Diverse, Global and Amortised Counterfactual Explanations for Uncertainty Estimates

Dan Ley, Umang Bhatt, and Adrian Weller

University of Cambridge

Abstract

To interpret uncertainty estimates from differentiable probabilistic models, recent work has proposed generating a single Counterfactual Latent Uncertainty Explanation (CLUE) for a given data point where the model is uncertain, identifying a single, on-manifold change to the input such that the model becomes more certain in its prediction. We broaden the exploration to examine δ-CLUE, the set of potential CLUEs within a δ ball of the original input in latent space. We study the diversity of such sets and find that many CLUEs are redundant; as such, we propose DIVerse CLUE (∇-CLUE), a set of CLUEs which each propose a distinct explanation as to how one can decrease the uncertainty associated with an input. We then further propose GLobal AMortised CLUE (GLAM-CLUE), a distinct and novel method which learns amortised mappings on specific groups of uncertain inputs, taking them and efficiently transforming them in a single function call into inputs for which a model will be certain. Our experiments show that δ-CLUE, ∇-CLUE, and GLAM-CLUE all address shortcomings of CLUE and provide beneficial explanations of uncertainty estimates to practitioners.

Introduction

Recent work [1] proposes CLUE (Counterfactual Latent Uncertainty Explanations), a method for finding an explanation of a model's predictive uncertainty of a given input by searching in the latent space of an auxiliary deep generative model (DGM), identifying a single change to the input such that the model becomes more certain in its prediction. However, there are limitations to CLUE, including the lack of a framework to deal with a potential diverse set of plausible explanations, despite proposing methods to generate them. CLUE introduces a latent variable DGM with decoder $\mu_\theta(\mathbf{x}|\mathbf{z})$ and encoder $\mu_\phi(\mathbf{z}|\mathbf{x})$. $\mathcal{H}$ refers to any differentiable uncertainty estimate of a prediction $\mathbf{y}$. CLUE minimises: $\mathcal{L}(\mathbf{z}) = \mathcal{H}(\mathbf{y}|\mu_\theta(\mathbf{x}|\mathbf{z})) + d(\mu_\theta(\mathbf{x}|\mathbf{z}), \mathbf{x}_0)$ to yield $\mathbf{x}_{\text{CLUE}} = \mu_\theta(\mathbf{x}|\mathbf{z}_{\text{CLUE}})$ where $\mathbf{z}_{\text{CLUE}} = \text{argmin}_{\mathbf{z}} \mathcal{L}(\mathbf{z})$. We propose δ-CLUE, ∇-CLUE and GLAM-CLUE, full details of which can be found in the paper.

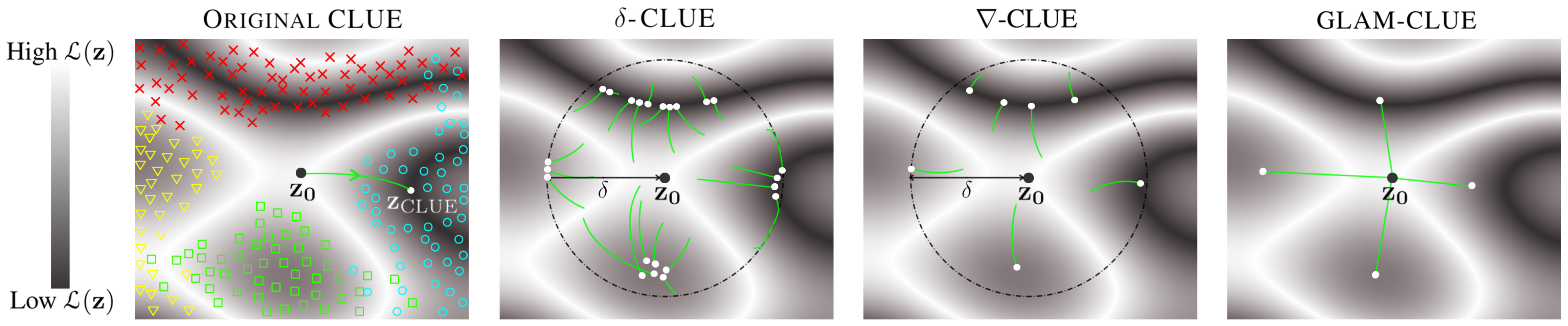

Figure 1: Conceptual colour map of objective function $\mathcal{L}(z)$ with $\mathbf{z}_0$ located in high cost region. White circles indicate explanations found. Left: Gradient descent to region of low cost [1]. Training points in colour. Left Centre: Gradient descent constrained to δ-ball. Diverse starting points yield diverse local minima, albeit with many redundant solutions. Right Centre: Direct optimisation for diversity (∇-CLUE). Right: Efficient mappings without gradient descent- each mapping applies to groups of inputs (GLAM-CLUE).

DIVerse CLUE

Our method δ-CLUE introduces a way of generating a set of CLUEs by restricting the search in latent space to a ball of radius δ and by randomly initialising within this ball. However, many CLUEs found therein are redundant. We introduce metrics D (detailed in Table 1) to measure the diversity in sets of CLUEs such that we can optimise for it directly: we term this DIVerse CLUE (∇-CLUE). By optimising simultaneously over k counterfactuals,we minimise $\mathcal{L}(\mathbf{z}_1, ..., \mathbf{z}_k) = -\lambda_D D(\mathbf{z}_1, ..., \mathbf{z}_k) + \frac{1}{k}\sum_{i=1}^{k} \mathcal{L}(\mathbf{z}_i)$ where $\mathcal{L}(\mathbf{z}_i) = \mathcal{H}(\mathbf{y}|\mu_\theta(\mathbf{x}|\mathbf{z}_i)) + d(\mu_\theta(\mathbf{x}|\mathbf{z}_i), \mathbf{x}_0)$, to yield $X_{\text{CLUE}} = \mu_\theta(X|Z_{\text{CLUE}})$ where $Z_{\text{CLUE}} = \underset{\mathbf{z}_1,...,\mathbf{z}_k}{\text{argmin}} = \mathcal{L}(\mathbf{z}_1, ..., \mathbf{z}_k)$. Note that we apply the diversity function in the latent space $\mathbf{z}$; it could equally be applied in input or prediction space.

Diversity Metric	Function (D)
Determinantal Point Processes	$\det(\mathbf{K})$ where $\mathbf{K}_{i,j} = \dfrac{1}{1 + d(\mathbf{x}_i, \mathbf{x}_j)}$
Average Pairwise Distance	$\frac{1}{\binom{k}{2}} \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} d(\mathbf{x}_i, \mathbf{x}_j)$
Coverage	$\frac{1}{d'} \sum_{i=1}^{d'} \left(\max_j (\mathbf{x}_j - \mathbf{x}_0)_i + \max_j (\mathbf{x}_0 - \mathbf{x}_j)_i \right)$
Prediction Coverage	$\dfrac{1}{c'} \sum_{i=1}^{c'} \max_j [(\mathbf{y}_j)_i]$
Distinct Labels	$\dfrac{1}{c'} \sum_{j=1}^{c'} \mathbf{1}_{[\exists i \,:\, y_i = j]}$
Entropy of Labels	$-\dfrac{1}{\log c'} \sum_{j=1}^{c'} p_j(k) \log p_j(k)$

Table 1: Diversity metrics D with arbitrary distance metric d.

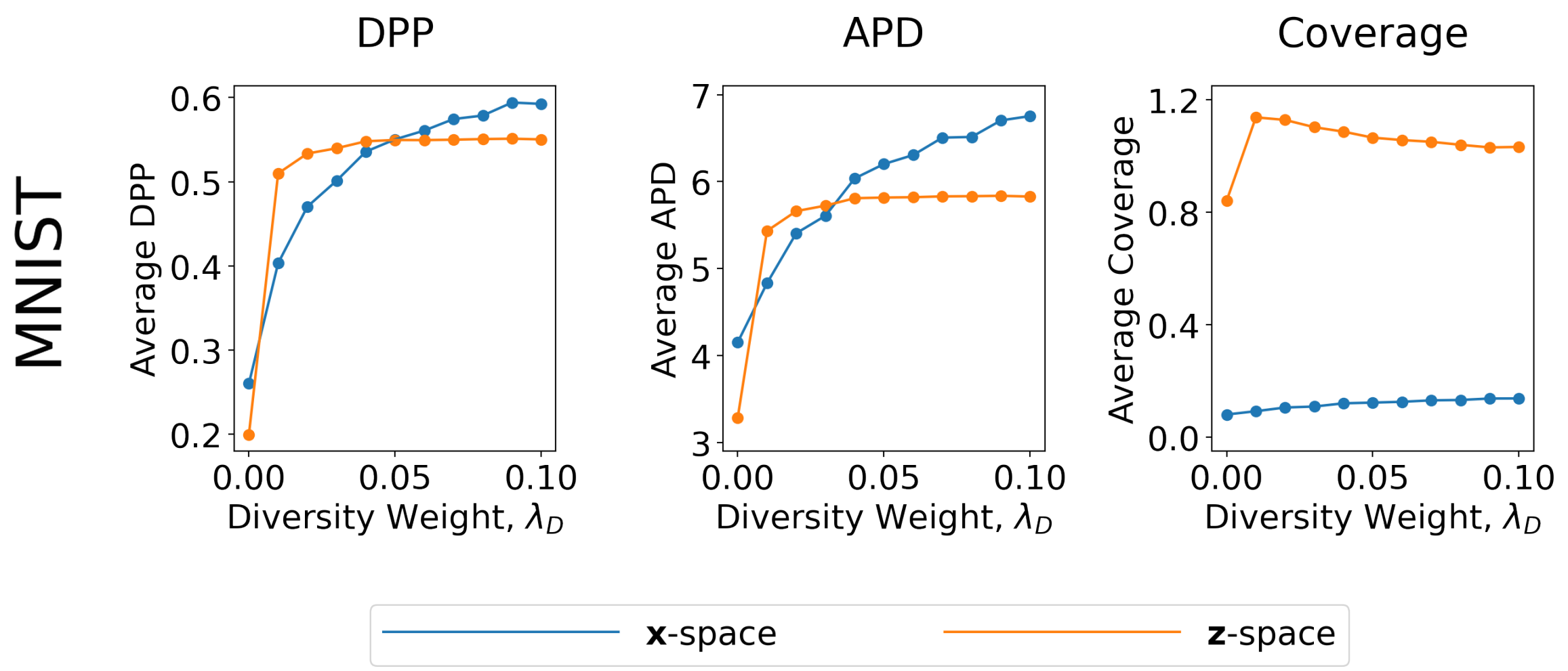

Figure 2: Effect of λ_D on diversity. DPP, APD and Coverage metrics evaluated on one set of $k = 10$ ∇-CLUEs (D = DPP).

GLobal AMortised CLUE

We desire a computationally efficient method that only requires a finite portion of the dataset from which global properties of uncertainty can be learnt, in the hope that we could apply these properties to unseen test data with a high degree of reliability. We therefore propose GLobal AMortised CLUE (GLAM-CLUE), which achieves such levels of reliability with considerable speedups. Summarising global properties of uncertainty can also be important in identifying areas in which the model does not perform as expected or the training data is potentially sparse.

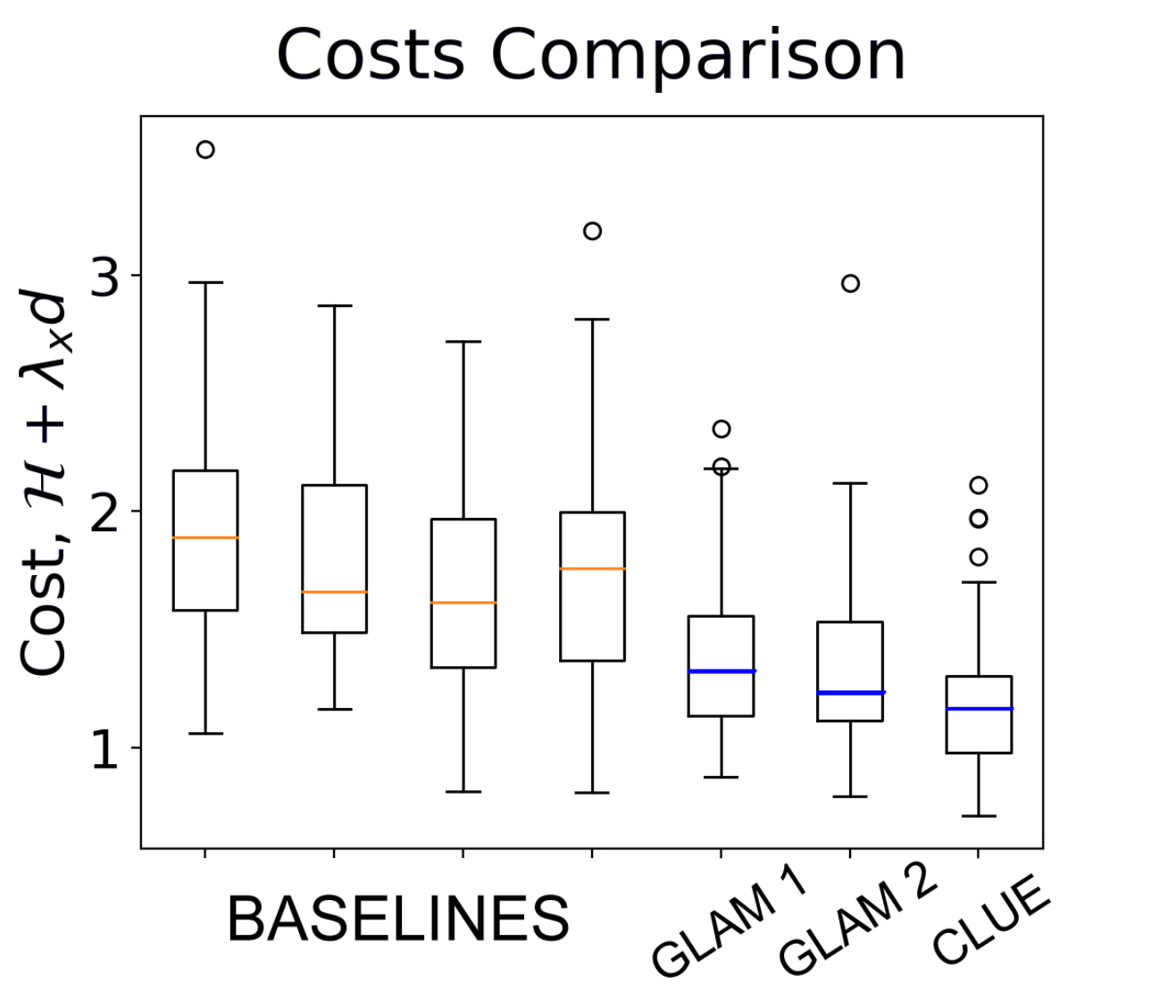

Figure 3: GLAM-CLUE vs baselines when mapping uncertain 7s to certain 7s in MNIST. Total costs $\mathcal{H} + \lambda_x d$.

High certainty points are taken from the training data to learn such mappings (GLAM 1), but we demonstrate improvements by instead using CLUEs generated from uncertain points in the training data (GLAM 2). At inference time, GLAM-CLUE performs significantly faster than CLUE by average CPU time (Table 2). For all uncertain 7s in MNIST, CLUE required 220 seconds to converge; GLAM-CLUE computed in around 1 second. While the baseline schemes achieve lower uncertainties, they do so at the expense of moving further from the original input (Figure 4), reducing the chance of yielding an actionable suggestion.

Input DBM	Latent DBM	Input NN
0.0306	0.0262	0.0236
Latent NN	GLAM-CLUE	CLUE
0.0245	0.0238	4.68

Table 2: Avg. CPU time in seconds to compute one MNIST counterfactual.

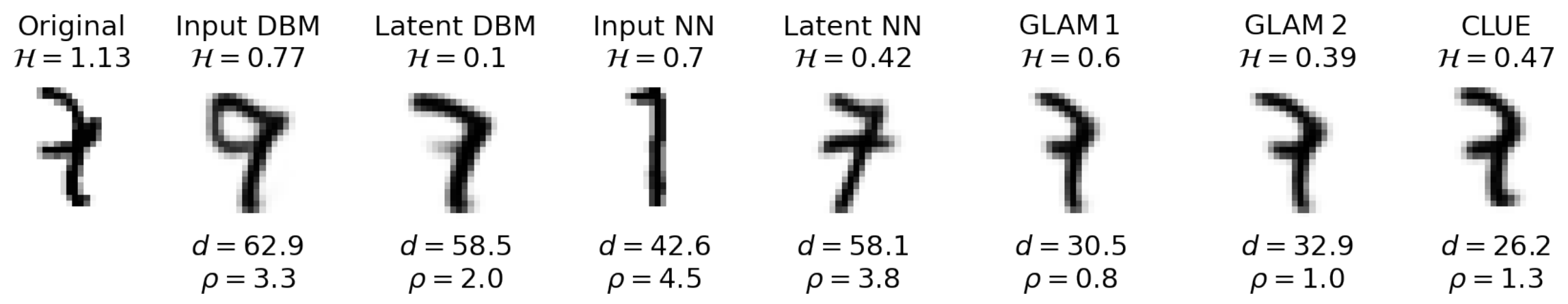

Figure 4: Comparison of explanations for an uncertain input (left) by the baselines, GLAM-CLUE, and CLUE. $\mathcal{H}$ is uncertainty, d is input space distance, ρ is latent space distance. Low $\mathcal{H}$ in baselines have unrealistically high d from the input.

Overall, we show that δ-CLUE, ∇-CLUE provide richer summaries for local explanations, whilst GLAM-CLUE address the seldom tackled problem of global counterfactuals.

References

[1] Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, and José Miguel Hernández-Lobato.
Getting a CLUE: A method for explaining uncertainty estimates.
In *International Conference on Learning Representations*, 2021.

Acknowledgements

UB acknowledges support from DeepMind and the Leverhulme Trust via the Leverhulme Centre for the Future of Intelligence (CFI) and from the Mozilla Foundation. AW acknowledges support from a Turing AI Fellowship under grant EP/V025379/1, The Alan Turing Institute, and the Leverhulme Trust via CFI. The authors thank Javier Antorán for his help.