Theor Chim Acta (1993) 85: 271–284

Theoretica Chimica Acta

A statistical treatment of large configuration interaction eigenvectors

Tests on model Hamiltonians and on classical MO-CI calculations

Daniel Maynau

Laboratoire de Physique Quantique (URA 505 du CNRS), IRSAMC, Université Paul Sabatier, 118, route de Narbonne F-31062 Toulouse Cedex, France

Received January 30, 1992/Accepted June 5, 1992

Summary. A statistical treatment for very large configuration interaction (CI) calculations is presented. The energy is written as a sum of elementary contributions, one per determinant, and the statistical choice is made among these elementary contributions. Two types of practical tests determine what conditions must be verified to get reliable results. It is verified that, if N is the dimension of the CI problem, the size of the statistical samples must grow as $\sqrt{N}$ to keep the same accuracy in the results.

Key words: Large configuration interaction eigenvectors – Elementary contributions – MO-CI calculations

1 Introduction

Several methods allow us to include correlation in quantum chemistry calculations. The full CI approach [1–3] is the most simple to describe since, for a given atomic orbital (AO) basis set, there is no further approximation. The size of the full CI matrix goes so rapidly with the basis set that only small systems can be studied, in spite of very efficient diagonalization techniques, direct CI programs and approximate CI schemes [4, 5] on one hand, and in spite of very large improvements in the rapidity of computers, on the other hand. The full CI calculations are therefore most widely used as benchmarks and a real chemical problem requires further, sometimes drastic approximations. Quantum chemistry calculations use various approximated methods among which we can mention, without being exhaustive, contracted CI, MRCI [6], the more sophisticated coupled cluster approaches [7], the pure perturbational treatments such as MP2 [8], MP4 [9–11], MP5 [12] and the methods mixing variation and perturbation [13]. None of these methods is universal, and depending on whether a large system is studied, or a great precision is needed, a more or less sophisticated (and expensive) method will be employed.

If one would like to seek what all these methods have in common, one could find that all these calculations deal with very large matrices and eigenvectors. The time-consuming steps are always matrix multiplications and represent the first limitation, and the second limitation is due to memory and disk storage: the problem has been solved by direct CI for several years [2] only for full CI calculations, and some works are in progress in selected direct CI [14, 15]. Since the final energy is obtained after very large calculations involving huge matrices and vectors, it is the result of a large number of very small contributions.

In such a situation, a statistical approach seems possible. Zarrabian et al. [16] propose a method similar to a truncated CI, in which the truncated part of the CI is chosen randomly. The process is repeated several times, and there is a new set of randomly chosen configurations for each calculation. They have unfortunately proposed no practical tests for the moment.

The method proposed here is completely different. We shall only consider cases for which each one of the small contributions can be calculated separately. The final energy will be written as a sum of N additive contributions, N being the size of the CI matrix. Each elementary contribution corresponds to a determinant of the CI basis, and the importance of its contribution will depend on the weight of the determinant in the wave function. Since one can anticipate whether this weight will be large or small, it is possible to make the statistical treatment only on small contributions.

It can already be said that this choice has two consequences: (i) the computational time is proportional to the percentage of determinants kept in the calculation. (ii) no iterative process can be used, since the independence of the various contributions would be lost. Thus, the statistical treatment will not concern a diagonalization process, but only the calculation of expectation values or perturbative energy.

This paper proposes two different examples of statistical methods on CI calculations. The first application deals with a semi-empirical valence bond (VB) Hamiltonian and the second one concerns more classical MO-CI calculations. As it will be demonstrated in these two tests, the crude application of statistical methods does not yield satisfactory results. After implementation of the model, the results become more reliable. The examples presented here are quite small, since the complete (i.e. non-statistical) calculation can be performed without difficulty, and the time ratio between statistical and non-statistical calculation is not very small. However, they show that, once having verified some conditions, they give reliable results and that the computational effort grows as $\sqrt{N}$, where N is the dimension of the CI matrix. These results are quite encouraging, and the model should be used in larger calculations.

2 General method and first test

In all work, we shall consider that the final energy can be written as a sum of elementary contributions ΔE_i.

$$E = \sum_{i=1}^{N} \Delta E_i \tag{1}$$

N is the dimension of the CI matrix and each corresponds to a determinant $|i\rangle$. This is obviously an arbitrary choice for the application of a statistical method to a CI calculation, and many other possibilities should exist. In particular, the sum could run on the matrix elements instead of the determinants. Zarrabian et al. [16] propose a statistical approach of CI calculations, in a completely different way leading to opposite conclusions, but they have unfortunately proposed no practical tests for the moment. A discussion tries to justify the choice made in this work in Sect. 5.

In a perturbational calculation of E, like MP for example, the energy appears already as a sum of elementary contributions. For MP_2, ΔE_i is given by:

$$\Delta E_i = \sum_{j=1}^{N} \frac{\langle 0|H|j\rangle^2}{\Delta_j} \tag{2}$$

where $|0\rangle$ is the single reference Hartree–Fock and $|j\rangle$ the doubly excited determinants. D_j is the Møller–Plesset denominator corresponding to $|j\rangle$.

As a first test of statistical calculation of E, it is possible to choose only n values of ΔE_i among the N independent contributions ($n \ll N$). Instead of selecting the most important contributions, as usual in truncation methods, the $n\,\Delta E_i$ will be randomly chosen. This choice gives a "statistical" value of the energy, which will be noted $\tilde{E}$:

$$\tilde{E} = \frac{N}{n} \sum_{j=1,n}^{N} {}^{s}\Delta E_i \tag{3}$$

where $\sum_{j=1,n}^{N}{}^{s}$ means that the n values of j result from a random choice among the N possible values.

Of course, MP_2 calculations do not require any particular computational effort, and such an approximation could only be justified in more complex processes like MP_4 for example. However, MP_2 will yield a first simple test.

Figure 1 gives the results of calculations on the N_2 molecule using the triple zeta basis set of Huzinaga [17] contracted by Dunning and Hay ($9s5p$ plus two d polarization functions with exponents 0.15 and 0.05 [18]). A first MP_2 calculation gives the exact MP_2 energy E and all the ΔE_i. In a second step, many different sets of $n\,\Delta E_i$ values are chosen in a random way.

In this paper, all the various histograms are obtained in the following way: a large number N_s of statistical energies $\tilde{E}$ are obtained using Eq. (3) (typically $N_s = 200$ in most of the figures). The histograms depend then on two parameters that are N_s and the number n of ΔE_i values in the calculation of each $\tilde{E}$. $N_t = N_s \cdot n$ values of ΔE_i must therefore be calculated. This two-parameter approach can seem somewhat complicated, but it is the only way to get some information about the quality of result. By computing already N_t values of ΔE_i, one would get the same mean energy, but no histogram could be obtained. On the other hand, it would

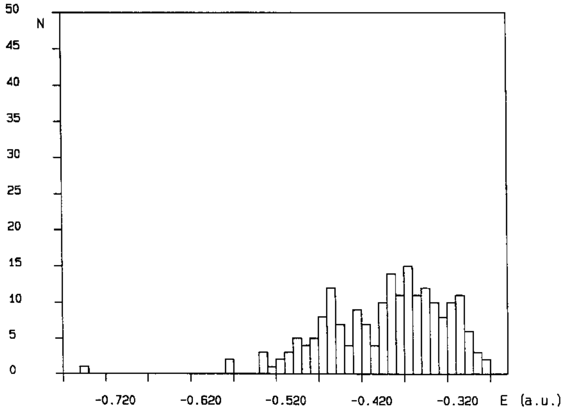

Fig. 1. Statistical MP for the N_2 molecule (Sect. 2): 1500 values ($n = 1500$) of ΔE_i are calculated and form a sample characterized by an energy $\tilde{E}$. If the statistical approach is correct, $\tilde{E}$ should be a good approximation of E. To have a good idea of the accuracy of the result, a large number ($Ns = 200$) of $\tilde{E}$ samples has been chosen, which allows us to draw a histogram. See the end of Sect. 2 for comments on the histograms of Figs. 1, 2, 3, 5, 6 and 10

be possible to calculate N_t values of $\tilde{E}$, considering that each value $N \cdot \Delta E_i$ is an approximation of the exact energy. This method would also give the same mean energy, but the histogram would be very flat, due to the very large dispersion of the various values of $N \cdot \Delta E_i$.

The results illustrated on Fig. 1 are completely unsatisfactory. The dispersion of the various results corresponding to different samples is very large. Obviously, such a crude approach is not sufficient, and more sophisticated models must be explored.

3 Tests on Pariser–Parr–Pople valence bond (PP-VB) Hamiltonian

Oujia and Malrieu [19] have developed a semi-empirical method yielding an approximate VB eigenvector for a PPP Hamiltonian of conjugated hydrocarbons. If ϕ_0 is the single-determinantal wave function of the molecule, an excellent correlated wave function may be obtained by decomposing ϕ_0 into its valence-bond components C_i and then multiplying the coefficient C_i of ϕ_i by an expression depending on the net charges (0, ± 1) on the various centers in ϕ_i. This calculation gives the coefficients c'_i of vector ψ'. The method gave at least 90% of the correlation energy on a series of structures involving up to 12 atoms. Each c'_i is calculated separately, and the computational time is proportional to the number of calculated coefficients.

If H is the PPP Hamiltonian, E and ψ are the exact eigenvalue and eigenvector:

$$H\psi = E\psi \tag{4}$$

and the approximate energy E' is obtained from ψ' by:

$$E' = \langle\psi'|H|\psi'\rangle \tag{5}$$

where $|\psi'\rangle$ is a normalized vector.

It is first necessary to write E' as a sum of elementary contributions $\Delta E'_i$, like in Eq. (1). If H_{ij} are the matrix elements of H and c'_i are the coefficients of $|\psi'\rangle$, one can write:

$$E' = \sum_{i=1}^{N} c'_i \sum_{j=1}^{N} H_{ij}c'_j \tag{6}$$

and $\Delta E'_i$ is then equal to:

$$\Delta E'_i = c'_i \sum_{j=1}^{N} H_{ij}c'_j \tag{7}$$

In the same way as in Sect. 2, the statistical energy is given by:

$$\tilde{E}' = \frac{N}{n} \sum_{i=1,n}^{N} {}^{s}\Delta E_i \tag{8}$$

One must notice that, to compute $\tilde{E}'$ in Eq. (8), it is necessary to know all the c'_i of the chosen sample, but also all the c'_j in interaction with the c'_i (i.e. for which H_{ij} is not zero) (Eq. (7)). Let ψ'^s be the vector in which the n coefficients corresponding to the determinants $|i\rangle$ of the sampling have the value c'_i of ψ', the other being zero. ψ'^s may be considered as the projection of ψ' onto the

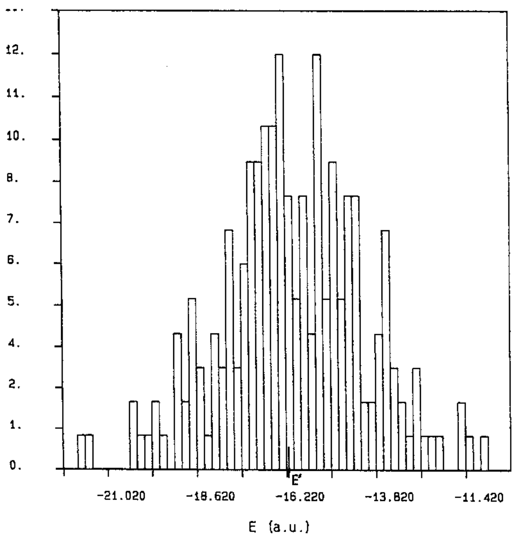

Fig. 2. Statistical PPP without renormalization of the wave function: see comments on Fig. 1. $n = 1000$, $Ns = 200$. E' is the non-statistical energy value of Eq. (5)

subspace of the n determinants $|i\rangle$ of the sampling. E' is therefore not calculated as an expectation value, but is given, after renormalization of ψ'^s, by:

$$\tilde{E}' = \langle \psi'^s | H | \psi' \rangle \tag{9}$$

and can obviously be lower than the exact energy E. The time ratio between the statistical and the exact calculation is therefore not equal to n/N, but $m \cdot n/N$, where m is the average number of determinants interacting with a given $|i\rangle$ of the sample. The choice of a PPP Hamiltonian, for which the density of non-zero elements in the matrix is very small (the number of non-zero elements on each line of the matrix is lower or equal to the number of bonds on the chemical graph), is very relevant here.

This approach is completely similar to the approach of Sect. 2 and gives very bad results too. The uncertainty on the value of E' is much larger than the correlation energy $E_{\mathrm{CI}} - E_{\mathrm{SCF}}$, as illustrated on Fig. 2.

However, calculating the energy as an expectation value, and not perturbationally as in Sect. 2 allows us to take account of a supplementary information: the vector $|\psi'\rangle$ must be normalized.

It is worth emphasizing that the use of the norm will be unusual in this work. It has nothing to do with the question on whether the perturbative wave function must be renormalized or not. The norm will be only used in the following way: (i) the sample of c_i' coefficients gives a statistical value for the energy, but also for the norm; (ii) the norm must be equal to unity, and the discrepancy between the statistically evaluated norm and 1 gives very useful information about the quality of the sample.

Calculating E' with the formula:

$$E' = \frac{\langle \psi' | H | \psi' \rangle}{\langle \psi' | \psi' \rangle^2} \tag{10}$$

one can approximate the norm $\mathcal{N}'$ of vector $|\psi'\rangle$ in the same statistical way as for the numerator of Eq. (10).

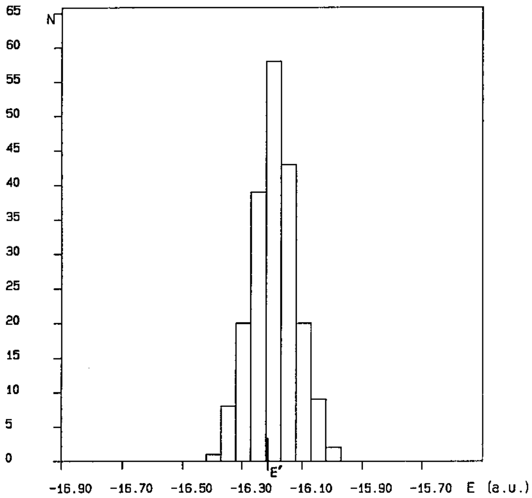

Fig. 3. Statistical PPP without renormalization of the wave function: see comments on Fig. 1. $n = 1000$, $Ns = 200$. E' corresponds to Eq. (5)

$$\mathcal{N}' = \frac{N}{n}\left(\sum_{i=1,n}^{N} {}^{s}c_i'^2\right)^{1/2} \tag{11}$$

The coefficients c_i' of ψ' must be modified, and Eq. (8) becomes:

$$E' = \sum_{i=1,n}^{N} {}^{s}\Delta E_i' \Bigg/ \sum_{i=1,n}^{N} {}^{s}c_i'^2 \tag{12}$$

where $\sum^s$ at numerator and denominator correspond to the same sample. The N/n coefficient has disappeared.

Figure 3 shows that this new formula gives very good results. Equation (10) takes account of the fact that if the sample of $\Delta E_i'$ contributions contains too important determinants, on the contrary, numerator and denominator vary in the same direction, which corrects an eventual bad sampling.

Such a large amelioration can however seem rather surprising. A better explanation can be given by the following considerations: Let us suppose that some coefficients c_i of the exact eigenvector ψ of H are known (Eq. (4)). Accepting this absurd supposition (it is impossible to know one exact coefficient without solving the whole program), a new relation between the coefficients can be obtained from Eq. (4):

$$\sum_{i=1}^{N} H_{ij}c_j = Ec_i \tag{13}$$

Equations (7) and (11) give:

$$\Delta E_i = Ec_i^2 \tag{14}$$

and any sample, even containing only one contribution ΔE_i gives the exact energy!

$$\tilde{E} = E\left(\sum_{i=1,n}^{N} {}^{s}c_i^2\right) \Bigg/ \left(\sum_{i=1,n}^{N} {}^{s}c_i^2\right) = E \tag{15}$$

The uncertainty in calculating $\tilde{E}'$ will be therefore dependent on the quality of the few c_i' required in Eq. (12). The quality of a statistical treatment will then

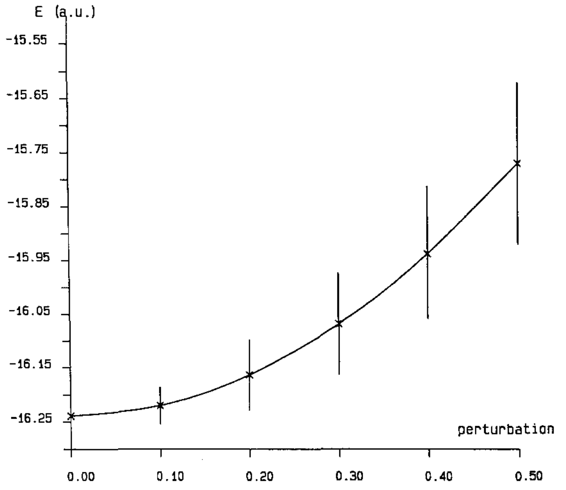

Fig. 4. Statistical PPP without renormalization of the wave function: various tests corresponding to wave function of different accuracy. The curve represents the energy as a function of the exact wave function perturbation (see text in Sect. 3). The vertical bars lengths corresponds to twice the root mean square for each perturbation.

depend on the quality of the formalism which gives the approximated c_i' coefficients.

In Fig. 4, various tests are presented, corresponding to samplings of different qualities. For these tests, the exact eigenvector ψ was first obtained by diagonalization, and more or less approximated vectors were obtained by pollution of the coefficients c_i of ψ by multiplying them by a random coefficient α_i for which $|1-\alpha_i| < \alpha_M$. The value α_M gives a measure of the quality of ψ'. The dispersion of the results due to the various samplings obviously increases ψ' as it is less accurate. Figure 4 shows the dependence of the dispersion in various samples as a function of the importance of the perturbation which damages the wave function.

4 Application to MO-CI calculations

The attempt presented in Sect. 2 to use a statistical method to get an approximated value of the MP_2 energy failed. As said in Sect. 3, the main reason for this failure was that no information was available about the norm of the perturbative coefficients of the MP_2 vector. In Sect. 3, the approximated vector ψ' should be normalized. In the random choice of some coefficients of ψ', this information allows us to correct an eventually atypical sample using Eq. (12). In applying here this statistical approach to MO-CI problems, the first condition will therefore be to compute the energy E' as an expectation energy:

$$E' = \frac{\langle\psi'|H|\psi'\rangle}{\langle\psi'|\psi'\rangle} \tag{16}$$

Contrary to the VB test, the MO-CI matrix contains much more non-zero elements. As pointed out in Sect. 3, not only the coefficients of the chosen sample must be calculated, but also all their "neighbors", i.e. all the determinants interacting with them. The MO-CI matrix is very dense, thus the neighbors are very numerous. Some further approximations must be introduced.

In the present paper, a statistical approach of the MO-CI problem will be applied on a previous work [20], concerning an approximation of SD-CI. Let us resume the theory of [20]: in a subspace S_0 spanned by the N most important determinants the Hamiltonian H was diagonalized, giving a zeroth-order function:

$$H_0 = \langle P_0 | H | P_0 \rangle \tag{17}$$

$$H_0 | P_0 \rangle = E_0 | \psi_0 \rangle \tag{18}$$

where P_0 is the projection operator onto the S_0 subspace. The remaining $N - N_0$ coefficients formed a vector ψ_1 and were calculated as follows:

$$c'_i = \frac{1}{E_0 - H_{ii}} \sum_{j=1}^{N_0} H_{ij} c_j^0 \quad i > N_0 \tag{19}$$

where the c_j^0's are the coefficients of ψ_0.

The approximate eigenvector ψ' that will be used in the statistical calculations will be given by:

$$\psi' = \frac{\psi_0 + \psi_1}{(1 + \langle \psi_1 \mid \psi_1 \rangle)^{1/2}} \tag{20}$$

In Eq. (19), the sum runs from 1 to N_0 instead of 1 to N, which will considerably diminish the computational time for calculating the c'_i coefficients on one hand, and avoid the calculation of the coefficients of the neighbors on the other hand, since they correspond to S_0 determinants which are already known.

After computing all the c'_i coefficients using Eq. (19), the energy in [20] was given by:

$$E' = \left(E_0 + \sum_{i=N_0+1}^{N} (E_0 - H_{ii}) c_i^2 \right) \Big/ \mathcal{N}^2 \tag{21}$$

where $\mathcal{N}$ is the norm of vector ψ'.

The energy E' obtained in Eq. (21) (ref. [20]) appears as a sum of elementary contributions $\Delta E'_i$:

$$E' = \frac{1}{\mathcal{N}^2} \left(E_0 + \sum_{i=N_0+1}^{N} \Delta E'_i \right) \tag{22}$$

In a statistical treatment of the sum of Eq. (22), if only n elementary contributions $\Delta E'_i$ are calculated instead of the total $N - N_0$, the statistical energy $\tilde{E}'$ is given by:

$$\tilde{E}' = \left(E_0 + \sum_{i=1,n}^{N-N_0} {}^s \Delta E'_i \right) \Big/ \left(\mathcal{N}_0 + \sum_{i=1,n}^{N-N_0} {}^s c'^2_i \right) \tag{23}$$

where $\mathcal{N}_0$ is the norm of ψ_0.

Of course, $\tilde{E}'$ will be more accurate if S_0 is of larger size, since the non-statistical contribution E_0 is more important. To get reliable results, S_0 must therefore be as large as possible. One may notice that if S_0 becomes larger, only the non-statistical part of $\tilde{E}$ will be more time consuming. The statistical part does not depend on N_0, but only on n. In fact, there are two reasons for which a larger S_0 makes easier calculation of the statistical part of the energy. The first one is that its contribution is less important with respect to $\tilde{E}'$ and the second reason is that all the largest contributions to the energy will be included in E_0, and the statistical contributions will be more homogeneous.

In this paper, we want to test the statistical part of $\tilde{E}$, and this test will be severe if N_0 is small. In the following calculations, the dimension of S_0 will therefore be as small as possible. The method is tested on two examples: the N_2 molecule and the benzene molecule.

For the N_2 molecule, the dimension of the S_0 subspace is 85 and the dimension of the total space is 16610 ($n = 85, N = 16610$). The energy E_0 is -0.1396 a.u., while the total correlation energy given by the approximate SD-CI is -0.3618 for an exact SD-CI energy of -0.3545 a.u.

The statistical study of N_2 is summarized on Fig. 5. 1500 values ($n = 1500$) of $\Delta E'_i$ and c'_i are calculated and form a sample α characterized by an energy E'_α. If the statistical approach is correct, $\tilde{E}'_\alpha$ should be a good approximation of E'. If one wishes more accurate information, a large number (200) of α samples must be selected, giving a figure like Fig. 5. The figure shows that all the energies give a result between -0.43 and -0.32 a.u., while the E_0 starting point corresponding to the 85 determinants is -0.14 a.u., for a calculation involving only 10% of the $N - N_0$ determinants of Eq. (22).

If one wishes more accurate results, it is possible to take account of all the samples of Fig. 5. The figure clearly indicates that the exact value lies around -0.36 a.u. But in this case, the computational effort is much larger (in the example of Fig. 5, $1500 \times 200 = 300000$ values of $\tilde{E}'_\alpha$ were calculated, which corresponds to 20 times the cost of the complete exact calculation). One must notice that this kind of methods should be applied only on very large systems. The poor efficiency of a SD-CI calculation on N_2 does not mean that the method is hopeless. The following test on the benzene molecule is more reasonable.

For the benzene molecule, the total dimension is 493574, and the dimension of S is 19. The energy E_0 is -0.0459 a.u., the approximated SD-CI energy E' is -0.6219 a.u., and the full SD-CI energy should lie around -0.6413 a.u. (see ref. [20]). The test presented on Fig. 6 is the same as for N_2, with $n = 9000$. All the $\tilde{E}'_\alpha$ energies lie between 0.58 a.u. and 0.70 a.u., one calculation of E'_α involves 1.8% of the $N - N_0$ determinants giving the energy E'. Figure 6, like for N_2, is given by a calculation of 200 values of $\tilde{E}'_\alpha$ and put very clearly into evidence a maximum at the correct value -0.62 a.u.

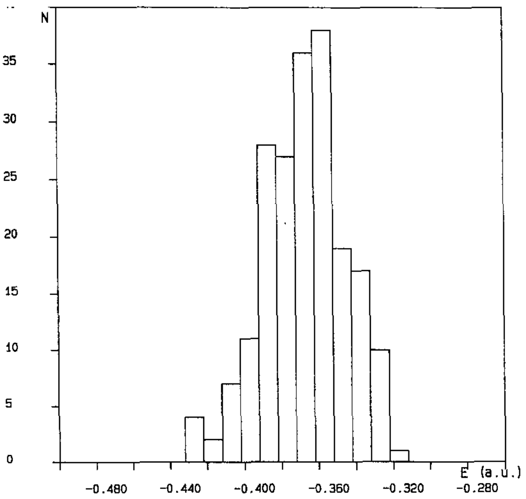

Fig. 5. *Ab-initio* SD-CI calculation on the N_2 molecule: ($n = 1500$, $Ns = 200$). E' given by Eq. (16) has the value -0.3618 a.u., see comments on Fig. 1

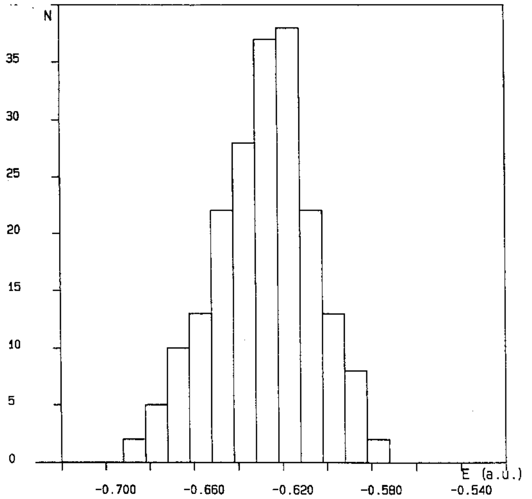

Fig. 6. *Ab-initio* SD-CI calculation on the benzene molecule: ($n = 9000, Ns = 200$). E' given by Eq. (16) has the value -0.6219 a.u., see comments on Fig. 1

The number of $\tilde{E}'_\alpha$ values to get Fig. 6 is 180000, corresponding to 36% of the total calculation of E'. Compared with the N_2 molecule (2000%), this clearly demonstrates the interest of such methods in very large calculations.

It would be desirable to start the calculation with small samples and to improve the accuracy until the distribution of Fig. 5 or 6 has a satisfactory shape. In Fig. 7, for the N_2 molecule, instead of 1500 values for each of the 200 samples, only 100 were chosen, giving a first iteration. For each iteration 100 values are added to each sample, the number of samples (200) remaining constant. At each iteration, the shape of the distribution becomes more reliable, till the root mean square reaches a given threshold value.

Figures 8 and 9 present several curves similar to Fig. 7 for N_2 and the benzene.

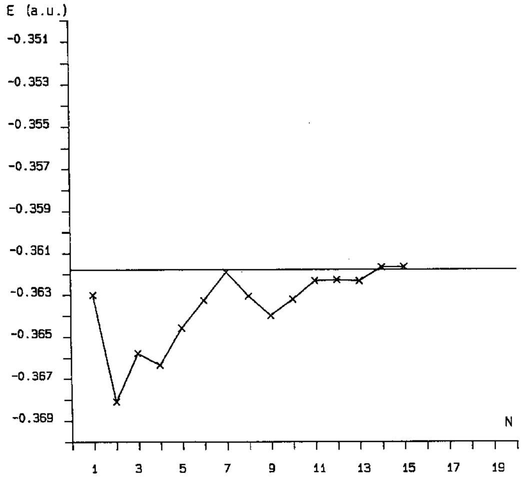

Fig. 7. *Ab-initio* SD-CI calculation on the on the N_2 molecule: ($Ns = 200$) Convergence of the mean value of the energy with the improvement of the samples. 100 values are added to each sample at each iteration. At iteration N, the size of each one of the Ns samples is then $100 \times N$ (see text in Sect. 4)

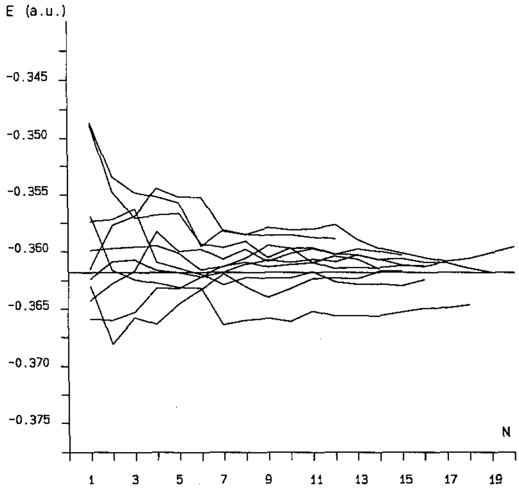

Fig. 8. *Ab-initio* SD-CI calculation on the N_2 molecule: various possible curves similar to the curve of Fig. 7

The curves for N_2 and benzene are made in the same conditions (200 samples of 100 values for each iteration, and the same convergence threshold of 0.02). One may notice that, while for N_2 the process converges after a number of iterations around 18, for benzene, this number is near 100. The ratio 100/18 = 5.6 should be compared to the square root of the ratio between the dimension of the matrices of benzene and N_2 ($\sqrt{493574/16610} = 5.5$). To keep a constant accuracy, the number of samples must grow like the square root of the dimension of the problem, which confirms that this kind of approach is particularly interesting in very large problems. A SD-CI problem is not the most favorable test for this method, and we would like to apply it to larger CI problems.

Looking at Figs. 8 and 9, some remarks can be made: (i) The convergence is very rapid at the beginning, but after a certain number of iterations, no improvement can be observed. An amelioration of the results seems therefore

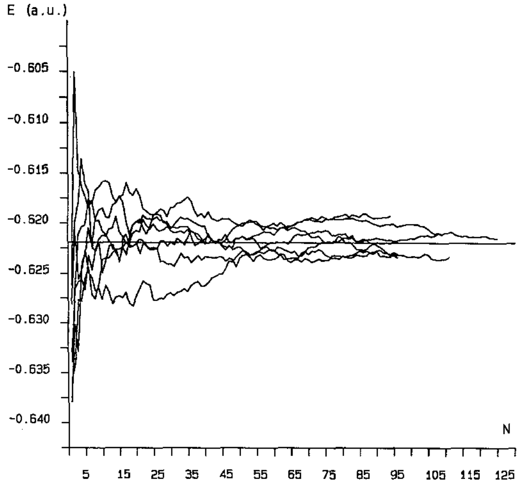

Fig. 9. Various curves for the benzene molecule. See Figs. 7 and 8

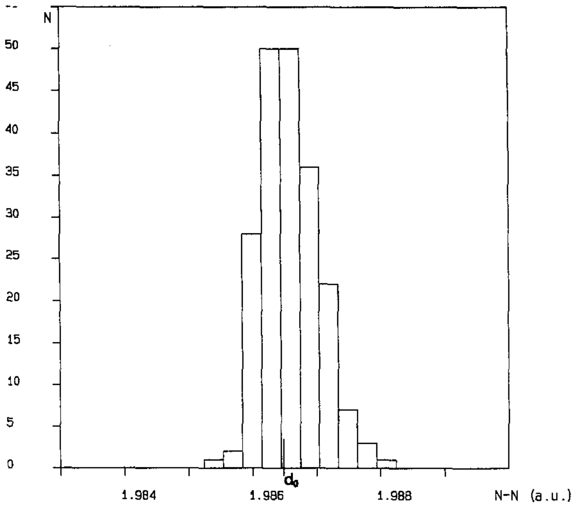

Fig. 10. Histogram of 200 ($Ns = 200, n = 1500$) values of N_2 equilibrium distance. d_0 is the equilibrium distance obtained with a non-statistical calculation

impossible by lowering the convergence threshold, but more probably by including some physical information in the model. It would be preferable to introduce some hierarchy among the determinants, while they are democratically treated in the tests presented here. (ii) The threshold itself is not very satisfactory and is not a correct criterion of the accuracy of the result. In Figs. 8 and 9, some curves remain constant near the exact value, and the threshold does not stop the process, while in other cases it stops just after the curves leave the correct value. Some other curves remain far from exact energy, and the duration of the process is not larger in these cases.

All the tests performed in the preceding paragraphs were concerning the energy. In quantum chemistry calculations, many other quantities are much more interesting and easier to obtain than the total energy. For example, a molecular geometry can be optimized at SCF, MP_2 and many other approximated methods, with reduced basis sets. All these methods will give completely different total energies, but the optimized geometries will be much more similar.

In performing a geometry optimization using the preceding statistical approach, one can hope to be in a similar situation: different samples will give quite different energies, but for each sample (i.e. for the same set of determinants), one can expect that the energy difference between two geometries will be rather constant. Figure 10 presents a statistical study of the equilibrium distance of N_2, in the same situation as Fig. 5: a sample of n ($n = 1500$) determinant is selected, and three values of the energy are calculated for three different internuclear distances with the same sample, giving a statistical equilibrium distance d. N_s ($N_s = 200$) values of d are calculated, which gives the histogram of Fig. 10. The result is quite encouraging, since the dispersion of the different equilibrium distances is very small.

5 Discussion

In this work, some characteristics of the various treatments have been kept constant in both model Hamiltonian and *ab-initio* tests.

(i) The energy has been written as sum of elementary contributions, each contribution corresponding to a determinant of the CI space. Only a small proportion of these elementary contributions are actually calculated, which leads to the statistical estimation of the total calculation.

(ii) Consequently, the statistically chosen vector may be written $P_s\psi'$, P_s being a projector and the statistical energy can be written as follows:

$$\tilde{E}' = \frac{\langle\psi'|H|P_s\psi'\rangle}{\langle\psi'|P_s\psi'\rangle} \tag{24}$$

The energy written in Eq. (24) is completely different from an expectation value. $\tilde{E}'$ can therefore be lower than E', which was illustrated in the various histograms. One may notice as a curious property that the overlap between the renormalized and $P_s\psi'$ vectors:

$$\frac{\langle\psi'|P_s\psi'\rangle}{(\langle\psi'|\psi'\rangle \cdot \langle P_s\psi'|P_s\psi'\rangle)^{1/2}} \tag{25}$$

vanishes if the density of the statistically chosen determinants is small enough, i.e. for a real statistical treatment.

(iii) No iterative process can be developed. In Eq. (25) the number of vector coefficients to be computed for ψ' is much larger than for $P\psi'$. A further iteration would increase the number of required coefficients of ψ' in the same proportion, like in a diagonalization process, for which the first trial vector is for example the single determinantal reference, the second iteration vector has the dimension of the singly and doubly excited determinants, and the third iteration takes account of all the determinants of a SDTQ-CI calculation. If one wants to avoid this explosion of the dimension, it is necessary to perform all the calculations within the statistically selected subspace. Instead of Eq. (24), intermediate calculations will use formulas like:

$$\langle P_s\psi'|H|P_s\psi'\rangle \tag{26}$$

Among the H_{ij} matrix elements of H, only the elements for which determinants i and j belong to the statistical subspace are taken into account. The density of the H diagonal elements kept will be equal to the density d_s of determinants kept in the statistical choice, while the density for H_{ij} $i \neq j$ will be proportional to d_s^2, which means that almost all the non-diagonal elements of H are lost!

(iv) Concurrently to the calculation of $\tilde{E}$, a statistical norm $\tilde{\mathcal{N}}$ is evaluated. This approach avoids a too large dispersion of the results due to some very atypical samples.

(v) In this paper, the sampling is made among the determinants of the CI basis. One could imagine other possibilities, like a sampling among the matrix elements for example. The process would be completely different, but other possibilities might arise (see ref. [11]).

6 Conclusion

A statistical approach of very large CI calculations is presented and tested in two different cases: a semi-empirical valence bond model Hamiltonian, and approximate SD-CI calculations. It appears that not all the statistical approaches give reliable results. The aim of this paper is not to establish what kind of conditions must be satisfied, but much more to exemplify in some tests giving good results that this unusual approach can be promising in some cases.

For both model Hamiltonian and *ab-initio* calculations, it was possible to obtain reliable results. It however appears that computational time can be saved only in very large large calculations, which is quite satisfactory. To keep a constant level of accuracy, the size of the sample must grow like $\sqrt{N}$, where N is the dimension of the CI problem.

References

1. Siegbahn PEM (1984) Chem Phys Lett 109:417
2. Knowles PJ, Handy NC (1984) Chem Phys Lett 111:315
3. Olsen J, Roos BO, Jorgensen P, Jensen HJA (1988) J Chem Phys 89:2185
4. Siegbahn PEM (1980) Int J Quant Chem 18:1229
5. Werner HJ, Reinsch EA (1982) J Chem Phys 76:3144
6. Buenker RJ, Perimhoff S (1974) Theor Chim Acta 35:33; Buenker RJ, Perimhoff S, Butscher W (1978) Mol Phys 35:771
7. Cizek J (1969) Adv Chem Phys 14:35
8. Pople JA, Nesbet RK (1954) J Chem Phys 22:571
9. Bartlett RJ, Purvis III GD (1978) Int J Quantum Chem 14:561
10. Frisch MJ, Krishnan R, Pople JA (1980) Chem Phys Lett 75:66; Krishnan R, Firsch MJ, Pople JA (1980) J Chem Phys 72:4244
11. Guest MF, Wilson S (1980) Chem Phys Lett 73:607
12. Kucharski SA, Noga J, Bartlett RJ (1990) J Chem Phys 90:7282
13. Spiegelmann F, Malrieu JP (1984) J Phys B 17:1259; Huron B, Rancurel P, Malrieu JP (1973) J Chem Phys 58:5745; Harrison RJ (199) J Chem Phys 94:5021
14. Caballol R, Malrieu JP (1992) Chem Phys Lett 188:543
15. Povill A, Rubio J, Illas F (1992) Theor Chim Acta 82:229
16. Zarrabian S, Kazempour KM, Estebez GA (1991) Chem Phys Lett 178:55
17. Huzinaga S (1965) J Chem Phys 53:1293
18. Dunning TH, Hay PJ (1977) in: Schaeffer HF (ed) Modern theoretical chemistry. Plenum, NY, Vol 3, p 1–27
19. Oujia B, Malrieu JP (1991) Phys Rev B 44:1480
20. Maynau D, Heully JL (1991) Chem Phys Lett 187:295