Reduced scaling in electronic structure calculations using Cholesky decompositions

Henrik Koch,[a] Alfredo Sánchez de Merás, and Thomas Bondo Pedersen
Institute of Molecular Science, University of Valencia, E-46100 Burjassot, Valencia, Spain

(Received 3 April 2003; accepted 10 April 2003)

We demonstrate that substantial computational savings are attainable in electronic structure calculations using a Cholesky decomposition of the two-electron integral matrix. In most cases, the computational effort involved calculating the Cholesky decomposition is less than the construction of one Fock matrix using a direct $O(N^2)$ procedure. © *2003 American Institute of Physics.*
[DOI: 10.1063/1.1578621]

INTRODUCTION

The notion of decomposing the two-electron integral matrix was first suggested by Beebe and Linderberg.[1] However, the idea does not seem to have received much attention in the quantum chemistry literature. Although some applications have been seen, the potential of the method has not been fully explored. Most noteworthy of these applications is the developments by Røeggen and co-workers.[2] However the most recent integral-direct implementation[3] is limited to family basis sets and as such of limited applicability. There are several reasons for this limited interest. First, the implementation for large general basis sets and large systems is by no means straightforward. Second and more important, the usefulness of the decomposition in subsequent computations is not transparent, thus rendering the advantages inconclusive.

However, there is a need in state of the art quantum chemistry to pursue different routes to reduce the computational requirements involved in accurate studies of large molecular systems. One approach to this problem is the so-called linear scaling techniques that make frequent use of the multipole expansion of the two-electron Coulomb interaction in order to reduce the computational scaling. However, these methods deteriorate as the size of the basis set on each atom increases and becomes more diffuse. Thus, we must seek methods that combine the sparsity for large systems and exploit the linear dependence in the product space of atomic orbitals. We believe the Cholesky approach is a viable attempt to attain this goal.

The Cholesky decomposition of the atomic orbital (AO) two-electron integrals may be written as

$$(\alpha\beta|\gamma\delta) = \sum_{J=1}^{M} L^J_{\alpha\beta} L^J_{\gamma\delta}, \tag{1}$$

where Greek letters denote atomic orbitals and M is the number of Cholesky vectors L. This representation is only useful if the number of Cholesky vectors needed in order to numerically represent the integrals is significantly less than the full dimension $N(N+1)/2$, where N is the number of atomic orbitals. Furthermore, the calculation of the decomposition must be carried out in an efficient integral direct manner, which avoids both storage of the precalculated integrals and recalculation of these integrals.

A related idea is used in the resolution of identity (RI) approach, also put forward by Beebe and Linderberg and later developed by Feyereisen and co-workers.[4,5] In this approach the two-electron integrals are written as an inner projection in terms of an auxiliary basis set labeled by P and Q

$$(\alpha\beta|\gamma\delta) = \sum_{PQ} (\alpha\beta|P)(P|Q)^{-1}(Q|\gamma\delta). \tag{2}$$

However, the procedure does not prescribe the construction of the auxiliary basis and this may typically be obtained by preoptimization. The clear drawback of this approach is the matrix inversion entering the expression together with the fact that errors scale with the size of the molecular system and these are statistical in nature.[6]

ALGORITHMS

The problem we face implementing the Cholesky decomposition is that the integral matrix is not positively definite but rather semidefinite. Actually, the integral matrix has most likely a slightly negatively definite part due to round off errors in the integral calculations, as we have shown by direct diagonalization for small cases, and there is no reason to believe this should be any different for larger systems and larger basis sets. The decomposition of a positive semidefinite matrix does not enjoy the stability of the procedure for strictly positive definite matrices. Round off errors are closely related to the dimension of the matrix and will increase with the dimension. Even employing full pivoting the Cholesky procedure has been shown to fail for semidefinite matrices.[7]

The decomposition to an accuracy Δ proceeds in the following manner.[1] Initially we calculate the diagonal elements $M_{pp}=(\alpha\beta|\alpha\beta)$, where p and later q will be used to denote compound AO indices. Based on the information in the diagonal we perform a prescreening and zero out elements that are smaller than $\Delta^2/X_{\max}$, where $X_{\max}$ is the maximum diagonal element. Once this initial screening has been carried out further improvements of the accuracy be-

[a]Permanent address: Department of Chemistry, Norwegian University of Science and Technology, N-7491 Trondheim, Norway. Electronic mail: koch@phys.chem.ntnu.no

yond Δ are not possible. In order to proceed we find the largest diagonal element and calculate the integrals $(**|\mathrm{AB})$, where AB is the shell pair that contains the diagonal element in question. We may now calculate the associated Cholesky vector entering the equation for the updated matrix

$$\tilde{M}_{pq}=M_{pq}-\left(\frac{M_{pJ}}{M_{JJ}^{1/2}}\right)\left(\frac{M_{qJ}}{M_{JJ}^{1/2}}\right)=M_{pq}-L_p^J L_q^J, \tag{3}$$

where the vector is implicitly defined. In an algorithm employing full pivoting we would have to discard the rest of the integrals in this shell pair unless the largest diagonal element of the $\tilde{M}$ matrix belongs to the same shell pair. This is however very unlikely and would lead to a prohibitively large number of integral recalculations. A more sound approach would be to decompose the remaining integrals in the shell pair. However, treating all diagonals larger than Δ makes the decomposition unstable even for small systems. Thus, we must control the size of diagonal elements treated in the shell pair and tailor these to the largest diagonal element at any given step in the decomposition. We have simply required that only diagonals larger than $X_{\max}/1000$ are decomposed. This will of course lead to some recalculation of integrals but as we shall see later these are actually negligible. The process now continues until all diagonal elements are smaller than Δ.

The SCF implementation is facilitated by a modified Fock matrix construction algorithm. We express the two-electron part of the AO Fock matrix in terms of the Cholesky decomposed integrals

$$F_{\alpha\beta}=\sum_J \left(2L_{\alpha\beta}^J\left(\sum_{\gamma\delta} D_{\gamma\delta}L_{\gamma\delta}^J\right)-\sum_k L_{\alpha k}^J L_{\beta k}^J\right), \tag{4}$$

where the AO density matrix is given as $D_{\alpha\beta}=\Sigma_k C_{\alpha k}C_{\beta k}$, in terms of the molecular orbital coefficients $C_{\beta k}$, where k label occupied orbitals. The implementation of Eq. (4) is straightforward resulting in the computational scaling $2MN^2O$, where O is the number of occupied orbitals.

We now proceed to the canonical orbital MP2 energy expression

$$E_{\mathrm{MP2}}=\frac{1}{2}\sum_{aibj}\frac{(2(ai|bj)-(aj|bi))(ai|bj)}{\varepsilon_a+\varepsilon_b-\varepsilon_i-\varepsilon_j}. \tag{5}$$

This is implemented in a batched loop over the a and b virtual orbital indices, making the algorithm virtually open ended with minimal storage requirements. The first part of the calculation involves the construction of transformed Cholesky vectors L_{ai}^J. From these we generate the integrals in the MP2 expression

$$(ai|bj)=\sum_{J=1}^{M} L_{ai}^J L_{bj}^J, \tag{6}$$

and process these by direct summation of the contributions in Eq. (5). We obtain the computational scaling MV^2O^2, where V is the number of virtual orbitals. We may reduce the scaling further directly decomposing the $(ai|bj)$ integrals using the transformed Cholesky vectors in the process. This gives a significantly smaller M in Eq. (6) leading to an overall reduction in computational requirements.

A few remarks about scaling and screening are now appropriate. Screening by the Cauchy–Schwarts inequality is an integral part of the Cholesky decomposition as the update matrix in Eq. (3) is positive semidefinite. Thus, at each step of the decomposition

$$|\tilde{M}_{pq}|\leq\sqrt{\tilde{M}_{pp}\tilde{M}_{qq}}\leq\sqrt{\tilde{M}_{pp}\tilde{X}_{\max}}, \tag{7}$$

assuming negligible round off errors. However, round off errors occur and we use a weaker criteria normally dividing by 1000. The inequality may be used for the individual diagonal elements as well as at shell level in the calculation of the $(**|\mathrm{AB})$ integrals. In the current implementation the Cholesky vectors are stored and read from disk, and one would be inclined to believe this is a limiting factor. However, in the limit of large basis sets the number of elements needed to be stored scale as N^2 much less than the potential N^4 number of raw two-electron integrals or the N^3 scaling suggested in Ref. 1. Performing a method specific decomposition preselecting or dynamically selecting the relevant parts of the two-electron integral matrix can facilitate linear scaling in the number of elements to be stored. In this sense the current implementation delivers an all purpose decomposed integral matrix. Exploiting the sparsity of the individual Cholesky vectors is an important goal as this will reduce the scaling of the SCF and MP2 algorithms discussed above. For instance, in the limit of a large system and assuming linear

TABLE I. Absolute errors in SCF energies reported in units of the particular decomposition threshold Δ. The numbers of Cholesky vectors are given in parentheses. The total dimension of the two-electron integral matrix is reported as $M_{\max}$ and the number of atomic orbitals as N. The number of orbitals that needed to be projected out of the basis is reported in parentheses in the last column. Errors for benzene aug-cc-pV6Z is with respect to the energy calculated using a threshold of 10^{-10}.

System (Basis set)	$\Delta=10^{-4}$	$\Delta=10^{-6}$	$\Delta=10^{-8}$	$\Delta=10^{-10}$	$M_{\max}$	N
TCO (aug-cc-pVDZ)	0.64(1284)	0.58 (2163)	0.42 (3687)	6.00 (5427)	48 205	310 (0)
TCO (aug-cc-pVTZ)	0.54(2794)	2.08 (4584)	2.39 (7255)	2.00(10547)	238 395	690 (0)
Benzene (aug-cc-pVDZ)	0.04 (933)	0.35 (1584)	0.12 (2548)	2.17 (3479)	18 528	192 (0)
Benzene (aug-cc-pVTZ)	0.34(1931)	0.05 (3238)	0.64 (4891)	0.08 (6714)	85 905	414 (1)
Benzene (aug-cc-pVQZ)	0.11(3270)	0.20 (5375)	0.74 (8014)	1.27(11178)	286 146	756 (2)
Benzene (aug-cc-pV5Z)	0.02(5437)	0.81 (8756)	0.01(12441)	0.06(16455)	771 903	1242(10)
Benzene (aug-cc-pV6Z)	0.66(8415)	0.10(12757)	0.50(17836)	– (23600)	1 798 356	1896(27)

TABLE II. Absolute errors in MP2 energies reported in units of the particular decomposition threshold Δ. The numbers of Cholesky vectors required to decompose $(ai|bj)$ integrals are given in parentheses. The total dimension of the $(ai|bj)$ integral matrix is reported as $M_{\max}$.

System (Basis set)	$\Delta=10^{-4}$	$\Delta=10^{-6}$	$\Delta=10^{-8}$	$M_{\max}$
TCO (aug-cc-pVDZ)	1.29	3.56	4.08	8 649
TCO (aug-cc-pVTZ)	1.12	15.7	15.1	20 429
Benzene (aug-cc-pVDZ)	2.6 (427)	2.9 (816)	3.2(1420)	3 591
Benzene (aug-cc-pVTZ)	9.5 (700)	6.3(1323)	8.2(2139)	8 232
Benzene (aug-cc-pVQZ)	20.3(1061)	10.3(1976)	12.1(3015)	15 393
Benzene (aug-cc-pV5Z)	23.6(1530)	53.9(2790)	10.4(4109)	25 431
Benzene (aug-cc-pV6Z)	34.3(2147)	31.2(3766)	76.2(5482)	38 808

scaling in the AO density, the evaluation of the exchange part of the Fock matrix will scale as N^2. Method specific decomposition should reduce this even further.

APPLICATIONS

First we would like to address the accuracy of the decomposition. For this purpose we have chosen to use trans-cyclooctene (TCO) and benzene as illustration. In Table I the errors in calculated SCF energies are reported for different levels of accuracy in the decomposition. The main features to notice in these errors are that the accuracy obtained in the SCF energy is mainly determined by the threshold in the decomposition. The errors are seen to be consistent and stable with respect to the threshold.

In Table I we also report the number of Cholesky vectors needed to decompose the integral matrix to a given threshold. There are several aspects to consider and let us initially focus on $\Delta=10^{-8}$ for benzene. We observe that the number of Cholesky vectors compared to full dimension decrease rapidly with the size of the basis; such that for aug-cc-pVDZ we obtain a reduction of 7.3 and this increase to 101 for aug-cc-pV6Z. For a standard application basis set like aug-cc-pVTZ the factor is 17.6. The factor between the number of Cholesky vectors and basis functions does not show such large variation. For aug-cc-pVDZ and aug-cc-pVTZ the factor is 13.3 and 11.8 respectively, and decreases further to 9.4 for aug-cc-pV6Z. We then look at the number of Cholesky vectors for different thresholds. Evidently, increasing the accuracy from 10^{-8} to 10^{-10} has a fairly high cost; around 30–40% in the number of vectors. The difference between 10^{-6} and 10^{-8} thresholds is around 40–60% with larger variation depending on the basis. In general we recommend using 10^{-8} or maybe 10^{-6} depending on the properties that need to be calculated. As a final remark about the benzene data we should note the amazing fact that only 27 orbitals had to be projected out of the aug-cc-pV6Z basis due to small eigenvalues in the overlap matrix. This clearly demonstrates that even this large basis set is far from saturation or what we might term numerical completeness. Turning to TCO in Table I, we see that errors are slightly higher than in benzene. Otherwise we note that the reduction factor for aug-cc-pVTZ is 32.8 for the 10^{-8} threshold.

We now turn our attention to the errors in the calculated MP2 energies in Table II. Two different computational strategies have been employed. For TCO we used the full set of Cholesky vectors and for benzene we decomposed the $(ai|bj)$ integral matrix before summing the contributions. We first observe that the errors are larger then for SCF, this has a natural explanation as the MP2 energies depend linearly on the errors in the orbitals. The MP2 accuracy also depends on the error in the $(ai|bj)$ integrals scaled by the orbital energy denominators, potentially increasing the MP2 error beyond the decomposition threshold. For benzene the errors refer to 10^{-10} threshold, as for basis sets larger than aug-cc-pVTZ it was impossible to carry out the calculations using our integral-direct coupled cluster code. Regarding the decomposition of the $(ai|bj)$ integral matrix we note that the reductions are not as pronounced as for the atomic orbital decomposition; however the maximum rank is also several orders of magnitude smaller. Even though we use a canonical basis the reductions are significant spanning from 2.5 to 7.1 for the threshold equal 10^{-8}.

Table III summarizes the decomposition for 1,4-bis[2-(4-diphenylamino-phenyl)-vinyl] benzene denoted stilbene-DD(6) in the table, C_{60} and benzene. The ratio between the number of Cholesky vectors and the number of basis functions ranges from 9.4 for benzene to 12.5 for stilbene-DD(6). In all cases M is less than 5% of $M_{\max}$, notably 1.8% for C_{60} and 1.0% for benzene, showing that the amount of integrals that must be calculated to numerically represent the integral matrix is minimal. For all systems and basis sets, the actual number of integral distributions is less than 8%, of which a negligible number of shell pairs must be recalculated as a consequence of the algorithm. The time spent recalculating integrals is minimized by employing segmented basis sets. Finally, we note that the total time required for the Cholesky decomposition algorithm is comparable to that of a single integral-direct Fock matrix build using density and integral prescreening. The total number of elements in the Cholesky vectors scale as N^2 although the prefactor can be large. For benzene the prefactor increases 95% going from aug-cc-pVDZ to aug-cc-VTZ, but when going from aug-cc-pV5Z to

TABLE III. The number of Cholesky vectors M and the maximum dimension. Furthermore, the number of shell pairs $(**|AB)$ that need to be calculated and the maximum (SP and $SP_{\max}$). The last column report timing ratio between one Fock matrix construction and the total time for the Cholesky decomposition threshold $\Delta=10^{-8}$.

System	Basis set	M	$M_{\max}$	SP	$SP_{\max}$	T_{DSCF}/T_{CD}
Stilbene-DD(6) ($C_{46}N_2H_{36}$)	6-31G	6301	127 260	2159 (21)	48 828	1.05
	6-31G++	7574	268 278	2640 (29)	98 790	1.35
Buckminster fullerene (C_{60})	aug-cc-pVDZ	17 390	952 890	253 (0)	3 321	0.35[a]
Benzene	aug-cc-pV6Z	17 836	1 798 356	238 (10)	7 750	1.75

[a]Integral threshold in decomposition 10^{-40}.

aug-cc-pV6Z the increase is only 40%. A method specific decomposition should make this sequence converge fast.

CONCLUSIONS

We have demonstrated the small numerical rank of the two-electron integral matrix for large molecular systems and large basis sets. The proposed algorithm is stable and can be used as a black box generator of the Cholesky vectors. The current implementation still requires some improvements as the calculations done in the inner most loop of the decomposition do not exploit the sparsity in the Cholesky vectors.

With respect to the practical applicability of the presented method an efficient approach to geometrical derivatives is imperative. Such an approach is obtained including certain derivative product functions and decomposing an expanded integral matrix. To be more explicit we write the first derivative integrals as

$$(\alpha\beta|\gamma\delta)^{(1)} = (\alpha^{(1)}\beta|\gamma\delta) + (\alpha\beta^{(1)}|\gamma\delta) + (\alpha\beta|\gamma^{(1)}\delta) + (\alpha\beta|\gamma\delta^{(1)}) \quad (8)$$

and observe that including the product functions $\alpha^{(1)}\beta$ in the decomposition we may express the derivative integrals in terms of Cholesky vectors. Higher derivative integrals can be calculated in a similar manner. This might greatly improve the cost of higher derivatives as the inclusion of additional product function will incur an even larger degree of linear dependence in the product space.

Integral-direct techniques for highly correlated *ab initio* models have expanded the application range for coupled cluster methods. These methods are still very demanding and are considered a serious bottleneck. We anticipate the Cholesky approach will remove this limitation and the future developments of these methods will focus on reducing the scaling, as well as an embarrassingly parallel implementation of the Cholesky decomposition will make applications virtually open ended.

ACKNOWLEDGMENTS

Computer resources from DCSC at University of Southern Denmark are acknowledged. This work was supported by the European Research and Training Network ''Molecular Properties and Molecular Materials'' (MOLPROP), Contract No. HPRN-CT-2000-00013. We acknowledge support from the Spanish MCT (Plan Nacional I+D+I) and European FEDER funds (Project No. BQU2001-2935-C02-01).

[1] N. H. F. Beebe and J. Linderberg, Int. J. Quantum Chem. **7**, 683 (1977).
[2] I. Røeggen and E. Wisløff-Nilssen, Chem. Phys. Lett. **132**, 154 (1986).
[3] I. Røeggen, private communication.
[4] M. W. Feyereisen, G. Fitzgerald, and A. Komornicki, Chem. Phys. Lett. **208**, 359 (1993).
[5] O. Vahtras, J. E. Almlöf, and M. W. Feyereisen, Chem. Phys. Lett. **213**, 514 (1993).
[6] F. Weigend, A. Köhn, and C. Hättig, J. Chem. Phys. **116**, 3175 (2002).
[7] N. J. Higham in *Reliable Numerical Computation*, edited by M. G. Cox and S. J. Hammarling (Oxford University Press, Oxford, UK, 1990), pp. 161–185.