Exponential Rarity of Multiple Likelihood Modes in Bivariate Seemingly Unrelated Regression

Alper Ferudun*

29 September 2026

Abstract

The likelihood of the crossed bivariate Gaussian seemingly unrelated regression model can have several local modes. Its almost-sure eventual uniqueness is known. We give an explicit finite-sample bound: for fixed noncollinear regressor vectors of correlation r, with $d = 1-|r|$ and $k = n-2 \geq 2$, the probability of more than one stationary point is at most $\exp\{-kd^3/[16(1+|r|)]\}+2\exp(-k/16)$. The bound is uniform over the slopes and all positive-definite error covariances. A determinant normal form gives a directly checkable unimodality certificate, which combines with elementary Gaussian tails. Common regressors are allowed by replacing k with $n-p-2$ after projection. For Gaussian random predictors of population correlation r_0, an explicit bound is $11\exp\{-(n-2)(1-|r_0|)^3/512\}$. A change-of-measure argument shows that the probability of multiple modes is $\exp\{-\Theta(n)\}$ at each fixed nonsingular random-design parameter. The constants are not claimed optimal. We also record the classical algebraic MANOVA certificate to distinguish the two parts of the motivating AIM question.

1 The question and quantitative statements

Consider the crossed bivariate SUR model

$$y_1 = \beta_1 x_1 + \varepsilon_1, \qquad y_2 = \beta_2 x_2 + \varepsilon_2, \qquad (\varepsilon_{i1}, \varepsilon_{i2}) \overset{\text{iid}}{\sim} N_2(0, \Sigma), \quad \Sigma \succ 0. \tag{1}$$

Here $x_1, x_2 \in \mathbb{R}^n$ are linearly independent, and both slopes and the entire error covariance are unrestricted parameters. A mode means a local maximum in this open parameter space, not a maximum restricted to an algorithmic search region. We count stationary points of the full likelihood, or equivalently of the covariance-profile likelihood.

Drton and Richardson [2] demonstrate multimodality and prove almost-sure eventual uniqueness for Gaussian random predictors. Question 11 of the AIM workshop list [1], recorded as AIM-COMPUTATION-0025 in [5], pairs an algebraic MANOVA question with Takemura's question about the speed of convergence to unimodality in bivariate SUR. We give explicit sampling-probability bounds for the crossed model (1).

Theorem 1 (Fixed design)**.** *Let $n \geq 4$ and define*

$$r = \frac{x_1^T x_2}{\|x_1\|\|x_2\|}, \qquad d = 1 - |r| > 0, \qquad k = n - 2.$$

If N_n is the number of stationary points of the likelihood in (1), then

$$\mathbb{P}\{N_n > 1\} \leq \min\left\{1, \exp\left(-\frac{kd^3}{16(1+|r|)}\right) + 2e^{-k/16}\right\}. \tag{2}$$

*Mercury Software GmbH. Email: alper@mercurycodelab.com. GitHub: AlperTheKing.

Off an event bounded by the right-hand side, there is exactly one stationary point, and it is a strict local and global maximum. The bound is uniform over β_1, β_2 and all $\Sigma \succ 0$. If the two equations also contain the same full-rank $n \times p$ design matrix Z, replace x_j by their projections onto $\mathrm{col}(Z)^\perp$ and k by $n-p-2$, requiring $k \geq 2$ and independence of the two projected regressors.

Theorem 2 (Gaussian random design)**.** *Suppose the predictor pairs are iid centered nonsingular bivariate Gaussian, independent of the errors, with population correlation r_0. Put $d_0 = 1 - |r_0|$. For $n \geq 4$,*

$$\mathbb{P}\{N_n > 1\} \leq \min\{1, 11e^{-(n-2)d_0^3/512}\}. \tag{3}$$

For each fixed nonsingular parameter of this model, the probability $\mathbb{P}(M_n)$ of at least two local modes satisfies

$$e^{-Cn} \leq \mathbb{P}(M_n) \leq 11e^{-(n-2)d_0^3/512} \tag{4}$$

for some finite C and all sufficiently large n.

Thus, for example, $n \geq 2 + 512 d_0^{-3} \log(11/\alpha)$, rounded upward and at least four, suffices for the upper failure probability α. These are conservative bounds, not estimates of the optimal exponent or the exact finite-sample probability. Arbitrary systems with several private regressors per equation, misspecification and non-Gaussian errors are not covered.

2 A determinant normal form and certificate

Write $X = [x_1, x_2]$ and $G = X^T X/n$. Fit the unrestricted regression of each response on both predictors:

$$y_1 = \alpha_1 x_1 + h x_2 + z_1, \qquad y_2 = j x_1 + \alpha_2 x_2 + z_2, \qquad X^T z_1 = X^T z_2 = 0.$$

Put $R = (z_i^T z_j/n)_{i,j=1}^2$ and assume $R \succ 0$. For candidate slopes let $a = \beta_1 - \alpha_1$ and $b = \beta_2 - \alpha_2$. Introduce

$$\begin{aligned} r &= G_{12}/\sqrt{G_{11}G_{22}}, & s &= R_{12}/\sqrt{R_{11}R_{22}}, \\ u &= h\sqrt{G_{22}/R_{11}}, & v &= j\sqrt{G_{11}/R_{22}}, \\ A &= a\sqrt{G_{11}/R_{11}}, & B &= b\sqrt{G_{22}/R_{22}}. \end{aligned}$$

Finally set

$$\tau = 1 - r^2, \quad \kappa = rs + \tau uv, \quad P = sv - ru, \quad Q = su - rv. \tag{5}$$

The residual covariance at (a, b) is $R + L^T G L$, where $L = \begin{bmatrix} -a & j \\ h & -b \end{bmatrix}$. Its determinant, divided by $R_{11}R_{22}$, is exactly

$$\begin{gathered} F(A, B) = \tau A^2 B^2 + A^2 + B^2 - 2\kappa AB + 2PA + 2QB + C, \\ C = 1 - s^2 + u^2 + v^2 - 2rsuv + \tau u^2 v^2. \end{gathered} \tag{6}$$

Indeed, expanding $\det(R + L^T G L)$ gives $\det R$, the terms quadratic in L and linear in R, and $\det G\,(ab - hj)^2$. The translation by the unrestricted own-regressor coefficients has removed both cubic terms. This is an identity for all real A, B.

Lemma 3 (A stationary-point certificate)**.** *For (6), suppose $\tau \geq 0$, $\mu = 1 - |\kappa| > 0$, and*

$$\tau(P^2 + Q^2) < 2\mu^3. \tag{7}$$

Then F has exactly one stationary point, its strict global minimum.

Proof. Let $w = (A, B)$ and $t = (P, Q)$. At a stationary point,

$$0 = \tfrac{1}{2} w \cdot \nabla F = 2\tau A^2 B^2 + A^2 + B^2 - 2\kappa AB + t \cdot w.$$

The quadratic form is at least $\mu \|w\|^2$, so every stationary point lies in the closed ball of radius $\|t\|/\mu$. The quartic Hessian is

$$2\tau \begin{pmatrix} B^2 & 2AB \\ 2AB & A^2 \end{pmatrix}.$$

Adding $|AB|I$ inside the brackets gives a positive semidefinite matrix: its determinant is $|AB|(A^2 + B^2) - 2A^2B^2 \geq 0$ and its diagonal is nonnegative. Therefore

$$\nabla^2 F(w) \succeq (2\mu - \tau \|w\|^2) I.$$

By (7) this is positive definite throughout the ball containing all stationary points. The nonnegative quartic and positive-definite quadratic make F coercive, so a global minimum exists. Strict convexity on the ball excludes a second stationary point and makes the minimum strict. □

Corollary 4. *In the SUR normal form, with $d = 1 - |r|$, a sufficient condition is*

$$u^2 + v^2 \leq \frac{d^3}{4\tau}. \tag{8}$$

Proof. Since $|s| < 1$ and $|uv| \leq (u^2 + v^2)/2$,

$$\mu \geq d - \frac{\tau}{2}(u^2 + v^2) \geq d - d^3/8 \geq 7d/8.$$

Also $P^2 + Q^2 \leq (|r| + |s|)^2 (u^2 + v^2) \leq 4(u^2 + v^2)$. The left side of (7) is at most d^3, while its right side is at least $(343/256)d^3 > d^3$. □

Remark 5. Global convexity is neither used nor generally true. When $r = s = u = v = 0$, $F = (1 + A^2)(1 + B^2)$ has one stationary point, but its Hessian at $A = B = 2$ has eigenvalue -6. The certificate confines all stationary points to a region where strict convexity does hold.

For fixed slopes, write the normalized residual covariance as $S(\beta)$. It is positive definite if $R \succ 0$. In the precision coordinate $\Omega = \Sigma^{-1}$ the covariance log likelihood is

$$\tfrac{n}{2} \log\det \Omega - \tfrac{n}{2} \operatorname{tr}(\Omega S(\beta)).$$

This is strictly concave and has its only stationary point at $\Omega = S(\beta)^{-1}$. Hence the profile is a constant minus $(n/2)\log\det S(\beta)$. Full and profiled stationary points correspond bijectively. The strict global determinant minimum certified above gives a strict global likelihood maximum, with no other full-space stationary point.

3 Proof of the fixed-design bound

Invertible predictor and response scalings preserve stationary-point counts, so assume $G_{11} = G_{22} = \Sigma_{11} = \Sigma_{22} = 1$. Let $\rho = \Sigma_{12}$ be the true error correlation. Under the correctly specified model the unrestricted cross coefficients satisfy

$$H = \begin{pmatrix} h \\ j \end{pmatrix} \sim N_2\left(0, \frac{1}{n\tau}\begin{pmatrix} 1 & -r\rho \\ -r\rho & 1 \end{pmatrix}\right). \tag{9}$$

This follows directly from the covariance of unrestricted least squares, $\mathrm{Cov}(\widehat{B}_{ij}, \widehat{B}_{k\ell}) = \Sigma_{j\ell}(G^{-1})_{ik}/n$. The largest eigenvalue in (9) is at most $(1+|r|)/(n\tau)$. Coupling H to a standard two-dimensional Gaussian vector gives

$$h^2 + j^2 \le \frac{1+|r|}{n\tau} W, \qquad W \sim \chi_2^2. \tag{10}$$

Each nR_{ii} has marginal distribution χ_k^2, $k = n-2$. No independence of the two diagonal entries is required. The elementary bound

$$\mathbb{P}\{\chi_k^2 < k/2\} \le e^{-k/16} \tag{11}$$

follows by applying Markov's inequality to $e^{-\chi_k^2/2}$: $e^{k/4}2^{-k/2} \le e^{-k/16}$. Except on an event of probability at most $2e^{-k/16}$, both $R_{ii} \ge k/(2n)$. On this event (10) gives

$$u^2 + v^2 \le \frac{2n}{k}(h^2+j^2) \le \frac{2(1+|r|)}{k\tau} W.$$

Thus Corollary 4 applies if $W \le kd^3/[8(1+|r|)]$. Since $\mathbb{P}\{\chi_2^2 > t\} = e^{-t/2}$, a union bound proves (2). Independence between the coupled Gaussian event and the residual event is not needed. For $k \ge 2$, $R \succ 0$ almost surely.

For common regressors Z, their coefficient score is $Z^T\text{residual} = 0$ for every positive-definite covariance. Profiling these coefficients is therefore exact orthogonal projection, not an approximation. An orthonormal basis of $\mathrm{col}(Z)^\perp$ leaves $n-p$ independent Gaussian error pairs. The unrestricted residual rank becomes $k = n-p-2$. Keeping divisor n for the covariance estimates leaves every preceding calculation unchanged. This proves the common regressor assertion of Theorem 1.

4 Gaussian random predictors

Standardize the predictor variances to one and put $\eta = d_0/4$. The four scalar Gaussian directions

$$X_1, \quad X_2, \quad (X_1+X_2)/\sqrt{2}, \quad (X_1-X_2)/\sqrt{2}$$

have variance at most two. For a centered Gaussian variable U of variance $v \le 2$ and $0 < \eta \le 1$,

$$\mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^n U_i^2 - v\right| > \eta\right\} \le 2e^{-n\eta^2/32}. \tag{12}$$

For completeness the usual chi-square Chernoff estimate is

$$\mathbb{P}\{|\chi_n^2/n - 1| > x\} \le 2\exp\{-n\min(x^2,x)/8\}.$$

For the upper tail use $x - \log(1+x) \ge \min(x^2,x)/4$; for the lower tail use $-x-\log(1-x) \ge x^2/2$ when $0<x<1$. Since $\min((\eta/v)^2, \eta/v) \ge \eta^2/4$, (12) follows.

With probability at least $1 - 8e^{-nd_0^2/512}$ all four empirical second moments differ from their expectations by at most η. Their differences show that every entry of the two-dimensional predictor sample covariance has error at most η. Its correlation therefore obeys

$$|r| \le \frac{|r_0|+\eta}{1-\eta}, \qquad 1-|r| \ge d_0/2.$$

Conditioning on the predictors, Theorem 1 yields

$$\mathbb{P}\{N_n > 1\} \le 8e^{-nd_0^2/512} + e^{-(n-2)d_0^3/256} + 2e^{-(n-2)/16}. \tag{13}$$

Using $0 < d_0 \le 1$ and comparing exponents gives (3).

5 A lower bound on the exponential scale

We prove the second assertion of Theorem 2 without assuming a large-deviation theorem. There is a nonempty open set of positive-definite empirical four-variable covariance matrices producing two strict modes. For an explicit point in this set, take

$$G = R = I_2, \qquad B_* = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}, \qquad V_* = \begin{pmatrix} I_2 & B_* \\ B_*^T & B_*^T B_* + I_2 \end{pmatrix} \succ 0.$$

The normal form is

$$F(A, B) = A^2 B^2 + A^2 + B^2 - 8AB + 25.$$

At $(\sqrt{3}, \sqrt{3})$ and $(-\sqrt{3}, -\sqrt{3})$ its Hessian has eigenvalues 4 and 12, while the origin is a saddle. These strict minima persist under small covariance perturbations by the implicit function theorem; the residual Schur complement remains positive. Consequently both give strict full-likelihood modes on an open neighborhood of V_*.

Let P be the true centered nonsingular Gaussian law of (X_1, X_2, Y_1, Y_2), and let $Q = N_4(0, V_*)$. Both densities are positive everywhere. Under Q^n, the empirical second-moment matrix enters the preceding neighborhood with probability tending to one. Also

$$\frac{1}{n} \log \frac{dQ^n}{dP^n} \longrightarrow D(Q\|P) < \infty \quad Q\text{-almost surely,}$$

by the law of large numbers. For all sufficiently large n, the intersection of the multimodal event and the event that this ratio is at most $D(Q\|P) + 1$ has Q^n-probability at least 1/2. Changing measure gives

$$P^n(M_n) \geq \tfrac{1}{2} e^{-n(D(Q\|P)+1)}.$$

This proves (4), for example with $C = D(Q\|P) + 2$. It does not specify the optimal exponent or a finite threshold for this lower bound, and it is not a lower bound for arbitrary fixed design sequences.

6 The companion MANOVA question

The following familiar argument records the algebraic part of the source question and is not claimed as a new MANOVA result. In a common-design model $Y = XB + E$, with X full rank, let $\widehat{B}$ be least squares and $S_0 = (Y - X\widehat{B})^T (Y - X\widehat{B})$. Then

$$S(B) = S_0 + (B - \widehat{B})^T X^T X (B - \widehat{B}).$$

If $S_0 \succ 0$, whitening yields

$$\frac{\det S(B)}{\det S_0} = \det(I + C^T C), \qquad C = (X^T X)^{1/2} (B - \widehat{B}) S_0^{-1/2}.$$

By Cauchy–Binet this is one plus the sum of squares of all nonempty square minors of C. The coefficient score for any fixed covariance forces $B = \widehat{B}$, and profiling gives $\widehat{\Sigma} = S_0/n$. There is one stationary point and mode. If S_0 is singular, the likelihood is unbounded above by shrinking a covariance eigenvalue in a zero-residual direction. For nonsingular Gaussian errors, $S_0 \succ 0$ almost surely exactly when $n - \operatorname{rank}(X)$ is at least the number of responses.

7 Verification, relation to prior work, and limitations

The proof is analytic and applies to all parameters satisfying its hypotheses. A standard-library exact-arithmetic checker supplements it: 3,969 rational (r, s, u, v) cases verify the determinant identity and the eliminated quintic by coefficient comparison and count its real roots by Sturm division. All 869 cases satisfying Lemma 3 have one real root; all 105 cases satisfying (8) satisfy that lemma. Another 500 cases check nonunit normalizations. Together with boundary and constant checks this gives 16,440 assertions per run, with identical mathematical reports under Python 3.9.6 and 3.13.5. These finite checks are not the general proof.

To reproduce the elimination, set $A = (\kappa B - P)/(1 + \tau B^2)$ in the second score equation. Its numerator is

$$\tau^2 B^5 + \tau^2 Q B^4 + 2\tau B^3 + \tau(2Q - \kappa P)B^2 \\ + (1 - \kappa^2 + \tau P^2)B + (Q + \kappa P).$$

The denominator is always positive. The coarse parameter grid contains one- and three-real-root examples, not five-real-root examples; it makes no claim to exclude the latter, which are established in [2].

The new quantitative argument is distinct from the qualitative limiting root argument of [2]. Sundberg [3] relates flatness and multiple modes to lack of fit in curved exponential families, explicitly without deriving sampling-probability statements. Related high-probability likelihood geometry for linear Gaussian covariance models is studied in [4]; the crossed mean-restriction model here has a different parameter space. The present bounds use classical least squares, covariance profiling, Gaussian tails and change of measure. We assert neither that these methods are new nor that a bounded literature search certifies absolute priority.

For the standard correctly specified Gaussian model of the motivating question, the probability has exponential order, with an explicit finite-sample upper bound. Exact probabilities, optimal exponents, larger unequal-regressor systems and non-Gaussian models remain beyond the results proved here. The source's MANOVA statement is classical and is included only to make the scope clear.

This unrefereed manuscript was developed with AI assistance for derivation, literature retrieval, implementation and exposition. The author is responsible for its claims. The retained originating researcher self-audit is not independent review or formal verification.

References

[1] American Institute of Mathematics, *Computational Algebraic Statistics*, workshop problem list, version 27 March 2004, Question 11, p. 6. https://aimath.org/WWN/compalgstat/compalgstat.pdf.

[2] M. Drton and T. S. Richardson, Multimodality of the likelihood in the bivariate seemingly unrelated regressions model, *Biometrika* **91** (2004), 383–392. doi:10.1093/biomet/91.2.383. Public precursor: University of Washington Technical Report 410 (2002), Theorems 2 and 3. https://stat.uw.edu/sites/default/files/files/reports/2002/tr410.pdf.

[3] R. Sundberg, Flat and multimodal likelihoods and model lack of fit in curved exponential families, *Scand. J. Stat.* **37** (2010), 632–643. doi:10.1111/j.1467-9469.2010.00703.x. Public precursor: Stockholm University Research Report 2009:1. https://www2.math.su.se/matstat/reports/seriea/2009/rep1/report.pdf.

[4] P. Zwiernik, C. Uhler and D. Richards, Maximum likelihood estimation for linear Gaussian covariance models, *J. R. Stat. Soc. B* **79** (2017), 1269–1292. https://arxiv.org/abs/1408.5604.

[5] ulamai, *UnsolvedMath*, version 1.6.0, record AIM-COMPUTATION-0025, frozen snapshot c6e7f41b4eb6. Hugging Face dataset. The frozen research aid supplies the classical MANOVA certificate and identifies the missing explicit SUR rate; it does not prove the rate theorems of this paper.