, let $\hat{\mathbf{z}}^{(i)} = \hat{\mathbf{z}}_k(\mathbf{v}^{(i)})$ denote its covariate-only MAP latent state for unit i (Eq. (11)). The criterion is the training-set IV-moment residual

$$S_k = \frac{1}{n} \sum_{i=1}^{n} \Big\{ y^{(i)} - \int \mu_{\omega_k}\big(x, \hat{\mathbf{z}}_0^{(i)}, \hat{\mathbf{z}}_1^{(i)}\big)\, p_{\phi_k}\big(x \mid w^{(i)}, \hat{\mathbf{z}}_0^{(i)}, \hat{\mathbf{z}}_2^{(i)}\big)\, dx \Big\}^2, \tag{23}$$

that is, Eq. (22) with $g(x, \mathbf{v}) = \mu_{\omega_k}(x, \hat{\mathbf{z}}_0(\mathbf{v}), \hat{\mathbf{z}}_1(\mathbf{v}))$, $\widehat{F} = p_{\phi_k}(\cdot \mid w, \hat{\mathbf{z}}_0(\mathbf{v}), \hat{\mathbf{z}}_2(\mathbf{v}))$, and $\mathcal{I}$ the full training sample; the integral is computed by Monte Carlo with $M = 1{,}000$ draws from p_{ϕ_k}. The

[2] https://github.com/liyuan9988/DeepFeatureIV

[3] https://github.com/shaodaqian/DML-IV

[4] https://github.com/meta-inf/fil

selected model is $\hat{k} = \arg\min_k S_k$, and all point estimates and predictive intervals are computed from this single model. The criterion measures how well a candidate satisfies the IV moment restriction on the data it was fitted to; it uses no validation split, held-out sample, evaluation grid, or structural truth, and the latent states entering it are inferred from $\mathbf{v}$ alone.

DeepIV. The first stage is a mixture density network $\widehat{F}_{\mathrm{MDN}}(x \mid w, \mathbf{v})$ and the second stage a response network g. The criterion is the sampled second-stage loss that DeepIV minimizes,

$$\widehat{R}_{\mathrm{DeepIV}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{S}\sum_{s=1}^{S}\left\{\tilde{y}^{(i)} - g(a^{(i,s)}, \mathbf{v}^{(i)})\right\}^2, \qquad a^{(i,s)} \sim \widehat{F}_{\mathrm{MDN}}(\cdot \mid w^{(i)}, \mathbf{v}^{(i)}), \tag{24}$$

with $S = 100$ draws per unit (training uses two). Its expectation over the draws equals Eq. (22) plus the average first-stage variance $n^{-1}\sum_i \mathrm{Var}_{\widehat{F}_{\mathrm{MDN}}}\{g(X, \mathbf{v}^{(i)}) \mid w^{(i)}, \mathbf{v}^{(i)}\}$, so it is the IV-moment residual in the form DeepIV optimizes.

DFIV. DFIV learns treatment features $\varphi(x)$, instrument features $\psi(w, \mathbf{v})$, and covariate features $\chi(\mathbf{v})$; it fits the stage-1 ridge regression of $\varphi(x)$ on $\psi(w, \mathbf{v})$ on one half $\mathcal{I}_1$ of its training split and, on the other half $\mathcal{I}_2$, fits $g(x, \mathbf{v}) = \hat{\mathbf{w}}^\top\{\varphi(x) \otimes \chi(\mathbf{v})\}$ by ridge regression of y on the predicted features $\widehat{\Phi}^{(i)} = \widehat{\mathbb{E}}[\varphi(X) \mid w^{(i)}, \mathbf{v}^{(i)}] \otimes \chi(\mathbf{v}^{(i)})$. The criterion is its stage-2 objective at the solution,

$$\widehat{R}_{\mathrm{DFIV}} = \frac{1}{|\mathcal{I}_2|}\sum_{i \in \mathcal{I}_2}\left\{y^{(i)} - \hat{\mathbf{w}}^\top\widehat{\Phi}^{(i)}\right\}^2 + \lambda_2\|\hat{\mathbf{w}}\|^2, \tag{25}$$

where λ_2 is its stage-2 ridge parameter. The first term is Eq. (22) on $\mathcal{I}_2$, because $\hat{\mathbf{w}}^\top\widehat{\Phi}^{(i)}$ is the linear-in-features estimate of $\int g(x, \mathbf{v}^{(i)})\, d\widehat{F}(x \mid w^{(i)}, \mathbf{v}^{(i)})$.

DML-IV. DML-IV splits the training sample into ten folds $\mathcal{I}_1, \ldots, \mathcal{I}_{10}$ and, for each fold k, fits on the other nine an outcome regression $\hat{m}_{-k}(w, \mathbf{v})$ of Y on (W, V) and a mixture-of-Gaussians treatment density $\widehat{F}_{-k}(x \mid w, \mathbf{v})$; the response network g is then fit on all folds. The criterion is its cross-fitted second-stage loss,

$$\widehat{R}_{\text{DML-IV}} = \frac{1}{n}\sum_{k=1}^{10}\sum_{i \in \mathcal{I}_k}\left\{\hat{m}_{-k}(w^{(i)}, \mathbf{v}^{(i)}) - \frac{1}{S}\sum_{s=1}^{S} g(a^{(i,s)}, \mathbf{v}^{(i)})\right\}^2, \tag{26}$$

with $a^{(i,s)} \sim \widehat{F}_{-k}(\cdot \mid w^{(i)}, \mathbf{v}^{(i)})$ and $S = 100$ (training uses one draw per unit). It is Eq. (22) with $y^{(i)}$ replaced by the cross-fitted estimate of $\mathbb{E}[Y \mid W, V]$, the double-machine-learning form of the same moment equation.

FIL. FIL splits the training sample into two halves, learns on one half a random-feature embedding $\varphi(x)$ of the treatment and a stage-1 regression estimate $\widehat{\mathbb{E}}[\varphi(X) \mid w, \mathbf{v}]$, and on the other half fits $g(x, \mathbf{v}) = \hat{\mathbf{w}}(\mathbf{v})^\top\varphi(x)$, with $\hat{\mathbf{w}}(\cdot)$ a network of the covariates, by minimizing $\{\tilde{y} - \hat{\mathbf{w}}(\mathbf{v})^\top\widehat{\mathbb{E}}[\varphi(X) \mid w, \mathbf{v}]\}^2$ with early stopping on the half $\mathcal{I}_{\mathrm{val}}$ it holds out from that fit. The criterion is the validation loss of the state FIL returns,

$$\widehat{R}_{\mathrm{FIL}} = \min_{\text{epochs}}\ \frac{1}{|\mathcal{I}_{\mathrm{val}}|}\sum_{i \in \mathcal{I}_{\mathrm{val}}}\left\{\tilde{y}^{(i)} - \hat{\mathbf{w}}(\mathbf{v}^{(i)})^\top\widehat{\mathbb{E}}[\varphi(X) \mid w^{(i)}, \mathbf{v}^{(i)}]\right\}^2, \tag{27}$$

which is Eq. (22) on $\mathcal{I}_{\mathrm{val}}$; the minimum over epochs is FIL's own early-stopping rule, so the criterion is the score by which FIL itself selects its model. Table 7 summarizes the five criteria.

The predictive intervals of BGM-IV are the quantiles of the posterior predictive distribution in Eq. (13), computed for the selected model of each repetition at every point of the evaluation grid. Their coverage is the probability that a new outcome at an evaluation point, drawn around the true structural value with the unit-variance disturbance of the design, falls inside the interval, averaged over the grid, and their width is measured in outcome units. None of the baseline implementations returns interval estimates on these benchmarks. FIL admits quasi-Bayesian uncertainty quantification for its kernelized second stage, but its extension to exogenous covariates, the variant applicable to the demand designs used here, uses a neural second stage and returns point estimates only, and its authors note that uncertainty quantification would in principle require a Bayesian neural network second stage, which is not part of their algorithm (Wang et al., 2022, Appendix G). Interval estimates are therefore reported for BGM-IV alone.

Table 7: Training-side selection criteria used in the multi-start protocol. Each criterion is each baseline's own second-stage objective, recomputed on training data after training, and the criterion of BGM-IV is the IV-moment residual of Eq. (23); all are versions of the IV-moment residual in Eq. (22).

Method	Estimate of $\int g \, d\widehat{F}$	Target	Sample $\mathcal{I}$
BGM-IV	Monte Carlo, 1,000 draws from p_ϕ at the covariate-only MAP latents	y	training sample
DeepIV	Monte Carlo, 100 draws from the mixture density network (sampled loss)	$\tilde{y}$	training sample
DFIV	Linear in the stage-1 predicted features, plus the ridge penalty	y	second-stage half
DML-IV	Monte Carlo, 100 draws from the cross-fitted mixture density	$\hat{m}_{-k}(w, \mathbf{v})$	training sample, ten folds
FIL	Stage-1 regression of the treatment embedding	$\tilde{y}$	validation half

G Benchmark Details

G.1 Low-Dimensional Demand Design

The low-dimensional demand benchmark follows the synthetic airline demand design used in DeepIV, KIV, and DFIV (Hartford et al., 2017; Singh et al., 2019; Xu et al., 2021). Let S denote customer group, T time of year, C the cost-shifter instrument, and U an unobserved demand shock. We draw

$$S \sim \mathrm{Unif}\{1, \ldots, 7\}, \qquad T \sim \mathrm{Unif}[0, 10], \qquad C, U, \eta \sim \mathcal{N}(0, 1),$$

and set

$$\psi(T) = 2\left\{\frac{(T-5)^4}{600} + \exp[-4(T-5)^2] + \frac{T}{10} - 2\right\}.$$

For a confounding level ρ, the structural disturbance and price are

$$\epsilon = \rho U + \sqrt{1-\rho^2}\eta, \qquad P = 25 + (C+3)\psi(T) + U.$$

The structural demand function and observed outcome are

$$f_0(P, T, S) = 100 + (10 + P)S\psi(T) - 2P, \qquad Y = f_0(P, T, S) + \epsilon.$$

Thus P is endogenous through U, while C shifts price and is mean independent of the structural disturbance conditional on the observed covariates. The observed covariates are $V = (T, S)$. Evaluation uses the standard grid with 20 prices in $[10, 25]$, 20 time points in $[0, 10]$, and all seven customer groups, giving 2800 structural evaluation points.

G.2 High-Dimensional Vector-Proxy Demand Design

The vector-proxy benchmark keeps the treatment, instrument, outcome, structural function, and evaluation grid fixed. Instead of observing the customer group directly, we create type-specific prototypes $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_7 \in \mathbb{R}^{784}$ using a fixed feature seed, and a unit with customer group S observes the proxy vector

$$\mathbf{r} = \boldsymbol{\mu}_S + \sigma_{\mathrm{rep}}\boldsymbol{\eta}, \qquad \boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I}_{784}),$$

with $\sigma_{\mathrm{rep}} = 0.5$. The same prototypes are used for all methods, and the evaluation grid attaches proxies to the same 20-by-20-by-7 combinations of price, time, and customer group. In each repetition the proxy is reduced to its first k principal components, fitted on the training proxies only, with a deterministic sign convention and one common rescaling of the scores (Appendix E); the identical transformation is applied to the training data and to the evaluation grid and is given to every method as the covariate $(T, \mathrm{PCA}_k(\mathbf{r}))$. Each method uses its demand-design network and hyperparameters with the input layer widened to $1 + k$ covariates. We use $\rho = 0.5$, $k \in \{1, 2, 4, 6, 8, 16, 32, 64, 128\}$, and $n \in \{5{,}000, 10{,}000, 20{,}000, 50{,}000\}$, with 20 repetitions per configuration.

G.3 IMAGE-COVARIATE DEMAND DESIGN

The image-covariate benchmark keeps the same treatment, instrument, outcome, structural function, and evaluation grid as the low-dimensional demand design. The only change is the observation of customer group: instead of exposing S as a categorical covariate, we attach a randomly sampled MNIST digit image with label S (LeCun et al., 1998). The observed covariate is therefore $\mathbf{v} = [T, \mathrm{image}_{784}]$ with $v_{\mathrm{dim}} = 785$. Training images are sampled from the MNIST training split and evaluation images from the test split, while the structural grid remains the same 20-by-20-by-7 demand grid.

H COMPLETE EXPERIMENT RESULTS

H.1 DEMAND DESIGN

Table 8 reports the structural MSE summarized in Figure 2 for all six methods, three training sample sizes, and five confounding levels.

Table 8: Structural MSE on the low-dimensional demand-design benchmark, grouped by training sample size n; entries are divided by 10^2. Each cell reports the mean $\pm$ sample standard deviation over 20 repetitions, in which all methods use the same training data. Every method except KIV is selected among ten runs by its own training-side criterion (Appendix F). The lowest mean in each row is in bold.

ρ	DeepIV	KIV	DML-IV	FIL	DFIV	BGM-IV
$n = 1{,}000$						
0.1	51.52 ± 3.63	79.89 ± 20.02	36.23 ± 3.75	10.60 ± 5.18	10.22 ± 3.78	**4.15 ± 0.91**
0.25	50.92 ± 3.70	73.74 ± 22.01	37.09 ± 3.94	11.11 ± 4.99	9.42 ± 1.87	**3.83 ± 0.69**
0.5	51.04 ± 3.32	75.95 ± 18.77	35.88 ± 3.85	10.70 ± 4.82	10.05 ± 2.89	**3.74 ± 0.47**
0.75	51.20 ± 3.60	73.62 ± 23.44	36.61 ± 4.26	10.36 ± 4.95	10.47 ± 4.41	**4.36 ± 0.70**
0.9	51.14 ± 3.44	72.79 ± 21.36	36.06 ± 3.26	10.29 ± 4.90	10.25 ± 3.99	**3.87 ± 0.69**
$n = 5{,}000$						
0.1	15.29 ± 1.89	51.16 ± 10.63	13.88 ± 1.73	2.50 ± 1.19	1.93 ± 0.32	**1.72 ± 0.25**
0.25	15.53 ± 1.56	49.99 ± 13.53	13.92 ± 1.96	2.54 ± 1.10	1.95 ± 0.37	**1.72 ± 0.37**
0.5	15.65 ± 1.96	47.49 ± 8.97	13.90 ± 2.09	2.37 ± 0.81	1.92 ± 0.29	**1.81 ± 0.38**
0.75	15.59 ± 1.88	53.31 ± 11.02	14.62 ± 1.90	2.36 ± 0.81	1.84 ± 0.37	**1.73 ± 0.39**
0.9	15.72 ± 2.00	49.87 ± 9.17	14.73 ± 1.85	2.32 ± 0.80	1.99 ± 0.31	**1.70 ± 0.22**
$n = 10{,}000$						
0.1	11.42 ± 1.28	41.92 ± 8.58	11.27 ± 1.78	2.01 ± 0.49	1.50 ± 0.28	**1.35 ± 0.18**
0.25	11.57 ± 1.53	39.51 ± 8.27	11.83 ± 1.70	2.01 ± 0.50	1.56 ± 0.26	**1.38 ± 0.25**
0.5	11.65 ± 1.28	43.55 ± 11.29	11.20 ± 1.26	1.96 ± 0.46	1.51 ± 0.30	**1.38 ± 0.26**
0.75	11.73 ± 1.53	41.04 ± 12.01	11.69 ± 2.26	1.93 ± 0.47	1.43 ± 0.34	**1.38 ± 0.25**
0.9	11.80 ± 1.14	41.07 ± 7.23	11.62 ± 1.84	1.93 ± 0.46	1.44 ± 0.26	**1.39 ± 0.31**

H.2 VECTOR-PROXY BENCHMARK

Table 9 reports the structural MSE of all methods for all four training sample sizes, and Figure 6 shows the three sample sizes not displayed in Figure 3.

H.3 IMAGE-COVARIATE BENCHMARK

The structural MSE of all methods on the image-covariate benchmark is reported in Table 1 of the main text. Figure 4 is drawn for one repetition of BGM-IV at $n = 50{,}000$ and $\rho = 0.5$, customer group $S = 4$, and the grid price $P = 17.11$, with time T varying over the evaluation grid and each grid point carrying its own MNIST test image; the curve is the posterior mean of the structural function, the band is the 95% predictive interval, and demand is normalized by the mean and standard deviation of the training outcomes. The intervals are quantiles of the posterior predictive distribution

Table 9: Structural MSE on the vector-proxy benchmark ($\rho = 0.5$), grouped by training sample size n; rows give the number k of principal components of the 784-dimensional proxy given to every method, and entries are divided by 10^2. Each cell reports the mean $\pm$ sample standard deviation over 20 repetitions, in which all methods use the same training data; every method is selected among ten runs by its own training-side criterion (Appendix F). The lowest mean in each row is in bold.

k	DeepIV	DML-IV	FIL	DFIV	BGM-IV
$n = 5{,}000$					
1	45.07 ± 28.57	46.95 ± 27.06	42.11 ± 29.31	38.60 ± 29.44	**35.58 ± 27.87**
2	14.19 ± 7.75	15.82 ± 8.34	6.55 ± 7.85	6.61 ± 8.56	**4.30 ± 6.86**
4	12.46 ± 1.10	15.32 ± 1.85	5.88 ± 1.56	5.44 ± 0.63	**2.71 ± 0.35**
6	12.73 ± 1.40	16.32 ± 1.39	6.04 ± 2.04	5.81 ± 0.77	**2.45 ± 0.50**
8	12.89 ± 1.20	16.04 ± 1.55	8.04 ± 2.85	7.16 ± 0.47	**3.12 ± 0.53**
16	13.57 ± 1.32	17.29 ± 1.67	55.97 ± 23.99	13.24 ± 1.02	**4.36 ± 0.89**
32	15.71 ± 2.08	22.61 ± 1.74	72.65 ± 5.47	24.83 ± 1.80	**6.36 ± 0.71**
64	21.66 ± 1.14	54.84 ± 4.02	94.98 ± 5.22	44.76 ± 3.43	**11.75 ± 2.55**
128	**40.40 ± 2.83**	100.07 ± 2.37	114.04 ± 4.49	77.53 ± 3.02	46.19 ± 8.27
$n = 10{,}000$					
1	51.41 ± 30.37	51.12 ± 29.75	52.50 ± 31.33	46.52 ± 30.75	**46.05 ± 30.36**
2	9.41 ± 1.16	10.25 ± 1.62	3.62 ± 0.92	3.11 ± 0.70	**1.91 ± 0.36**
4	9.58 ± 1.15	12.02 ± 1.27	4.17 ± 0.89	3.74 ± 0.34	**2.14 ± 0.22**
6	9.84 ± 0.94	13.15 ± 1.37	4.13 ± 1.23	3.68 ± 0.52	**1.97 ± 0.20**
8	10.22 ± 1.04	13.12 ± 1.27	4.70 ± 1.34	4.44 ± 0.54	**2.40 ± 0.31**
16	10.80 ± 0.98	12.90 ± 1.13	19.67 ± 15.61	7.56 ± 0.45	**3.41 ± 0.43**
32	11.48 ± 1.10	14.85 ± 0.94	70.63 ± 4.52	16.36 ± 1.70	**4.73 ± 0.56**
64	14.41 ± 1.87	27.81 ± 2.17	81.39 ± 5.75	35.77 ± 3.25	**7.09 ± 0.83**
128	21.04 ± 1.91	70.67 ± 4.48	104.19 ± 3.31	75.26 ± 3.03	**12.88 ± 1.38**
$n = 20{,}000$					
1	48.20 ± 33.94	47.52 ± 33.97	48.91 ± 35.41	43.80 ± 34.08	**43.25 ± 33.99**
2	6.87 ± 1.00	7.17 ± 1.01	3.72 ± 1.18	2.34 ± 0.42	**1.59 ± 0.44**
4	7.60 ± 0.97	8.99 ± 1.12	4.38 ± 1.26	2.74 ± 0.57	**1.75 ± 0.25**
6	7.70 ± 0.89	9.23 ± 1.02	5.03 ± 1.49	2.50 ± 0.40	**1.61 ± 0.31**
8	8.42 ± 0.80	9.60 ± 1.05	4.14 ± 1.79	2.90 ± 0.58	**1.86 ± 0.24**
16	8.40 ± 0.87	10.03 ± 0.93	6.91 ± 1.88	4.32 ± 0.41	**2.92 ± 0.40**
32	9.01 ± 1.11	10.49 ± 1.05	26.14 ± 6.10	8.50 ± 0.56	**5.20 ± 2.56**
64	10.28 ± 1.25	16.24 ± 1.06	76.74 ± 7.00	24.84 ± 2.54	**5.49 ± 0.65**
128	12.39 ± 1.26	44.70 ± 2.35	94.47 ± 6.59	64.51 ± 6.69	**8.73 ± 1.13**
$n = 50{,}000$					
1	53.88 ± 25.16	52.06 ± 24.82	53.48 ± 24.89	49.74 ± 24.43	**49.27 ± 25.15**
2	5.04 ± 0.47	5.03 ± 0.89	2.77 ± 0.99	2.01 ± 0.41	**1.11 ± 0.17**
4	5.42 ± 0.62	5.60 ± 0.74	3.74 ± 1.14	2.26 ± 0.64	**1.51 ± 0.22**
6	6.36 ± 0.67	5.86 ± 0.65	3.80 ± 1.33	2.04 ± 0.50	**1.22 ± 0.18**
8	6.45 ± 0.66	6.13 ± 0.62	3.69 ± 1.20	2.29 ± 0.35	**1.52 ± 0.26**
16	6.63 ± 0.87	6.44 ± 0.71	4.85 ± 1.31	2.78 ± 0.51	**2.41 ± 0.58**
32	7.09 ± 0.85	7.37 ± 0.73	9.40 ± 4.75	3.83 ± 0.45	**3.16 ± 0.32**
64	7.34 ± 0.78	9.69 ± 0.96	21.75 ± 4.89	7.37 ± 0.53	**4.30 ± 1.13**
128	7.51 ± 0.72	19.36 ± 1.20	94.78 ± 6.58	28.62 ± 3.16	**5.33 ± 1.00**

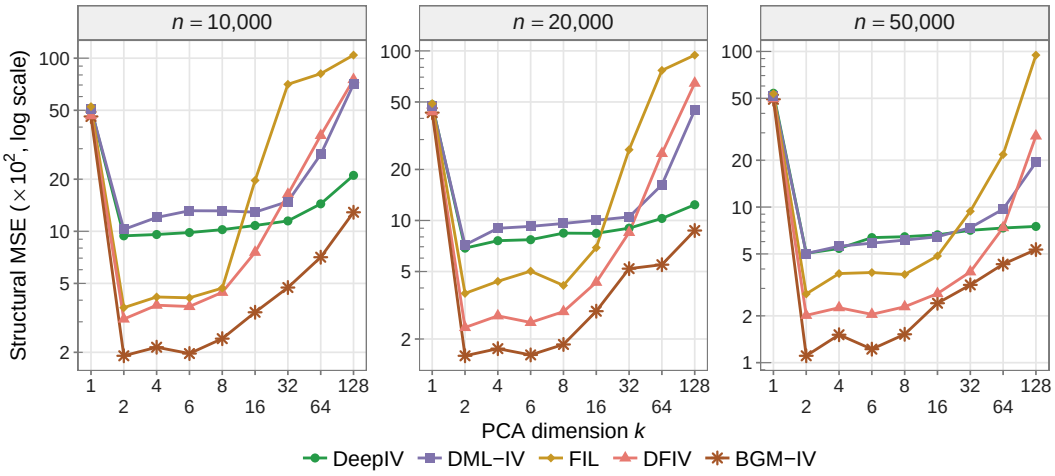


Figure 6: Structural MSE on the vector-proxy benchmark ($\rho = 0.5$) at $n = 10{,}000$, $20{,}000$, and $50{,}000$ as a function of the number k of principal components, placed at equal spacing; each point is the mean over 20 repetitions, on a logarithmic axis in units of 10^2. The corresponding panel for $n = 5{,}000$ is Figure 3(a); numerical values, including standard deviations, are in Table 9.

in Eq. (13) from 4 HMC chains with 1,000 warm-up and 1,000 retained draws each. Over the 20 repetitions of this setting, the 90%, 95%, and 99% predictive intervals have mean coverage 90.7%, 93.8%, and 96.8% and mean width 84.0, 99.5, and 129.7 outcome units, with standard deviations below 1 point for coverage and below 8 outcome units for width.

H.4 Weak Instruments on the Image-Covariate Benchmark

Table 10 varies the strength of the instrument on the image-covariate benchmark.

Table 10: Structural MSE on the image-covariate demand-design benchmark as the instrument strength λ varies, grouped by training sample size n; entries are divided by 10^2. The price equation is $P = 25 + (\lambda C + 3)\psi(T) + U$, so $\lambda = 1$ is the standard benchmark and smaller λ weakens the instrument; the confounding level is $\rho = 0.9$. Following Shao et al. (2024), the MSE is computed on 10,000 test points drawn from the same data-generating process, so that the test points follow the observed distribution of price and covariates rather than the fixed price grid. Each cell reports the mean $\pm$ sample standard deviation over 10 repetitions, in which all methods use the same training data; every method is selected among ten runs by its own training-side criterion (Appendix F). The lowest mean in each row is in bold.

λ	DeepIV	DML-IV	FIL	DFIV	BGM-IV
$n = 5{,}000$					
1	52.73 ± 3.22	53.39 ± 1.71	70.38 ± 6.91	69.30 ± 3.32	**19.85 ± 1.69**
0.8	45.09 ± 2.31	49.80 ± 2.10	63.51 ± 8.26	71.01 ± 2.76	**20.40 ± 1.40**
0.6	39.69 ± 2.21	44.77 ± 2.28	56.59 ± 2.80	68.95 ± 3.58	**20.91 ± 1.62**
0.4	33.99 ± 1.60	40.71 ± 2.21	49.37 ± 4.94	68.30 ± 4.38	**22.57 ± 2.44**
0.2	29.02 ± 1.34	37.72 ± 2.52	43.15 ± 4.14	67.54 ± 4.00	**26.74 ± 3.68**
$n = 10{,}000$					
1	41.27 ± 2.04	44.70 ± 2.07	60.52 ± 2.92	55.76 ± 3.78	**16.26 ± 1.50**
0.8	35.44 ± 1.87	42.08 ± 1.81	57.97 ± 2.89	54.48 ± 4.69	**15.14 ± 1.31**
0.6	29.71 ± 1.96	38.60 ± 1.31	49.66 ± 2.81	52.96 ± 4.76	**16.29 ± 1.32**
0.4	24.73 ± 1.20	34.37 ± 1.59	42.91 ± 4.13	53.06 ± 6.12	**18.33 ± 2.51**
0.2	21.34 ± 1.61	31.81 ± 1.43	35.85 ± 3.40	52.12 ± 4.49	**18.95 ± 1.90**

I PAIRED COMPARISONS AGAINST THE STRONGEST BASELINE

Table 11: Paired comparison of BGM-IV with the strongest baseline on the vector-proxy benchmark ($\rho = 0.5$), grouped by training sample size n; k is the number of principal components. In every configuration the strongest baseline is the one with the lowest mean structural MSE (Table 9). "Baseline" and "BGM-IV" are mean structural MSEs over 20 repetitions divided by 10^2; "Difference" is the mean paired difference, BGM-IV minus baseline, in the same units, so negative values favor BGM-IV; "Lower in" is the number of repetitions in which BGM-IV has the lower error. The last column is the two-sided exact Wilcoxon signed-rank p-value, Holm-adjusted over the 36 configurations of the table.

k	Strongest baseline	Baseline	BGM-IV	Difference	Lower in	p (Holm)
$n = 5{,}000$						
1	DFIV	38.60	35.58	-3.02	17/20	0.005
2	FIL	6.55	4.30	-2.25	19/20	< 0.001
4	DFIV	5.44	2.71	-2.73	20/20	< 0.001
6	DFIV	5.81	2.45	-3.36	20/20	< 0.001
8	DFIV	7.16	3.12	-4.04	20/20	< 0.001
16	DFIV	13.24	4.36	-8.88	20/20	< 0.001
32	DeepIV	15.71	6.36	-9.35	20/20	< 0.001
64	DeepIV	21.66	11.75	-9.91	20/20	< 0.001
128	DeepIV	40.40	46.19	$+5.79$	3/20	0.036
$n = 10{,}000$						
1	DFIV	46.52	46.05	-0.47	13/20	0.231
2	DFIV	3.11	1.91	-1.20	19/20	< 0.001
4	DFIV	3.74	2.14	-1.60	20/20	< 0.001
6	DFIV	3.68	1.97	-1.71	20/20	< 0.001
8	DFIV	4.44	2.40	-2.04	20/20	< 0.001
16	DFIV	7.56	3.41	-4.16	20/20	< 0.001
32	DeepIV	11.48	4.73	-6.74	20/20	< 0.001
64	DeepIV	14.41	7.09	-7.32	20/20	< 0.001
128	DeepIV	21.04	12.88	-8.16	20/20	< 0.001
$n = 20{,}000$						
1	DFIV	43.80	43.25	-0.55	15/20	0.160
2	DFIV	2.34	1.59	-0.75	18/20	0.001
4	DFIV	2.74	1.75	-0.98	20/20	< 0.001
6	DFIV	2.50	1.61	-0.89	18/20	< 0.001
8	DFIV	2.90	1.86	-1.04	19/20	< 0.001
16	DFIV	4.32	2.92	-1.40	20/20	< 0.001
32	DFIV	8.50	5.20	-3.30	17/20	0.001
64	DeepIV	10.28	5.49	-4.79	20/20	< 0.001
128	DeepIV	12.39	8.73	-3.66	19/20	< 0.001
$n = 50{,}000$						
1	DFIV	49.74	49.27	-0.47	16/20	0.061
2	DFIV	2.01	1.11	-0.91	20/20	< 0.001
4	DFIV	2.26	1.51	-0.74	19/20	< 0.001
6	DFIV	2.04	1.22	-0.82	20/20	< 0.001
8	DFIV	2.29	1.52	-0.77	20/20	< 0.001
16	DFIV	2.78	2.41	-0.37	15/20	0.160
32	DFIV	3.83	3.16	-0.67	20/20	< 0.001
64	DeepIV	7.34	4.30	-3.04	18/20	< 0.001
128	DeepIV	7.51	5.33	-2.18	19/20	< 0.001

For the two high-dimensional benchmarks, we compare BGM-IV in every configuration with the baseline that has the lowest mean structural MSE in that configuration. Because all methods are fit to the same training sample within a repetition, the 20 repetitions form pairs, and we test the

paired differences with the two-sided exact Wilcoxon signed-rank test, adjusting the p-values by Holm's procedure (Holm, 1979) within each benchmark. Tables 11 and 12 report the results. On the vector-proxy benchmark, BGM-IV is significantly more accurate than the strongest baseline at the 5% level in 31 of the 36 configurations. Four configurations show no significant difference: $k = 1$ for $n \geq 10{,}000$, where no method recovers the customer type, and ($n = 50{,}000$, $k = 16$). In the remaining configuration, ($n = 5{,}000$, $k = 128$), DeepIV is significantly more accurate. On the image-covariate benchmark, BGM-IV is significantly more accurate in all 15 configurations, and it has the lower error in every repetition for $n \geq 5{,}000$.

Table 12: Paired comparison of BGM-IV with the strongest baseline on the image-covariate benchmark, grouped by training sample size n. In every configuration the strongest baseline is the one with the lowest mean structural MSE (Table 1). "Baseline" and "BGM-IV" are mean structural MSEs over 20 repetitions divided by 10^2; "Difference" is the mean paired difference, BGM-IV minus baseline, in the same units, so negative values favor BGM-IV; "Lower in" is the number of repetitions in which BGM-IV has the lower error. The last column is the two-sided exact Wilcoxon signed-rank p-value, Holm-adjusted over the 15 configurations of the table.

ρ	Strongest baseline	Baseline	BGM-IV	Difference	Lower in	p (Holm)
$n = 1{,}000$						
0.1	DFIV	102.65	80.04	-22.62	17/20	< 0.001
0.25	DFIV	102.92	79.18	-23.74	18/20	< 0.001
0.5	DFIV	103.20	75.85	-27.36	19/20	< 0.001
0.75	DFIV	102.90	80.50	-22.40	20/20	< 0.001
0.9	DFIV	101.58	78.36	-23.22	20/20	< 0.001
$n = 5{,}000$						
0.1	DFIV	69.11	26.50	-42.61	20/20	< 0.001
0.25	DFIV	71.22	25.32	-45.90	20/20	< 0.001
0.5	DFIV	71.10	25.54	-45.56	20/20	< 0.001
0.75	DFIV	71.12	24.91	-46.21	20/20	< 0.001
0.9	DFIV	69.33	25.94	-43.39	20/20	< 0.001
$n = 10{,}000$						
0.1	DFIV	55.38	17.12	-38.25	20/20	< 0.001
0.25	DFIV	55.82	17.16	-38.66	20/20	< 0.001
0.5	DFIV	55.73	17.14	-38.58	20/20	< 0.001
0.75	DFIV	55.68	17.13	-38.54	20/20	< 0.001
0.9	DFIV	54.69	17.14	-37.54	20/20	< 0.001

J Computational Cost

All BGM-IV models were trained on CPU nodes of a shared computing cluster; no GPU was used for training on any benchmark. The HMC sampler of Section 3.6 ran on CPU for the vector-proxy benchmark, where one sampling job took between 5 and 9 minutes per repetition, and on a GPU for the image-covariate benchmark. The baselines ran on the devices supported by their implementations, which differ across methods and sample sizes. Table 13 and Figure 7 report the median wall-clock time of one repetition, that is, the ten training runs of the multi-start protocol together with the selection step and the evaluation, and, for BGM-IV, the separate wall-clock time of one HMC job. We report $n = 10{,}000$ as a representative setting. On the image-covariate benchmark the predictive intervals are computed at $n = 50{,}000$ (Figure 4), where training took 1,244 minutes on CPU and one HMC job 67 minutes on a GPU per repetition. On the vector-proxy benchmark the times at $k = 128$ are close to those at $k = 8$ (49 instead of 59 minutes for BGM-IV). Because the devices differ, these timings document the cost of the reported experiments and are not a controlled speed comparison.

Table 13: Median wall-clock time in minutes of one repetition (ten training runs of the multi-start protocol, selection, and evaluation) and, for BGM-IV, of one HMC job (4 chains; 2,000 warm-up and 5,000 retained draws per chain), with the device in parentheses. Medians are taken over the repetitions of all confounding levels at the given sample size. HMC is reported for the vector-proxy benchmark; on the image-covariate benchmark it was run at $n = 50{,}000$ (see the text), and it is not reported for the demand design. KIV has no multi-start protocol, so its entry is the time of a single fit; KIV was not run on the vector-proxy benchmark.

Method	Demand $n = 10{,}000$	Vector proxy ($k = 8$) $n = 10{,}000$	Image $n = 10{,}000$
KIV	40.7 (CPU)	–	52.1 (CPU)
DeepIV	17.1 (CPU)	17.5 (CPU)	379 (CPU)
DML-IV	284 (CPU)	329 (CPU)	377 (GPU)
FIL	7.7 (GPU)	9.2 (CPU)	34.1 (CPU)
DFIV	3.1 (GPU)	4.0 (CPU)	660 (CPU)
BGM-IV, training	42.4 (CPU)	58.5 (CPU)	457 (CPU)
BGM-IV, HMC	–	6.0 (CPU)	–

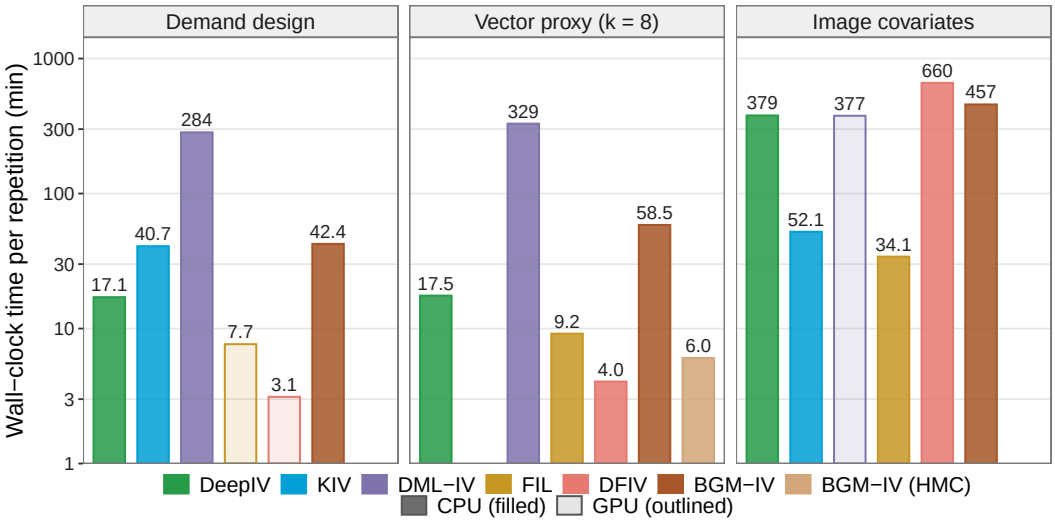


Figure 7: Median wall-clock time of one repetition at $n = 10{,}000$ on the three benchmarks, on a logarithmic axis in minutes, with the value printed above each bar. Filled bars were run on CPU and outlined bars on GPU. BGM-IV (HMC) is the separate sampling job for the predictive intervals, shown for the vector-proxy benchmark; the corresponding job of the image-covariate benchmark was run at $n = 50{,}000$ (Table 13).

K Ablation Details

This appendix describes the settings and variants of the ablation study of Section 4.5, reports the paired comparisons, and compares BGM-IV with the two strongest baselines in every setting. All runs use the low-dimensional demand design of Appendix G.1 at $n = 1{,}000$ and $\rho \in \{0.1, 0.5, 0.9\}$, with 20 repetitions per configuration. Each variant changes one component of BGM-IV and keeps everything else unchanged, namely the hyperparameters of Appendix E including the tempering exponent $\gamma = 0.01$, the $K = 10$ candidates and the selection rule of Eq. (23), the covariate-only MAP readout, and the training sample and initialization seeds of each repetition. Every variant is therefore paired with the full model by repetition. The structural MSE is evaluated on the 2800-point grid of Appendix G.1. Table 2 of the main text reports the mean and standard deviation over repetitions, and Table 14 the ratio of the mean structural MSE of each variant to that of the full model, together with the number of repetitions in which the full model has the lower error.

Settings. Each component is examined in a setting in which the difficulty it addresses is present. *Standard*: the design of Appendix G.1 without modification. *Amplified confounding*: the structural disturbance is amplified by a factor of 100, $\epsilon = 100\,(\rho U + \sqrt{1-\rho^2}\,\eta)$. The price equation, the instrument, the structural function f_0, and the evaluation grid are unchanged, and the correlation between ϵ and the confounder U remains ρ. *Nuisance covariates*: five variables $N_1, \ldots, N_5 \sim \mathcal{N}(0,1)$, independent of each other and of all other variables, are appended to the observed covariates, $V = (T, S, N_1, \ldots, N_5)$. They enter neither the price nor the outcome, so the structural function is unchanged. In each repetition they are drawn after all other variables from the same random stream, so that the remaining variables coincide with those of the standard design, and the evaluation grid receives a single fixed draw. The covariate dimension is $\dim(\mathbf{v}) = 7$. In the standard and amplified-confounding settings the latent partition is $[\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3] = [3, 2, 1, 2]$ as in Appendix E; the partition of the nuisance-covariate setting is given below.

Choice of the amplification. The amplification is chosen so that the confounding bias is clearly visible at $n = 1{,}000$. In the demand design, $P - 25 - 3\psi(T) = C\psi(T) + U$ with C and U independent standard normal, so $\mathbb{E}[U \mid P, T, S] = (P - 25 - 3\psi(T))/(1 + \psi(T)^2)$, and a regression of the outcome on the observed price and the covariates converges to $f_0 + \mathbb{E}[\epsilon \mid P, T, S]$ with $\mathbb{E}[\epsilon \mid P, T, S] = c\rho\,\mathbb{E}[U \mid P, T, S]$, where c is the amplification factor. The squared confounding bias of such a regression therefore scales with $c^2\rho^2$. Amplifying the disturbance by $c = 100$ makes this bias large relative to the estimation error at $n = 1{,}000$, so that the two likelihoods can be clearly distinguished, and the bias grows with ρ^2, in the order of the increases in Table 2.

Without EGM initialization (standard setting). The variant reduces the EGM initialization of Algorithm 1 to a single minibatch update, so the generators and the encoder enter the stochastic optimization training essentially at their random initialization, and the latent states are set from this untrained encoder. The stochastic optimization training, the selection, and the readout are unchanged. Removing the initialization increases the mean structural MSE by a factor of 17.7 to 18.9, and the full model has the lower error in all 60 paired repetitions.

Observed-treatment likelihood (amplified-confounding setting). The variant replaces the IV-integrated pseudo-likelihood $p_{\mathrm{IV}}(y \mid w, \mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2)$ by the outcome likelihood $p_\omega(y \mid x, \mathbf{z}_0, \mathbf{z}_1)$ evaluated at the observed treatment, in the update of the outcome generator in Eq. (8) and in the latent update in Eq. (10), with the same tempering exponent $\gamma = 0.01$. Correspondingly, the integrated outcome moment of the EGM initialization in Eq. (21) is replaced by the outcome mean at the observed treatment. The instrument still enters the treatment model, which is trained as in the full model, and the latent partition, the selection, and the readout are retained. The structural MSE increases by a factor of 1.18, 1.98, and 3.63 at $\rho = 0.1$, 0.5, and 0.9, and the full model has the lower error in 15, 20, and 20 of the 20 paired repetitions.

Unstructured latent space (nuisance-covariate setting). In this setting the block sizes of the full model are specified from the design of the covariates: one dimension for $\mathbf{z}_0$, since the time of year enters both price and demand, one for $\mathbf{z}_1$, since the customer group enters demand only, none for $\mathbf{z}_2$, since no covariate enters price only, and five for the nuisance block $\mathbf{z}_3$, one per nuisance covariate, giving $[\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3] = [1, 1, 0, 5]$. Only the block sizes are specified; the assignment of covariates to blocks is learned. The unstructured variant assigns all seven latent dimensions to a single block that enters every generator, so that the covariate generator takes $\mathbf{z}$, the treatment generator takes $(w, \mathbf{z})$, and the outcome generator takes $(x, \mathbf{z})$. The total latent dimension, the IV-integrated pseudo-likelihood, the EGM initialization, and all other components are retained. The unstructured latent vector increases the structural MSE by a factor of 1.40, 1.44, and 1.43 at $\rho = 0.1$, 0.5, and 0.9, and the full model has the lower error in 17, 18, and 17 of the 20 paired repetitions.

Location of the nuisance information. To locate the information about the nuisance covariates in the trained latent space, we regress each nuisance covariate on the latent states of the training observations at the end of training, for the selected candidate of each repetition, and report the five-fold cross-validated R^2 of a linear regression, averaged over the five covariates and the 60 repetitions. In the partitioned model, the nuisance block $\mathbf{z}_3$ attains $R^2 = 0.96$, whereas the coordinates read by the treatment and outcome generators, $\mathbf{z}_0$ and $(\mathbf{z}_0, \mathbf{z}_1)$, attain $R^2 \approx 0$ (-0.01). In the unstructured

model, every generator reads all coordinates, which attain $R^2 = 0.96$. The partition therefore keeps the nuisance variation out of the treatment and outcome generators.

Comparison with baselines. To put the error levels of the modified settings in context, we also run DFIV and FIL, the two strongest baselines in the standard design at $n = 1{,}000$ (Table 8), in all three settings with the protocol of Appendix F. In the nuisance-covariate setting, the five covariates are appended to the covariates of each baseline, and to its instrument vector where that vector contains the covariates, and the input layers are widened accordingly. Table 15 reports the results. Both modifications make the problem harder for every method; for example, relative to the standard design, the structural MSE of DFIV increases by a factor of 7.4 to 9.9 under amplified confounding and 7.9 to 8.3 with the nuisance covariates, and that of BGM-IV by a factor of 8.2 to 9.0 and 7.8 to 8.8. BGM-IV attains the lowest mean structural MSE in every setting and at every confounding level, and has the lower error than each baseline in at least 19 of the 20 paired repetitions of every configuration.

Table 14: Paired comparison of each ablation variant with the full model on the demand design at $n = 1{,}000$. Each cell reports the ratio of the mean structural MSE of the variant to that of the full model, and in parentheses the number of repetitions, out of 20, in which the full model has the lower error.

Setting	Variant	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
Standard	without EGM initialization	17.67 (20)	18.85 (20)	17.88 (20)
Amplified confounding	observed-treatment likelihood	1.18 (15)	1.98 (20)	3.63 (20)
Nuisance covariates	unstructured latent space	1.40 (17)	1.44 (18)	1.43 (17)

Table 15: Structural MSE (divided by 10^2, mean $\pm$ standard deviation over 20 repetitions) of BGM-IV and of DFIV and FIL, the two strongest baselines in the standard design, in the three ablation settings on the demand design at $n = 1{,}000$. BGM-IV is the full model of Table 2 in each setting, and the lowest mean in each setting and column is in bold.

Setting	Method	$\rho = 0.1$	$\rho = 0.5$	$\rho = 0.9$
Standard	BGM-IV	**4.15 $\pm$ 0.91**	**3.74 $\pm$ 0.47**	**3.87 $\pm$ 0.69**
	DFIV	10.22 $\pm$ 3.78	10.05 $\pm$ 2.89	10.25 $\pm$ 3.99
	FIL	10.60 $\pm$ 5.18	10.70 $\pm$ 4.82	10.29 $\pm$ 4.90
Amplified confounding	BGM-IV	**35.21 $\pm$ 8.03**	**33.82 $\pm$ 9.01**	**31.65 $\pm$ 6.23**
	DFIV	101.23 $\pm$ 44.32	94.64 $\pm$ 38.53	76.02 $\pm$ 17.60
	FIL	59.40 $\pm$ 23.75	58.50 $\pm$ 23.07	64.26 $\pm$ 33.37
Nuisance covariates	BGM-IV	**32.34 $\pm$ 9.98**	**32.86 $\pm$ 9.89**	**30.53 $\pm$ 8.69**
	DFIV	82.15 $\pm$ 11.26	78.98 $\pm$ 12.03	85.15 $\pm$ 12.82
	FIL	77.84 $\pm$ 14.84	78.05 $\pm$ 14.96	76.69 $\pm$ 16.41