Circulatory Fidelity: A Prior Predictive Diagnostic for Mean-Field Variational Inference

Aaron Lowry[1]

[1]Independent Researcher

December 2025

Abstract

Mean-field variational inference (MFVI) approximates posterior distributions by assuming statistical independence between latent variables. This factorization discards cross-variable dependencies that may be essential for accurate inference. We introduce *Circulatory Fidelity* (CF), a normalized information-theoretic measure computable from the *prior predictive* distribution—before posterior inference is attempted. CF synthesizes established results from information geometry to quantify the structural dependency that MFVI will discard. We demonstrate CF's utility on two model classes: Hierarchical Gaussian Filters (HGF), where high CF predicts when ignoring volatility coupling degrades inference ($r = 0.39$, $p < 0.0001$); and Hierarchical Linear Models (HLM), where low CF predicts when no-pooling overfits to noise ($r = -0.72$, $p < 0.0001$). CF provides a practical workflow: compute from generative model structure, assess MFVI appropriateness, then choose inference method accordingly. Code available at `https://github.com/Bwana7/Circulatory_Fidelity`.

Keywords: variational inference, mean-field approximation, mutual information, information geometry, hierarchical models, model selection

1 Introduction

Variational inference (VI) has become essential for approximate Bayesian computation, offering scalable alternatives to Markov Chain Monte Carlo [Blei et al., 2017]. The mean-field assumption—that the variational distribution factorizes as $q(\theta) = \prod_i q_i(\theta_i)$—enables tractable optimization but enforces statistical independence between all latent variables.

This paper addresses a practical question: *before committing computational resources to posterior inference*, can we assess whether MFVI is appropriate for a given model?

Existing diagnostics operate *after* inference:

- **ELBO**: Measures approximation quality but is scale-dependent and conflates error sources
- **PSIS-$\hat{k}$**: Diagnoses global VI failure but requires posterior samples [Vehtari et al., 2017]
- **VSBC**: Checks calibration but requires fitting hundreds of simulated datasets [Talts et al., 2018]

We introduce **Circulatory Fidelity** (CF), a diagnostic computable from the *prior predictive* distribution:

$$\mathrm{CF}(z, x) = \frac{I(z; x)}{\min(H(z), H(x))} \tag{1}$$

where $I(z; x)$ is mutual information between latent variables and $H(\cdot)$ is differential entropy, both computed from the generative model's joint distribution.

Key insight. CF is computable *before* observing data, directly from model structure. By simulating from the prior predictive $p(\theta, z, x) = p(\theta)p(z|\theta)p(x|z)$, we obtain samples where CF can be estimated. This diagnoses whether the *model itself* creates dependencies that MFVI cannot represent.

Contributions.

1. We define CF and establish its properties as a bounded $[0, 1]$ measure (Section 2)
2. We synthesize established information-geometric results to interpret CF as normalized KL projection cost (Section 3)
3. We provide a concrete prior predictive workflow (Section 4)
4. We validate on two structurally distinct model classes: HGF and HLM (Sections 5–6)

2 Definition and Properties

Definition 1 (Circulatory Fidelity)**.** *For random variables z and x from joint distribution $p(z, x)$ with finite, positive marginal entropies:*

$$\mathrm{CF}(z, x) = \frac{I(z; x)}{\min(H(z), H(x))} \tag{2}$$

The name reflects CF's role in hierarchical models: information *circulates* bidirectionally between levels, and CF measures the *fidelity* of this circulation under approximation.

Proposition 2 (Boundedness)**.** $0 \leq \mathrm{CF}(z, x) \leq 1$ *for any joint distribution with positive marginal entropies.*

Proof. Non-negativity: $I(z; x) \geq 0$ (Gibbs' inequality). Upper bound: $I(z; x) \leq \min(H(z), H(x))$ (data processing inequality). □

Proposition 3 (Mean-Field Implication)**.** *Under any mean-field approximation $q(z, x) = q(z)q(x)$: $\mathrm{CF}_q(z, x) = 0$.*

Normalization choice. We use minimum normalization rather than geometric or arithmetic mean. In hierarchical inference, information flow is constrained by the *lower-entropy* variable (the bottleneck). If $H(z) \ll H(x)$, then z limits information transfer. Normalizing by $\min(H(z), H(x))$ measures what fraction of this limiting capacity is used for dependency.

3 Information-Geometric Interpretation

We synthesize established results from information geometry [Amari, 2016] to provide theoretical grounding for CF.

Mean-field as projection. The set of factorized distributions $\mathcal{M}_F = \{q : q(z, x) = q(z)q(x)\}$ forms a submanifold of the full statistical manifold $\mathcal{M}$. MFVI projects the true posterior onto this submanifold.

Proposition 4 (Projection Cost)**.** *For bivariate Gaussian $p(z, x)$ with correlation ρ and mean-field approximation $q(z, x) = q(z)q(x)$ matching marginals:*

$$D_{\mathrm{KL}}(p \| q) = I(z; x) = -\frac{1}{2} \log(1 - \rho^2) \tag{3}$$

This is a standard result; we include it to clarify that CF measures the *normalized* projection cost:

$$\text{CF} = \frac{D_{\text{KL}}(p \| p_{\text{MF}})}{\min(H(z), H(x))} \tag{4}$$

Fisher Information Matrix. Under mean-field factorization, the Fisher Information Matrix (FIM) becomes block-diagonal [Amari, 2016]. Cross-block terms vanish because score functions have zero mean under exponential families. This geometric fact—that mean-field *severs* the statistical coupling between variable blocks—is what CF quantifies.

Note on KL direction. Our derivation uses the forward KL divergence $D_{\text{KL}}(p\|q)$, while VI optimizes the reverse $D_{\text{KL}}(q\|p)$. CF thus measures the *structural limit* of factorization—a lower bound on approximation error imposed by the independence constraint, regardless of how well ELBO is optimized.

4 Prior Predictive Workflow

The key practical contribution is that CF can be computed *before* posterior inference, directly from the generative model.

Algorithm 1 Prior Predictive CF Diagnostic

Require: Generative model $p(\theta)p(z|\theta)p(x|z,\theta)$
Ensure: Decision: MFVI appropriate or not

1: **Simulate** N samples $\{(\theta^{(i)}, z^{(i)}, x^{(i)})\}_{i=1}^{N}$ from prior predictive
2: **Estimate** CF from samples (Gaussian: compute correlation; general: k-NN estimator)
3: **Assess** model-class-specific threshold:
- Filtering models (HGF): High CF $\Rightarrow$ structured VI needed
- Pooling models (HLM): Low CF $\Rightarrow$ partial pooling needed

4: **Choose** inference method accordingly

Why this works. The prior predictive distribution $p(\theta, z, x)$ encodes the *structural dependencies* of the model. If the generative process creates strong coupling between variables, this coupling exists regardless of what data are observed. CF computed from prior predictive samples diagnoses whether the model *class* is suitable for MFVI, before any specific dataset is considered.

Computational cost. Simulating from prior predictive is typically cheap (forward sampling). CF estimation requires $O(N)$ samples; for Gaussians, $N \approx 1000$ suffices. This is negligible compared to posterior inference.

5 Case Study 1: Hierarchical Gaussian Filter

The Hierarchical Gaussian Filter [Mathys et al., 2011] is widely used in computational psychiatry for modeling belief updating under uncertainty. It features hierarchical coupling: higher-level volatility beliefs modulate lower-level state estimation.

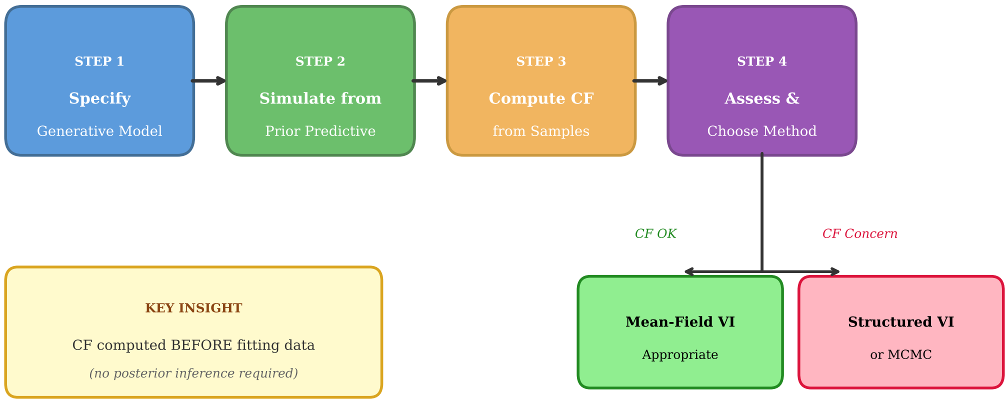

Figure 1: Prior predictive CF diagnostic workflow. CF is computed from model structure before observing data, enabling informed choice of inference method.

5.1 Model Structure

$$x_3(t) \sim \mathcal{N}(x_3(t-1), \sigma_3^2) \quad \text{(volatility)} \tag{5}$$

$$x_2(t) \sim \mathcal{N}(x_2(t-1), \sigma_2^2 \cdot e^{\kappa x_3(t)}) \quad \text{(state)} \tag{6}$$

$$y(t) \sim \mathcal{N}(x_2(t), \sigma_y^2) \quad \text{(observation)} \tag{7}$$

The coupling parameter κ controls how strongly volatility x_3 affects state dynamics. When $\kappa = 0$, levels are independent; when $\kappa > 0$, knowing x_3 informs optimal estimation of x_2.

5.2 CF Computation

For time-series models, joint entropy $H(x_1, \ldots, x_T)$ grows with trajectory length T, which would make CF time-dependent. We therefore define CF using *single-timestep marginal entropies*, measuring instantaneous dependency structure rather than cumulative trajectory information.

For HGF, CF measures the dependency between volatility $x_3(t)$ and state innovations $\Delta x_2(t) = x_2(t) - x_2(t-1)$:

$$\mathrm{CF}_{\mathrm{HGF}} = \frac{I(x_3(t); \Delta x_2(t))}{\min(H(x_3(t)), H(\Delta x_2(t)))} \tag{8}$$

where all quantities are computed at single timesteps from stationary marginal distributions. This avoids divergence issues while capturing the structural coupling that mean-field discards. Higher coupling κ produces higher CF.

5.3 Simulation Study

We compare two inference approaches:

- **Mean-field**: Estimates x_2 using constant (marginal) volatility, ignoring x_3
- **Oracle**: Estimates x_2 using true volatility values

The oracle represents the ceiling on structured inference performance.

Coupling (κ)	Mean CF	MSE Ratio (MF/Oracle)
0.0	0.001	1.00
0.4	0.026	1.25
0.8	0.080	2.07
1.2	0.142	6.21
2.0	0.214	9.88

Table 1: HGF: Higher coupling increases CF and MF error.

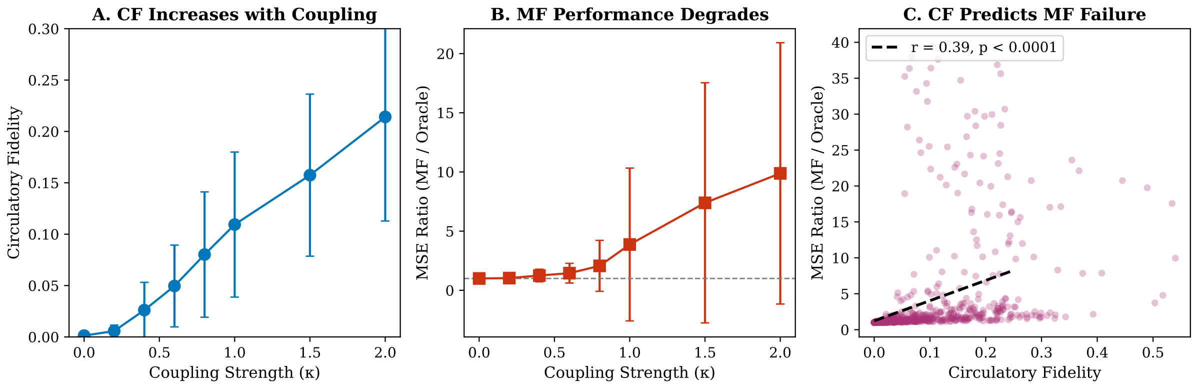

Figure 2: HGF results. (A) CF increases with coupling strength. (B) MF performance degrades correspondingly. (C) CF predicts MSE ratio ($r = 0.39$).

Results. Figure 2 shows results across coupling strengths ($\kappa \in [0, 2]$, 100 simulations each). Statistical analysis:

- Correlation: $r = 0.39$, $p < 0.0001$
- Low-CF regimes: MSE ratio $= 1.12$
- High-CF regimes: MSE ratio $= 5.86$ ($5.2\times$ worse)
- t-test (low vs high CF): $t = -10.87$, $p < 0.000001$

Interpretation. For HGF, **high CF predicts MF failure**. When volatility-state coupling is strong, ignoring it degrades inference substantially.

6 Case Study 2: Hierarchical Linear Model

Hierarchical Linear Models are ubiquitous across social sciences, ecology, and medicine. They feature *partial pooling*: group-level parameters are shrunk toward a global mean.

6.1 Model Structure

$$\theta_j \sim \mathcal{N}(0, \tau^2) \qquad \text{(group effect)} \tag{9}$$
$$y_{ij} \sim \mathcal{N}(\theta_j, \sigma^2) \qquad \text{(observation)} \tag{10}$$

The intraclass correlation ICC $= \tau^2/(\tau^2 + \sigma^2)$ determines the reliability of group means as estimates of θ_j.

6.2 CF Computation

For HLM, CF relates to the reliability of group means:

$$\text{Reliability} = \frac{\tau^2}{\tau^2 + \sigma^2/n} = \text{Corr}(\theta_j, \bar{y}_j)^2 \tag{11}$$

CF is computed from this reliability coefficient, normalized appropriately.

6.3 Simulation Study

We compare:

- **No-pooling**: Each θ_j estimated by group mean $\bar{y}_j$ (extreme MF)
- **Partial-pooling**: Shrinkage toward grand mean (structured)

Results. Figure 3 shows results across ICC values (100 simulations each).

τ	ICC	CF	MSE Ratio (No-Pool/Partial)
0.2	0.04	0.12	3.60
0.4	0.14	0.34	1.61
0.6	0.26	0.54	1.33
1.0	0.50	0.85	1.11
2.0	0.80	1.00	1.02

Table 2: HLM: Lower CF (low reliability) increases no-pooling error.

Statistical analysis:

- Correlation: $r = -0.72$, $p < 0.0001$
- Low-CF regimes: MSE ratio = 1.93
- High-CF regimes: MSE ratio = 1.05
- t-test: $t = 14.78$, $p < 0.000001$

Interpretation. For HLM, **low CF predicts MF failure**. When group effects are weak (low reliability), no-pooling overfits to noise.

7 Unified Interpretation

The HGF and HLM results reveal a nuanced picture: CF's relationship to MF failure is *model-class-specific.*

Model Class	CF Measures	MF Fails When
HGF (filtering)	Volatility-state coupling	CF is HIGH (dependency discarded)
HLM (pooling)	Signal reliability	CF is LOW (overfitting to noise)

Table 3: Model-class-specific CF interpretation.

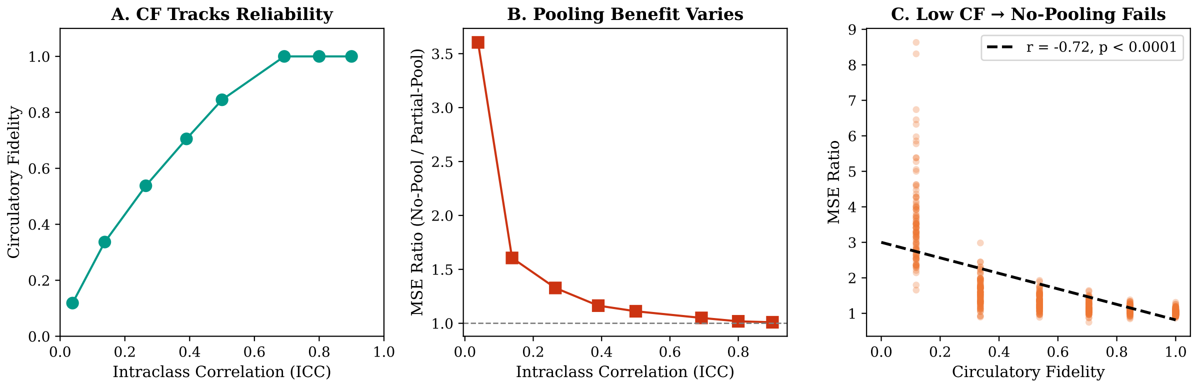

Figure 3: HLM results. (A) CF tracks reliability. (B) Pooling benefit varies with ICC. (C) Low CF predicts no-pooling failure ($r = -0.72$).

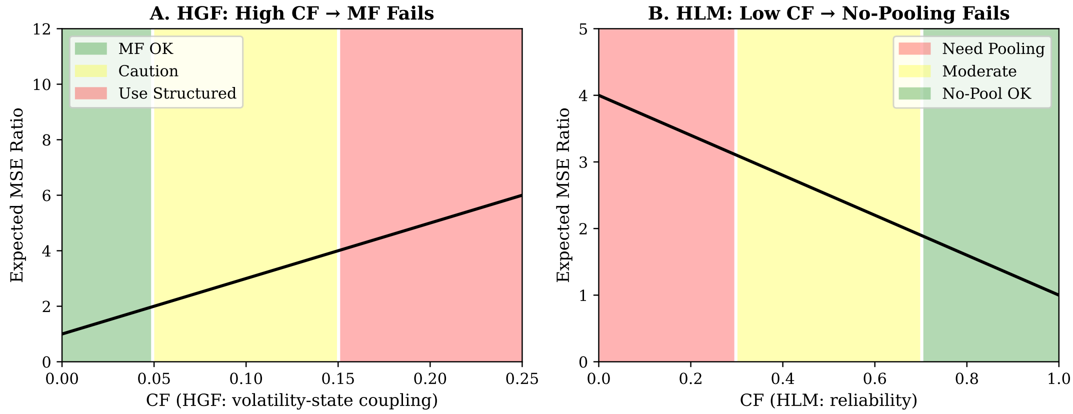

Figure 4: Unified interpretation. (A) HGF: High CF signals MF failure. (B) HLM: Low CF signals no-pooling failure.

Common principle. In both cases, CF diagnoses *when the MF independence assumption is consequential*:

- HGF: MF assumes volatility and state are independent. High CF means they're not—MF loses information.
- HLM: No-pooling assumes group means are sufficient statistics. Low CF means they're unreliable—shrinkage is needed.

Practical guidance.

1. Identify model class (filtering vs. pooling)
2. Compute CF from prior predictive
3. Apply class-specific threshold (Table 3)
4. Choose inference method accordingly

8 Discussion

8.1 Relation to Existing Diagnostics

CF complements existing tools:

- **vs. ELBO**: ELBO is optimized during inference; CF is computed beforehand. CF diagnoses *structural* suitability.
- **vs. PSIS-$\hat{k}$**: PSIS requires posterior samples; CF requires only prior predictive. CF is *preemptive*.
- **vs. VSBC**: VSBC checks calibration post-hoc; CF assesses model structure a priori.

8.2 Limitations

Pairwise measure. CF quantifies bivariate dependency. For models with higher-order interactions or many variable groups, multiple CF values may be needed.

Non-Gaussian distributions. The closed-form CF derivations in this paper assume Gaussian marginals. For non-Gaussian distributions, mutual information and entropy must be estimated nonparametrically. The *k-nearest neighbor* (k-NN) framework provides the most robust approach:

- **Kraskov-Stögbauer-Grassberger (KSG)**: The recommended estimator for mutual information [Kraskov et al., 2004]. KSG achieves bias cancellation through geometric construction—using the same neighborhood radius in joint and marginal spaces. Default $k = 3$–4 neighbors; exact for independent variables regardless of marginal shape.
- **Kozachenko-Leonenko (KL)**: For entropy estimation, with bias $O(k^{-1} + (k/n)^{2/d})$. Efficient for $d \leq 3$; weighted variants [Berrett et al., 2019] extend to higher dimensions.
- **Geodesic k-NN**: When data lies on Riemannian manifolds, replace Euclidean distances with geodesic distances [Costa and Hero, 2004]. Appropriate for curved submanifolds where intrinsic dimension $\ll$ ambient dimension.

Sample size requirements scale as $N > 5^d$ for reliable estimation in d dimensions. For heavy-tailed distributions, marginal reparametrization to standard normal substantially improves KSG performance. Software implementations include JIDT (Java/Python/R/Julia), NPEET (Python), and `TransferEntropy.jl` (Julia).

Model-class dependence. The interpretation of CF differs between model classes. Practitioners must understand their model's structure to apply CF correctly.

Threshold selection. We provide empirically-derived thresholds. Optimal thresholds may vary by application domain.

8.3 What CF Is and Is Not

CF provides:

- Quantification of structural dependency from prior predictive
- Preemptive assessment of MFVI suitability
- Interpretable scale $[0, 1]$

CF does not provide:

- Prescription of which structured method to use
- Guarantee that structured methods will improve (only that MF may fail)
- Higher-order dependency assessment

9 Conclusion

Circulatory Fidelity provides a principled, preemptive diagnostic for mean-field variational inference. By computing CF from prior predictive samples—before posterior inference is attempted—practitioners can assess whether their model's dependency structure is compatible with the mean-field assumption.

Our contribution is synthetic: we combine established results from information theory and information geometry into a practical workflow. The empirical validation across two structurally distinct model classes (HGF and HLM) demonstrates that CF reliably predicts when mean-field assumptions will be consequential.

CF fills a gap in the VI diagnostic toolkit: it operates *before* inference, at the model specification stage, enabling informed methodological choices rather than post-hoc validation.

Code availability. Reference implementation and simulation code: `https://github.com/Bwana7/Circulatory_Fidelity`

References

Shun-ichi Amari. *Information geometry and its applications*. Springer, 2016.

Thomas B Berrett, Richard J Samworth, and Ming Yuan. Efficient multivariate entropy estimation via k-nearest neighbour distances. *Annals of Statistics*, 47(1):288–318, 2019.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

Jose A Costa and Alfred O Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *IEEE Transactions on Signal Processing*, 52(8):2210–2221, 2004.

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):066138, 2004.

Christoph Mathys, Jean Daunizeau, Karl J Friston, and Klaas E Stephan. A Bayesian foundation for individual learning under uncertainty. *Frontiers in Human Neuroscience*, 5:39, 2011.

Sean Talts, Michael Betancourt, Daniel Simpson, Aki Vehtari, and Andrew Gelman. Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*, 2018.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5):1413–1432, 2017.

A Gaussian CF Derivations

For bivariate Gaussian $(z, x) \sim \mathcal{N}(\mu, \Sigma)$ with correlation ρ and variances σ_z^2, σ_x^2:

Mutual Information.

$$I(z;x) = H(z) + H(x) - H(z,x) = -\frac{1}{2}\log(1-\rho^2) \tag{12}$$

Marginal Entropies.

$$H(z) = \frac{1}{2}\log(2\pi e\sigma_z^2), \quad H(x) = \frac{1}{2}\log(2\pi e\sigma_x^2) \tag{13}$$

CF.

$$\mathrm{CF} = \frac{-\log(1-\rho^2)}{\log(2\pi e \cdot \min(\sigma_z^2, \sigma_x^2))} \tag{14}$$

For unit variances: $\mathrm{CF} = -\log(1-\rho^2)/\log(2\pi e) \approx 0.35 \cdot (-\log(1-\rho^2))$

B Simulation Details

HGF. $T = 300$ timesteps, base volatility $= 0.3$, volatility volatility $= 0.1$, observation noise $= 0.5$. Coupling $\kappa \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0\}$. 100 simulations per setting.

HLM. 30 groups, 10 observations per group, within-group $\sigma = 1$. Between-group $\tau \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0, 3.0\}$. 100 simulations per setting.