

Choosing PCA Components in High Dimensions: The Calibration–Power Trade-off Between Parallel Analysis and Spectral Thresholding

Heberto G. Salinas

Abstract

When the number of variables is not small relative to the number of observations, what false-positive rate and detection power result from parallel analysis and Marchenko–Pastur spectral thresholding? The comparison uses synthetic Gaussian covariance models where the latent structure is known exactly. The evidence comes from 13,500 pure-noise datasets across 27 sample-size and aspect-ratio cells and 673,000 single-spike datasets across the same 27 panels and 18 signal strengths, 686,500 simulated datasets in all. Under pure Gaussian noise, parallel analysis retained at least one component with mean probability 0.049 across the 27 cells, numerically close to 0.05 under the frozen finite- B threshold profiles. The known-variance MP edge retained at least one component with mean probability 0.138, and the estimated-variance MP rule, retained as a sensitivity analysis, 0.176. The higher false-retention rate bought greater finite-sample detection power: at signal strength $\gamma = 1.5$, parallel analysis detected the planted component with equal-cell mean probability 0.702, compared with 0.810 for the known-variance MP edge and 0.773 for the estimated-variance rule. The asymptotic boundary at $\gamma = 1$ was not a finite-sample switch; detection increased gradually, and near-certain detection required signals substantially beyond the theoretical boundary. Neither method dominated: parallel analysis was the more conservative choice, and the raw MP edge was more sensitive, but its edge is not a 5% significance threshold and should not be interpreted as one. These conclusions are limited to centered covariance PCA under Gaussian identity noise with one planted spike.

1 The component-retention decision

1.1 Why the decision matters

PCA always returns ordered sample components, including when the data contain no low-rank signal. The practical question is therefore not whether components can be computed. It is which observed eigenvalues should be interpreted as evidence of structure rather than finite-sample noise.

Two common approaches answer this question differently:

- **Parallel analysis** compares observed eigenvalues with a reference spectrum generated from random data of the same shape. Horn’s original formulation used mean null eigenvalues [5]. Later work developed quantile-based versions [3], including the 95th-percentile rule used here [4].
- **Spectral edge thresholding** retains eigenvalues above the upper edge predicted by the Marchenko–Pastur law for a high-dimensional noise spectrum [7].

Parallel analysis is calibrated through an explicit null simulation. The MP edge is a limiting spectral boundary, not a tail-probability cutoff. The experiment tests the consequence of that distinction.

1.2 Scope

The report evaluates detection of zero or one latent component. It does not compare all common PCA rules, evaluate component compression, or analyze real-data interpretability. Cumulative explained variance, Kaiser’s rule, scree procedures, cross-validation, multi-spike recovery, assumption violations, and drilling applications are outside the manuscript’s active scope.

2 Minimal mathematical model

2.1 Pure noise

Let X contain n independent Gaussian observations of p variables. Under the null model, the population covariance is the identity matrix. Columns are centered, and the sample covariance is

$$S = \frac{X_c^\top X_c}{n-1}. \quad (1)$$

When p/n is not negligible, the sample eigenvalues spread even though every population eigenvalue equals one. The limiting spectrum and its support are described by the Marchenko–Pastur law [7]. This high-dimensional spread is the reason a large observed eigenvalue need not imply a latent factor.

2.2 One planted component

The signal model has population covariance

$$\Sigma = I_p + (\lambda - 1)vv^\top, \quad (2)$$

where v is a unit direction. Signal strength is parameterized as

$$\lambda = 1 + \gamma\sqrt{c}, \quad c = p/n. \quad (3)$$

For this parameterization, the asymptotic separation boundary occurs at $\gamma = 1$, equivalent to the population-spike threshold $\lambda = 1 + \sqrt{c}$ in the spiked covariance model for positive limiting aspect ratio [1]. Paul gives the corresponding real Gaussian eigenstructure results in the $c < 1$ orientation [8]. This belongs to the broader phase-transition phenomenon associated with Baik, Ben Arous, and P       [2]. The boundary predicts limiting spectral separation, not the finite-sample probability that a retention rule selects the component.

2.3 Methods

Parallel analysis. For every (n, p) shape, the implementation builds a seeded Gaussian null profile with $B = 1000$ draws. Leading observed eigenvalues are compared sequentially with their position-matched 95th-percentile null thresholds, following the quantile-based extension of parallel analysis [4, 3].

MP-known. The known-variance rule retains sample eigenvalues above the finite-shape form of the Marchenko–Pastur upper edge [7],

$$\left(1 + \sqrt{p/(n-1)}\right)^2. \quad (4)$$

The simulated noise variance is exactly one.

MP-estimated. The sensitivity rule multiplies the same edge by a noise-variance estimate derived from the median of the nonzero empirical spectrum and the corresponding bulk-conditional MP median.

3 Controlled computational design

3.1 Parameter grid

The experiments cross sample sizes $n = 100, 250, 500, 1000$ with allowed aspect ratios from $c = 0.05$ through $c = 5$, producing 27 (n, c) panels. Integer p uses round-half-up. The single-spike experiment evaluates 18 values of γ from 0.25 to 8, with dense coverage near the boundary.

3.2 Replication and aggregation

The pure-noise experiment uses 500 replications in every cell. The single-spike experiment begins with 250 replications and uses a predeclared precision-only top-up for selected transition cells, up to 2,500 replications.

Every cross-cell mean in this report gives equal weight to each (n, c) panel. Replication-weighted pooling is avoided because Stage 3 replication counts depend on the precision of the MP-known detection estimate and are therefore endogenous to that monitored outcome.

Monte Carlo uncertainty is propagated from the cell-level binomial rates through each equal-cell mean. The reported MCSE is conditional on the fixed design grid; it quantifies simulation noise, not uncertainty over a population of possible (n, c) conditions. For PA it is also conditional on the one frozen $B = 1000$ null profile per shape and excludes uncertainty from estimating those 95th-percentile thresholds. Because each profile is shared across replications of its shape, that omitted component is not reduced by adding Stage 2 or Stage 3 datasets. At the three headline signal strengths ($\gamma = 1, 1.5, 2$), all reported conditional plug-in 95% Monte Carlo half-widths are below 0.010.

3.3 Reproducibility

The completed artifacts contain deterministic seed identities, cell-level checkpoints, method outputs, profile hashes, configuration hashes, and source manifests. Report tables and figures are regenerated from the immutable Stage 2 and Stage 3 Parquet artifacts by `report/build_assets.py`.

4 Null calibration

Figure 1 shows the cell-level false-retention probabilities. Under pure noise, parallel analysis is centered near 0.05 for the frozen finite- B profiles. The two MP rules retain false components

more often. The distinction is structural: the MP upper edge describes the limiting edge of the bulk spectrum; it is not a 95th-percentile bound for the largest sample eigenvalue. Under real Gaussian identity covariance, the centered and scaled largest eigenvalue has a nondegenerate Tracy–Widom limit rather than collapsing deterministically below the edge [6].

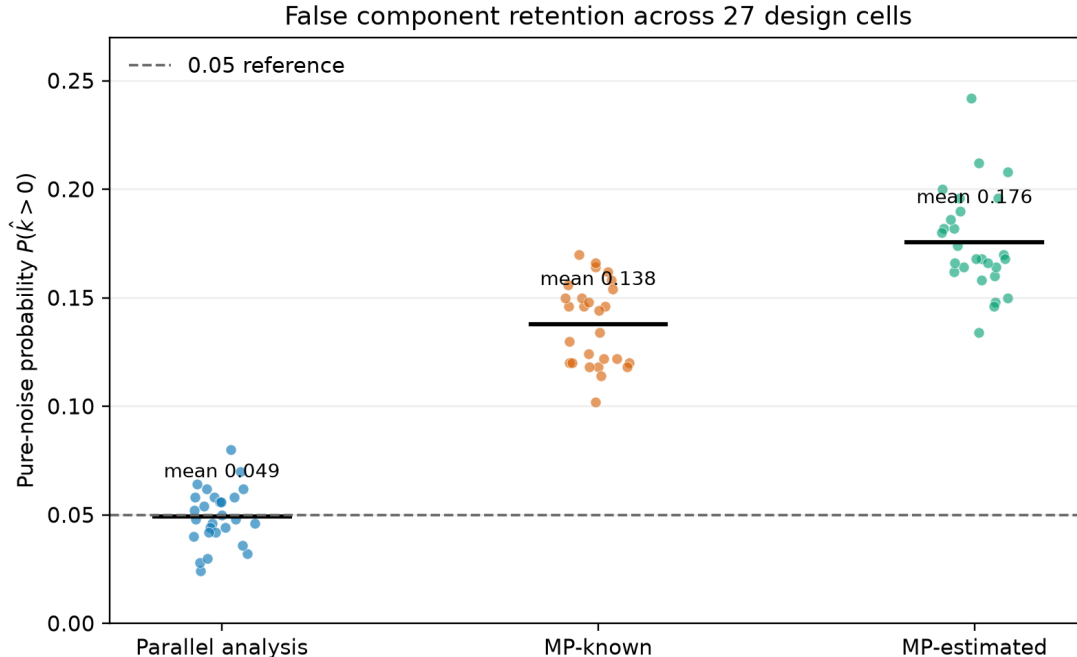


Figure 1: Cell-level false-retention probabilities $P(\hat{k} > 0)$ under pure Gaussian noise for each of the 27 (n, c) design cells. Horizontal bars mark the equal-cell mean for each rule; the dashed line marks the 0.05 reference.

Table 1: Pure-noise false-retention summary over 27 design cells (13,500 datasets).

Rule	Equal-cell mean	Conditional plug-in 95% MC half-width	Cell range
Parallel analysis	0.049	0.004	0.024–0.080
MP-known	0.138	0.006	0.102–0.170
MP-estimated	0.176	0.006	0.134–0.242

The cell range is descriptive heterogeneity over the 27 tested designs, not an uncertainty interval. For PA, the listed half-width excludes finite- B threshold-profile uncertainty and must not be read as total uncertainty about nominal calibration. The machine-readable source is `report/generated/null_false_retention_summary.csv`.

5 Detection through the asymptotic boundary

Figure 2 shows equal-cell detection against signal strength. Detection rises continuously rather than jumping from zero to one at $\gamma = 1$. MP-known is more sensitive throughout the transition

region, while parallel analysis retains its more conservative null behavior. Table 2 lists the three headline values.

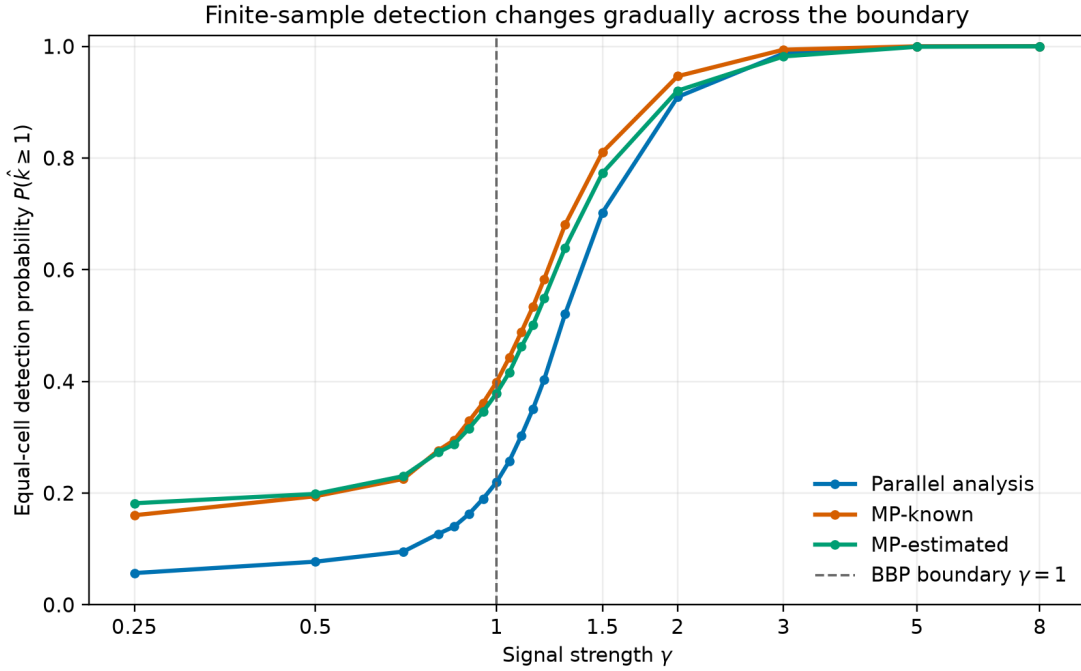


Figure 2: Equal-cell detection probability $P(\hat{k} \geq 1)$ against signal strength γ for the three rules, averaged with equal weight over the 27 (n, c) panels. The dashed line marks the asymptotic separation boundary $\gamma = 1$.

The maps in Figure 3 average equally over the tested sample sizes within each (c, γ) cell. They are descriptive summaries of the frozen grid, not claims for untested dimensions.

Table 2: Equal-cell mean detection probability at the three headline signal strengths (27 panels per value of γ).

Signal strength	Parallel analysis	MP-known	MP-estimated
$\gamma = 1.0$	0.220	0.399	0.379
$\gamma = 1.5$	0.702	0.810	0.773
$\gamma = 2.0$	0.909	0.947	0.921

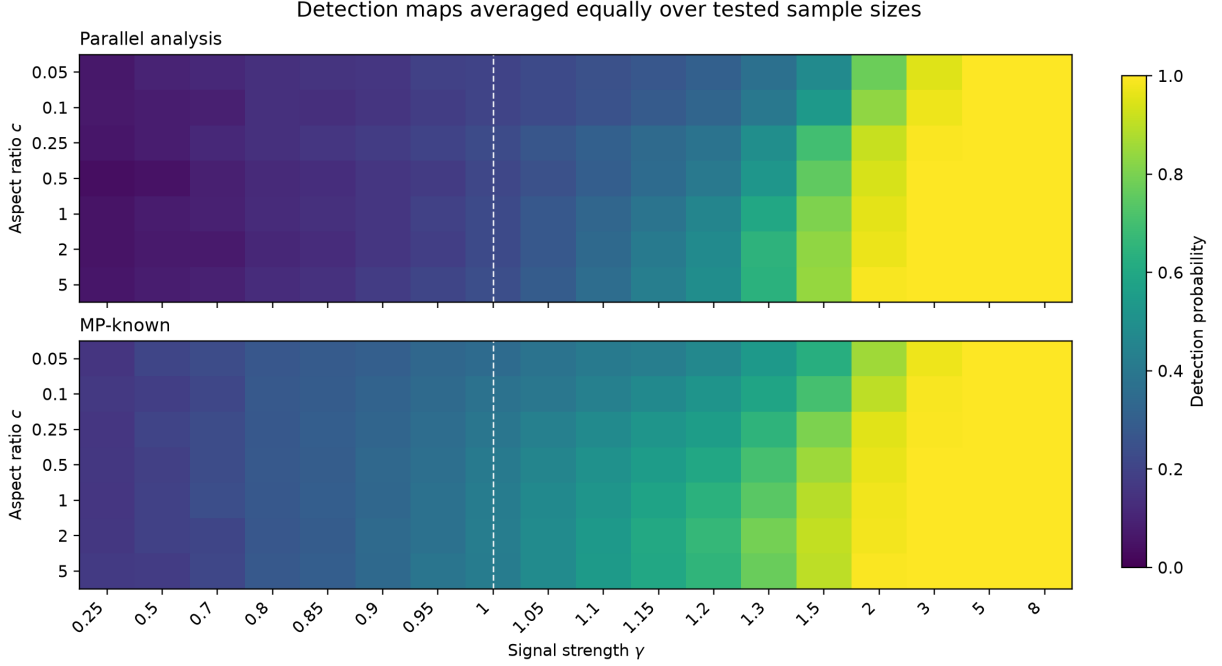


Figure 3: Detection probability over (c, γ) for parallel analysis (top) and MP-known (bottom), averaged with equal weight over the tested sample sizes within each cell. The dashed line marks $\gamma = 1$.

6 The calibration–power trade-off

The central comparison is not “correct method versus incorrect method.” It is a trade-off:

- PA more effectively controls false retained components under the matched Gaussian null.
- MP-known detects weaker planted structure more often.
- MP-estimated does not simply interpolate between the two; estimating the noise scale changes the threshold randomly and requires separate interpretation.

A practitioner should therefore state the error preference before selecting the rule. If false components are especially costly, PA is favored by these experiments. If missing a weak component is more costly and additional false retention can be investigated downstream, MP-known offers greater sensitivity.

7 Limits

1. The primary generator is Gaussian with identity noise.
2. There is at most one planted component in the report evidence.
3. The study uses centered covariance PCA, not standardized correlation PCA.
4. The tested dimensions follow a fixed grid and do not justify universal thresholds.

5. No real dataset provides known latent rank, so no drilling or other domain example is used as correctness evidence.
6. The study compares frozen implementations, including a seeded finite- B parallel-analysis profile. PA uncertainty summaries condition on that profile and do not quantify threshold-estimation uncertainty.
7. The report does not establish a mechanism for findings from the excluded multi-spike study.

8 Reproducibility and source record

The active evidence base is:

- `results/stage2/stage2.parquet`
- `results/stage2/stage2_gate_decision.json`
- `results/stage3/stage3.parquet`
- `results/stage3/stage3_gate_decision.json`
- `STAGE2_MEMO.md`
- `STAGE3_MEMO.md`

The report-only generated manifest is written to `report/generated/report_summary.json`. Primary theory and method citations are listed below. Implementation-specific claims trace to the frozen project artifacts rather than to external literature.

References

- [1] J. Baik and J. W. Silverstein, Eigenvalues of large sample covariance matrices of spiked population models, *Journal of Multivariate Analysis* **97** (2006), no. 6, 1382–1408, [doi:10.1016/j.jmva.2005.08.003](https://doi.org/10.1016/j.jmva.2005.08.003).
- [2] J. Baik, G. Ben Arous, and S. Péché, Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices, *Annals of Probability* **33** (2005), no. 5, 1643–1697, [doi:10.1214/009117905000000233](https://doi.org/10.1214/009117905000000233).
- [3] A. Buja and N. Eyuboglu, Remarks on parallel analysis, *Multivariate Behavioral Research* **27** (1992), no. 4, 509–540, [doi:10.1207/s15327906mbr2704_2](https://doi.org/10.1207/s15327906mbr2704_2).
- [4] L. W. Glorfeld, An improvement on Horn parallel analysis methodology, *Educational and Psychological Measurement* **55** (1995), no. 3, [doi:10.1177/0013164495055003002](https://doi.org/10.1177/0013164495055003002).
- [5] J. L. Horn, A rationale and test for the number of factors in factor analysis, *Psychometrika* **30** (1965), 179–185, [doi:10.1007/BF02289447](https://doi.org/10.1007/BF02289447).
- [6] I. M. Johnstone, On the distribution of the largest eigenvalue in principal components analysis, *Annals of Statistics* **29** (2001), no. 2, 295–327, [doi:10.1214/aos/1009210544](https://doi.org/10.1214/aos/1009210544).

- [7] V. A. Marchenko and L. A. Pastur, Distribution of eigenvalues for some sets of random matrices, *Mathematics of the USSR-Sbornik* **1** (1967), no. 4, 457–483, [doi:10.1070/SM1967v001n04ABEH001994](https://doi.org/10.1070/SM1967v001n04ABEH001994).
- [8] D. Paul, Asymptotics of sample eigenstructure for a large dimensional spiked covariance model, *Statistica Sinica* **17** (2007), 1617–1642, <https://www3.stat.sinica.edu.tw/statistica/oldpdf/A17n418.pdf>.