Cross-Correlation in cricket data and RMT

Manu Kalia * and Saugata Ghosh

The Creative School, E-791, C. R. Park, New Delhi-110019, India

February 10, 2015

Abstract

We analyze cross-correlation between runs scored over a time interval in cricket matches of different teams using methods of random matrix theory (RMT). We obtain an ensemble of cross-correlation matrices C from runs scored by eight cricket playing nations for (i) test cricket from 1877 -2014 (ii)one-day internationals from 1971 -2014 and (iii) seven teams participating in the Indian Premier league T20 format (2008-2014) respectively. We find that a majority of the eigenvalues of C fall within the bounds of random matrices having joint probability distribution $P(x_1 \ldots, x_n) = C_{N\beta} \prod_{j<k} w(x_j) \left|x_j - x_k\right|^{\beta}$ where $w(x) = x^{N\beta a} \exp\left(-N\beta b x\right)$ and β is the Dyson parameter. The corresponding level density gives Marchenko-Pastur (MP) distribution while fluctuations of every participating team agrees with the universal behavior of Gaussian Unitary Ensemble (GUE). We analyze the components of the deviating eigenvalues and find that the largest eigenvalue corresponds to an influence common to all matches played during these periods.

PACS numbers: 05.45.Tp, 05.40.-a

*Corresponding author: manukalia24@gmail.com

1 Introduction

Analyzing correlations among cricket teams of different era has been a topic of interest for sports experts and journalists for decades. In this paper we study such influence (or interaction) by constructing cross-correlation matrix C [1–6] formed by runs scored by teams over different time intervals, formally called a time series.We consider the time series of batting scores posted per innings by a team in all official ICC International Test matches played. Then we construct an ensemble of cross-correlation matrices corresponding to Test data for that cricket team. We repeat the process for One Day International (ODI) and Indian Premier League (IPL) T20 cricket matches. We assume the correlations to be random and compare the fluctuating properties of C with that of random matrices. Within the bounds imposed by the RMT model, fluctuations of C show brilliant agreement with the "universal" results of GUE [7–9], while the level density corresponds to the MP distribution [10]. This implies that interactions in C are random, or in simple words not governed by any causality principal. However outside the bounds, eigenvalues of C show departure from RMT predictions, implying influence of external non-random factors common to all matches played during this period. To understand this effect, we remove k extreme bands from C and perform the Kolmogorov-Smirnov (KS) Test. We observe a better agreement with RMT predictions.

We organize the paper as follows: After a brief description of the data analyzed in sub-section [1.1], we define cross-correlation matrix in sub-section[1.2]. Section[2] introduces our RMT model along with a brief proof of MP distribution. We analyze our results and its corresponding RMT model in Section [3]. This is followed by concluding remarks.

1.1 Data analysed

We construct three ensembles, corresponding to runs scored in Tests, ODIs and Indian Premier League (IPL).

- The ODI ensemble comprises of cross-correlation matrices constructed from runs scored by India, England, Australia, West Indies, South Africa, New Zealand, Pakistan and Sri Lanka for all official ICC One Day International matches played between 1971 and 2014. For each country we have a sequence of runs scored in both home and away

matches. An ensemble of fifty one 90×90 matrices are constructed from the time series data.

- The Test ensemble comprises of cross-correlation matrices constructed from runs scored by India, England, Australia, West Indies, South Africa, New Zealand, Pakistan and Sri Lanka. For each country we have a sequence of runs scored per innings (each match has a maximum of two innings) in both home and away matches. The Test scores have been taken for all matches played between England, Australia and South Africa between 1877 and 1909 and all official ICC Test matches thereafter, till 2014. An ensemble of seventy 90×90 matrices are constructed from the time series data.

- The IPL ensemble comprises of cross-correlation matrices constructed from runs scored by Chennai Super Kings, Rajasthan Royals, Royal Challengers Bangalore, Delhi Daredevils, Kings XI Punjab, Kolkata Knight Riders and Mumbai Indians for all official BCCI IPL T20 matches played between 2008 and 2014. For each team we have a sequence of batting scores posted per match. An ensemble of twenty eight 20×20 matrices are constructed from the time series data.

1.2 Cross-correlation matrix

Cross-correlation matrix C is constructed from a given time series $X = \{X(1), X(2), \ldots\}$ by defining subsequences $X_i = \{X(i), X(i+1), \ldots, X(N)\}$ and $X_j = \{X(j), X(j+1), \ldots, X(N - \Delta t)\}$, separated by a "lag" $\Delta t = i - j$, $j < i$ and $i, j \in \mathbb{N}$. We then normalize the subsequences by defining

$$Y_i = \frac{X_i - \mu_{X_i}}{\sigma_{X_i}}. \tag{1}$$

Finally, cross-correlation matrix C [1] is defined as

$$C_{i,j} = \langle Y_i Y_j \rangle \,, \tag{2}$$

where μ_{X_i} and σ_{X_i} are sample mean score and standard deviation of the subsequence X_i respectively, and $\langle \ldots \rangle$ denotes a time average over the period studied. This is the correlation coefficient between the subsequences Y_i and Y_j and help us understand the correlation between runs scored by a given

team at different time intervals. The matrix elements lie between -1 and 1 and the matrices so constructed are Hermitian.

Now, we construct multiple matrices on a single time series, giving rise to an ensemble of matrices. Letting $C^{(1)} = C$ (as constructed above), we construct another matrix $C^{(2)}$ by removing first N elements of the time series considered, and constructing the cross-correlation matrix with the method described above. We continue this process of construction till the length of the truncated time series becomes less than N.

2 Random Matrix Model

Unitary Ensemble of random matrices is invariant under unitary transformation $H \rightarrow W^T H W$ where the ensemble is defined in the space T_{2G} of Hermitian matrices and W is any unitary matrix. Also, the various linearly independent elements of H, must be statistically independent [7].

Joint probability distribution function of eigenvalues $\{x_1, x_2, ..., x_N\}$ is given by,

$$P_{N\beta}(x_1, .., x_N) = C_{N\beta}. \prod_{j<k} x_j^{N\beta a} \exp\left(-N\beta b \sum_1^N x_j\right) |x_j - x_k|^\beta , \qquad (3)$$

where $\beta = 1, 2$ and 4 correspond to orthogonal (OE), unitary (UE) and symplectic (SE) ensembles respectively and $C_{N\beta}$ is the normalization constant [7]. We define n-point correlation function by

$$R_n^{(\beta)}(x_1, .., x_n) = \frac{N!}{(N-n)!} \int dx_{n+1} \ldots \int dx_N P_{N\beta}(x_1, .., x_N). \qquad (4)$$

This gives a hierarchy of equations [9] given by

$$\beta R_1(x) \int \frac{R_1(y)}{(x-y)} dy + \frac{w\prime(x)}{w(x)} R_1(x) = 0, \qquad (5)$$

where

$$w(x) = x^{N\beta a} \exp[-N\beta b x]. \qquad (6)$$

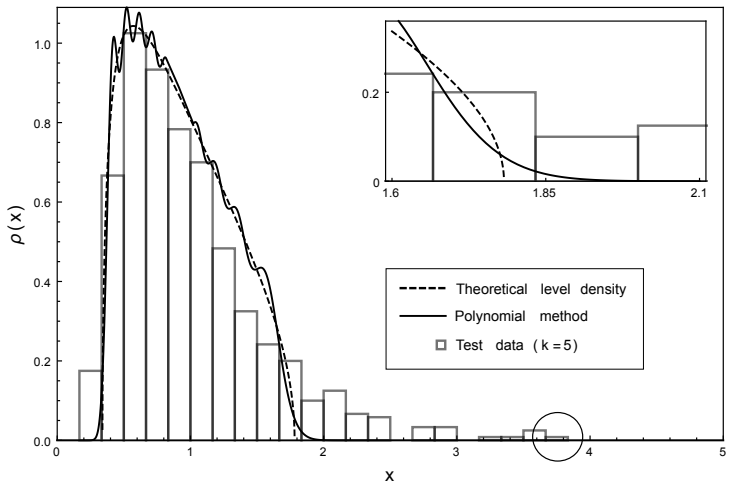


Figure 1. Level Density for averaged Test data with $k = 5$. The solid line refers to Marchenko-Pastur result (9) and the dashed line refers to the finite N result, obtained by the polynomial method described in Section 2. Here, $a = 2.75$, $b = 3.535$, $X_- = 0.339601$ and $X_+ = 1.78204$ in (9). The largest eigenvalue is circled towards the end of the spectrum.

We solve the integral equation using the resolvent

$$G(z) = \int \frac{R_1(y)}{z-y} dy, \tag{7}$$

which satisfies

$$G(x+i0) = \int \frac{R_1(y)}{x-y} dy - i\pi R_1(x). \tag{8}$$

Multiplying Eq.(5) by $x/(z-x)$ and integrating over x we get after some elementary calculation

$$\begin{aligned} \rho(x) \equiv \frac{R_1(x)}{N} &= \frac{b}{\pi x}\sqrt{(x-X_-)(X_+-x)}; \qquad X_- < x < X_+, \\ &= 0, \qquad\qquad \text{otherwise.} \end{aligned} \tag{9}$$

where

$$X_\pm = \frac{a+1}{b} \pm \frac{\sqrt{2a+1}}{b}. \tag{10}$$

For finite N, following Dyson-Mehta method [7], we use

$$\rho(x) = \frac{1}{N}\sum_{j=0}^{N-1} \phi_j^2(x), \qquad \phi_j(x) = \sqrt{w(x)} P_j(x), \tag{11}$$

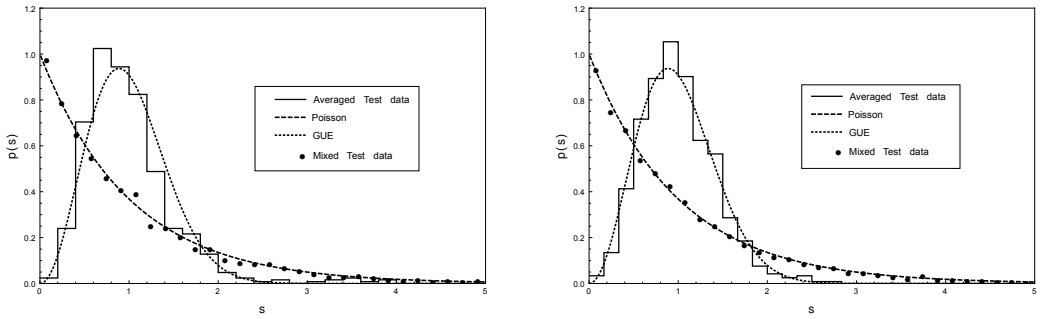
where $P_j(x)$ are orthonormal polynomials which satisfy

$$\int_{X_-}^{X_+} P_j(x)P_k(x)w(x)dx = \delta_{j,k}, \qquad j,k \in \mathbb{N}. \tag{12}$$

To understand the correlation in the system, we first need to unfold the eigenvalues to eliminate global effect over fluctuation. The sequence of scores for each country is unfolded independently. The corresponding unfolded eigenvalues y_k are given by [11],

$$y_k = \int_{X_-}^{x_k} \rho(x)\mathrm{d}x, \tag{13}$$

and the mean spacing of the unfolded eigenvalues y_k is 1. We perform unfolding using both (i) the theoretical level density (9) and (ii) numerical integration of the data and obtain the best-fit over the integrated density.



(a) Theoretical unfolding (b) Numerical unfolding

Figure 2. Nearest neighbour spacing distribution for mixed and averaged Test data obtained via numerical and theoretical unfolding (using Marchenko-Pastur result (9) with $a = 2.75$, $b = 3.535$, $X_- = 0.339601$ and $X_+ = 1.78204$). The solid line refers to spacing distribution of experimental data with $k = 5$, the dotted line refers to GUE result and the dashed line refers to the Poisson case.

For $\{S_i|S_i = y_{i+1} - y_i\}$, $s_i = S_i/D$ where y_i denote successive unfolded levels and D is the average spacing, the level spacing distribution $p(s)ds$ is defined as the probability of finding an s_i between s and $s + ds$ [7]. For no correlations between the levels, we have the Poisson distribution

$$p(s) = \exp[-s], \tag{14}$$

while for GUE, we get the Wigner's surmise

$$p(s) = \frac{32s^2}{\pi^2} \exp\left[-\frac{4}{\pi}s^2\right]. \tag{15}$$

We consider 8 sequences of eigenvalues for Test data obtained by ensemble averaging over each country. We unfold these sequences individually and average over the 8 sequences of spacings. The result shows remarkable agreement with GUE predictions (Fig. 2). Upon mixing of the eigenvalues of the Test data we observe Poisson distribution (Fig. 2).

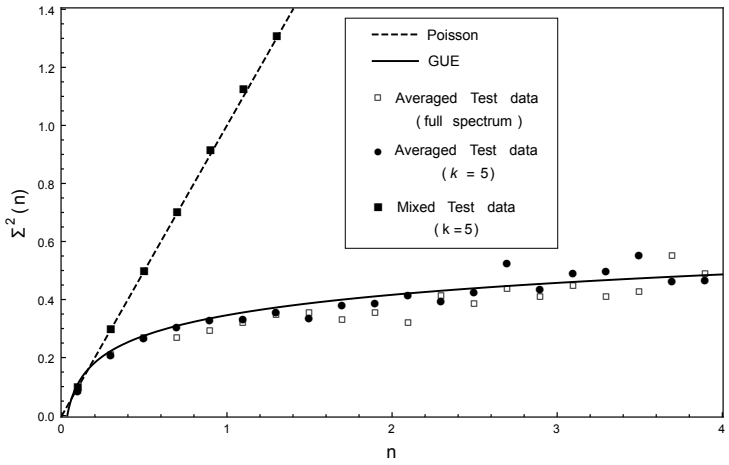


Figure 3. Number variance for the averaged and mixed Test data obtained via numerically unfolding over the spectra. The solid line refers to GUE result (18) and the dashed line refers to Poisson case. The figure plots three cases: (i) Averaged Test data with $k = 5$ extreme diagonals removed (ii) Mixed Test data with $k = 5$ extreme diagonals removed and (iii) Mixed Test data for the entire spectrum when no diagonals are removed from the matrices.

Another statistic considered is the linear statistic or the number variance. For n_k unfolded levels in consecutive sequences of length n, we define the moments [11],

$$M_p(n) = \frac{1}{N}\sum_{k=1}^{N} n_k^p, \tag{16}$$

where N is the number of sequences considered, each of length n covering the entire spectrum. Then the number variance $\Sigma^2(n)$ is given by

$$\Sigma^2(n) = M_2(n) - n^2. \tag{17}$$

For GUE, number variance is given by [7],

$$\Sigma^2(n) = \frac{1}{\pi^2}\left(\ln(2\pi n) + \gamma + 1\right), \tag{18}$$

where γ is the well known Euler constant. Number variance is known to be very sensitive for larger values of n on account of spectral rigidity. Fig 3 shows a very good agreement of the experimental number variance result of the Test data to that of the GUE result for cases when $k = 0$ and $k = 5$ extreme diagonals are removed from both ends of the matrices involved in calculation.

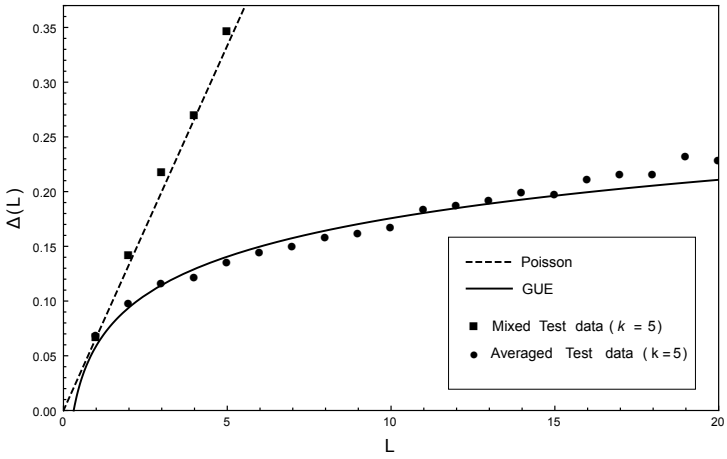


Figure 4. The Dyson-Mehta least squares statistic for the averaged and mixed Test data with $k = 5$ extreme diagonals removed from both ends of the matrices involved in calculation obtained via numerically unfolding the spectrum . The solid line refers to the GUE result (20) and the dashed line refers to the result for the Poisson case (21).

The other statistics considered is the Dyson-Mehta least square statistic or the spectral rigidity statistic [7] which measures the long-range correlations and irregularity in the level series in the system by calculating the least square deviation of the unfolding function from a straight line $y = aE + b$ over different ranges L. The statistic $\Delta(L)$ for $L = L_2 - L_1$ is given by the integral,

$$\Delta(L) = \frac{1}{L}\int_{L_1}^{L_2}(N(E) - aE - b)^2 dE, \tag{19}$$

where $N(E)$ is the unfolding function. The mean value of the statistic for

the GUE case is given by [7],

$$\langle \Delta \rangle = \frac{1}{2\pi^2}(\ln(2\pi L) + \gamma - 5/4). \tag{20}$$

For Poisson case, the least square statistics is given by

$$\langle \Delta \rangle = \frac{s}{15}. \tag{21}$$

3 Analysis

The problem that one encounters in analysis of such data are

1. The finite length of time series available introduces measurement noise.

2. A bigger time series will introduce more contributions from non-random events which will affect the "universality" result but will provide information about the correlations among different time series.

We study the RMT model defined by Eq.(3). We obtain MP distribution (9) for the level density as $N \to \infty$. We observe that the level density of eigenvalues of C in the bulk shows a remarkable agreement with the MP distribution for all Test, ODI and IPL data. However, some large eigenvalues exist outside the bounds $[X_-, X_+]$. To ensure that these eigenvalues are not due to finite N effect, we obtain level-density for finite N. For this, we develop the corresponding orthonormal polynomials using Gram-Schmidt method and using Eq.(11) for $N = 10$ obtain the level density and compare that with ensembles of cricketing data. (Fig. 1). We observe that the large eigenvalues still remain outside the bounds.

The next question is if these large eigenvalues non random, in which case our RMT model will not only show disagreement with the level density but also "spoil" the RMT predictions. To verify this, we make RMT analysis over the entire spectrum and compare its results with the truncated sparse matrix, which removes the large eigenvalues. KS test shows that our level density and spacing distribution analysis is considerably hampered by the presence of these large eigenvalues, thereby conforming the existence of non random long range correlations.

To track the level of non-randomness, we remove k, ($k << N$) extreme bands out of $2N-1$ bands of the $N \times N$ matrices C and perform the KS test. We perform numerical unfolding over the eigenvalues where the integrated

density of states are fitted with a polynomial. For ODI, where $N = 90$ we obtain a p-value of 0.640311 for the full spectrum and a p-value of 0.9025 for spectrum of the matrix with $k = 15$. For the Test data (again $N = 90$), we obtain a p-value of 0.49 for unfolding the full spectrum and a p-value of 0.855394 when unfolding the spectrum of the matrix with $k = 5$.Thus by creating a sparse matrix, which removes the large eigenvalues, our results converge to RMT predictions by $\approx 30\%$. This proves the existence of non randomness in the system introduced by elements C_{ij}, with $|i - j| \approx N$. We observe that as we increase the value of k, the largest eigenvalue in the spectrum gradually reduces and converges towards the bound imposed by the RMT model as shown in Fig. 4. We then do theoretical unfolding on the new data and observe similar agreement on KS test.

For the number variance calculation, we first unfold the spectrum and calculate number variance both within bounds and over the entire spectrum. The former gives a good agreement with GUE while the latter, as expected, shows deviation, pointing towards the presence of large eigenvalues which are due to correlation coefficients between runs scored over a long time gap.

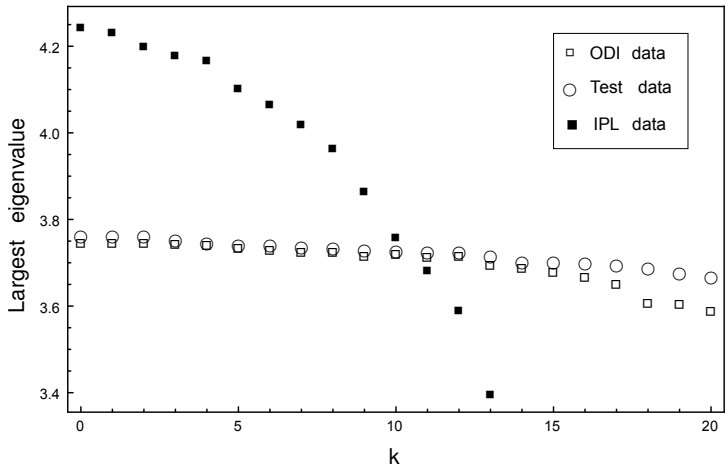


Figure 5. Largest eigenvalue in the averaged spectrum vs. k for the Test, ODI and IPL data

Finally, theoretical unfolding is performed over the spectra using Eqs.(13) and (9). The MP distribution parameters for the Test data ($k = 5$) are given in Fig. 2. For the ODI data ($k = 15$), we have $a = 2.475$, $b = 3.15$, $X_- = 0.328806$ and $X_+ = 1.87754$ as the optimal parameters for Eq. 9.

Lastly, we mix levels obtained from the time series of all teams and observe a Poisson distribution (Fig. 2).

4 Conclusion

From the statistical analysis of test, ODI and IPL data, we conclude that the eigenvalues of cross-correlation matrices display GUE universality. The Test and ODI data are the only sets of data we found to be large enough to give results of the nature produced in this paper. Thus even though the T20 results of the BCCI IPL matches are also considered the small N effect is visible in our GUE results.

We observe Wigner surmise when we study the ensembles of different countries (in tests and ODI s)/teams (IPL) separately. However, upon mixing the data of all countries, we get Poisson statistics, both for spacing and number variance. Here we may recall that while studying nuclear data statistics [12], eigenvalues with same spin show GOE but mixed data gives Poisson.

To ensure that the large eigenvalue which lies outside the bounds are not due to the size of the matrices, we obtain the level density using the polynomial method for finite N. We observe that the large eigenvalues were still lying well outside the bounds. Also while numerical unfolding over the whole spectra (and not under the MP bound), we observe that the number variance show departure from GUE. However, by removing the long-range interaction terms from C, we observe a better agreement with RMT predictions, both for level density as well as spacing distribution and number variance.

We believe that eigenvalues close to the upper bound still maintains randomness and any deviation is due to temporal effect. For example, scores getting affected due to a sudden burst of performance of an individual player over a tournament or bilateral series. However, the larger eigenvalues are probably caused due to more stable, non random influence like the effect on cricketing performance due to the advent of new technology. However this needs a thorough investigation. We wish to come back to this in a later publication.

Acknowledgement

We acknowledge ESPN Cricinfo for providing us with the cricket data.

References

[1] Vasiliki Plerou, Parameswaran Gopikrishnan, Bernd Rosenow, Luís A. Nunes Amaral, Thomas Guhr, and H. Eugene Stanley. Random matrix approach to cross correlations in financial data. *Phys. Rev. E*, 65:066126, Jun 2002.

[2] Tayeb Jamali, Hamed Saberi, and G.R. Jafari. Fractional gaussian noise: a random-matrix-theory inspired perspective. 2013.

[3] Akihiko Utsugi, Kazusumi Ino, and Masaki Oshikawa. Random matrix theory analysis of cross correlations in financial markets. *Phys. Rev. E*, 70:026110, Aug 2004.

[4] Laurent Laloux, Pierre Cizeau, Jean-Philippe Bouchaud, and Marc Potters. Random matrix theory and financial correlations. *Int. J. Theor. Appl. Finance 3*, page 391, 2000.

[5] C. Biely and S. Thurner. Random matrix ensembles of time-lagged correlation matrices: Derivation of eigenvalue spectra and analysis of financial time-series. *ArXiv Physics e-prints*, September 2006.

[6] C. W. J. Beenakker. Random-matrix theory of quantum transport. *Rev. Mod. Phys.*, 69:731–808, Jul 1997.

[7] M.L. Mehta. *Random Matrices*. Pure and Applied Mathematics. Elsevier Science, 2004.

[8] Saugata Ghosh. Long-range interactions in the quantum many-body problem in one dimension: Ground state. *Phys. Rev. E*, 69:036118, Mar 2004.

[9] Saugata Ghosh. *Skew-orthogonal polynomials and Random matrix theory*. CRM Monograph Series. AMS, 2009.

[10] V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[11] Jac Verbaarschot. Topics in random matrix theory. url: http://tonic.physics.sunysb.edu/ verbaarschot/lecture/.

[12] R. U. Haq, A. Pandey, and O. Bohigas. Fluctuation properties of nuclear energy levels: Do theory and experiment agree? *Phys. Rev. Lett.*, 48:1086–1089, Apr 1982.