

Estimating Rare Collective Absence Events Under Social Dependence and Latent External Forcing

Tamas Bodai

bodait@yahoo.com

Hungarian University of Agriculture and Life Sciences <https://orcid.org/0000-0002-3049-107X>

Research Article

Keywords: Ising model, social dependence, latent variable model, Monte Carlo, rugged landscape

Posted Date: May 25th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9787134/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Estimating Rare Collective Absence Events Under Social Dependence and Latent External Forcing

Tamás Bódai

May 18, 2026

Department of Applied Statistics, Institute of Mathematics and Basic Sciences,
Hungarian University of Agriculture and Life Sciences, Budapest, Hungary
Correspondence to: bodai.tamas@uni-mate.hu, bodait@yahoo.com

Abstract

We show that estimating the probability of collective absence is highly nontrivial once both social interactions and latent external forcing are present. While observing a proxy variable related to the latent forcing substantially improves inference, practical identifiability remains weak in finite samples. Numerical experiments reveal a rugged likelihood landscape with multiple competing minima and approximate compensation directions between latent forcing and social coupling strength. In particular, different parameter combinations may yield very similar likelihood values while implying substantially different estimates of the latent forcing probability. Extensive optimisation experiments demonstrate that the resulting ambiguity is strongly sample-dependent and persists even for moderately large sample sizes. Nevertheless, sufficiently informative proxy observations substantially improve estimation accuracy. Our results highlight the distinction between structural identifiability in principle and practical identifiability in finite samples for latent-variable network models.

1 Introduction

Suppose a lecturer wishes to estimate the probability

$$P_{aa} := \mathbb{P}(\text{all students absent}),$$

where “aa” stands for “all absent”. The problem appears simple at first sight, but in practice it is highly nontrivial. If student attendance decisions were independent Bernoulli random variables, one could estimate individual attendance probabilities and multiply them. However, such an assumption is plainly unrealistic. Students interact socially, influence each other, coordinate behaviour, and respond collectively to external circumstances such as weather, examinations, concerts, transportation disruptions, or holidays.

Empirically, lecturers occasionally observe entire classes failing to appear. Such events occur much more frequently than predicted under independence. Thus, independence is not merely an approximation; it is qualitatively contradicted by observation. At the same time, direct empirical estimation of P_{aa} is unreliable. Even over many semesters, one may observe only a handful of complete absence events. If n_{aa} denotes the number of all-absence occurrences in T observations, then the empirical estimator

$$\hat{P}_{aa} = \frac{n_{aa}}{T}$$

has standard error

$$\text{SE}(\hat{P}_{aa}) = \sqrt{\frac{P_{aa}(1 - P_{aa})}{T}}.$$

When P_{aa} is small and n_{aa} is only 2–3, relative uncertainty becomes enormous. Taking a “master” sample of $T = 50$ generated by a realistic model of Sec. 4 with $N = 15$ students making up the class, we have $n_{aa} = 3$, and, therefore, $\hat{P}_{aa} = 0.06$ and $\text{SE}(\hat{P}_{aa}) = 0.034$. That is, the half-width of the confidence interval of 95% confidence would be just about the size of the estimate itself, implying a small but very uncertain probability, indeed. This motivates the search for structured parametric models capable of leveraging dependence information throughout the full attendance configuration space. The present work investigates both the possibilities and limitations of such an approach.

2 Independent Model

Let the binary or Bernoulli “indicator” variable

$$X_{i,t} \in \{0, 1\}$$

denote the attendance state of student i at lecture t , where

$$X_i = 1$$

means absent. Under independence,

$$\mathbb{P}(X_1, \dots, X_N) = \prod_{i=1}^N p_i^{X_i} (1 - p_i)^{1-X_i},$$

where

$$p_i = \mathbb{P}(X_i = 1).$$

The all-absence probability becomes

$$P_{aa} = \prod_{i=1}^N p_i.$$

This estimate is computationally trivial but grossly underestimates observed collective absences whenever social dependence exists. To appreciate the scale of the effect, consider the realistic model of Sec. 4, but temporarily disregard inter-student variation in baseline attendance tendency by taking a uniform value $\alpha = -1.2$. In isolation, this would imply marginal absence probabilities of approximately

$$p_i = \mathbb{P}(X_i = 1) \approx 0.3.$$

Under independence, the probability of all $N = 15$ students being absent would then become the minuscule value

$$P_{aa} \approx 0.3^{15} \approx 1.5 \times 10^{-8}.$$

Yet, in the synthetic data generated from the full interacting model, the marginal absence probabilities are closer to 0.65 on average. Clearly, it matters enormously which number is raised to the power $N = 15$: the difference amounts to a factor of

$$\left(\frac{0.65}{0.3}\right)^{15} \approx 10^5.$$

Even so, the independent estimate,

$$P_{aa} \approx 0.0017,$$

remains far below the true probability

$$P_{aa} \approx 0.08,$$

known from long simulations of the full model (Sec. 4). Thus, the incorrect independence assumption would not only substantially underestimate the probability of collective absence, but also cast students in considerably worse light than warranted: without peer pressure and social reinforcement or some kind of temptation, they would in fact be much more likely to do the right thing. Consequently, a statistics teacher experiencing no student turning up only three years into his teaching career would be justified in suspecting “foul play”, namely, that students did not decide independently to skip class.

3 Social Dependence Models

Chain-Based Models

A preliminary attempt to incorporate dependence is to assume a directed influence structure:

$$X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_N.$$

This leads to the following approximation of the joint probability:

$$\mathbb{P}(\mathbf{X}) = \mathbb{P}(X_1, \dots, X_N) \approx \mathbb{P}(X_1) \prod_{i=2}^N \mathbb{P}(X_i \mid X_{i-1}).$$

Such models can generate larger estimates of P_{aa} than independence. However, they rely on highly restrictive assumptions:

- a fixed ordering of students,
- asymmetric influence,
- a unique propagation pathway,
- absence of collective latent shocks.

From attendance data, we can estimate the pair-wise conditional probabilities $S_{i,j} = \mathbb{P}(X_j \mid X_i)$. We can refer to $S_{i,j}$ as the matrix encoding (low-order, pair-wise) social dependence. The estimate is then

$$\hat{S}_{i,j} = \frac{\langle x_{i,t} x_{j,t} \rangle_t}{\langle x_{i,t} \rangle_t},$$

where the angle brackets denote sample mean, e.g.

$$\hat{p}_i = \langle x_{i,t} \rangle_t = \frac{1}{T} \sum_{t=1}^T x_{i,t}.$$

Form our master sample, we can estimate P_{aa} based on the above chain influence structure, or, equally “well”, based on a “hub dependence structure”, as:

$$\hat{P}_{aa} \approx \hat{p}_1 \prod_{j=2}^N \hat{S}_{1,j} = 0.015.$$

Even averaging over all possible “ringleaders” – prompted by the estimate not being invariant under the change of the hub – leaves the structural ordering/arrangement assumption fundamentally artificial. This estimate:

$$\hat{P}_{aa} \approx \left\langle \hat{p}_i \prod_{j \neq i} \hat{S}_{i,j} \right\rangle_i = 0.009,$$

is still an order of magnitude smaller than the true probability. However, it is not clear whether the problem is more with the naivety of the approach of taking social dependence into account or that it is disregarding exogenous effects ($\beta > 0$).

4 Latent-Variable Social Interaction Model

The most plausible or realistic model would combine:

1. endogenous social interaction [1], and
2. exogenous latent forcing.

Model Definition

Introduce a latent binary or Bernoulli variable

$$Z_t \in \{0, 1\},$$

$Z \sim \text{Bernoulli}(\pi)$, where:

- $Z_t = 0$: ordinary day,
- $Z_t = 1$: external shock day.

Conditional on Z_t , attendance configurations follow an Ising-type distribution [2, 3]:

$$\mathbb{P}(\mathbf{X}_t \mid Z_t) \propto \exp \left(\sum_i h_i(Z_t) X_{i,t} + \sum_{i < j} \theta_{ij} X_{i,t} X_{j,t} \right),$$

with

$$h_i(Z_t) = \alpha_i + \beta_i Z_t.$$

The parameters have the following interpretation:

- α_i : baseline absence tendency,
- β_i : sensitivity to external shocks,
- θ_{ij} : social interaction strength,

Marginally,

$$\mathbb{P}(\mathbf{X}) = (1 - \pi) \mathcal{P}_0(\mathbf{X}) + \pi \mathcal{P}_1(\mathbf{X}),$$

where

$$\mathcal{P}_0(\mathbf{X}) = \mathbb{P}(\mathbf{X} \mid Z = 0) = \frac{1}{\mathcal{Z}_0} \exp \left(\sum_i \alpha_i X_i + \sum_{i < j} \theta_{ij} X_i X_j \right),$$

with normalisation constant called “partition function”

$$\mathcal{Z}_0 = \sum_{\mathbf{x} \in \{0,1\}^N} \exp \left(\sum_i \alpha_i x_i + \sum_{i < j} \theta_{ij} x_i x_j \right), \text{ and}$$

$$\mathcal{P}_1(\mathbf{X}) = \mathbb{P}(\mathbf{X} \mid Z = 1) = \frac{1}{\mathcal{Z}_1} \exp \left(\sum_i (\alpha_i + \beta_i) X_i + \sum_{i < j} \theta_{ij} X_i X_j \right),$$

with partition function

$$\mathcal{Z}_1 = \sum_{\mathbf{x} \in \{0,1\}^N} \exp \left(\sum_i (\alpha_i + \beta_i) x_i + \sum_{i < j} \theta_{ij} x_i x_j \right),$$

are two interacting Ising systems differing only through the shifted fields. For this article, we adopt the following setup: $\pi = 0.2$ and $\alpha_i, \beta_i, \theta_i$ are samples from normal random variables $\mathcal{N}(-1.2, 0.15)$, $\mathcal{N}(1.2, 0.2)$, $\mathcal{N}(\mu_{w,b}, \sigma_{w,b})$, in that order, where the within-group mean dependence strength is $\mu_w = 0.5$ with a variation of $\sigma_w = 0.1$, and, for the between-group dependences, $\mu_b = 0.08$, $\sigma_b = 0.04$. That is, we divide the cohort of 15 students into three groups of 5 “friends” each group. We treat this model as the “synthetic truth”, something that generates data. Note that the “master sample” is generated by using a random seed of `rng(2)` in Matlab. Given that we know the exact model underlying the data, we will be able to assess the effectiveness of the method of estimation and the quality of the estimates. This is the central objective of our work.

5 Likelihood-Based Estimation

The approximation-based estimation methods considered earlier proved unreliable for the combined latent-variable interaction model. We therefore turned to full likelihood estimation [4, 5]. Given observations

$$\{(\mathbf{x}_t, w_t)\}_{t=1}^T,$$

such that the information carried by the proxy W about Z depends on the separation between ρ_0 and ρ_1 ,

$$\rho_0 = \mathbb{P}(W = 1 \mid Z = 0), \quad \rho_1 = \mathbb{P}(W = 1 \mid Z = 1),$$

the likelihood function is

$$\mathcal{L}(\Theta) = \prod_{t=1}^T [(1 - \pi)\mathcal{P}_0(\mathbf{x}_t) f_0(w_t) + \pi\mathcal{P}_1(\mathbf{x}_t) f_1(w_t)], \quad (1)$$

where

$$\Theta = (\alpha, \beta, \theta, \pi)$$

collects all model parameters to be inferred and

$$f_0(w) = \mathbb{P}(W = w \mid Z = 0) = \rho_0^w (1 - \rho_0)^{1-w},$$

$$f_1(w) = \mathbb{P}(W = w \mid Z = 1) = \rho_1^w (1 - \rho_1)^{1-w}.$$

In the present study, the proxy characteristics, ρ_0, ρ_1 , were treated as known quantities in order to control the informativeness of the proxy variable W . Interpretationally, W_t may represent weather, transport conditions, local events, or other partially informative signals. If we do not observe Z or its proxy W , then the likelihood function simplifies to

$$\mathcal{L}(\Theta) = \prod_{t=1}^T [(1 - \pi)\mathcal{P}_0(\mathbf{x}_t) + \pi\mathcal{P}_1(\mathbf{x}_t)].$$

Since the system size considered here is moderate ($N = 15$), the partition functions $\mathcal{Z}_0, \mathcal{Z}_1$ were evaluated exactly by explicit summation over all 2^N configurations [2].

Parameter estimation was performed by numerical minimisation of the negative log-likelihood (NLL),

$$-\log \mathcal{L}(\Theta).$$

Two derivative-free optimisation approaches were investigated. First, many of the parameter-sweep experiments, including those shown in Fig. 2, employed MATLAB’s `patternsearch` algorithm, a direct-search optimisation method from the Global Optimization Toolbox. Second, to investigate optimisation dynamics directly, we implemented a custom coordinate-search algorithm. In this scheme, parameters were updated sequentially by testing positive and negative

perturbations along individual coordinate directions and accepting only likelihood-improving moves. Adaptive step sizes were used to accelerate or damp the search depending on local improvement. The custom coordinate-search experiments were run for thousands of iterations while recording the evolving values of

$$\hat{\pi}, \quad \hat{P}_{aa}, \quad \text{and} \quad -\log \mathcal{L},$$

yielding the optimisation trajectories shown in Fig. 3.

6 Emergence of Non-Identifiability

6.1 Empirical Observation

The most striking finding was that, even when based on **attendance data** $x_{i,t}$ **alone**, likelihood optimisation consistently produced highly variable estimates of π , even while accurately reproducing the observable probability P_{aa} ; see Fig. 1. In particular, our Monte Carlo experiments [6] suggest that:

- estimates of π became quantised,
- multiple distinct parameter combinations yielded nearly identical likelihood,
- optimisation trajectories drifted slowly along extended flat regions,
- increasing sample size reduced estimation error in P_{aa} more strongly than in π .

6.2 Likelihood Ridge

This behaviour indicates weak practical identifiability [7]. The reason is intuitive. The observable statistics can be reproduced either by:

1. strong social coupling with rare shocks, or by
2. weak social coupling with frequent shocks.

These mechanisms compensate each other. Formally, the marginal model depends on the mixture

$$(1 - \pi)\mathcal{P}_0 + \pi\mathcal{P}_1.$$

Suppose we increase π while simultaneously decreasing the contrast between $\mathcal{P}_0$ and $\mathcal{P}_1$. Then the observable mixture distribution changes only weakly. This creates an approximate invariance direction in parameter space:

$$(\pi, \beta, \theta) \mapsto (\pi + \delta\pi, \beta + \delta\beta, \theta + \delta\theta),$$

along which the likelihood varies minimally. Hence the negative log-likelihood develops elongated valleys in parameter space. These valleys correspond to compensation directions along which increases in latent forcing can be partially offset by decreases in social interaction strength, while leaving many observable statistics nearly unchanged. In finite samples, such valleys become rugged and fragmented by sampling fluctuations, producing multiple competing local minima with similar likelihood values.

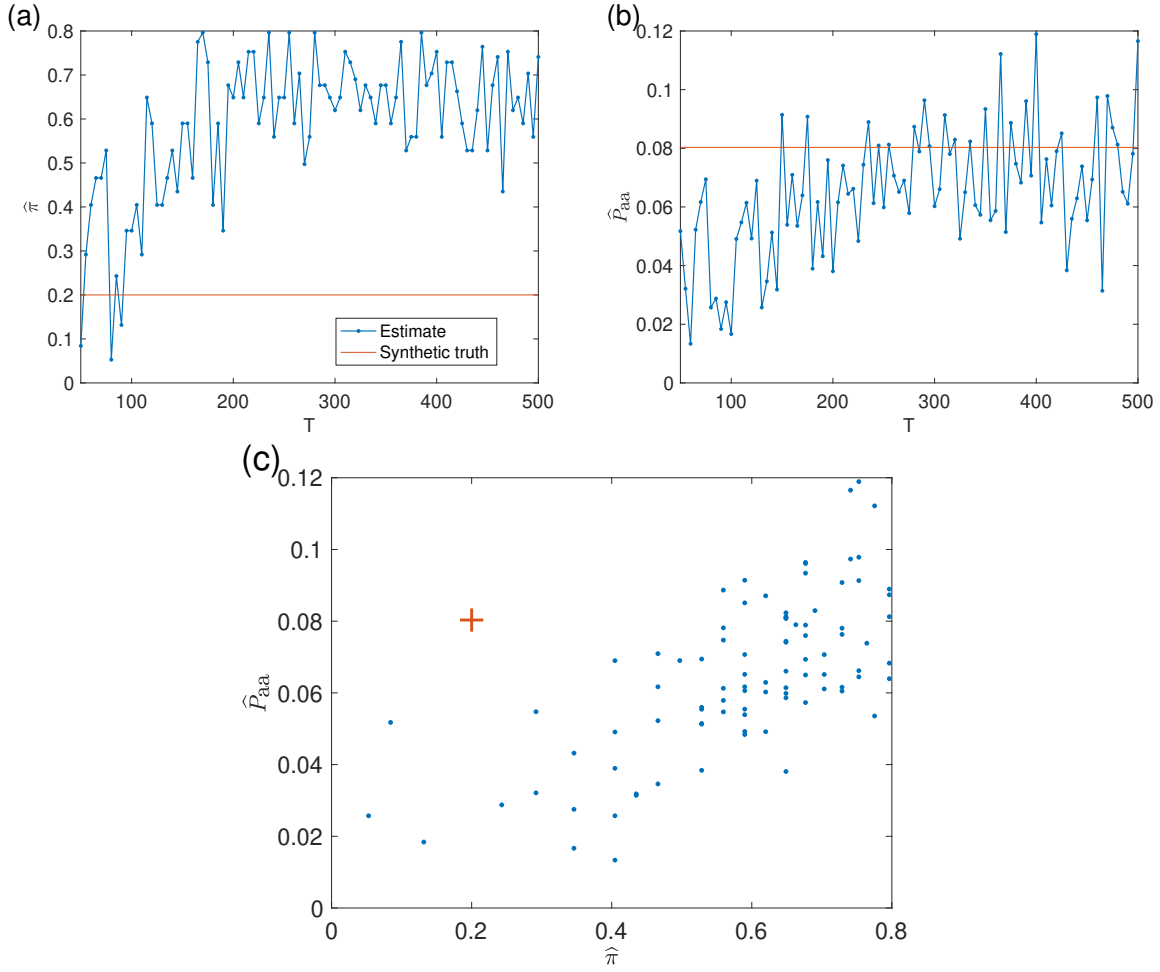


Figure 1: Model inference based on attendance data alone: dependence on sample size T . Although increasing sample size improves inference only slowly, substantial finite-sample ambiguity persists even for relatively large T . The number of iterates by `patternsearch` are limited to a maximum count of 100, which was always reached, the optimisation therefore terminating “prematurely”.

6.3 Regularisation Attempts

We investigated two regularisation strategies [4]. We employed a “Quadratic Prior”

$$\lambda_{\pi}(\pi - \pi_0)^2$$

and, in turn, a “Beta Prior”:

$$-(a-1)\log \pi - (b-1)\log(1-\pi).$$

These regularisation terms were added to the negative log-likelihood. However, these merely forced the estimates toward the prior means (results not shown). I.e., neither approach restored identifiability. Actually, this could be expected when a likelihood contains an extended flat ridge (Sec. 6.2): weak regularisation has negligible effect, whereas strong regularisation effectively determines the estimate by fiat.

7 Evidence for practical non-identifiability

7.1 Empirical Observation

In the face of non-identifiability based on attendance data alone, we will next investigate whether the observation of a proxy W of the latent variable Z would restore identifiability. Figure 2 shows the dependence of estimates on sample size and the proxy informativeness parameter

$$\Delta_W = \rho_1 - \rho_0.$$

Clearly, an increasingly better proxy improves identifiability in terms of a smaller variance and bias of $\hat{\pi}$. However, estimates of the coveted absence probability P_{aa} remain highly variable even with a strongly informative proxy, i.e., high Δ_W . Given that optimisation always stopped prematurely, one might suspect that poor estimates of P_{aa} arise just because of that.

However, Figure 3 (a) strongly suggests that it is not a numerical issue but structural. In this diagram, we see a number of instances of different optimisation runs for nearby Δ_W to settle into distinct parameter configurations despite prolonged optimisation over thousands of iterations [7]. Importantly, for the realised finite samples, the negative log-likelihood at these competing solutions was frequently lower than the value corresponding to the true generating parameters (not shown). This suggests that the ambiguity cannot be attributed solely to premature optimisation termination. Instead, finite-sample fluctuations generate spurious minima capable of overfitting the observed data better than the true generating parameters for the finite realised sample.

The resulting ambiguity was found to depend strongly on the realised attendance sample. Different synthetic datasets generated from identical underlying parameters often produced substantially different optimisation outcomes. Increasing the sample size T reduced this variability only gradually. Furthermore, the sample dependence is found to diminish considerably for estimates of π but not P_{aa} , as seen in panels (b) and (c) of Fig. 3, respectively.

7.2 Likelihood geometry and compensation directions

To better understand the source of the observed ambiguity, we examined slices of the likelihood surface. First, we evaluated the negative log-likelihood as a function of

$$(\pi, \Delta_W),$$

while keeping all remaining parameters fixed at their true values. Averaging over many realisations of the proxy variable W yielded a smooth likelihood surface whose minimum remained

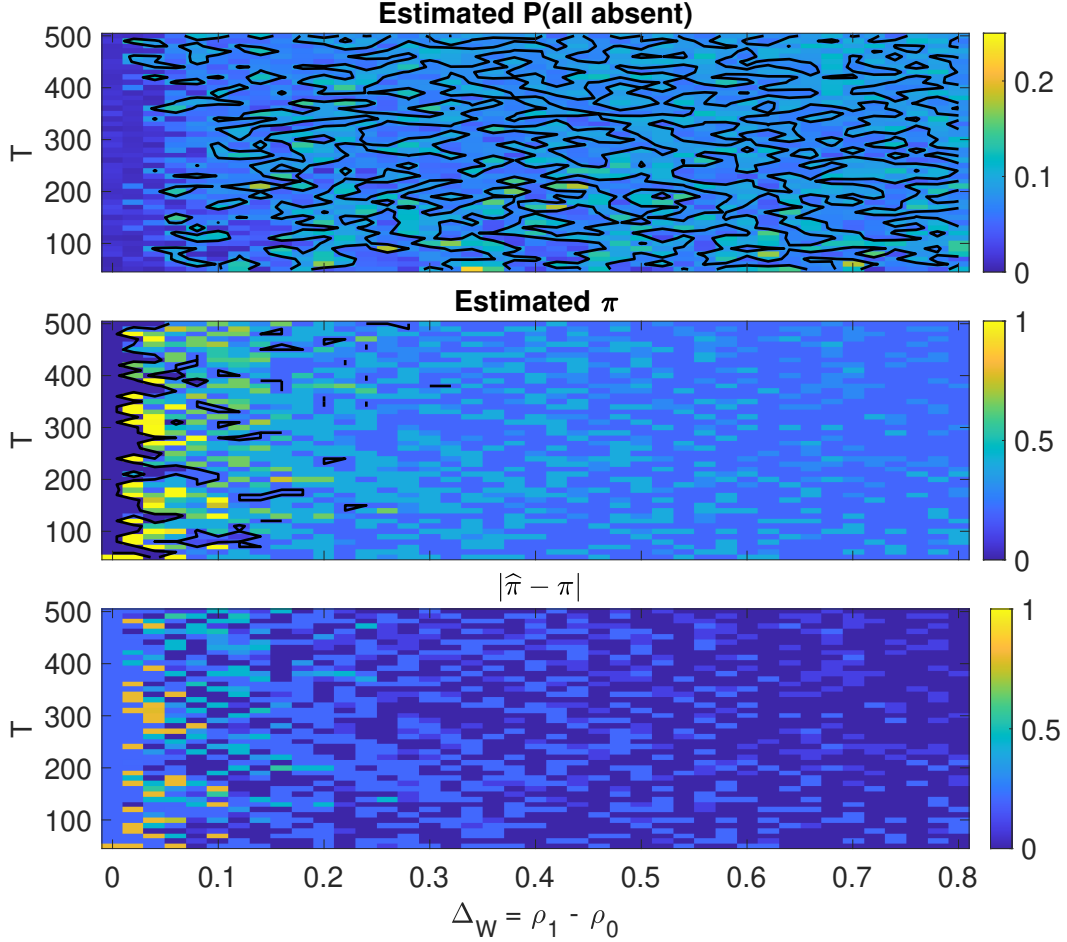


Figure 2: Dependence of identifiability on sample size T and goodness of proxy Δ_W . The number of iterates by `patternsearch` are limited in a maximum count of 250, which was always reached, the optimum search, thereby, terminating “prematurely”. Black contour lines indicate the true values.

close to the true value of π across all values of Δ_W ; see Fig. 4 (a). This demonstrates that the ambiguity does not originate from a flat likelihood in the π -direction alone.

The situation changes fundamentally once the remaining parameters are allowed to vary. Figure 4 (b) shows slices of the likelihood surface in the plane spanned by

$$(\pi, \lambda_\theta),$$

where λ_θ rescales the interaction matrix,

$$\theta \mapsto \lambda_\theta \theta.$$

A clear valley emerges also in this parameter plane. Increasing latent forcing strength can be partially compensated by decreasing social interaction strength, and vice versa, while preserving many observable statistical features of the attendance data. Thus, the model exhibits approximate compensation manifolds in parameter space rather than exact structural non-identifiability. These compensation directions generate rugged finite-sample likelihood landscapes containing multiple competing minima with comparable likelihood values [8]. The numerical evidence therefore suggests that the principal inferential difficulty is not the absence of information in principle, but rather weak practical identifiability caused by finite-sample variability and compensatory parameter trade-offs.

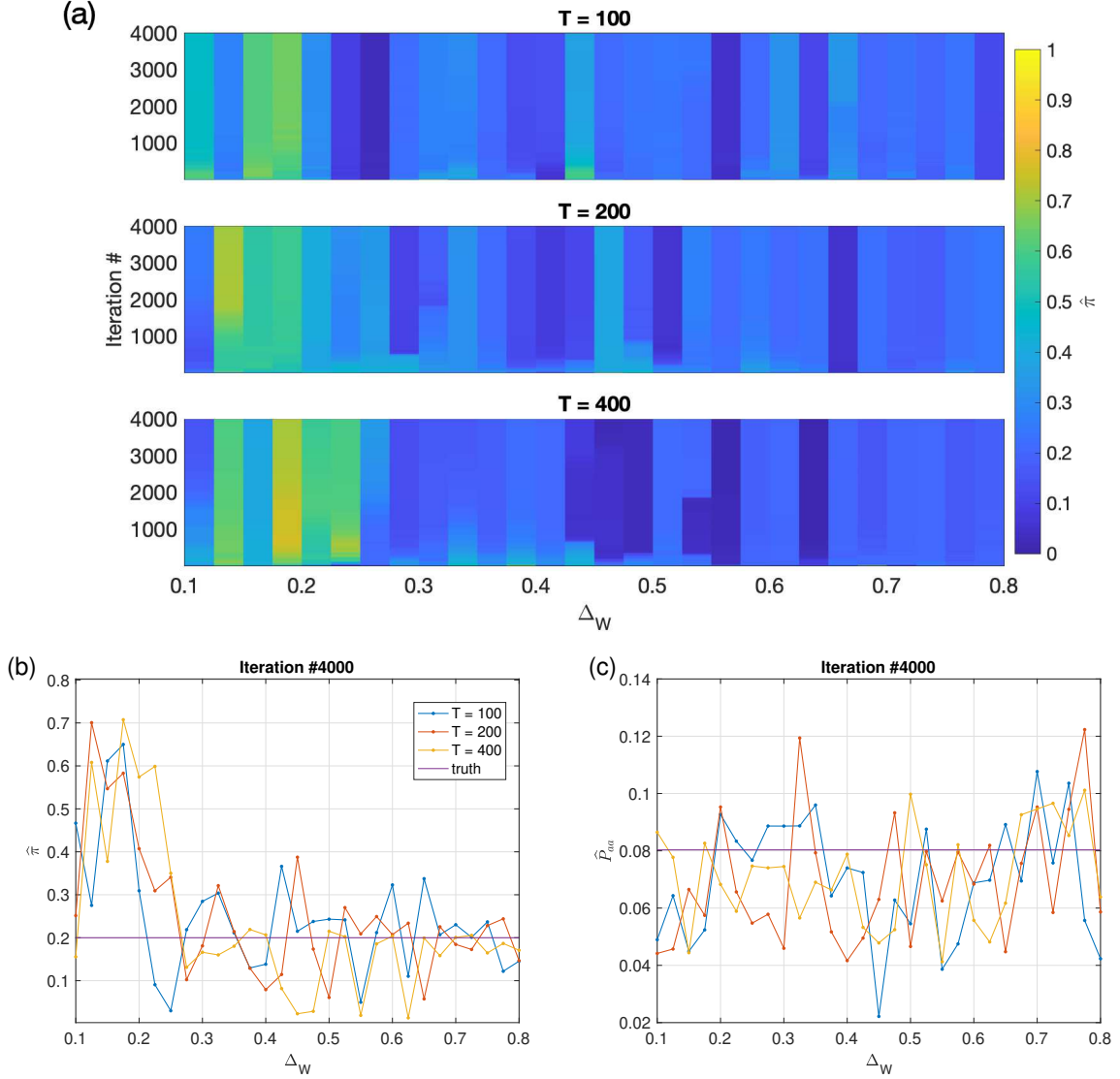


Figure 3: Dependence of identifiability on sample size T and the number of iterates of optimum search.

8 Conclusion

We investigated the estimation of the probability of collective absence in the presence of both social interactions and latent external forcing. Simple independence-based approaches severely underestimate the probability of collective absence whenever social dependence is present. Interaction-based models substantially improve estimation of extreme collective events, but latent external forcing introduces major inferential challenges. Approximation-based estimation methods proved unreliable for the combined latent-variable interaction model, necessitating full likelihood-based inference with exact partition-function evaluation.

The resulting numerical experiments revealed strong evidence of practical non-identifiability. Although informative proxy observations substantially improve inference, finite-sample likelihood landscapes remain rugged and highly multimodal. Different parameter configurations can achieve nearly identical likelihood values while implying substantially different decompositions between latent forcing and endogenous social interaction.

Importantly, the ambiguity was found to arise not from a flat likelihood in the latent forcing parameter alone, but from approximate compensation directions involving multiple parameters

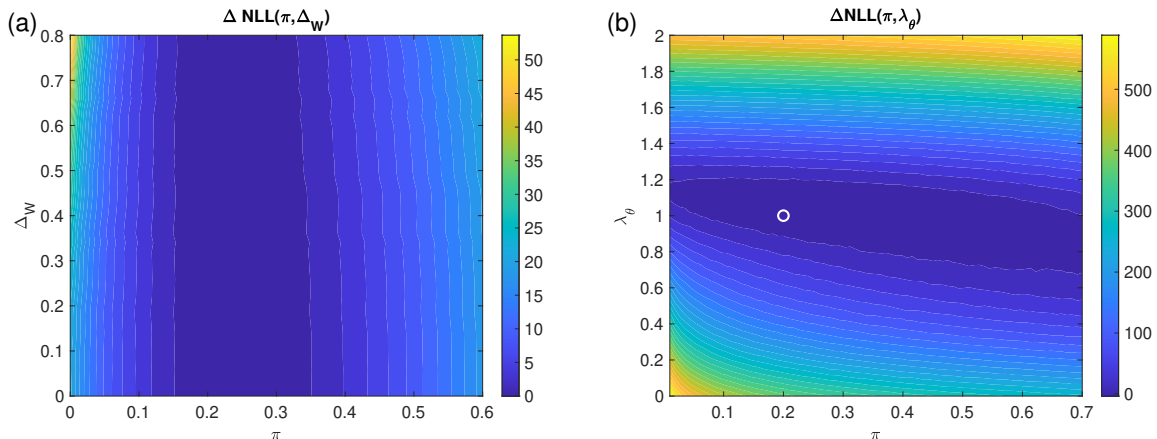


Figure 4: Identifiability in view of dependences of the “cost function”, the “negative log-likelihood”, NLL. (a) Fixing all parameters but π and the influence of the goodness of the proxy Δ_W . (b) Compensation between social dependence and exogenous effects affecting absence for fixed $\Delta_W = 0.4$.

simultaneously. In particular, latent forcing intensity and social coupling strength can partially compensate each other while preserving observable attendance statistics.

The study illustrates a broader statistical phenomenon: even when a model is structurally identifiable in principle, practical identifiability may remain weak in realistic finite samples due to finite-sample fluctuations, optimisation trapping, and compensatory parameter trade-offs.

Beyond the pedagogical setting motivating this work, similar inferential difficulties are likely to arise in many latent-variable network systems involving interacting agents subject to unobserved common forcing. Inference in interacting binary systems is known to become statistically challenging in high-dimensional settings [9].

Acknowledgements

I would like express my gratitude to Anna Szilvia Láng for the idea of estimating the probability of collective absence. I thank Gábor Drótos for a useful discussion upon asking him if he can guess whether the Ising-type model is identifiable based on attendance data alone. I greatly acknowledge OpenAI Inc. for making ChatGPT available for all, which I used throughout working on this problem [10]. This is how I obtained much of the model formulation, simulation codes, interpretation- and at times the explanation of results, references from the literature on solid state physics, statistical physics and numerical methods. I had ChatGPT also to produce the first draft of this paper, in latex syntax. Except for the original idea, as said, the majority of the ideas on the direction of this study, the numerical experiments to conduct, etc., are from myself. I provided very detailed prompts to ChatGPT to be able to benefit from its generative capacity. Much of the interpretation of numerical results were performed by myself. My work was supported by the Research Excellence Programme of the Hungarian University of Agriculture and Life Sciences (<https://research.uni-mate.hu/hu/2026cohort>). I’m grateful to HUN-REN for access to computing resources on the HUN-REN Cloud, on the account of the project (<https://science-cloud.hu/projektek/data-driven-seasonal-weather-forecast>).

References

- [1] Julian Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.
- [2] Linda E. Reichl. *A Modern Course in Statistical Physics*. Wiley, 1998.

- [3] Rodney J. Baxter. *Exactly Solved Models in Statistical Mechanics*. Academic Press, 1982.
- [4] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.
- [5] George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, 2 edition, 2002.
- [6] M. E. J. Newman and G. T. Barkema. *Monte Carlo Methods in Statistical Physics*. Oxford University Press, 1999.
- [7] Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.
- [8] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.
- [9] Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1 -regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.
- [10] OpenAI. ChatGPT (gpt-5.5). <https://chatgpt.com/share/69f7c7d0-036c-832c-a229-a5acfe4dd521>, 2026. Large language model used for discussion, code development, mathematical derivations, and manuscript editing assistance.