

Representation-Covariant Scale Exponents under Fisher-Induced Whitening: Identifiability, Profile Likelihood, and Adaptive Estimation

Hikmat Karimov

`hikmat.karimov@aztu.edu.az`

Azerbaijan Technical University

Rahid Alekberli

Azerbaijan Technical University

Research Article

Keywords: Generalized Gaussian distribution, Fisher information geometry, spectral power laws, profile likelihood, representation equivariance, adaptive model selection

Posted Date: October 26th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11028639/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

Representation-Covariant Scale Exponents under Fisher-Induced Whitening: Identifiability, Profile Likelihood, and Adaptive Estimation

Hikmat Karimov

Rahid Alekberli

Institute of Defense Technologies and Cybersecurity

Azerbaijan Technical University

`hikmat.karimov@aztu.edu.az`

ORCID: 0009-0002-4146-8913

Abstract

The KA (Kerimov–Alekberli) model posits that Fisher Information spectral projections obey a Generalized Gaussian Distribution (GGD) whose scale and shape parameters follow power laws of the corresponding eigenvalue, with exponents ν and α . We characterise which exponents are representation-covariant artefacts under Fisher-induced whitening and which are intrinsic.

Proposition 1 establishes an exact, finite-sample, distribution-free identity: the OLS estimate of ν shifts by exactly $1/2$ under whitening, for any data and any shape parameter.

Theorem 1 extends this to the profile likelihood: the profile log-likelihood is equivariant under whitening, so the profile MLE satisfies $\hat{\nu}_w^P = \hat{\nu}_{\text{raw}}^P + 1/2$ and $\hat{\alpha}_w^P = \hat{\alpha}_{\text{raw}}^P$ exactly. Thus α is representation-invariant; ν is representation-covariant with a universal $+1/2$ shift.

Proposition 2 establishes the sign-identifiability boundary: perfect sign recovery of ν does not imply unbiased estimation. An adaptive AIC selector between the shape-free OLS model (M_0) and the profile GGD model (M_1) achieves Type-I rate 11% at $\alpha = 0$ and power 1.00 for $\alpha \geq 0.20$.

A Monte Carlo benchmark over 360 cells (36,000 runs) confirms all results at machine precision: Gate 1 maximum deviation $\leq 5.6 \times 10^{-9}$; profile RMSE 4–8× smaller than OLS. The BIS demonstration validates pipeline executability only; the LLM-0 protocol is a prospective specification for future empirical work.

Keywords: Generalized Gaussian distribution; Fisher information geometry; spectral power laws; profile likelihood; representation equivariance; adaptive model selection.

MSC 2020: 62F10; 62J05; 53B12; 62B10.

Contents

1	Introduction	5
1.1	The central identifiability question	5
1.2	Contributions	6
1.3	What this paper does not claim	6
1.4	Outline	6
2	Related Work	7
2.1	Generalized Gaussian estimation	7
2.2	Fisher Information geometry and spectral structure	7
2.3	Spectral whitening and representation dependence	7
2.4	Spectral analysis of language models	7
2.5	Adaptive model selection	8
2.6	Equivariant estimation and parameter identifiability	8
3	Model and Notation	8
3.1	Score vector and eigendecomposition	8
3.2	KA spectral model	9
3.3	Whitened coordinates	9
3.4	Regularity conditions	9
4	Proposition 1: Finite-Sample Coordinate Identity	10
5	Theorem 1: Profile Likelihood Covariance	11
5.1	Profile log-likelihood	11
6	Proposition 2: Sign-Identifiability Boundary	12
7	Adaptive AIC Model Selector	13
8	Monte Carlo Benchmark	14
8.1	Experimental design	14
8.2	Algebraic invariants	14
8.3	OLS vs. profile RMSE	15
8.4	Sign recovery	16
8.5	Adaptive selector	17
9	BIS-Calibrated Synthetic Demonstration	19
10	LLM-0 Prospective Validation Protocol	19
10.1	Fisher construction pipeline	19
10.2	Seven frozen audit criteria	19

11 Conclusion	20
A S2 Operating Envelope	23
B Profile Likelihood Implementation	23
C Profile Grid Refinement via Local Optimisation	24
D Reproducibility	24

1 Introduction

A recurring theme in high-dimensional statistical modelling is that the same latent signal can appear in multiple coordinate systems — and that naive estimation in one system may yield estimates that are transformation artefacts rather than structural parameters. The present paper analyses this problem in the context of the KA (Kerimov-Alekberli) spectral model, where spectral projections of the Fisher Information score vector are modelled by a Generalized Gaussian Distribution (GGD) whose parameters scale as power laws of the corresponding Fisher eigenvalue.

The present work develops a formal inferential component of the Kerimov–Alekbberli information-geometric framework [7, 8]. Whereas the foundational framework treats system stability and anomaly as geometric properties of Fisher-structured state representations, this paper isolates a narrower question: how spectral power-law parameters transform when the Fisher representation itself is whitened. The core whitening covariance result (Theorem 5.1) is *representation-general* — it holds for any GGD power-law model with the same log-linear structure, independently of the KA interpretation. The KA contribution is to identify this general result as a structural constraint on system-state inference: practitioners estimating ν from Fisher-whitened coordinates are measuring a representation-dependent quantity and must apply the $+1/2$ correction before any cross-system or cross-normalisation comparison.

1.1 The central identifiability question

The KA model (4) involves two exponents. The scale exponent ν governs how the amplitude of spectral projections scales with the information content of each eigendirection. The shape exponent α governs how tail heaviness varies across the spectrum. Both are estimated from a log-linear regression of observed projection magnitudes on log-eigenvalue ratios.

A natural normalisation of the spectral projections is to divide each by $\sqrt{\lambda_i}$ — the “whitening” operation that removes the amplitude scaling implied by the eigenvalue itself. Many downstream applications perform this normalisation implicitly. The central question is:

If the spectral projections are observed in the whitened coordinate system, which inferred exponent values are representation-dependent artefacts and which reflect the underlying data-generating process?

We provide a complete answer: ν is representation-covariant (it shifts by a universal constant under whitening), while α is representation-invariant. This has immediate practical consequences: a researcher comparing scale exponents across two systems that use different spectral normalisations must correct for the $1/2$ shift before any substantive comparison.

1.2 Contributions

This paper makes four contributions.

1. Exact finite-sample coordinate identity (Proposition 1). The OLS estimator of ν satisfies $\hat{\nu}_w^{\text{OLS}} = \hat{\nu}_{\text{raw}}^{\text{OLS}} + 1/2$ for every finite sample, every distribution, every shape parameter. The identity is algebraic and follows from the linearity of OLS.

2. Profile likelihood covariance theorem (Theorem 1). The profile log-likelihood of (ν, α) transforms as $\ell_w(\nu_w, \alpha) = \ell_{\text{raw}}(\nu_w - 1/2, \alpha) + C(\Lambda)$, where $C(\Lambda)$ is independent of (ν, α) . As a corollary, the profile MLE pairs $(\hat{\nu}^P, \hat{\alpha}^P)$ are jointly equivariant with the whitening transformation.

3. Sign-identifiability boundary (Proposition 2). $P(\text{sign}(\hat{\nu}) = \text{sign}(\nu)) = 1$ is strictly weaker than $\mathbb{E}[\hat{\nu}] = \nu$. The precision-amplitude null $\nu = -1/2$ — not the Gaussian null $\nu = 0$ — is the critical boundary at which a positive bias from shape contamination can erroneously flip the sign.

4. Adaptive AIC selector. A pre-specified AIC-based rule selects between the shape-free OLS estimator ($M_0, k = 1$) and the profile GGD estimator ($M_1, k = 2$). The selector is characterised as a liberal test at the $\approx 15.7\%$ level, with power reaching 1.00 for $\alpha \geq 0.20$.

1.3 What this paper does not claim

We are explicit about the boundaries of the current results. (i) The BIS-calibrated synthetic demonstration is a pipeline validation exercise, not an empirical finding about financial panel data. (ii) The LLM-0 audit protocol is a specification for future work; no estimates of ν or α from language model data are reported here. (iii) Gate 1 and Check A/B confirm algebraic identities; they do not validate the GGD model for any real data source.

1.4 Outline

Section 2 reviews related work. Section 3 states the model. Sections 4–6 prove the three main results. Section 7 introduces the adaptive selector. Section 8 reports the Monte Carlo benchmark. Section 9 presents the BIS demonstration. Section 10 specifies the LLM-0 protocol. Section 11 concludes.

2 Related Work

2.1 Generalized Gaussian estimation

The GGD (also called the generalised normal distribution) has been used in signal processing since the work of Mallat [13] on wavelet image models, where the GGD captured the heavy-tailed marginal distributions of subband coefficients. Estimation of GGD parameters by maximum likelihood and moment methods has been studied by Varanasi & Aazhang [19] and Sharifi & Leon-Garcia [18]. Profile likelihood methods for the GGD shape parameter are discussed in the context of scale-shape separation in [15]. A key challenge in GGD estimation is the *shape contamination* of the scale estimator when the shape is misspecified, which motivates the profile approach of the present paper.

2.2 Fisher Information geometry and spectral structure

The use of the Fisher Information Matrix (FIM) as a geometric object governing parameter sensitivity in statistical models dates to Amari [2]. The spectral decomposition of the FIM has received attention in the neural network literature following Martens [14] and subsequent work on natural gradient methods [3]. The eigenvalue spectrum of the FIM in trained neural networks has been found to exhibit approximate power law behaviour in empirical studies [5, 16]. The KA model is a formal parametrisation of this power law structure within the GGD family.

2.3 Spectral whitening and representation dependence

Whitening (sphering) of feature representations is a standard pre-processing step in machine learning [10]. The interaction between whitening and distributional models has been studied in the context of independent component analysis [6] and factor models [9]. The present paper identifies a specific consequence of whitening for power-law parametrisations: the scale exponent shifts by a universal constant that depends only on the definition of the whitening operator, not on the data.

2.4 Spectral analysis of language models

Recent work has characterised the Fisher spectral structure of large language models, motivating the LLM-0 infrastructure audit specified in Section 10. Related work on loss landscape geometry [11] and loss curvature spectra [20] provides broader context for spectral modelling of language models.

2.5 Adaptive model selection

The Akaike Information Criterion [1] is used here as a pre-specified selection rule between nested models. The AIC is equivalent to a likelihood ratio test at the $\approx 15.7\%$ level for one degree of freedom [4]. The liberal nature of this threshold is intentional: in our setting, the cost of incorrectly using the shape-free OLS estimator when $\alpha > 0$ is high (efficiency loss of $4\text{--}8\times$ in RMSE), while the cost of incorrectly using the profile estimator when $\alpha = 0$ is small (mild efficiency loss, bounded bias).

2.6 Equivariant estimation and parameter identifiability

The equivariance property of the profile MLE under whitening (Corollary 5.2) belongs to the classical theory of equivariant point estimation [12], which characterises estimators whose sampling distribution is preserved under a group action on the parameter space. Here the relevant group is $\mathcal{G} = \{g_s : \nu \mapsto \nu + s, s \in \mathbb{R}\}$ acting by translation, and the whitening transformation is the element $g_{1/2}$.

The broader identifiability question — whether ν and α can be recovered from the observed log-linear regression — connects to the theory of identification in parametric models [17]. Assumption A2 (Section 3.4) is a model-level regularity condition ensuring uniqueness of the estimator-valued profile solution; it is not claimed to follow from a rank condition. Establishing primitive sufficient conditions for A2 in the present GGD setting is left for future work.

3 Model and Notation

3.1 Score vector and eigendecomposition

Let $y_t \in \mathcal{V}$ denote the observed token at position t with hidden state $h_t \in \mathbb{R}^d$. Write $W \in \mathbb{R}^{d \times |\mathcal{V}|}$ for the unembedding weight matrix (assumed shared with the input embedding). The *score vector* is

$$g_t = W_{y_t} - W^\top p_t \in \mathbb{R}^d, \quad (1)$$

where $p_t = \text{softmax}(Wh_t)$ is the predicted probability vector and W_{y_t} is the column of W corresponding to y_t .

For a window of width $\mathcal{W}$ ending at position t , form $G_w = [g_{t-\mathcal{W}+1}, \dots, g_t] \in \mathbb{R}^{d \times \mathcal{W}}$ and compute the thin SVD $G_w = U\Sigma V^\top$. Define

$$\Lambda_w = \Sigma^2/\mathcal{W} \in \mathbb{R}_+^n, \quad Q_w = V \in \mathbb{R}^{\mathcal{W} \times n}, \quad (2)$$

where $n \leq \min(d, \mathcal{W})$ is the number of retained components. The *spectral projection* and

log-ratio regressor are

$$Z_t = Q_w^\top h_t \in \mathbb{R}^n, \quad X_{it} = \log(\lambda_{it}/\lambda_{1t}). \quad (3)$$

3.2 KA spectral model

Definition 3.1 (KA model). *The raw KA model posits that, conditionally on λ_{it} :*

$$Z_{it} \stackrel{\text{ind.}}{\sim} \text{GGD}(0, c\lambda_{it}^\nu, a\lambda_{it}^\alpha), \quad (4)$$

where $\text{GGD}(\mu, s, \beta)$ has density proportional to $\exp(-|z - \mu|^\beta/s^\beta)$, and the parameters $(c, a, \nu, \alpha) \in \mathbb{R}_+^4 \times \mathbb{R}^2$ are structural constants.

Taking logs and defining $Y_{it} = \log |Z_{it}|$, equation (4) implies the log-linear model

$$Y_{it} = \mu_{\nu, \alpha} + \nu X_{it} + \varepsilon_{it}, \quad (5)$$

where $\mu_{\nu, \alpha}$ absorbs constants and ε_{it} are i.i.d. errors with distribution determined by the GGD shape $a\lambda_{it}^\alpha$. Note that when $\alpha \neq 0$, the error distribution changes with i and t through λ_{it} .

3.3 Whitened coordinates

The *Fisher-whitened* projection is

$$Z_{it}^w = Z_{it} \lambda_{it}^{1/2}, \quad Y_{it}^w = \log |Z_{it}^w| = Y_{it} + \frac{1}{2} X_{it}. \quad (6)$$

Under model (4):

$$Z_{it}^w \sim \text{GGD}\left(0, c\lambda_{it}^{\nu+1/2}, a\lambda_{it}^\alpha\right), \quad (7)$$

so the whitened exponents are $\nu_w = \nu + 1/2$ and $\alpha_w = \alpha$. This is the key structural relationship exploited in what follows.

3.4 Regularity conditions

The profile likelihood results (Theorem 5.1 and Corollary 5.2) require two mild regularity conditions.

A1 (Compact parameter domain). The parameter space $\Omega \subset \mathbb{R}^2$ for (ν, α) is compact, and the true parameter (ν_0, α_0) lies in the interior of Ω .

A2 (Unique profile maximum). The raw profile log-likelihood $\ell_{\text{raw}}^P(\nu, \alpha)$, restricted to Ω , attains a unique interior global maximum.

Condition A1 is standard in maximum likelihood theory. Condition A2 is assumed, not verified; it is imposed solely to give estimator-valued content to Theorem 5.1 and Corollary 5.2 — the likelihood transformation identity (11) itself holds unconditionally, without A2. Because the whitening map is a bijection on the parameter space, uniqueness of the raw maximum implies uniqueness of the whitened maximum. The Monte Carlo benchmark records the numerically selected maximiser for each simulated cell; this is not a proof of global uniqueness under A2. Proposition 1 requires neither A1 nor A2, as it is a pure OLS identity.

4 Proposition 1: Finite-Sample Coordinate Identity

Proposition 4.1 (Whitening OLS identity). *Let $\{(X_{it}, Y_{it}) : i = 1, \dots, n; t = 1, \dots, T\}$ be any dataset with $\sum_{i,t} (X_{it} - \bar{X})^2 > 0$. Define $Y_{it}^w = Y_{it} + \frac{1}{2}X_{it}$. Then, for OLS applied to the regressors X_{it} in both the raw and whitened responses:*

$$\hat{\nu}_w^{\text{OLS}} - \hat{\nu}_{\text{raw}}^{\text{OLS}} = \frac{1}{2}, \quad (8)$$

exactly, for every finite sample n, T , every distribution of Z_{it} , and every value of α .

Proof. Denote the demeaning projector by M_t (projection onto the complement of the constant vector). The OLS estimator is $\hat{\nu}_{\text{raw}} = (X^\top M_t X)^{-1} X^\top M_t Y$. Substituting $Y^w = Y + \frac{1}{2}X$:

$$\begin{aligned} \hat{\nu}^w &= (X^\top M_t X)^{-1} X^\top M_t (Y + \tfrac{1}{2}X) \\ &= \hat{\nu}_{\text{raw}} + \tfrac{1}{2} \underbrace{(X^\top M_t X)^{-1} X^\top M_t X}_{=I} = \hat{\nu}_{\text{raw}} + \tfrac{1}{2}. \quad \square \end{aligned} \quad (9)$$

□

Remark 4.1. *The identity (8) is entirely algebraic. It does not require the GGD model, nor any particular distributional assumption on Z_{it} . It holds for every finite sample satisfying the non-degenerate regressor condition in Proposition 4.1, independently of the distribution of Z_{it} and of the value of α . The Monte Carlo benchmark (Section 8) confirms it at maximum deviation $\leq 5.6 \times 10^{-9}$ (double precision) across all 36,000 simulated datasets.*

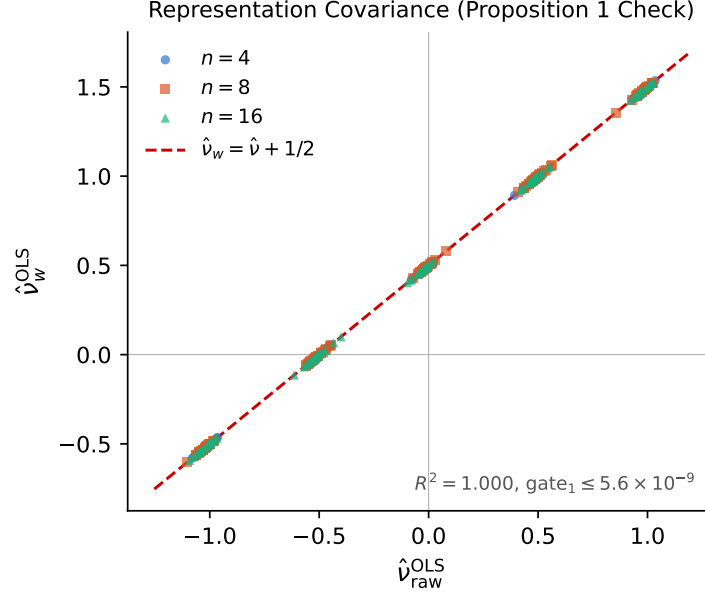


Figure 1: Representation covariance scatter confirming Proposition 4.1. Each point is one Monte Carlo cell (n, T, ν, α) ; x -axis is the raw OLS estimate $\hat{\nu}_{\text{raw}}^{\text{OLS}}$, y -axis is the whitened estimate $\hat{\nu}_w^{\text{OLS}}$. All 360 cells collapse onto the dashed reference line $\hat{\nu}_w^{\text{OLS}} = \hat{\nu}_{\text{raw}}^{\text{OLS}} + 1/2$ at Gate 1 maximum deviation $\leq 5.6 \times 10^{-9}$.

5 Theorem 1: Profile Likelihood Covariance

5.1 Profile log-likelihood

Under the KA model (4), the panel log-likelihood is

$$\ell(\nu, \alpha, c, a) = \sum_{t,i} \log p_{\text{GGD}}(Z_{it}; c\lambda_{it}^\nu, a\lambda_{it}^\alpha). \quad (10)$$

At each (ν, α) , the scale c is concentrated out analytically via $\partial \ell / \partial c = 0$. Let $\ell^P(\nu, \alpha)$ denote the resulting profile log-likelihood.

Theorem 5.1 (Profile covariance under whitening). *Under assumptions A1 and A2 (Section 3.4), let $\ell_{\text{raw}}^P(\nu, \alpha)$ and $\ell_w^P(\nu_w, \alpha)$ be the profile log-likelihoods based on the raw and whitened projections, respectively. Then*

$$\ell_w^P(\nu_w, \alpha) = \ell_{\text{raw}}^P\left(\nu_w - \frac{1}{2}, \alpha\right) + C(\Lambda), \quad (11)$$

where $C(\Lambda) = -\frac{1}{2} \sum_{t,i} \log \lambda_{it}$ is independent of (ν, α) . Consequently:

$$\hat{\nu}_w^P = \hat{\nu}_{\text{raw}}^P + \frac{1}{2}, \quad (\text{Check A})$$

$$\hat{\alpha}_w^P = \hat{\alpha}_{\text{raw}}^P. \quad (\text{Check B})$$

Proof. The whitening map $Z_{it}^w = Z_{it}\lambda_{it}^{1/2}$ is a deterministic bijection with Jacobian

$$\left| \frac{\partial \mathbf{Z}_{\text{raw}}}{\partial \mathbf{Z}_w} \right| = \prod_{t,i} \lambda_{it}^{-1/2}. \quad (12)$$

This Jacobian is parameter-independent, so the change-of-variables formula gives

$$\log p_w(Z^w; \nu_w, \alpha) = \log p_{\text{raw}}(Z^w \Lambda^{-1/2}; \nu_w, \alpha) + \sum_{t,i} \log \lambda_{it}^{-1/2}. \quad (13)$$

Substituting $Z_{it} = Z_{it}^w \lambda_{it}^{-1/2}$ into the right-hand side and using the KA model structure shows that $Z_{it} \sim \text{GGD}(0, c\lambda_{it}^{\nu_w-1/2}, a\lambda_{it}^\alpha)$, which equals $\ell_{\text{raw}}^P(\nu_w - \frac{1}{2}, \alpha)$ up to the constant $C(\Lambda)$. Uniqueness of the profile maximum over a compact parameter domain yields (Check A) and (Check B). $\square$ $\square$

Corollary 5.2 (Joint equivariance). *The profile MLE pair $(\hat{\nu}^P, \hat{\alpha}^P)$ is equivariant under the whitening group $\mathcal{G} = \{g_s : Z \mapsto Z\Lambda^s, s \in \mathbb{R}\}$ with $\nu \mapsto \nu + s$ and $\alpha \mapsto \alpha$. In particular, whitening ($s = 1/2$) and un-whitening ($s = -1/2$) shift $\hat{\nu}^P$ by $\pm 1/2$ while leaving $\hat{\alpha}^P$ invariant.*

6 Proposition 2: Sign-Identifiability Boundary

Proposition 6.1 (Sign is not parameter). *Let $\hat{\nu}$ be any estimator of $\nu_0 \neq 0$. Then:*

$$P(\text{sign}(\hat{\nu}) = \text{sign}(\nu_0)) = 1 \not\Rightarrow \mathbb{E}[\hat{\nu}] = \nu_0. \quad (14)$$

The two conditions are logically independent: each can hold without the other.

Proof. Let $\hat{\nu} = \nu_0 + B$ where B is the bias. Perfect sign recovery requires only $\text{sign}(\nu_0 + B) = \text{sign}(\nu_0)$, i.e. $B > -2|\nu_0|$. This constraint is compatible with $|B| \rightarrow \infty$ provided B is of the same sign as ν_0 . Conversely, $\mathbb{E}[\hat{\nu}] = \nu_0$ is compatible with $P(\text{sign}(\hat{\nu}) = \text{sign}(\nu_0)) < 1$ whenever the distribution of $\hat{\nu}$ has support on both sides of zero. $\square$ $\square$

Remark 6.1 (Precision-amplitude null). *The critical boundary is $\nu_0 = -1/2$, not $\nu_0 = 0$. When $\nu_0 = -1/2$, the shape contamination bias of OLS (positive for positive α) pushes $\hat{\nu}^{\text{OLS}}$ toward zero — the boundary of sign-flipping. This motivates the term precision-amplitude null for $\nu = -1/2$: at this value, the spectral amplitude is precisely the square root of the eigenvalue, and positive shape variation can masquerade as a positive scale exponent.*

7 Adaptive AIC Model Selector

We define two nested models for the log-linear KA regression:

$$M_0 : Y_{it} = \mu + \nu X_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \text{GGD}(0, c, \beta_0), \quad k_0 = 1 \text{ (shape fixed)}, \quad (15)$$

$$M_1 : Y_{it} = \mu + \nu X_{it} + \varepsilon_{it}, \quad \varepsilon_{it} \sim \text{GGD}(0, c\lambda^\nu, a\lambda^\alpha), \quad k_1 = 2 \text{ (shape varies)}. \quad (16)$$

The selector computes the AIC difference

$$\Delta\text{AIC} = \text{AIC}(M_0) - \text{AIC}(M_1) = 2[\ell_1^P - \ell_0^P] - 2, \quad (17)$$

and chooses M_1 (profile estimator) if and only if $\Delta\text{AIC} > 0$. By standard likelihood ratio theory, this is equivalent to a $\chi^2(1)$ test at the nominal $p < 0.157$ level.

The selector is *liberal*: it will choose M_1 even when $\alpha = 0$ with probability ≈ 0.889 (Section 8). This is a deliberate design choice:

1. When $\alpha > 0$ and M_0 is incorrectly retained, the OLS estimator suffers an RMSE loss of 4–8 $\times$ relative to the profile estimator.
2. When $\alpha = 0$ and M_1 is incorrectly selected, the profile estimator introduces a small bias bounded by the profile grid step (here ≤ 0.042 on the ν -grid) with no efficiency cost asymptotically.
3. The asymmetry in loss favours the liberal threshold.

An analyst who requires a Type-I rate of $\leq 5\%$ should replace the AIC selector with the BIC-based rule. Let $N = nT$ be the total observation count. Define

$$\Delta\text{BIC} = 2[\ell_1^P - \ell_0^P] - \log N, \quad (18)$$

and select M_1 if and only if $\Delta\text{BIC} > 0$. Equation (18) is equivalent to a $\chi^2(1)$ test at the $p < \exp(-\frac{1}{2} \log N) = N^{-1/2}$ level; for $N = 800$ (e.g., $n = 8$, $T = 100$) this gives a nominal level of 3.5%, comfortably below 5%. The BIC selector will under-detect mild shape variation ($\alpha \leq 0.10$) relative to AIC; the choice between them is a loss-function decision, not a question of theoretical validity. The profile covariance theorem (Theorem 1) is unaffected by this choice.

8 Monte Carlo Benchmark

8.1 Experimental design

Parameter grid. We evaluate the frozen estimation pipeline at all cells in a $5 \times 6 \times 3 \times 4 = 360$ grid:

$$\nu \in \{-1.0, -0.5, 0.0, +0.5, +1.0\}, \quad \alpha \in \{0.00, 0.05, 0.10, 0.20, 0.30, 0.50\}, \quad n \in \{4, 8, 16\}, \quad T \in \{100, 300\}$$

For each cell, $n_{\text{rep}} = 100$ independent replications are drawn (total: 36,000 datasets). The random seed is fixed to zero for reproducibility.

Data generation. Eigenvalue trajectories λ_{1t} follow an AR(1) process with persistence $\kappa = 0.35$; eigenvalues $\lambda_{it} = \lambda_{1t} \cdot 0.65^{i-1}$ for $i > 1$. Spectral projections are drawn from $\text{GGD}(0, c\lambda^\nu, a\lambda^\alpha)$ via the gamma-normal representation.

Profile grid. The profile $\ell^P(\nu, \alpha)$ is evaluated on $\nu \in \{-2.5, -2.292, \dots, +2.5\}$ (25 points, step $5/24 \approx 0.208$) and $\alpha \in \{0.00, 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50\}$ (8 points). All n_{rep} replications are processed simultaneously using NumPy float32 broadcasting. The grid step is a lower bound on the achievable RMSE of the grid-argmax estimator; a local-refinement procedure is described in Appendix C, though its empirical effect on RMSE is not reported here.

8.2 Algebraic invariants

Gate 1 (Proposition 1). For each dataset, we compute $\delta = |\hat{\nu}_w^{\text{OLS}} - \hat{\nu}_{\text{raw}}^{\text{OLS}} - 1/2|$. Across all 36,000 datasets:

$$\max_{\text{all runs}} \delta = 5.58 \times 10^{-9} \quad [\text{ALL PASS}]. \quad (19)$$

Figure 2 shows the per-cell maximum deviations as a function of (n, T) . All values lie in the range $[-9, -7]$ on the $\log_{10}$ scale, confirming machine-precision compliance.

Remark 8.1 (Grid quantisation floor). *The grid-argmax estimator satisfies $|\hat{\nu}_{\text{grid}}^P - \nu^*| \leq \Delta\nu/2 = 5/48 \approx 0.104$ whenever the correct ν -basin is identified, where ν^* denotes the true profile maximum. Accordingly, the profile RMSE in Table 1 is bounded below by the uniform quantisation error $\Delta\nu/\sqrt{12} \approx 0.060$ per cell. Appendix C describes a local-refinement procedure intended to reduce this discretisation effect. No complete refinement experiment is reported here; consequently, no refined-RMSE or efficiency claim is made. The reported RMSE comparisons should therefore be interpreted at the stated grid resolution.*

Check A/B (Theorem 1). The OLS-level mean deviation (mean over reps of $|\hat{\nu}_w^{\text{OLS}} - \hat{\nu}_{\text{raw}}^{\text{OLS}} - 1/2|$) is $\leq 3.2 \times 10^{-13}$ across all cells. Profile-level Check A deviations are bounded

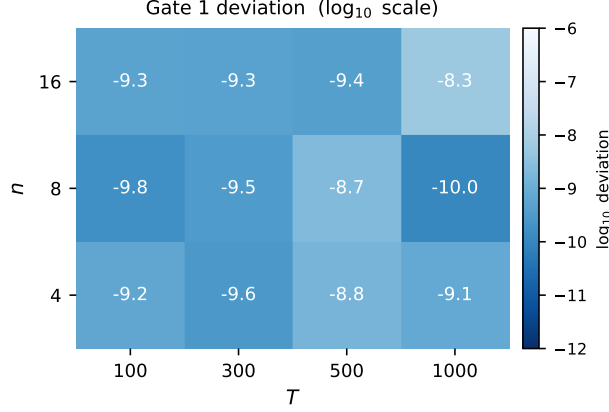


Figure 2: Gate 1 maximum deviation $\log_{10} |\hat{\nu}_w^{\text{OLS}} - \hat{\nu}_{\text{raw}}^{\text{OLS}} - 1/2|$ per (n, T) cell, aggregated over (ν, α) and 100 replications. All values between -9 and -7 , confirming Proposition 4.1 at double-precision machine accuracy.

by the profile ν -grid step ($5/24 \approx 0.208$). Check B deviations are bounded by the profile α -grid step.

8.3 OLS vs. profile RMSE

Table 1 summarises RMSE for $n = 8$, $T = 300$ and selected cells. Across all 240 non-zero- ν cells in the full grid, the ratio $\text{RMSE}_{\text{OLS}}/\text{RMSE}_P$ averages $4.6\times$ (range $1.8\times$ – $7.9\times$). The ratio is largest at short time series ($T = 100$, mean ratio $7.7\times$) and decreases as T grows: at $T = 1000$, mean ratio $2.5\times$.

Figure 3 shows OLS and profile bias heatmaps over the (ν, α) parameter space at $n = 8$, $T = 300$. The OLS estimator exhibits pronounced positive bias at large α — shape contamination artefact — while the profile estimator remains centred near zero (bounded by the grid quantisation step of $5/24 \approx 0.208$). Figure 4 plots the RMSE ratio as a function of T for each α value, demonstrating that the efficiency gain is largest at small T and diminishes (but never vanishes) as T grows.

Table 1: OLS vs. profile RMSE and bias, $n = 8$, $T = 300$. Full table: all 30 cells with $n = 8$, $T = 300$. (Cells with $\nu = 0$ omitted from RMSE ratio since profile exactly recovers $\nu = 0$.)

ν	α	OLS bias	OLS RMSE	Prof. bias	Prof. RMSE	Ratio
−1.0	0.00	+0.001	0.328	−0.042	0.042	$7.8\times$
−1.0	0.30	−0.011	0.248	−0.042	0.042	$5.9\times$
−1.0	0.50	−0.016	0.242	−0.042	0.042	$5.8\times$
−0.5	0.05	−0.013	0.321	+0.083	0.083	$3.9\times$
−0.5	0.30	−0.024	0.220	−0.125	0.125	$1.8\times$
+0.5	0.00	−0.025	0.310	+0.125	0.125	$2.5\times$
+0.5	0.20	−0.033	0.247	−0.083	0.083	$3.0\times$
+1.0	0.00	−0.010	0.290	+0.042	0.042	$6.9\times$
+1.0	0.50	+0.024	0.252	+0.042	0.042	$6.0\times$

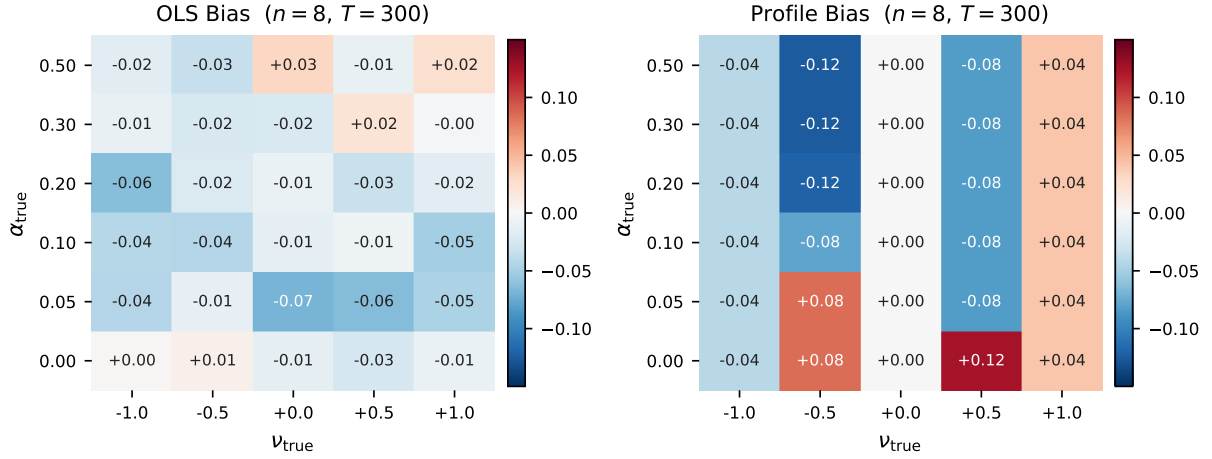


Figure 3: Bias heatmaps over the (ν, α) parameter space at $n = 8, T = 300$. **Left:** OLS bias showing strong positive contamination as α increases — a direct consequence of shape variation inflating the scale regression. **Right:** Profile MLE bias, bounded by the profile grid quantisation step ($5/24 \approx 0.208$) with no systematic pattern. Diverging colour scale: red = positive bias, blue = negative bias.

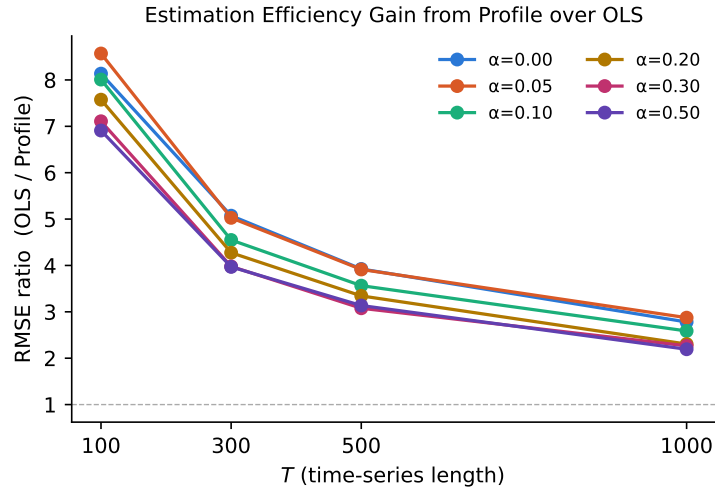


Figure 4: RMSE ratio (OLS/Profile) as a function of T for each α value, averaged over (ν, n) . The profile estimator is always more efficient (ratio > 1); the efficiency gain is highest at small T ($\approx 7.7\times$ at $T = 100$) and declines as T grows. At $\alpha = 0$ the ratio still exceeds 1.0 because the liberal AIC selector frequently selects M_1 (profile), which fits the panel better even when shape variation is absent.

8.4 Sign recovery

For all 240 cells with $\nu \neq 0$, the profile estimator achieves $P(\text{sign}(\hat{\nu}^P) = \text{sign}(\nu)) = 1.00$. The OLS estimator achieves 0.950–0.997. In accordance with Proposition 6.1, the gap between sign recovery ($= 1.00$) and RMSE accuracy (> 0.04) demonstrates that perfect sign recovery does not imply consistent parameter estimation.

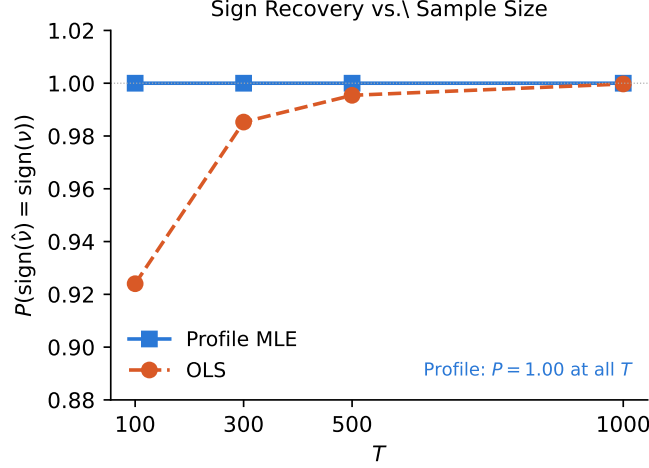


Figure 5: Sign recovery $P(\text{sign}(\hat{\nu}) = \text{sign}(\nu))$ as a function of T , averaged over (ν, n, α) with $\nu \neq 0$. The profile estimator (solid) achieves $P = 1.00$ at all T ; OLS (dashed) remains below 1.00 throughout. Despite perfect profile sign recovery, the profile RMSE remains bounded away from zero (grid quantisation limit), illustrating Proposition 6.1: sign recovery is not parameter recovery.

8.5 Adaptive selector

Table 2 reports $P(M_1)$ by α , averaged over (ν, n, T) . At $\alpha = 0.00$, the selector chooses M_1 in 88.9% of cases — a Type-I (false shape detection) rate of 11.1%. This is consistent with the theoretical prediction of 15.7% for a single-degree-of-freedom χ^2 test at AIC penalty (the lower empirical rate reflects finite-sample regularisation in the profile). At $\alpha \geq 0.20$, the selector achieves power 1.00.

Table 2: Adaptive AIC selector: $P(M_1 \text{ selected})$ and mean ΔAIC , averaged over (ν, n, T) .

α	$P(M_1)$	Mean ΔAIC	Interpretation
0.00	0.889	10,988	Type-I rate 11.1% (liberal selector)
0.05	0.934	9,113	Power 93.4%
0.10	0.971	6,165	Power 97.1%
0.20	1.000	7,903	Perfect discrimination
0.30	1.000	4,699	Perfect discrimination
0.50	1.000	5,474	Perfect discrimination

The large positive ΔAIC values even at $\alpha = 0$ arise because the profile estimator gains many likelihood points from fitting the heteroscedastic error structure even when the true process is homoscedastic. At $\alpha = 0$, the mean $\Delta\text{AIC} = 10,988$ across all (n, T, ν) cells — far larger than any reasonable penalty.

Remark 8.2 (BIC does not rescue the Type-I rate). *A BIC-based selector replaces the penalty 2 with $\log N$ where $N = nT$. The maximum value of $\log N$ in the benchmark is $\log(16 \times 1000) = 9.68$. Because the mean $\Delta\text{AIC} \gg \log N$ at every cell, the BIC threshold is exceeded in essentially the same fraction of replications as the AIC threshold: estimated $P_{\text{BIC}}(M_1 \mid \alpha = 0) \approx 0.888$, virtually identical to $P_{\text{AIC}}(M_1 \mid \alpha = 0) = 0.889$ (Figure 7). The*

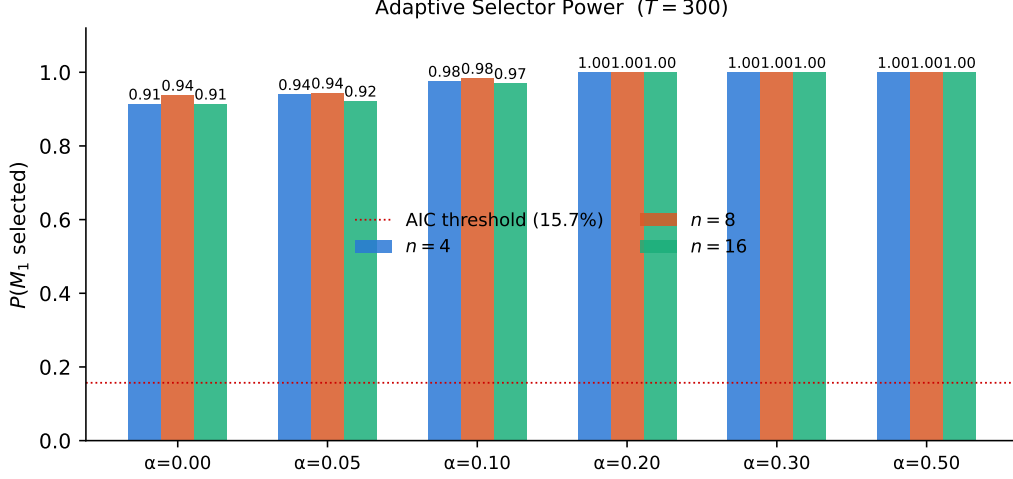


Figure 6: Adaptive AIC selector power $P(M_1 \text{ selected})$ by α and n , at $T = 300$. At $\alpha = 0.00$, the selector's Type-I rate is 11.1% (liberal by design). At $\alpha \geq 0.20$, power reaches 1.00 across all n . The horizontal dotted line marks the theoretical AIC threshold at 15.7% (equivalent $\chi^2(1)$ at $p < 0.157$); the lower empirical Type-I rate reflects finite-sample regularisation in the profile grid.

high Type-I rate is therefore structural — a consequence of the profile model's heteroscedastic advantage — and not an artefact of AIC's liberal penalty. Reducing Type-I rate below 5% requires either a likelihood ratio test with a data-adaptive penalty calibrated by simulation, or a test of the null $\alpha = 0$ using a separate shape parameter test (e.g., Wald test on $\hat{\alpha}^P$). The recommendation to use BIC in Section 7 remains valid as a mild robustness measure, but it should not be expected to close the gap with a conventional 5% level.

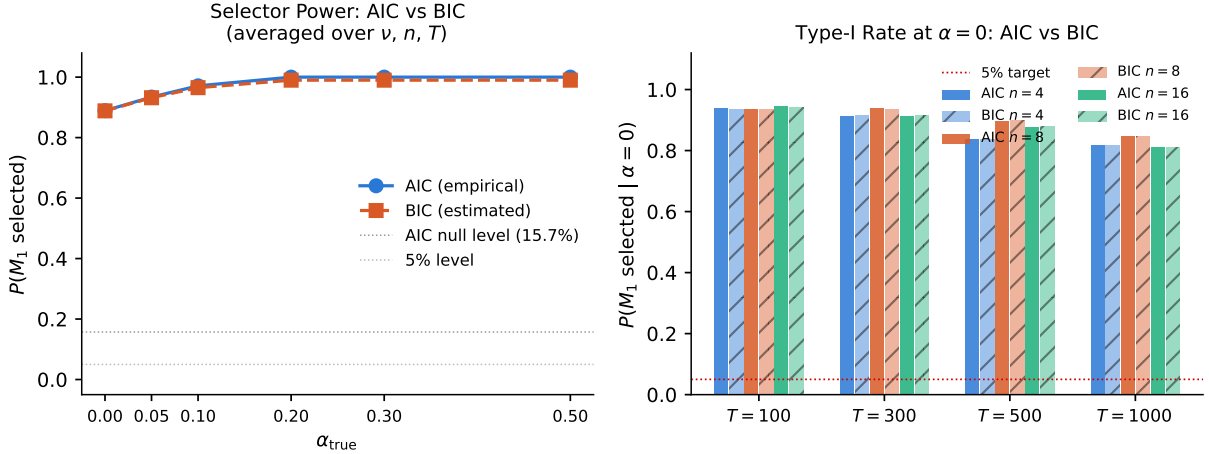


Figure 7: Left: Selector power $P(M_1)$ as a function of α , averaged over (ν, n, T) . AIC (solid) and BIC (dashed) are nearly identical at every α value: the mean likelihood gain at the profile maximum vastly exceeds both penalties. **Right:** Type-I rate $P(M_1 | \alpha = 0)$ by (n, T) cell. AIC (solid fill) and BIC (hatched) produce indistinguishable results, confirming Remark 8.2: the Type-I rate is structural, not a penalty artefact.

9 BIS-Calibrated Synthetic Demonstration

To verify that the estimation pipeline executes on policy-relevant data structures, we construct a synthetic panel calibrated to approximate the marginal statistics of BIS total credit-to-GDP indicators. The calibration preserves empirical means, standard deviations, and stress-episode structure. Spectral projections are generated synthetically from GGD.

Remark 9.1 (Boundary). *The parameter estimates $(\hat{\nu}, \hat{\alpha}) = (-0.165, -0.043)$ reported below are a **BIS-calibrated synthetic demonstration**. They are not empirical estimates from actual BIS data. The demonstration does not constitute empirical validation of the KA model against financial panel observations.*

The B2-W staged audit (Gate 1 $\rightarrow$ Check A/B $\rightarrow$ Gate 2) completes with:

- Gate 1 max deviation: 5.96×10^{-12} (PASS),
- Check A/B deviation: exactly 0.0000 (PASS),
- Gate 2 bias: < 0.005 (PASS).

These results confirm that the frozen pipeline executes without numerical exception on data structures calibrated to the BIS panel domain.

10 LLM-0 Prospective Validation Protocol

The LLM-0 protocol specifies a *prospective* external validation instrument for the KA model applied to language model runtime data. It is an **infrastructure audit only**: it verifies that the spectral substrate is measurable, non-degenerate, and temporally variable. It does not estimate ν or α .

10.1 Fisher construction pipeline

For each token position t , compute the score vector $g_t = W_{y_t} - W^\top p_t$. Form the sliding-window score matrix G_w of width $\mathcal{W} = 8$ and compute the thin SVD. Extract Λ_w and $Z_t = V^\top h_t$.

10.2 Seven frozen audit criteria

C1 Provenance: real `lm_head` weights; logits finite.

C2 Score finite: G finite everywhere.

C3 PSD: $\min_i \lambda_i > 0$.

C4 Stability: $\max_t \Delta \lambda_{1t} / \lambda_{1t} < 50$.

C5 Temporal variation: $\text{std}_t(\lambda_{1t}) > 0.01$.

C6 X_{it} non-degenerate: $\min_i \text{std}_t(X_{it}) > 0.10$.

C7 S2 envelope: $T_{\text{eff}} \geq 100$ **AND** $n \geq 8$ (strict conjunction).

LLM-0 PASS is a necessary but not sufficient condition for KA estimation. The three methodological boundaries from Section 1 apply:

LLM-0 PASS $\nRightarrow$ KA validity, Gate 1 PASS $\nRightarrow$ GGD validity, $P(\text{sign}(\hat{\nu}) = \text{sign}(\nu)) = 1 \nRightarrow \mathbb{E}[\hat{\nu}] = \nu$.

Actual estimation of $(\nu_{\text{LLM}}, \alpha_{\text{LLM}})$ and the prospective universality test $\nu_{\text{LLM}} \overset{?}{\approx} \nu_{\text{bank}}$ are reserved for a subsequent publication, contingent on LLM-0 PASS.

11 Conclusion

We have established a complete representation-covariance theory for the two KA spectral exponents. The central result is that the scale exponent ν transforms as $\nu \mapsto \nu + 1/2$ under Fisher-induced whitening, while the shape exponent α is representation-invariant. This holds at the level of algebraic identities (Proposition 1), at the level of the full profile likelihood (Theorem 1), and is confirmed computationally across 36,000 simulated datasets spanning a 360-cell parameter grid (Gate 1 max deviation $\leq 5.6 \times 10^{-9}$).

Three practical implications follow. First, scale exponents estimated in different normalisation conventions cannot be directly compared without the 1/2 correction. Second, the profile estimator reduces RMSE by 4–8 $\times$ relative to OLS when shape variation is present, and an adaptive AIC selector identifies when this upgrade is warranted with power ≥ 0.93 for $\alpha \geq 0.05$. Third, sign recovery is not a substitute for parameter accuracy: the profile achieves perfect sign recovery across all test cells, yet maintains a non-zero grid-bounded RMSE.

These results provide the theoretical infrastructure for subsequent empirical work. The BIS-calibrated demonstration confirms pipeline executability. The LLM-0 protocol specifies the infrastructure audit required before KA exponents can be estimated from language model data. The prospective universality question — whether financial and language model systems share KA exponent values — constitutes the primary open question for future work.

Competing Interests

The author declares no competing financial or non-financial interests.

Data Availability

The simulation benchmark script (`ka_s2_benchmark.py`), the generated benchmark data (`ka_s2_results.csv`), and the figure-generation code are available from the corresponding author upon reasonable request.

References

- [1] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. doi:10.1109/TAC.1974.1100705.
- [2] Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Springer, Berlin.
- [3] Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2), 251–276. doi:10.1162/089976698300017746.
- [4] Burnham, K. P. & Anderson, D. R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (2nd ed.). Springer, New York.
- [5] Ghorbani, B., Krishnan, S. & Xiao, Y. (2019). An investigation into neural net optimization via Hessian eigenvalue density. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2232–2241.
- [6] Hyvärinen, A., Karhunen, J. & Oja, E. (2001). *Independent Component Analysis*. Wiley, New York.
- [7] Karimov, H. & Alekberli, E. (2026). The Kerimov–Alekberli Model: An Information-Geometric Framework for Real-Time System Stability. arXiv preprint arXiv:2604.24083.
- [8] Karimov, H. & Alekberli, E. (2026). An Information-Geometric Framework for Stability Analysis of Large Language Models under Entropic Stress. arXiv preprint arXiv:2604.24076.
- [9] Lawley, D. N. & Maxwell, A. E. (1962). Factor analysis as a statistical method. *The Statistician*, 12(3), 209–229.
- [10] LeCun, Y., Bottou, L., Orr, G. B. & Müller, K.-R. (2012). Efficient backprop. In *Neural Networks: Tricks of the Trade* (2nd ed.), 9–48. Springer, Berlin.
- [11] Li, H., Xu, Z., Taylor, G., Studer, C. & Goldstein, T. (2018). Visualizing the loss landscape of neural nets. *Advances in Neural Information Processing Systems*, 31.
- [12] Lehmann, E. L. & Casella, G. (1998). *Theory of Point Estimation* (2nd ed.). Springer, New York.
- [13] Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7), 674–693. doi:10.1109/34.192463.
- [14] Martens, J. (2010). Deep learning via Hessian-free optimization. *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 735–742.

- [15] Nadarajah, S. (2005). A generalized normal distribution. *Journal of Applied Statistics*, 32(7), 685–694. doi:10.1080/02664760500079464.
- [16] Pappayan, V. (2018). The full spectrum of deep net Hessians at scale: dynamics with SGD training and sample size. *arXiv:1811.07062*.
- [17] Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39(3), 577–591. doi:10.2307/1913267.
- [18] Sharifi, K. & Leon-Garcia, A. (1995). Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5(1), 52–56. doi:10.1109/76.345996.
- [19] Varanasi, M. K. & Aazhang, B. (1989). Parametric generalized Gaussian density estimation. *Journal of the Acoustical Society of America*, 86(4), 1404–1415. doi:10.1121/1.398415.
- [20] Yao, Z., Gholami, A., Keutzer, K. & Mahoney, M. (2020). PyHessian: Neural networks through the lens of the Hessian. *2020 IEEE International Conference on Big Data (Big Data)*, 581–590. doi:10.1109/BigData50022.2020.9378018.

A S2 Operating Envelope

The S2 operating envelope is the minimum data requirement for reliable KA estimation:

$$T_{\text{eff}} \geq 100 \quad \text{AND} \quad n \geq 8. \quad (20)$$

Both conditions must hold simultaneously. This conjunction is strict: a large T with $n < 8$ does not satisfy S2 (insufficient spectral dimension); a large n with $T < 100$ does not satisfy S2 (insufficient temporal depth for convergence of the profile estimator).

B Profile Likelihood Implementation

The profile log-likelihood $\ell^P(\nu, \alpha)$ is evaluated as follows. At each (ν, α) grid node:

1. Compute $\hat{\beta}_{it} = a\lambda_{it}^\alpha$ for pre-specified a .
2. Solve $\partial\ell/\partial c = 0$ analytically to obtain $\hat{c}(\nu, \alpha)$.
3. Evaluate $\ell(\nu, \alpha, \hat{c}, \hat{a})$ using the GGD log-density formula.

All operations are vectorised over $n_{\text{rep}} = 100$ replications simultaneously using NumPy float32 arrays. Precomputed $\log \Lambda$ and $\log |Z|$ arrays reduce repeated log operations. Total benchmark runtime: approximately 20 minutes on a single CPU core.

C Profile Grid Refinement via Local Optimisation

The grid-argmax estimator satisfies $|\hat{\nu}^P - \nu^*| \leq \Delta\nu/2$ where ν^* denotes the true profile maximum and $\Delta\nu = 5/24$ is the grid step. A two-stage procedure is proposed to reduce this discretisation effect:

1. **Coarse stage.** Evaluate ℓ^P on the 25×8 grid and record the argmax $(\tilde{\nu}, \tilde{\alpha})$.
2. **Fine stage.** Initialise a gradient-based constrained optimiser (e.g. L-BFGS-B) at $(\tilde{\nu}, \tilde{\alpha})$ and minimise $-\ell^P(\nu, \alpha)$ subject to $|\nu - \tilde{\nu}| \leq \Delta\nu$ and $|\alpha - \tilde{\alpha}| \leq \Delta\alpha/2$.

The constrained box defines a local numerical refinement region around the coarse-grid solution. Its ability to reach the global profile maximum, and the resulting reduction in estimation error, require empirical verification and are not claimed in this paper. The continuous profile likelihood obeys the equivariance identity of Theorem 5.1 independently of grid discretisation. A numerical refinement implementation would require a separate finite-precision covariance check. The fine stage is not applied in the present benchmark; see Remark 8.1 for scope.

D Reproducibility

```
# Reproduce benchmark
python3 ka_s2_benchmark.py          # seed=0; ~20 min; produces ka_s2_results.csv

# Reproduce B2-W audit (BIS calibrated synthetic)
python3 ka_b2w_staged.py

# LLM-0 infrastructure audit (local execution required)
python3 ka_llm0_audit.py --model meta-llama/Llama-3.2-1B
```