Improving Cy's Beta Estimator

Paul Gutkovich and Zack West

1 Introduction

A cornerstone of Random Matrix Theory is the β-Hermite ensemble, which is defined by the parameter β, known as the Dyson index. This index dictates the degree of eigenvalue repulsion: $\beta = 1$ corresponds to the Gaussian Orthogonal Ensemble, $\beta = 2$ to the Gaussian Unitary Ensemble, and $\beta = 4$ to the Gaussian Symplectic Ensemble.

The local correlations between eigenvalues, particularly the distribution of nearest-neighbor spacings, are universal in the bulk of the spectrum and depend only on β. For a large $N \times N$ matrix with ordered eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_N$, the raw spacings are defined as $\delta_i = \lambda_{i+1} - \lambda_i$. To meaningfully analyze these local fluctuations, the data must first be unfolded. Unfolding is the crucial preprocessing step that removes the large-scale variations imposed by the global spectral density $\rho(\lambda)$ (e.g., the Wigner Semicircle Law for Hermite ensembles). The unfolded spacings s_i are then defined such that their mean is normalized to unity across the spectrum: $s_i \approx \rho(\lambda_i) \cdot \delta_i$.

The probability density function (PDF) for these unfolded spacings, $P(s)$, is captured with remarkable accuracy by the Wigner Surmise for general β:

$$P_\beta(s) = A_\beta s^\beta e^{-B_\beta s^2}, \quad A_\beta = 2\frac{\Gamma(\frac{\beta+2}{2})^{\beta+1}}{\Gamma(\frac{\beta+1}{2})^{\beta+2}},\ B_\beta = \frac{\Gamma(\frac{\beta+2}{2})^2}{\Gamma(\frac{\beta+1}{2})^2}.$$

The central objective of this work is the Maximum Likelihood Estimation of the unknown index β from a given sequence of raw eigenvalue spacings.

We focus on numerically comparing the performance and inherent limitations of three primary estimation strategies and their associated preprocessing techniques:

1. Wigner Surmise-based MLE: This approach applies the $P_\beta(s)$ distribution to the unfolded spacings. It necessitates the use of a preprocessing technique, such as local normalization, to approximate the density function $\rho(\lambda)$. We also use a Generalized Gamma distribution introduced in [2], which approaches the Wigner surmise for large β, but is more accurate for smaller β.

2. Gap Ratio-based MLE: This technique utilizes the distribution of the consecutive spacing ratios $r_i = \frac{\delta_{i+1}}{\delta_i}$. Crucially, this method eliminates the need for explicit unfolding because the local density $\rho(\lambda)$ cancels out in the ratio. Wigner surmise-like approximations for the probability of this ratio were introduced in [1].

3. Empirical Data-based MLE: We also construct MLEs based on empirical data, for both locally-normalized spacings and gap ratios.

We also investigate the impact of critical data cleaning steps, particularly the necessity of eliminating spacings from the spectral edges, where universality breaks down. We also created a website in which one can use these various estimators to find the Dyson index given a sequence of spacings, as a modernized version of Cy Chan's Beta Estimator. The code for our website can be found here.

2 Data Preprocessing

2.1 Spectrum Unfolding

The estimation methods that don't utilize the gap ratios are derived for unfolded eigenvalue spacings, which have a mean spacing of unity across the spectrum. The formal definition of spectrum unfolding is via the cumulative distribution function of the spectral density:

$$\lambda_k^{(u)} = F_{N,\beta}(\lambda_k) = N \int_{-\infty}^{\lambda_k} p_{\beta,N}(\lambda) d\lambda \approx N \int_{-\infty}^{\lambda_k} \frac{1}{\pi\beta N} \sqrt{2\beta N - \lambda^2} d\lambda,$$

where $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$ is the original spectrum, and we have $p_{\beta,N}(\lambda) \approx \frac{1}{\pi\beta N}\sqrt{2\beta N - \lambda^2}$ from the Wigner Semicircle Law. This unfolded spectrum has a uniform density of states, giving $\mathbb{E}[\lambda_{k+1}^{(u)} - \lambda_k^{(u)}] = 1$.

However, the raw spacings $\delta_i = \lambda_{i+1} - \lambda_i$ obtained directly from a random matrix do not satisfy this property. The global spectral density $\rho(\lambda)$ varies across the spectrum, leading to non-uniform mean spacings. Applying the maximum likelihood estimators directly to raw spacings would yield biased and unreliable estimates of β.

To address this issue, we employ a local normalization procedure to approximate the unfolding transformation. For each raw spacing δ_i, we compute the mean spacing within a sliding window centered at that position:

$$\bar{\delta}_i = \frac{1}{2w+1} \sum_{j=i-w}^{i+w} \delta_j$$

where w is the window half-width. The approximate unfolded spacing is then obtained by defining $s_i = \delta_i / \bar{\delta}_i$. This local normalization ensures that the mean spacing is approximately unity in each local region, effectively removing the large-scale variations imposed by $\rho(\lambda)$. The choice of window size w represents a trade-off: smaller windows better capture local density variations but are more susceptible to statistical fluctuations, while larger windows provide smoother estimates but may fail to adapt to rapid changes in density.

2.2 Edge Removal

Another preprocessing step is the removal of spacings near the spectral edges. The universal predictions of random matrix theory, including the Wigner surmise, are valid only in the bulk of the spectrum. Near the edges, eigenvalue statistics exhibit fundamentally different behavior that does not depend solely on β and is not captured by the distributions used in our estimators.

We remove a fixed number of spacings from both ends of the spectrum. Specifically, we discard the first and last 10% of spacings, though this fraction can be adjusted. The combination of local normalization and edge removal transforms the raw eigenvalue spacings into a dataset suitable for maximum likelihood estimation.

2.3 Gap Ratios

Another way to eliminate the large-scale variations imposed by $\rho(\lambda)$ is to consider the ratio of consecutive spacings $r_i = \frac{\delta_{i+1}}{\delta_i}$ as the local density cancels out in the ratio. For the spacing ratios, we also eliminate data from the edges as it doesn't follow the same statistics as the bulk.

3 Maximum Likelihood Estimators

All of our maximum likelihood estimators are defined by maximizing the sum of the log-likelihoods of various distributions for either the unfolded spacing data or the gap ratio data. The estimator uses the assumption that β is in $\{0.0, 0.1, \ldots, 10.0\}$.

3.1 Wigner Surmise and Generalized Gamma MLE

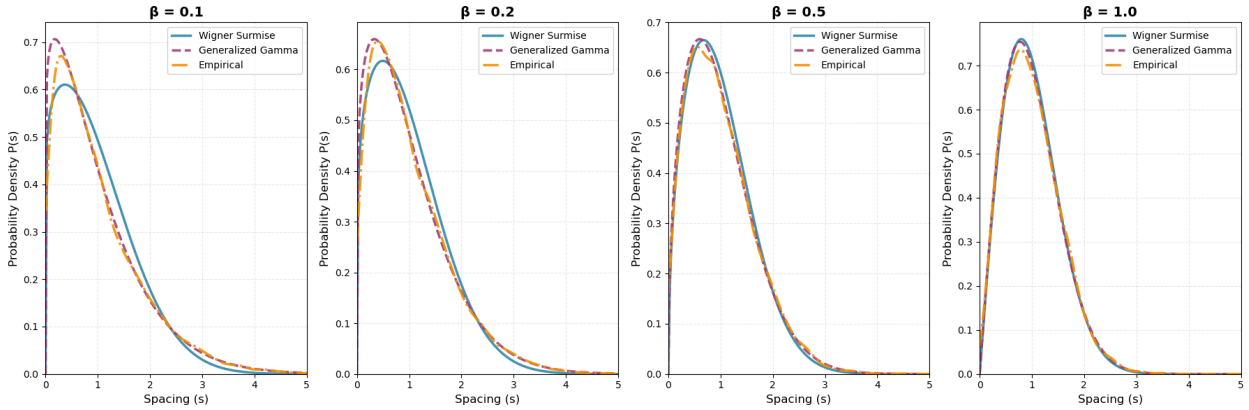

Figure 1: Comparison of Wigner and GG Distributions for Various Beta

The first likelihood function that we implemented was the Wigner Surmise, and a similar Generalized Gamma distribution. The Wigner Surmise is given by

$$P_\beta(s) = A_\beta s^\beta e^{-B_\beta s^2}, \quad A_\beta = 2\frac{\Gamma(\frac{\beta+2}{2})^{\beta+1}}{\Gamma(\frac{\beta+1}{2})^{\beta+2}}, \; B_\beta = \frac{\Gamma(\frac{\beta+2}{2})^2}{\Gamma(\frac{\beta+1}{2})^2}.$$

The Wigner Surmise distribution is a special case of the Generalized Gamma (GG) distribution, defined by

$$p_{\omega_1,\omega_2}(s) = a_{\omega_1,\omega_2} s^{\omega_1} \exp\left(-b_{\omega_1,\omega_2} s^{\omega_2}\right),$$

$$a_{\omega_1,\omega_2} = \frac{\omega_2 \left[\Gamma((2+\omega_1)/\omega_2)\right]^{\omega_1+1}}{\left[\Gamma((1+\omega_1)/\omega_2)\right]^{\omega_1+2}}, \quad b_{\omega_1,\omega_2}(s) = \left(\frac{\Gamma((2+\omega_1)/\omega_2)}{\Gamma((1+\omega_1)/\omega_2)}\right)^{\omega_2}.$$

where ω_1, ω_2 are two shape parameters. Note that the Wigner Surmise is obtained when $\omega_1 = \beta, \omega_2 = 2$. In [2], the authors approximate the spacings distribution with the GG distribution with two free parameters, giving $\omega_1 = \beta, \omega_2 = 2 - 2\exp(-2.12\beta^{0.75})$. Note that for large β, this approaches the Wigner Surmise as ω_2 approaches 2. A comparison of the Wigner Surmise, GG distribution, and the empirical distribution is shown for various values of β in Figure 3.1.

3.2 Ratio Surmise MLE

As for the case of unfolded spacings, the exact pdf of the bulk gap ratios is very difficult to find exactly, but a good approximation was found in [1] by considering the case of 3×3 matrices. This ratio surmise distribution is given by

$$\frac{1}{Z_\beta} \frac{(r + r^2)^\beta}{(1 + r + r^2)^{1+\frac{3}{2}\beta}}$$

where Z_β is the normalization constant, which we compute it numerically.

3.3 Data Based MLE

In addition to approximate functions, we also consider likelihood functions based on data samples. We sample unfolded spacings and gap ratios from $10,000 \times 10,000$ matrices, eliminate the edge values, and define smooth probability distribution functions using Scipy's gaussian kernel density estimator.

4 Results

Below we have the mean errors and mean squared errors of the various estimators for $N = 100, 250, 1000$ and $\beta = 0.5, 2.5$. Each data point is based on 1000 trials. We also include screenshots of the website.

Table 1: Comparison of Beta Estimation Methods ($N = 100$)

Method	**True Beta**	**Mean Error**	**MSE**
Wigner Surmise	0.5	-0.056	0.213
Generalized Gamma	0.5	0.045	0.192
Empirical (Spacings)	0.5	0.132	0.256
Gap Ratio Surmise	0.5	0.180	0.308
Empirical (Ratios)	0.5	0.431	0.567
Wigner Surmise	2.5	0.139	0.604
Generalized Gamma	2.5	0.156	0.601
Empirical (Spacings)	2.5	0.303	0.677
Gap Ratio Surmise	2.5	0.154	0.725
Empirical (Ratios)	2.5	0.653	1.044

Table 2: Comparison of Beta Estimation Methods ($N = 250$)

Method	**True Beta**	**Mean Error**	**MSE**
Wigner Surmise	0.5	-0.091	0.151
Generalized Gamma	0.5	0.020	0.111
Empirical (Spacings)	0.5	0.099	0.171
Gap Ratio Surmise	0.5	0.155	0.215
Empirical (Ratios)	0.5	0.396	0.450
Wigner Surmise	2.5	0.015	0.340
Generalized Gamma	2.5	0.032	0.338
Empirical (Spacings)	2.5	0.165	0.409
Gap Ratio Surmise	2.5	0.022	0.402
Empirical (Ratios)	2.5	0.495	0.689

Table 3: Comparison of Beta Estimation Methods ($N = 1000$)

Method	**True Beta**	**Mean Error**	**MSE**
Wigner Surmise	0.5	-0.102	0.120
Generalized Gamma	0.5	0.008	0.059
Empirical (Spacings)	0.5	0.085	0.119
Gap Ratio Surmise	0.5	0.150	0.169
Empirical (Ratios)	0.5	0.399	0.417
Wigner Surmise	2.5	-0.016	0.172
Generalized Gamma	2.5	0.001	0.171
Empirical (Spacings)	2.5	0.119	0.226
Gap Ratio Surmise	2.5	-0.005	0.200
Empirical (Ratios)	2.5	0.400	0.486

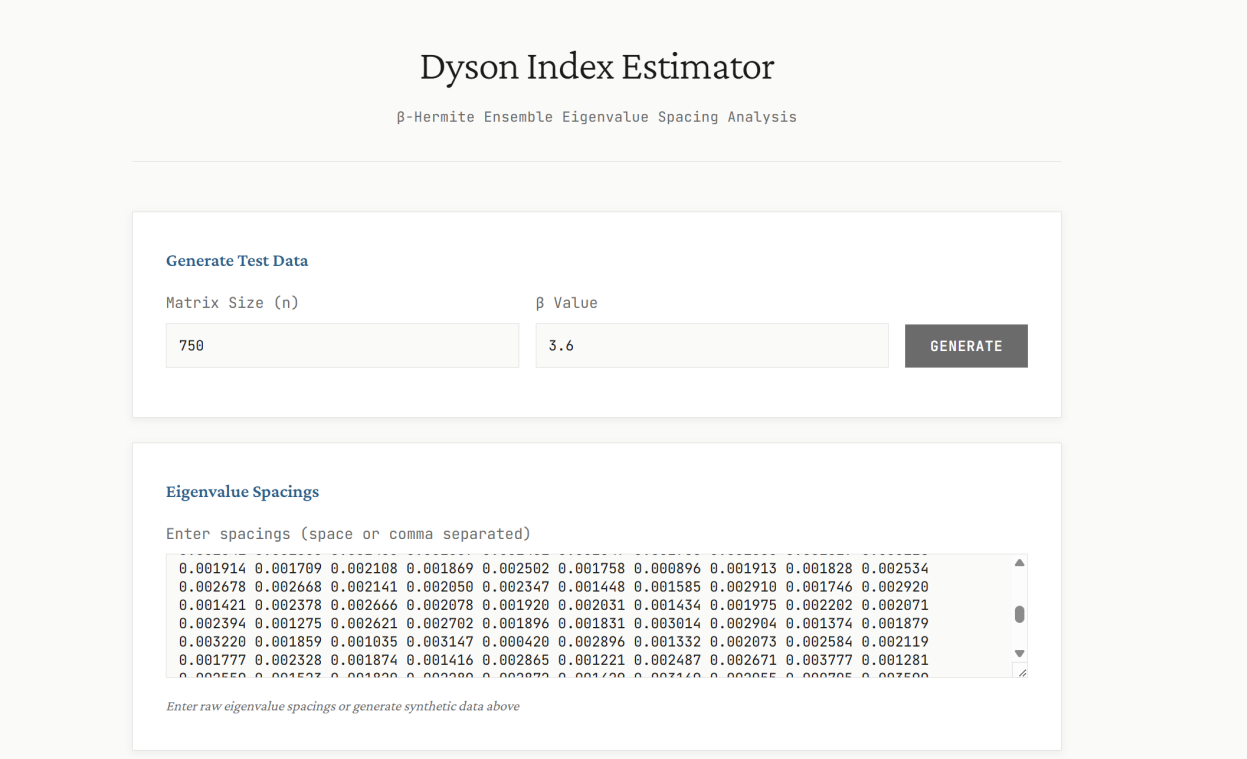

Figure 2: Spacings Input Form

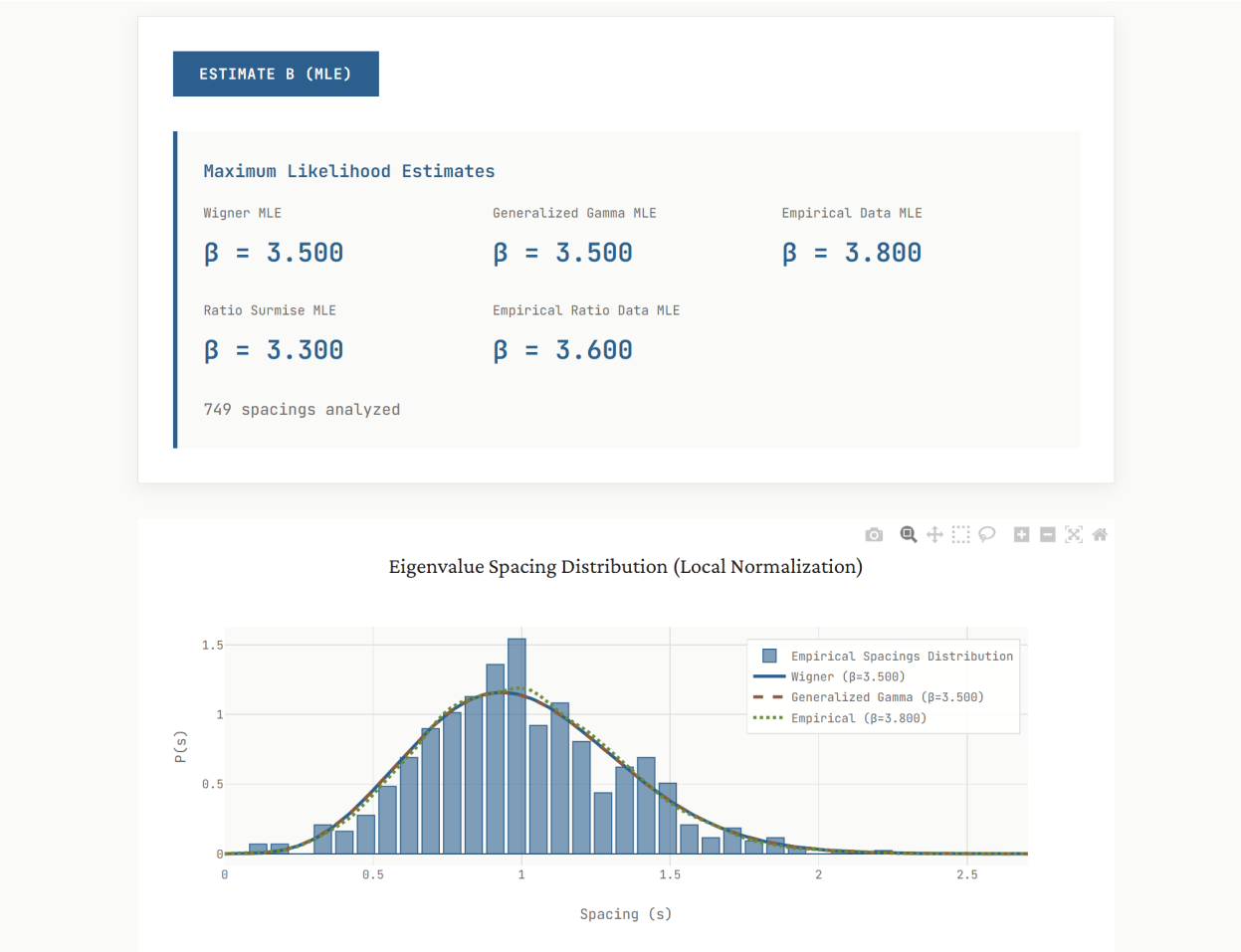

Figure 3: Estimated Betas and Spacing Distribution

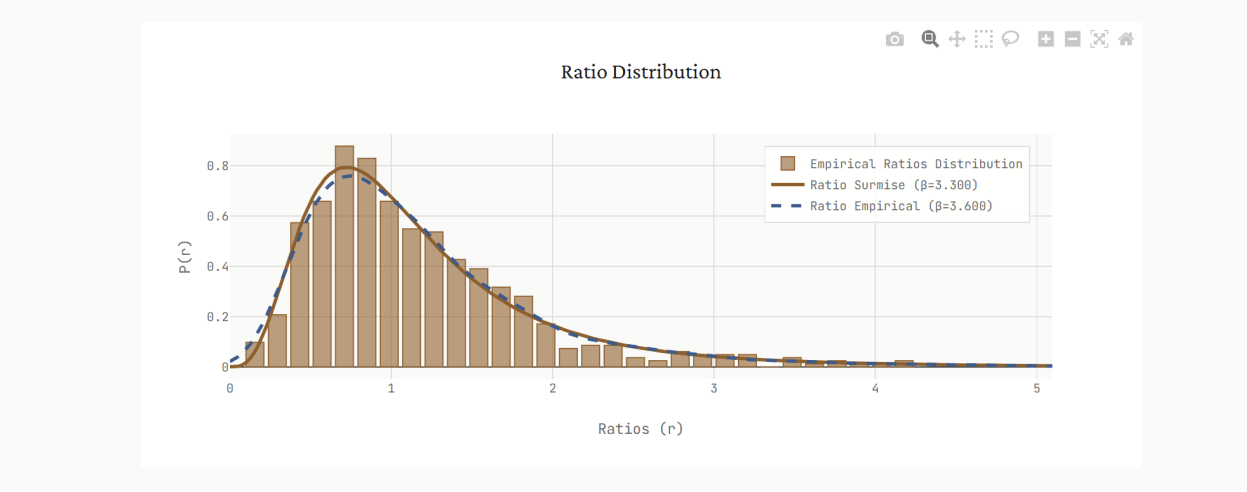

Figure 4: Ratio Distribution

5 Conclusion

As expected, estimation errors decrease monotonically with increasing matrix dimension N. While the dependence of mean error on β exhibits considerable variability across methods, the mean squared error (MSE) increases consistently as a function of β for all estimators. This behavior likely reflects the diminishing rate of change in the spacing distribution as β increases, making the estimation problem increasingly sensitive to statistical fluctuations.

The Generalized Gamma estimator demonstrates superior performance relative to the Wigner Surmise for small β, attributable to its more accurate representation of the empirical spacing distribution in this regime. However, the performance gap between these two methods diminishes as β increases, with both MSEs converging as expected from the theoretical convergence of the two underlying distributions when $\omega_2 \to 2$.

We observe a systematic positive bias across most estimators, the origin of which warrants further investigation. Additionally, the empirical data-based estimators exhibit degraded performance at larger values of β, particularly pronounced for the gap ratio method. We hypothesize that this deterioration stems from excessive smoothing in Scipy's Gaussian kernel density estimation procedure. Future work should explore adaptive bandwidth selection or alternative density estimation techniques to address this limitation.

References

[1] Y. Y. Atas, E. Bogomolny, O. Giraud, and G. Roux. Distribution of the ratio of consecutive level spacings in random matrix ensembles. *Phys. Rev. Lett.*, 110:084101, Feb 2013.

[2] G. Le Caër, C. Male, and R. Delannay. Nearest-neighbour spacing distributions of the beta-hermite ensemble of random matrices. *Physica A: Statistical Mechanics and its Applications*, 383(2):190–208, 2007.