Gibbs2: A new version of the quasi-harmonic model code. I. Robust treatment of the static data.

A. Otero-de-la-Roza *, and Víctor Luaña

Departamento de Química Física y Analítica, Facultad de Química, Universidad de Oviedo, 33006 Oviedo, Spain.
A contribution from the Malta-Consolider group
(`http://www.malta-consolider.com/`).

Abstract

We describe in this article the techniques developed for the robust treatment of the static energy *versus* volume theoretical curve in the new version of the quasi-harmonic model code [Comput. Phys. Commun. **158** (2004) 57]. An average of strain polynomials is used to determine, as precisely as the input data allow it, the equilibrium properties and the derivatives of the static $E(V)$ curve. The method provides a conservative estimation of the error bars associated to the fitting procedure. We have also developed the techniques required for detecting, and eventually removing, problematic data points and jumps in the $E(V)$ curve. The fitting routines are offered as an independent octave package, called AsturFit, with an open source license.

PACS: 07.05.Bx, 71.15.-m, 64.10.+h 65.40.-b 64.30.Jk

Key words: Equations of State in Solids, Equilibrium Properties of Solids, Data Analysis, Treatment of Noisy Data, Discontinuous Data, Thermal Effects.

PROGRAM SUMMARY/NEW VERSION PROGRAM SUMMARY

Authors: A. Otero-de-la-Roza and V. Luaña.
Program Title: AsturFit.
Journal Reference:
Catalogue identifier:
Licensing provisions: GPL, version 3.
Programming language: GNU Octave.

* Corresponding author
Email addresses: `alberto@carbono.quimica.uniovi.es` (A. Otero-de-la-Roza), `victor@carbono.quimica.uniovi.es` (Víctor Luaña).

Operating system: Unix, GNU/Linux.
Number of processors used: 1.
Supplementary material: a collection of datasets, test scripts, and model outputs.
Classification: 4.9 Minimization and fitting.
Nature of problem: Fit the total energy *versus* volume data of a solid to a continuous function and extract the equilibrium propeties and the derivatives of the energy, with an estimation of the error introduced by the fitting procedure.
Additional comments: The techniques discussed have been implemented in GIBBS2, to be included with the second part of this article. Included here is the OCTAVE implementation of the routines, useful for interactive work and also for the creation of independent scripts. Some representative examples are included as test cases.
No. of bytes in distributed program, including test data, etc: 2 MByte.
No. of lines in distributed program, including test data, etc: 3017 lines of source code, 606 lines of testing code, some 29000 lines of data and sample output, a user manual of 27 pages.
Distribution format: tar gzip file.
External routines: the GSL and OPTIM packages from the octaveforge site (`http://octave.sourceforge.net/`).

LONG WRITE-UP

1 Introduction

Published in 2004 in this journal [1] the quasi-harmonic Debye model code GIBBS has become a popular and inexpensive method for deriving thermal behavior out from energy *versus* volume data obtained from electronic structure crystal calculations. The experience in this years of using the code and answering to the questions and problems posed by the users has been the driving force behind the development of a new version of the program, more robust from a numerical point of view and with much improved capabilities.

We have divided the presentation of the new GIBBS2 into two articles. This first one specializes in the reliable treatment of the $E(V)$, and eventually $p(V)$, data. The second article discusses the thermodynamic models implemented in the new code and describes the actual structure and use of the program.

The poor properties of the input data have been the source of most troubles reported by the users to the authors of the original GIBBS code. Discontinuities, noise, and inadequate range of data being the three more common sources of concern. We have tried to implement techniques to deal with those problems or, at least, detect and document clearly their occurrence.

At difference from experimental errors, typically random, errors from a theoretical calculation tend to be reproducible and systematic and need a very different treatment. Some of the common origins of trouble in the calculation of the crystal cell energy versus cell shape and size include: (a) lack of convergence with respect to the calculation basis (plane waves, local orbitals, etc.) or conditions (Brillouin zone integration grid, for instance); (b) changes in the symmetry treatment between adjacent points (a typical problem in the calculation of elastic constants); (c) triggering a threshold distance that changes the number of neighbors or the number of integrals of some type included in the calculation; and (d) numerical instability, like failing to detect ill-conditioned matrices, for instance.

Whereas the real solution to the theoretical systematic errors is improving the calculation conditions or the codes to avoid the source of trouble, there are times when recovering as much information as possible from the data is desirable. In our case we need to determine a smooth $E(V)$ curve from possibly noisy or discontinuous data in such a way that the derivatives of the curve are as faithful as possible.

A number of theoretical models have been developed to represent the equation of state (EOS) of solids, be it in $E(V)$ or $p(V)$ form. Several of them have been included in GIBBS2. The advantage of these models is that using a reduced number of parameters they can cover a wide range of conditions. The disadvantage, however, is that most of the parameters are nonlinear and starting from a good set of values is usually a requirement for achieving the convergence of the models. Most of our work, however, has been based on the use of polynomials and augmented polynomial functions, easier to work with from a numerical point of view, and easier to adapt to the treatment of problematic data sets.

The methods described in this article have been developed and tested as OCTAVE routines before being incorporated into GIBBS2, which is written in FORTRAN90. The experience has been quite fruitful and offers the most useful possibility of working interactively with a data set. Accordingly, we have included the OCTAVE routines as an independent package called AsturFit.

The rest of the article is organized as follows. Section 2 describes the source of the data that will be used in most of our tests. Section 3 introduces some of the most popular analytical EOS as they have been incorporated into GIBBS2. The most important part of this article is probably contained in section 4, devoted to the linear fitting of strain EOS of arbitrary degree. The next section, 5, addresses the detection and, eventually, the solution of different types of problematic data. The set of OCTAVE routines created for the development of the new techniques is the subject of section 6. Finally, section 7 presents our conclusions.

2 Test datasets

We will illustrate most of our analysis in this work using a local density (LDA) calculation of the rock-salt phase of MgO. The calculation has been done with the Quantum espresso code [2], using plane-waves and ultrasoft pseudopotentials under very strict conditions: cutoff energy (80 Ry), $\boldsymbol{k}$-mesh (shifted $4 \times 4 \times 4$ Monkhorst-Pack grid) and $\boldsymbol{q}$-mesh ($6 \times 6 \times 6$). A fine linear grid of 129 points has been calculated in the volume range 72–143 bohr^3 and this will be our main reference. Some extra calculations have been done to illustrate particular aspects, but they will be explicitly described when used.

3 Analytical Equation of State forms

A large number of analytical forms have been proposed to represent the behavior of the $p(V)$ or the $E(V)$ curves under low temperature conditions. Most of these equations use a null pressure as the reference condition and contain a few number of nonlinear adjustable parameters: the energy, volume and bulk modulus at zero pressure, E_0, V_0 and B_0, respectively, and some pressure derivatives of the bulk modulus, also evaluated at the reference pressure, B_0', B_0'', ... In addition, different foundations and objectives lie behind each form, but a number of excellent sources describe their origin from a historical and physical point of view, including Zharkov and Kalinin, 1971 [3]; Stacey et al, 1981 [4]; Eliezer and Ricci, 1991 [5]; Anderson, 1995 [6]; Holzapfel, 1996 [7]; Poirier, 2000 [8]; Holzapfel, 2001 [9]; Eliezer et al, 2001 [10]; Stacey, 2005 [11]; and Peiris and Gump, 2008 [12].

In general, most of the traditional work on the EOS has been devoted to the design of a functional form with as few free parameters as possible that behaves correctly on a large range of volumes, in fact, all the way from $V \to 0$ to $V \to \infty$ if that is possible. Instead, our objective is extracting faithfully the information contained in the theoretical data, determining the equilibrium properties and the derivatives of the $E(V)$ curve, and obtaining an estimate of the error associated to the treatment of the data. Expanding the volume range of the EOS has for us the meaning of extending the electronic structure calculations to shorter or larger geometries, and the most important property that we demand from the EOS fitting is flexibility.

We have implemented nine different $E(V)$ forms, plus four $p(V)$ forms, chosen among the most popular EOS's used in the literature. The implementation is modular, so adding and using new forms is very easy.

3.1 Murnaghan EOS

Murnaghan historical paper [13] is based on the principle of conservation of mass combined with Hooke's law for an infinitesimal variation of stress in the solid. An equivalent result can be simply obtained by assuming a linear variation of the bulk modulus with pressure, $B(p) = B_0 + B_0'p$, to integrate $B = -V(\partial p/\partial V)_T$:

$$V(p) = V_0 \left(1 + \frac{B_0'}{B_0} p\right)^{-1/B_0'} \tag{1}$$

which can be inverted to

$$p(V) = \frac{B_0}{B_0'} \left[\left(\frac{V_0}{V}\right)^{B_0'} - 1 \right]. \tag{2}$$

The hydrostatic work equation, $dW = -pdV$, can then be used to integrate:

$$E = E_0 + \frac{B_0 V}{B_0'} \left[\frac{(V_0/V)^{B_0'}}{B_0' - 1} + 1 \right] - \frac{B_0 V_0}{B_0' - 1} \tag{3}$$

for the volume dependent energy under null temperature conditions.

The Murnaghan EOS is popular because of its simple functional form. For a variety of solids it has been proved to be coincident with the experimental measurements up to pressures of the order of $B_0/2$ [6]. Other sources cite a compression of some 10%, i.e. $V/V_0 < 0.9$, as the safe limit of application.

3.2 Birch-Murnaghan and Birch EOS family

The Birch-Murnaghan [13–15] family of EOS, perhaps the most commonly used in the fitting of experimental $p(V)$ data, is the result of assuming a polynomial form for the energy,

$$E(f) = \sum_{k=0}^{n} c_k f^k, \tag{4}$$

in terms of the finite Eulerian strain;

$$f = \frac{1}{2} \left[\left(\frac{V_r}{V}\right)^{2/3} - 1 \right], \tag{5}$$

where V_r is a reference volume. Enforcing the next limiting conditions:

$$\lim_{f \to 0} \left\{ V; E; p = -\frac{dE}{dV}; B = -V\frac{dp}{dV}; B' = \frac{dB}{dp}; ... \right\} = \{V_0; E_0; 0; B_0; B_0'; ...\} \tag{6}$$

is enough to determine

$$\begin{aligned}
V_r &= V_0, && (7)\\
c_0 &= E_0, && (8)\\
c_1 &= 0, && (9)\\
c_2 &= \frac{9}{2}V_0B_0, && (10)\\
c_3 &= \frac{9}{2}V_0B_0(B_0'-4), && (11)\\
c_4 &= \frac{3}{8}V_0B_0\{9[B_0B_0''+(B_0')^2]-63B_0'+143\}, ... && (12)
\end{aligned}$$

so the reference volume and the EOS parameters correspond to the properties at the equilibrium point. Some algebraic effort is required to extend these expressions to high order, and a computer algebra code like MAXIMA is of great help [16].

The simplest form, BM2 or second order Birch-Murnaghan takes the form

$$\begin{aligned}
E &= E_0 + \frac{9}{2}B_0V_0f^2 = E_0 + \frac{9}{8}B_0V_0(x^{-2/3}-1)^2, && (13)\\
p &= 3B_0f(1+2f)^{5/2} = \frac{3}{2}B_0\left(x^{-7/3}-x^{-5/3}\right), && (14)\\
B &= B_0(7f+1)(2f+1)^{5/2}, && (15)
\end{aligned}$$

where $x=(V/V_0)$. BM3 produces

$$\begin{aligned}
E &= E_0 + \frac{9}{2}V_0B_0f^2[1+(B_0'-4)f]\\
&= E_0 + \frac{9}{16}V_0B_0\frac{(x^{2/3}-1)^2}{x^{7/3}}\{x^{1/3}(B_0'-4)-x(B_0'-6)\}, && (16)\\
p &= \frac{3}{2}B_0f(2f+1)^{5/2}[2+3(B_0'-4)f]\\
&= \frac{3}{8}B_0\frac{x^{2/3}-1}{x^{10/3}}\{3B_0'x-16x-3x^{1/3}(B_0'-4)\}, && (17)\\
B &= \frac{1}{2}B_0(2f+1)^{5/2}\{(27f^2+6f)(B_0'-4)-4f+2\}\\
&= \frac{B_0}{8x^{10/3}}\{x^{5/3}(15B_0'-80)-x(42B_0'-196)+27x^{1/3}(B_0'-4)\}. && (18)
\end{aligned}$$

and BM4

$$\begin{aligned}
E &= E_0 + \frac{3}{8}V_0B_0f^2\{(9H - 63B_0' + 143)f^2 + 12(B_0' - 4)f + 12\}, && (19)\\
p &= \frac{1}{2}B_0(2f+1)^{5/2}\{(9H - 63B_0' + 143)f^2 + 9(B_0' - 4)f + 6\}, && (20)\\
B &= \frac{1}{6}B_0(2f+1)^{5/2}\{(99H - 693B_0' + 1573)f^3\\
&\quad + (27H - 108B_0' + 105)f^2 + 6(3B_0' - 5)f + 6\}, && (21)
\end{aligned}$$

where $H = B_0B_0'' + (B_0')^2$.

3.3 Poirier-Tarantola EOS family

The Poirier-Tarantola [17] EOS family is based of the expansion of the strain energy in terms of the *natural* or Hencky linear strain: $f_N = \ln(l/l_0)$, with l a characteristic cell length and l_0 its equilibrium value, or $f_N = \ln(V/V_0)^{1/3}$ under hydrostatic conditions. The first member would be produced by a second order expansion:

$$\begin{aligned}
E &= E_0 + \frac{9}{2}B_0V_0f_N^2 = E_0 + \frac{1}{2}B_0V_0\ln^2 x, && (22)\\
p &= -3B_0f_Ne^{-3f_N} = -\frac{B_0}{x}\ln x, && (23)\\
B &= B_0(1 - 3f_N)e^{-3f_N} = \frac{B_0}{x}(1 - \ln x). && (24)
\end{aligned}$$

Much better results should be obtained with the PT3 EOS,

$$\begin{aligned}
E &= E_0 + \frac{9}{2}B_0V_0f_N^2[(B_0' + 2)f_N + 1]\\
&= E_0 + \frac{1}{6}B_0V_0\ln^2 x[(B_0' + 2)\ln x + 3], && (25)\\
p &= -\frac{3}{2}B_0f_Ne^{-3f_N}[3(B_0' + 2)f + 1] = -\frac{B_0\ln x}{2x}[(B_0' + 2)\ln x + 2], && (26)\\
B &= -\frac{B_0}{2}e^{-3f_N}[9(B_0' + 2)f_N^2 - 6(B_0' + 1)f_N - 2]\\
&= -\frac{B_0}{2x}[(B_0' + 2)\ln x(\ln x - 1) - 2], && (27)
\end{aligned}$$

and the PT4 EOS:

$$\begin{aligned} E &= E_0 + 9B_0V_0f_N^2\{3(H+3B_0'+3)f_N^2 + 4(B_0'+2)f_N + 4\} \\ &= E_0 + \frac{1}{24}B_0V_0\ln^2 x\{(H+3B_0'+3)\ln^2 x + 4(B_0'+2)\ln x + 12\}, \end{aligned} \tag{28}$$

$$\begin{aligned} p &= -\frac{3}{2}B_0f_Ne^{-3f_N}\{3(H+3B_0'+3)f_N^2 + 3(B_0'+2)f_N + 2\} \\ &= -\frac{B_0\ln x}{6x}\{(H+3B_0'+3)\ln^2 x + 3(B_0'+6)\ln x + 6\}, \end{aligned} \tag{29}$$

$$\begin{aligned} B &= -\frac{1}{2}B_0e^{-3f_N}\{9(H+3B_0'+3)f_N^3 - 9(H+2B_0'+1)f_N^2 - 6(B_0'+1)f_N - 2\} \\ &= -\frac{B_0}{6x}\{(H+3B_0'+3)\ln^3 x - 3(H+2B_0'+1)\ln^2 x - 6(B_0'+1)\ln x - 6\}, \end{aligned} \tag{30}$$

where, again, $H = B_0''B_0 + (B_0')^2$.

3.4 Vinet EOS

A different way of deriving the EOS comes from assuming an interaction potential between neighbor atoms, as Mie did with the Lennard-Jones potential in 1903 [18]. A recent member of this family is the Vinet or *Universal* EOS [19,20], first proposed by Stacey *et al.* in 1981 [4]. The Vinet EOS is the result of using a Rydberg interatomic potential, $E(a) = -\Delta E(1+a)e^{-a}$, where $a = (r - r_0)/l$, r is the interatomic distance, r_0 the minimum energy distance and l is a scaling length. Following Cohen *et al.* [21] the crystal energy is given by

$$E(V) = E_0 + \frac{4B_0V_0}{(B_0'-1)^2} - \frac{2B_0V_0}{(B_0'-1)^2}[3(B_0'-1)(\eta-1)+2]\exp\left\{-\frac{3}{2}(B_0'-1)(\eta-1)\right\} \tag{31}$$

from which the following expressions can be obtained for the pressure and bulk modulus:

$$p(V) = 3B_0\frac{1-\eta}{\eta^2}\exp\left\{-\frac{3}{2}(B_0'-1)(\eta-1)\right\} \tag{32}$$

$$B(V) = -\frac{B_0}{2\eta^2}[3\eta(\eta-1)(B_0'-1) + 2(\eta-2)]\exp\left\{-\frac{3}{2}(B_0'-1)(\eta-1)\right\} \tag{33}$$

where $\eta = (V/V_0)^{1/3} = x^{1/3}$.

Cohen *et al.* [21] have shown that the Vinet EOS excels at reproducing the available experimental data of solids from the extremely soft noble gases or n-H_2 to the extremely hard metals, covalent or ionic compounds. This is likely the reason behind Vinet's EOS recent popularity.

3.5 Holzapfel EOS

Holzapfel [22,9] has been a strong advocate of the need that the solid EOS tend to the known behavior of a Fermi gas under infinite pressure conditions. Most EOS forms are not appropriate to impose this limit, but a generalization of the Vinet form can be shown to achieve it. This is the origin of Holzapfel's AP2 form:

$$p = 3B_0 \frac{1-\eta}{\eta^5} e^{c_0(1-\eta)} [1 + c_2\eta(1-\eta)], \tag{34}$$

where $\eta = (V/V_0)^{1/3}$. The c_0 and c_2 coefficients are given by

$$c_0 = -\ln(3B_0/p_{FG0}), \quad c_2 = \frac{3}{2}(B_0' - 3) - c_0, \tag{35}$$

where $p_{FG0} = a_{FG}(Z/V_0)^{5/3}$ is the pressure of the free electron gas for the atom of atomic number Z, and $a_{FG} = (3\pi^2)^{2/3}\hbar^2/(5m_e) \approx 0{,}023\,369\,\mathrm{nm}^5\mathrm{GPa}$ is the Fermi gas constant. The leading η^{-5} term in the pressure equation ensures that $\lim_{p\to\infty} B' = 5/3$, a result derived from the Thomas-Fermi theory.

The energy can be obtained by integrating $E - E_0 = -3V_0 \int_1^\eta p(\eta)\eta^2 d\eta$. The result is

$$\begin{aligned} E = E_0 + 9B_0V_0 \Big\{ & [\Gamma(-2, c_0\eta) - \Gamma(-2, c_0)]\, c_0^2 e^{c_0} \\ & + [\Gamma(-1, c_0\eta) - \Gamma(-1, c_0)]\, c_0(c_2 - 1)e^{c_0} \\ & - [\Gamma(0, c_0\eta) - \Gamma(0, c_0)]\, 2c_2e^{c_0} + \frac{c_2}{c_0}\left[e^{c_0(1-\eta)} - 1\right] \Big\} \end{aligned} \tag{36}$$

where

$$\Gamma(a, z) = \int_z^\infty t^{a-1}e^{-t}dt \quad (a \in \mathbb{R}, z \geq 0) \tag{37}$$

is the upper incomplete gamma function. Notice that this expression for the energy is different from the version previously published [22,9] that, according to our experience, produces wrong results.

3.6 Spinodal BCNT

The spinodal EOS by Baonza *et al.* [23,24] originates from the literature of EOS for liquids and it was later proved to work well on a collection of very compressible and of very stiff solids [25]. At difference from the previous EOS, that take the equilibrium point as the reference geometry, here the reference is the spinodal point, where the curvature of the $E(V)$ potential changes from positive to negative or, equivalently, $B(V_{\mathrm{sp}}) = 0$. The spinodal represents a critical situation of the crystal, so Baonza *et al.* adopt a critical exponent form:

$$\ln \frac{V}{V_{\rm sp}} = -\frac{K^\star}{1-\beta}(p - p_{\rm sp})^{1-\beta}, \tag{38}$$

where

$$B_0 = \frac{(-p_{\rm sp})^\beta}{K^\star}, \quad B_0' = \frac{\beta B_0}{(-p_{\rm sp})}, \quad \ln \frac{V_{\rm sp}}{V_0} = \frac{\beta}{B_0'(1-\beta)}, \quad \beta \approx 0.85. \tag{39}$$

Fitting parameters are $V_{\rm sp}$, $p_{\rm sp}$, $K^\star$ and, optionally, β, although a $\beta = 0.85$ value has been shown to work well on many cases.

A detailed description of this EOS, as implemented in the original GIBBS code, can be found in ref. [26]. The energy form [26] is

$$E(V) = E_0 + MV_0(-p_{\rm sp})I(y), \tag{40}$$

where

$$M = \frac{K^\star(-p_{\rm sp})^{1-\beta}}{1-\beta}, \quad y = 1 - \frac{1}{M}\ln\frac{V}{V_0}, \quad I(y) = \int_1^y e^{M(1-s)}(s^{1/(1-\beta)} - 1)ds, \tag{41}$$

and the integral $I(y)$ can be calculated numerically up to the desired precision.

3.7 Anton-Schmidt EOS

Anton and Schmidt proposed in 1997 [27] the next functional form that *adequately fitted* the theoretical calculations of a collection of intermetallic compounds. The interest of this empirical EOS is the use of E_∞, supposedly the energy at the dissociation limit, rather than the energy at equilibrium, as a fitting parameter. The formulas for pressure, energy, and bulk modulus are

$$p(V) = -\beta\left(\frac{V}{V_0}\right)^n \ln\left(\frac{V}{V_0}\right), \tag{42}$$

$$E(V) = \frac{\beta V_0}{n+1}\left(\frac{V}{V_0}\right)^{n+1}\left[\ln\left(\frac{V}{V_0}\right) - \frac{1}{n+1}\right] + E_\infty, \tag{43}$$

$$B(V) = \beta\left(\frac{V}{V_0}\right)^n\left[1 + n\ln\frac{V}{V_0}\right]. \tag{44}$$

The condition that $B(V)$ and $B'(V)$ become B_0 and B_0', respectively, when $V \to V_0$ can be used to derive that $B_0 = \beta$ and $B_0' = -2n$. On the other hand, when fitting data for several temperatures, Mayer *et al.* [28] propose using $n+1 = {}^5\!/\!_6 - \gamma_G$, where $\gamma_G = d\ln\theta_D/d\ln V$ is the Debye-Grüneisen parameter.

Table 1
Equilibrium properties of the rock-salt phase of MgO obtained by nonlinear fitting of different EOS.

EOS	V_0 (bohr^3)	B_0 (GPa)	B_0'	B_0'' (GPa^{-1})
Murn.	124.885	177.23	3.505	
Vinet	124.783	165.42	4.524	
AP2	124.797	167.52	4.345	
BM2	124.603	175.76	4.000	−0,022 1
PT2	122.775	267.78	2.000	−0,003 7
BM3	124.824	170.04	4.116	−0,023 6
PT3	124.637	160.80	5.212	−0,090 4
BM4	124.821	169.27	4.171	−0,025 4
PT4	124.847	169.24	4.089	−0,016 2

3.8 Nonlinear fitting

The above equations depend nonlinearly of a collection of parameters, E_0, V_0, B_0, B_0', ... that represent physical properties of the solid at equilibrium and can, in principle, be obtained experimentally by independent methods. The use of a given analytical EOS may have significant influence on the results obtained, particularly because the parameters are far from being independent. The number of parameters has to be considered in comparing the goodness of fit of functional forms with different analytical flexibility. The possibility of using too many parameters, beyond what is physically justified by the information contained in the experimental data, is a serious aspect that deserves consideration. All these aspects: the nonlinear fitting, the quality of the fitting, the possible overfitting as the number of parameters increase, the interdependence between the parameter estimates, and the confidence limits of the estimates are significant issues that raise important statistical concerns [29–31].

In GIBBS2, the nonlinear fitting of the analytical EOS forms is done by means of a standard Levenberg-Marquardt (LM) algorithm, as implemented in the `lmdif1()` routine of MINPACK [32]. The jacobian is calculated by a forward difference numerical approach, but an analytical form adapted to each EOS could be developed if required. Our tests show that the LM algorithm converges provided that the initial estimation of the parameters is good enough.

Table 1 shows the equilibrium properties of MgO obtained by fitting several of the previously described EOS to exactly the same set of theoretical $E(V)$ data.

Table 2
Linear and nonlinear fitting of BM3 and BM4 to the MgO rock-salt data. The fitting model is identical, no matter the procedure, but only the linear method is able to converge high order members, like BM10.

model	fit	V_0 (bohr^3)	E_0 (Ry)	B_0 (GPa)	B_0'	B_0'' (GPa^{-1})
BM3	nonlin.	124.823661	-34.344311	170.0414	4.115715	—
BM3	lin.	124.823661	-34.344311	170.041394	4.115715	-0.023630
BM4	nonlin.	124.821340	-34.344283	169.2683	4.171190	-0.025412
BM4	lin.	124.821340	-34.344283	169.268283	4.171190	-0.025412
BM12	lin.	124.824369	-34.344285	169.319792	4.151925	-0.024118

The most relevant result is the dispersion of values. Excluding the dubious PT2 result, the bulk modulus goes from 161 to 177 GPa, and B_0' from 3.5 to 5.2. The trend shows that the successive members of a family tend towards a less disperse solution, but BM4 and PT4 are still far from perfect and, more importantly, the convergence of the nonlinear fitting is a very delicate process for such high order members of the BM and PT families. Furthermore, there is a strong correlation between some of the nonlinear parameters, for instance 86–98% between B_0, B_0', and B_0'' in BM4.

4 Strain polynomials

The difficulty in using higher orders in the BM or PT family comes from the nonlinear character of the fitting procedure and yet, both families originate from a polynomial expansion of the energy. Robust and essentially exact methods have been developed for the fitting of polynomials up very high orders. Is the result of linear and nonlinear fitting the same? Table 2 shows that, as far as the nonlinear fitting converges, the BMn EOS is the same no matter the fitting procedure used, and the same happens for the PT and other strain polynomial families. The immediate advantage is that it is now possible to examine the results of BM or PT to any expansion order required by the data.

4.1 Linear fitting for the BM family

Before going to more physical concerns, like the convergence with the polynomial order, let us examine how the linear fitting is done in the case of the BM family, and how the solid properties are extracted. The first step is transforming the input volume data into the eulerian strain values

$$f = \frac{1}{2}\left[\left(\frac{V_r}{V}\right)^{2/3} - 1\right], \quad V = V_r(2f+1)^{-3/2}, \tag{45}$$

where V_r is a reference volume. Essentially, any finite positive value can be used as V_r if the strain polynomial includes the $c_1 f^1$ linear term and we determine the equilibrium position by minimizing the polynomial with respect to f.

The second step is the linear fitting to the polynomial form

$$E(f) = \sum_{k=0}^{n} c_k f^k. \tag{46}$$

It is important using a robust and efficient method, like QR or Cholesky decomposition [33], for solving the least squares equations. The failure to do so in the original GIBBS produces some of the worst causes of instability of the code.

Next, we need to determine the minimum of the $E(f)$ curve to obtain the equilibrium properties of the crystal phase. In practice this has shown to be a delicate step, and we have finally recurred to a safe combination of Newton and bisection search. Under some circumstances the polynomial will have no minimum in the range of fitted volumes (for instance high pressure phases) and this must be adequately detected and taken into account.

We also need obtaining the derivatives of the static energy with respect to the volume. Up to the third or fourth derivatives are used in the thermal models (e.g. in the Debye model). The best method that we have found is based on the chain rule: use the derivatives of the energy *versus* the strain, plus the derivatives of the strain *versus* the volume, to get the final desired properties. Therefore, the next step is determining

$$E_{mf} \equiv d^m E/df^m = \sum_{k=m}^{n} c_k k(k-1)...(k-m+1) f^{k-m}, \quad 0 \leq m \leq n. \tag{47}$$

The derivatives of the strain, $f_{mV} = d^m f/dV^m$, are easily determined by a simple recursive procedure

$$f_{(m+1)V} = -\frac{3m+2}{3V} f_{mV} \quad m = 1, 2, ... \tag{48}$$

where

$$f_{1V} = -(1/3V)(V_r/V)^{2/3}. \tag{49}$$

It must be stressed that this is the only part of the procedure that changes when the definition of the strain is modified.

The derivatives $E_{nV} = d^nE/dV^n$ can now be written in terms of the E_{mf} and f_{mV} components. This is a place where a computer algebra program is very helpful, and we have used MAXIMA to get

$$E_{1V} = E_{1f} f_{1V}; \tag{50}$$

$$E_{2V} = f_{1V}^2 E_{2f} + f_{2V} E_{1f}; \tag{51}$$

$$E_{3V} = f_{1V}^3 E_{3f} + 3 f_{1V} f_{2V} E_{2f} + f_{3V} E_{1f}; \tag{52}$$

$$E_{4V} = f_{1V}^4 E_{4f} + 6 f_{1V}^2 f_{2V} E_{3f} + (4 f_{1V} f_{3V} + 3 f_{3V}^2) E_{2f} + f_{4V} E_{1f}; ... \tag{53}$$

Some of the most important static properties of the solid can be obtained immediately in terms of the E_{mV} derivatives:

$$p = -E_{1V}; \tag{54}$$

$$B = V E_{2V}; \tag{55}$$

$$B' = -\frac{V E_{3V}}{E_{2V}} - 1; \tag{56}$$

$$B'' = E_{2V}^{-3} [V (E_{4V} E_{2V} - E_{3V}^2) + E_{3V} E_{2V}]; ... \tag{57}$$

4.2 Linear fitting for an indefinite number of strain polynomial families

The eulerian strain is just one form of an indefinite number of appropriate strain formulations [6]. As we have already described, the *natural* strain, $f = (1/3)\ln x$, produces the Poirier-Tarantola (PT) family [17]. The lagrangian strain, $f = (x^{2/3} - 1)/2$, was used in 1970 by Thomson [34] for a fourth order EOS, so we have called Thomson family to the lagrangian strain polynomials. The *infinitesimal* strain, $f = 1 - x^{-1/3}$, was proposed in 1938 by Bardeen [35]. All those strain forms evaluate to zero at the reference volume, but we can use many other functions of the volume ($x^{\pm 1}$, $x^{\pm 1/3}$, ... V) as an appropriate origin for an *strain* EOS family. In all cases, the techniques of the previous section can be generalized.

Table 3 shows the equations needed to generalize the procedure of section 4.1 to any strain type variable. Except for the f_{mV} derivatives, everything else remains unchanged.

4.3 Calculating error bars with the average of polynomials

The capability of using high order strain polynomials introduces two new problems: how to determine the best fitting and how to avoid using parameters statistically unsupported by the data, i.e. avoid *overfitting* [30]. The idea of fitting a set of polynomials to the $E(V)$ data and defining a polynomial

Table 3
Generalizing the polynomial strain fitting to arbitrary strain forms only requires determining the specific form of the f_{mV} derivatives.

strain	EOS	f	$f_{(m+1)V}$	f_{1V}	s
eulerian	BM	$\frac{1}{2}\left(x^{-2/3}-1\right)$	$(3m+2)sf_{mV}$	$-\frac{x^{-2/3}}{3V}$	$-\frac{1}{3V}$
lagrangian	Thomson	$\frac{1}{2}\left(x^{2/3}-1\right)$	$(3m-2)sf_{mV}$	$-\frac{x^{2/3}}{3V}$	$-\frac{1}{3V}$
natural	PT	$\frac{1}{3}\ln x$	msf_{mV}	$\frac{1}{3V}$	$-\frac{1}{V}$
infinitesimal	Bardeen	$1-x^{-1/3}$	$(3m+1)sf_{mV}$	$\frac{(1-f)^4}{3V_r}$	$-\frac{1}{3V}$
x^3		x^3	$(3m-1)sf_{mV}$	$\frac{x^{1/3}}{3V}$	$-\frac{1}{3V}$
V		V	0	1	0

average was already explored in the PhD work of M. A. Blanco [26] and it is a fundamental technique of the original GIBBS code [1]. Given a data set with N_i points and a strain polynomial of degree n_i, the quality of the fit is determined by

$$q_i = [N_i/n_i][\mathcal{S}_i/\mathcal{S}_{\min}], \tag{58}$$

where

$$\mathcal{S}_i = \sum_{k=1}^{N_i}[E_k - E(V_k)]^2 \tag{59}$$

is the square sum of residuals and $\mathcal{S}_{\min} = \min_i \mathcal{S}_i$. At difference from $\mathcal{S}_i$, the quality q_i gives more weight to those polynomials that fit better more data with less parameters, in the spirit of statistical measurements like Akaike's information criterion [36]. Based on the q_i values, a normalized probability distribution is defined as

$$\mathcal{P}_i = \frac{e^{-q_i}}{\sum_k e^{-q_k}} \qquad \text{(method1)}, \tag{60}$$

or

$$\mathcal{P}_i = \frac{e^{-q_i^2}}{\sum_k e^{-q_k^2}} \qquad \text{(method2)}. \tag{61}$$

The $\mathcal{P}_i$'s weights are then used to: (1) average the polynomial coefficients, thus defining an *average polynomial*; and (2) average the equilibrium properties of the individual polynomials, determining in this way the probability distribution of the properties (mean, standard deviation, skew, kurtosis, etc). The standard deviation, in particular, is used to provide an estimated error bar for the determined properties. Volume dependent properties like $p(V)$, $B(V)$, and $B'(V)$ are commonly determined from the average polynomial, although the $\mathcal{P}_i$'s weights can also be used to average the values of the individual polynomials thus providing an estimated error bar too.

The trick used in GIBBS of eliminating points from the extremes of the volume range is no longer used for normal data, being substituted by a far more general and useful bootstrap method when necessary (see section 5.1).

We are now in conditions to test the convergence of the strain fit polynomials and families. The first relevant test is contained in Fig. 1. The figure evidences that: (a) there are important differences between the equilibrium properties obtained from the low order polynomials, like BM3 and BM4, but the high order BM polynomials converge to rather similar values; (b) the different families have equilibrium properties that coincide within the estimated error bars; and (c) the eulerian strain tends to show slightly better error bars, but any strain polynomial family works nicely.

Fig. 2 presents another important evidence: once we reach some level the error bars do not change importantly if we continue adding polynomials with increasing degree to the average pool. When very high orders are reached the bars start to grow slowly, perhaps showing the first effects of overfitting. As an extreme test we have tried both, individual and average of polynomials fittings up to degree 40 and 50 with our main data set of 129 points and the results remained stable. From our experience with this and other data sets, there is a degree (around 10–14) for which the error bars of all equilibrium properties are small and this constitutes a reasonable and safe level to use.

We have performed many other tests, some of which are worth describing briefly. The results of the average fitting and, in particular, the size of the error bars have remained stable when: (a) we have reduced the size of the data set, without changing the range, to 65, 33, and 17 by selecting one every 2, 4, and 8 points, respectively from the original 129 points; (b) we have increased the data set, again keeping the volume range, to 257, 513, 1025, and 2049 by calculating new points regularly spaced between the originals; (c) we have produced a new set of 129 points for a very small range of 0.2 bohr^3 around the known equilibrium geometry. The last experiment deserves some comment. It is easy to assume, and it has been argued several times, that using a very fine grid closely around the equilibrium geometry is the best strategy to determine the equilibrium properties but, in practice, the limited precision of the calculations, set, for instance, by the convergence conditions, places a lower bound on the grid interval that contains a meaningful and independent information.

A final word of caution is important. The behavior of the polynomials is not guaranteed outside the range of volumes of the fitted data. In other words, polynomials should never be used for the extrapolation of data. According to our tests, however, it is safe using the equilibrium properties obtained with the average of strain polynomials method as the parameters of the analytical EOS, described in previous section, that have been designed with the purpose

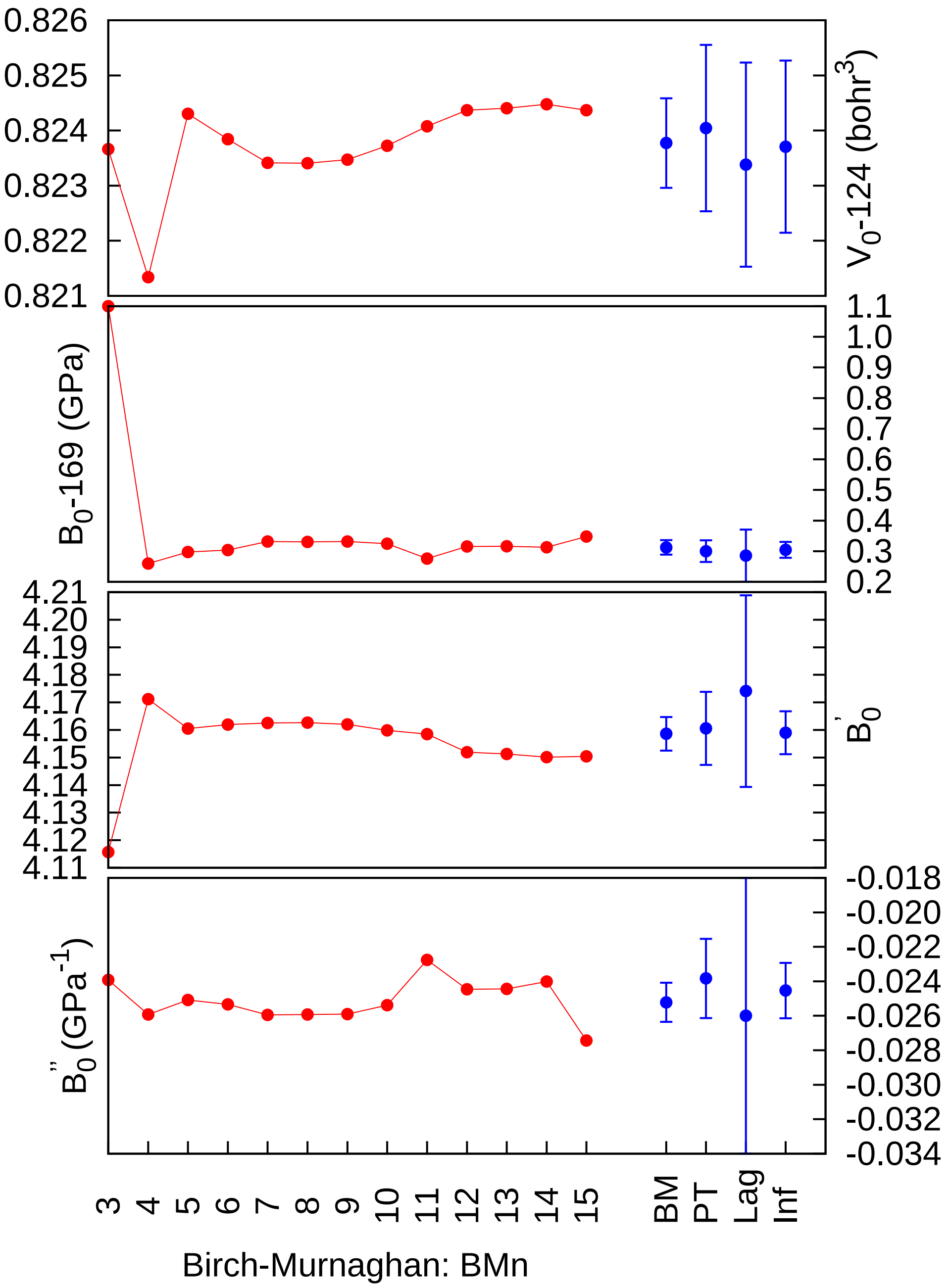

Fig. 1. Equilibrium properties of MgO (data set Y1) obtained from the fit of a Birch-Murnaghan polynomial of degree 3 to 15 (left points, represented as big red dots), and from the polynomial averages of the BM, PT, Thomas and Bardeen families (right points, represented as blue dots with error bars). The properties of the BM2 model are out of the used scale.

of doing a reasonable extrapolation.

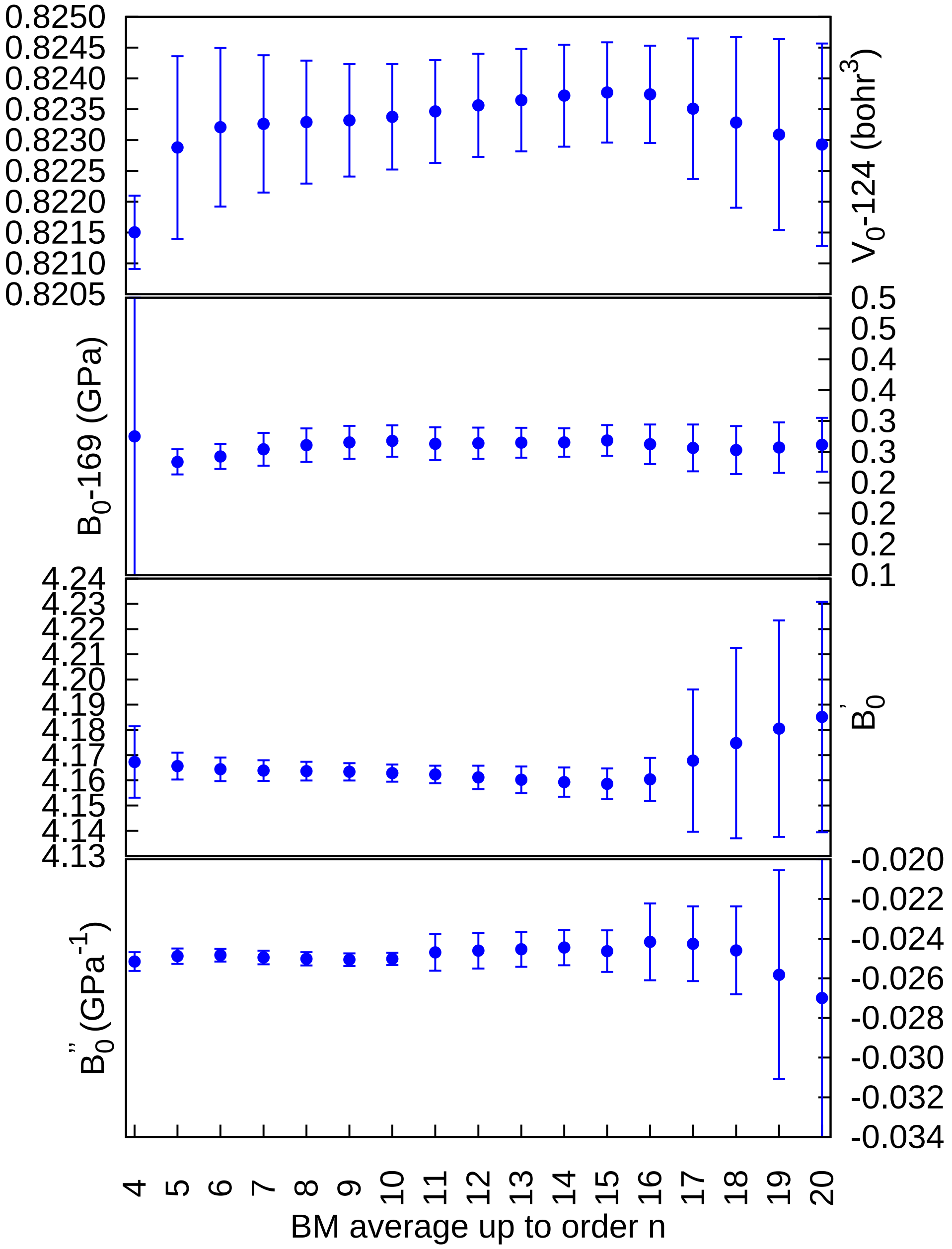

Fig. 2. Equilibrium properties of MgO (data set Y1) obtained from the average of BM strain polynomials of degree 2 to n, where n is represented as the abscissa.

4.4 Determining error bars with bootstrap resampling

The estimated error bars are an important asset of the average of strain polynomials method so, given the arbitrariness in the choice of weights, an independent confirmation of their relevance is convenient. Bootstrap resampling [30,31] is an established statistical technique that consist on extracting and analyzing random samples from a large dataset. We have worked with our original 129 points dataset for MgO, fitting each sample with a BM6 fixed

Table 4
BM6 fitting to a resampling of the original 129 data points for rock-salt MgO. The results of the average of BM polynomials are shown from comparison.

sample	points	V_0 (bohr^3)	B_0 (GPa)	B'_0	B''_0 (GPa^{-1})
10^2	50–76	124.82385(36)	169.3117(78)	4.1618(13)	-0.02489(17)
10^3	48–79	124.82381(34)	169.3105(82)	4.1620(12)	-0.02491(15)
10^4	45–84	124.82381(34)	169.3107(75)	4.1620(11)	-0.02491(15)
10^5	39–89	124.82381(34)	169.3107(75)	4.1620(11)	-0.02491(15)
10^6	39–91	124.82381(34)	169.3107(75)	4.1620(11)	-0.02491(15)
avg14BM		124.82372(83)	169.314(20)	4.1593(58)	-0.02461(85)

degree polynomial, and the results are presented in table 4.

The results make clear that some 10^3 samples are enough to converge the results when the data is truly smooth and consistent. It also shows that the bootstrap error bars are some one third to one fifth of those obtained with the average of polynomials method. Keeping in mind that the bootstrap method is considered to provide optimistic values for the standard deviations [31], we will keep the error bars described in previous section, 4.3, as a more conservative and safe approach.

In any case, the bootstrap resampling represents a good addition to the basic techniques in the GIBBS2 code, and we will see how it will help us to detect problematic points in the input dataset.

5 Detection and treatment of problematic data

Ordinary least squares fitting is very sensitive to the presence of noise [31,37], be it in the form of outliers, i.e. points that deviate wildly from the normal trend of the sample, jumps or steps, i.e. volume ranges that differ from the general trend by an additive energy, or random small errors that differ from point to point. The strain average method developed in section 4.3 is very effective in detecting those problems in the form of large error bars for the equilibrium properties. Tables 5 and 6 show the level of precision than can be expected on a collection of typical crystals and electronic structure calculations. We can see that smooth $E(V)$ curves can be expected to provide 4–7 significant figures of the equilibrium volume and bulk modulus, 3–5 significant figures of B'_0, and 2–3 in the case of B''_0. Much worse errors can be used to detect calculations with problems.

Table 5
Equilibrium parameters with estimated error bars for a collection of elements in the bcc, fcc, and diamond (A4) phases. All results correspond to FPLAPW (Full Potential Linear Augmented Plane Waves) calculations in the LDA approximation, and have been done with the RUNWIEN interface [38] of the WIEN2K code (release 10) [39,40]. The fitting errors in the total energy occur at the μRy level and have been ignored in the table.

	phase	points	V_0 (bohr^3)	B_0 (GPa)	B_0'	B_0'' (GPa^{-1})
Li	bcc	103	128.3203(83)	15.078(12)	3.3570(48)	-0.1820(23)
C	A4	50	37.22569(13)	468.44(13)	3.6636(23)	-0.00840(79)
Na	bcc	61	224.4463(60)	9.1635(84)	3.751(17)	-0.343(49)
Al	fcc	151	106.587(38)	83.45(18)	4.657(20)	-0.0879(21)
K	bcc	100	432.56(81)	4.508(51)	3.75(21)	-0.98(97)
Ca	fcc	61	256.073(44)	18.044(92)	2.74(14)	0.87(21)
V	bcc	97	84.690(10)	211.51(65)	3.796(79)	-0.024(11)
Cu	fcc	61	73.4879(66)	189.96(24)	5.003(21)	-0.0351(21)
Ge	A4	46	150.4627(32)	72.031(98)	4.883(19)	-0.065(29)
Rb	bcc	101	523.81(37)	3.610(23)	3.75(10)	-1.50(21)
Nb	bcc	80	115.6270(21)	191.36(16)	3.669(23)	-0.0222(32)
Mo	bcc	74	101.8167(14)	293.42(11)	4.181(15)	-0.0195(26)
Rh	fcc	61	89.2381(51)	319.80(47)	5.005(39)	-0.0256(21)
Pd	fcc	61	95.4633(22)	229.97(38)	5.362(21)	-0.0415(51)
Ag	fcc	61	108.0954(89)	139.683(86)	5.688(14)	-0.0601(27)
Sn	A4	56	229.312(10)	45.230(47)	4.914(26)	-0.117(26)
Pb	fcc	58	196.1186(64)	52.46(10)	4.846(24)	-0.115(44)

Tables 5 and 6 show that, in the case of elements or compounds with the same crystal structure we can observe an inverse correlation between the bulk modulus and the cell volume: the smaller V_0 the larger B_0 and vice versa. Diamond establishes the record of largest B_0, a property related, but not identical, to the hardness of the material. The possibility that osmium could beat the B_0 record of diamond originated a vivid controversy a few years ago [41,42]. A principal role in this debate can be attributed to the errors associated to the fitting of analytical EOS forms to experimental $p(V)$ or theoretical $E(V)$ data [43]. The use of statistical fitting techniques like the ones developed and discussed here could have revealed early large error bars in the initial results originating the controversy.

Table 6
Equilibrium parameters with estimated error bars for a collection of crystals in the rock-salt (B1), CsCl-type (B2), and perovskite phases ($E2_1$). All results correspond to FPLAPW (Full Potential Linear Augmented Plane Waves) calculations in the LDA approximation, and have been done with the RUNWIEN interface [38] of the WIEN2K code (release 10) [39,40]. The fitting errors in the total energy occur at the μRy level and have been ignored in the table.

	phase	points	V_0 (bohr^3)	B_0 (GPa)	B_0'	B_0'' (GPa^{-1})
LiF	B1	61	101.05483(67)	87.1125(60)	4.3042(13)	-0.04979(38)
LiCl	B1	61	207.16832(77)	41.1285(85)	4.4208(28)	-0.1193(23)
LiBr	B1	61	253.106(14)	33.380(23)	4.451(11)	-0.1435(63)
LiI	B1	61	330.4771(44)	25.7741(89)	4.4984(93)	-0.2107(76)
NaF	B1	61	154.40(13)	61.6(13)	4.75(61)	-0.16(36)
NaCl	B1	61	275.7674(14)	32.1958(72)	4.7846(37)	-0.1668(75)
NaBr	B1	60	326.82(13)	26.54(12)	4.85(23)	-0.299(93)
NaI	B1	61	414.2209(75)	20.8000(97)	4.8018(96)	-0.274(23)
KCl	B1	60	378.5993(34)	24.2768(73)	5.0081(95)	-0.2626(54)
KBr	B1	59	436.410(16)	20.275(16)	4.9879(70)	-0.298(18)
KI	B1	60	534.7816(47)	16.017(10)	4.9725(60)	-0.391(34)
RbCl	B1	61	437.6956(72)	21.7990(81)	5.140(13)	-0.3050(57)
RbBr	B1	61	500.083(22)	18.4292(81)	5.145(26)	-0.379(11)
RbI	B1	61	605.51(18)	14.73(12)	5.04(22)	-0.39(38)
CsF	B1	59	331.71(97)	34.0(12)	5.0(11)	-0.15(84)
CsCl	B2	61	422.97(52)	24.22(34)	4.86(40)	-0.07(33)
CsBr	B2	61	478.612(30)	21.153(15)	5.226(31)	-0.371(31)
CsI	B2	61	575.385(29)	16.994(14)	5.157(22)	-0.426(50)
GaP	B1	61	264.61(16)	89.9(16)	4.57(50)	-0.07(14)
GaAs	B1	61	297.219(32)	74.28(13)	4.666(52)	-0.095(22)
GaSb	B1	61	374.0041(50)	55.733(81)	4.814(11)	-0.099(17)
MgO	B1	61	121.680(17)	171.78(98)	4.21(14)	-0.0187(82)
$KMgF_3$	$E2_1$	50	403.55(11)	86.24(63)	4.46(30)	-0.09(13)
$KZnF_3$	$E2_1$	61	423.8305(91)	95.997(21)	4.7494(93)	-0.0581(16)

Bulk modulus increases almost linearly with pressure, and this was the foundation of Murnaghan EOS, quite successful for its simplicity. The slope at the equilibrium position, B_0', is between 4 and 5 for most materials and the curvature, B_0'', is very small and negative. Tables 5 and 6 are good examples of both trends. Diamond shows again as one of the records of low curvature. The large curvature of K and Rb, and the unusual positive value of $B_0''(\mathrm{Ca})$ also stand out, but the large error bars associated to those extremes put doubts on their consistency. The useful role of the error bars on B_0 and its derivatives on detecting problematic fittings is clearly evidenced by the shown results.

5.1 Detection of outliers

The sensitivity of ordinary least squares fitting to the presence of outliers is an inherited consequence and it can be avoided by using a more robust measurement of the difference between the analytical model and the input data set, for instance minimizing the sum of absolute value deviations [30,37,31]. Unfortunately, the standard robust regression techniques give rise to non linear equations, which means problems of convergence as the number of parameters of the model increases. We have included the least absolute deviation among the techniques available to the GIBBS2 user, but we will explore here a different method for detecting and removing the outliers.

Our preferred technique is a combination of bootstrap sampling and average of strain polynomials of different degree. The bootstrap produces samples of different composition and number of points, and eventually some of them will be free of the problematic points. Those clean samples will tend to yield the best fits and thus the main contributions to the average polynomial. Given the random nature of the resampling production it may happen that no clean sample is produced in a run, but those cases will be revealed by the large error bars of the equilibrium properties. A failed run can be stopped or repeated with a sample of increased size. In general, the larger the sample the more probable that the method succeeds at producing one or more clean fittings.

Fig. 3 and Table 7 show a typical data set with outliers, produced artificially by a random modification of 12 of the 129 point data set of MgO. Using a default 50% chance of including any given point, there is a probability of $(1/2)^{12} \approx 0.024\%$ that a random sample will be free of all the outliers. For a set with n outliers and a run of N samples, the probability of failure is $(1 - 1/2^n)^N$. Assuming $n = 12$, the probability of failure is 78.3% for $N = 10^3$ but only 8.7% if $N = 10^4$. The first run of 10^3 samples in Table 7 failed to produce a good fit, as the large error bars in B_0 and derivatives clearly show. The second run, also with 10^3 samples, and the third with 10^4 samples were successful and their results agree within their small error bars.

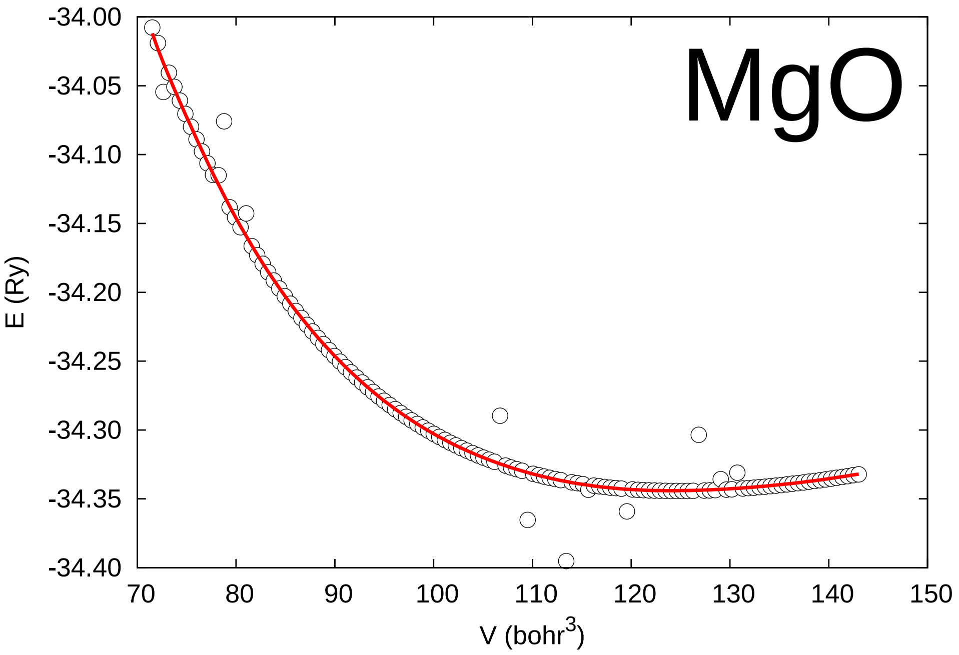

Fig. 3. Data sample for the MgO crystal containing a 10% of outlier points. The solid (red) line is, in fact, a superposition of two indistinguishable fitted lines that show a rather different behavior at the equilibrium point (see Table 7).

Table 7
Results of the bootstrap plus strain polynomial average method on a data sample of MgO containing some 10% outlier points. The first run with 10^3 samples was not able to detect and remove the outliers, and the large error bars provide the telltale evidence of the problem. The second and third run produced a satisfactory result.

Samples	V_0 (bohr^3)	B_0 (GPa)	B_0'	B_0'' (GPa^{-1})
10^3	124.12(86)	179(40)	7.3(6.6)	$-0.3(2.4)$
10^3	124.82323(98)	169.337(51)	4.166(19)	$-0.0258(53)$
10^4	124.8242(17)	169.294(56)	4.159(25)	$-0.0224(44)$

5.2 Step discontinuities

Jumps in the $E(V)$ curve, like those shown in the figure 4, are typical of calculations that proceed by adding up terms (integrals, interaction potentials, ...) coming from progressively distant neighbor shells. Different cell volumes can result in adding up different number of shells and then energy discontinuities appear. Techniques exist to avoid this problem: from special methods to converge the shell sums up to a high precision, to simply conserving the list of the shells added up and using it for all the volumes. We will examine the case of the calculations gone wrong and we will try to determine the derivative information from the discontinuous $E(V)$ curve.

The method starts by defining a functional form able to accommodate the discontinuities:

$$Q^m(x) = P(x) + S_m(x) = P(x) + \sum_{i=1}^{m} \delta_i H(x - t_i) \tag{62}$$

where $P(x)$ is a continuous function and

$$H(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x \geq 0 \end{cases} \tag{63}$$

is a Heaviside step function. In other words, $\delta_i H(x - t_i)$ produces a jump of δ_i to the right of the t_i abscissa value. The $S_m(x)$ discontinuous function describes m jumps.

Our method for detecting the position and height of a jump is based on linear interpolation. First, the energy of each point is compared with the value interpolated from its left and right neighbor points. The jumps stand out as a couple of successive points that differ markedly from their predicted values. Once the position of the jump is known, the height of the step is also determined by linear interpolation. To this end, the points to the left of the jump are used to extrapolate the value that it should have the point to the right, and vice versa. The differences between the predictions and the actual values produce two estimates of the step, and both are finally averaged. It must be noticed that at least two or three points must form each continuous interval, otherwise the linear interpolation will only produce erratic results.

We have also tried a different method based on a linear squares fitting of the $Q^m(x)$ function to the data. The δ_i step parameters can, in fact, be obtained by a linear least squares method, but the result will only be good if the smooth part, the $P(x)$ function, is also a good representation of the data once the jump is removed. Accordingly, we have devised a procedure that is based on two steps that are repeated successively until convergence. First, the approximate step function, $S_m(x)$, is substracted from the data and a smooth function $P(x)$ is fitted to the result using, for instance, the average of polynomials method. Second, the smooth function $P(x)$ is now removed from the data and a least squares fitting to the left part is done to refine the δ_i's. The method works but the convergence of this procedure is very slow and it does not improve, in general, upon the result obtained with the much simpler linear interpolation.

Fig. 4 represents a synthetic test case for the jump detection routines. Starting from the MgO data set described in section 2 we added two steps of $-0{,}01$ and $+0{,}015\,\mathrm{Ry}$ for $V > 80$ and $V > 120$ bohr^3, respectively. The linear interpolation method detected both steps and proposed a correction of $-0{,}009\,998\,025$ and $+0{,}014\,999\,49\,\mathrm{Ry}$, quite close to the exact values. In fact, once the proposed correction was applied, the recovered $E(V)$ curve was able

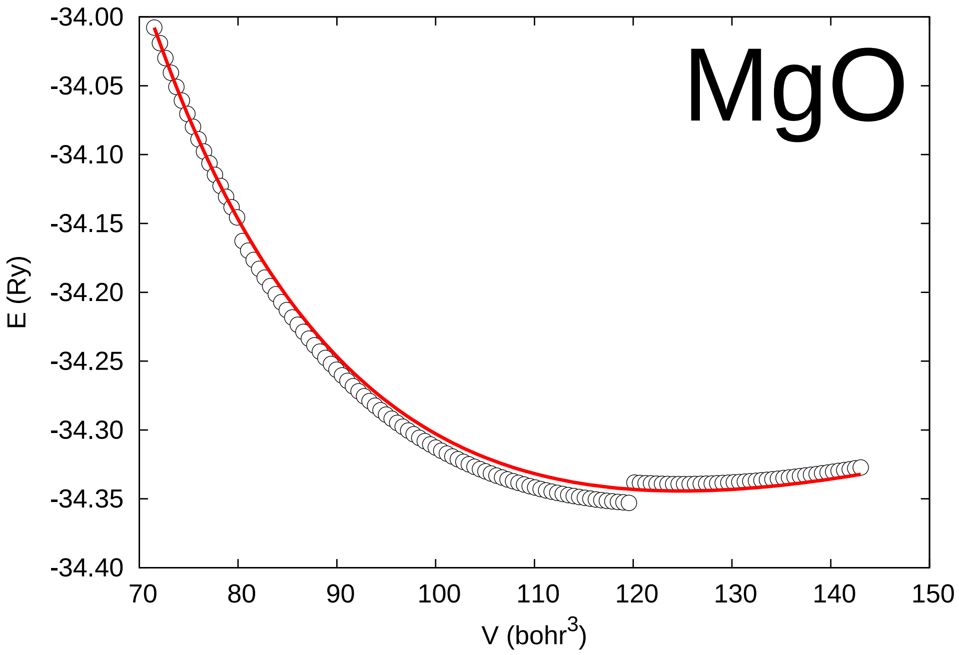

Fig. 4. Dots represent a data set with two jumps. The methods described in text corrected the jumps and produced the continuous curve.

to reproduce the equilibrium properties of the original data set almost exactly: for instance B_0'' was estimated as $-0.0240(23)$ GPa^{-1} compared to the original $-0.0243(19)$ GPa^{-1}. This was a particularly good case, given the extreme smoothness of the data once the jumps were discarded. In many real cases we have found that jumps appear together with other type of noise and the methods examined here can only do an approximate work at correcting the data.

6 The octave fitting routines

The fitting routines are designed in such a way that the user can, either, work interactively with the data or prepare a small script to perform the desired task. Datasets are contained in files with the following structure:

```
# File: mgo-sety1.dat
# MgO: Quantum-espresso, LDA.
# Mg -> von Barth-Car norm-conserving pseudopotential.
# O --> TM norm-conserving pseudopotential.
# z 1
# volume bohr^3
# energy ry
 71.519556379847018 -34.007753979999997
 72.078302914062903 -34.019008200000002
 72.637049448278788 -34.029926469999999
 73.195795982494673 -34.040520219999998
```

```
...
```

The comment lines describing the volume and energy units are very important. The fitting routines work internally in atomic units (bohr^3 and hartree), but the routine in charge of reading the data file can detect and transform different input units, and also the number of molecular formulas contained in a unit cell. This can be easily adapted to other useful data, like the total number of electrons in the molecular formula, required by the Holzapfel AP2 EOS.

Let us see a typical interactive session:

```
# Get help on the use of any routine:
help readEVdata
# Read in some data file:
[V,E] = readEVdata('mgo-sety1.dat');
# Any volume can be used as reference.
# For instance the mean or the median:
Vref = mean(V);      # or Vref = median(V);
# Fit to an average of BM polynomials:
c = avgstrainfit(V,E,Vref,:,:,:,1);
# Get and plot properties:
vfinegrid = linspace(min(V),max(V),201);
prop = straineval(c,Vref,vfinegrid,:);
plot(V,E,'ob', vfinegrid,prop.E,'-r');
xlabel('V (bohr^3)'); ylabel('E (hartree)'); grid('on');
print('plot-V-E.eps','-depsc');
plot(prop.p,prop.B,'-o');
xlabel('p (GPa)'); ylabel('B (GPa)'); grid('on');
print('plot-p-B.eps','-depsc');
...
```

This session illustrates some important features: (a) all the routines are fully commented and adapted to the online help system of OCTAVE; (b) many parameters in the routines have appropriate default values; (c) many routines have several levels of printing, determined from the input value of the variable `LOG`.

An alphabetic list of the routines, with a brief description of its function follows:

allfits.m — Given a single data set, perform all implemented types of fitting.

asturfit.m — Given a single data set, perform a standard task of fitting, producing the most useful plots, and writing files to further processing by other codes.

avgstrainfit.m — Fit the $E(V)$ data to an average of strain polynomials.

checknoise.m — Analysis of the noise (outliers and steps) of a suspicious datafile.

errformatf.m — Convert a number x and error dx to the `x(dx)` form.
figures.m — Determine the actual number of significant figures (trailing zeros discarded) of a number or a vector.
guessjumps3.m — Check for the existence of jumps in the data for a general $y(x)$ curve and, eventually, correct the values of y.
nlf.m — Non-linear fitting of several analytical EOS forms to the $E(V)$ data. This routine uses **leasqr** from the OPTIM package, and **gamma_inc** from the GSL package, both in OCTAVE-FORGE.
noisify.m — Add artificial noise to a set of data. This is only used to test the noise analysis routines.
readEVdata.m — Read a set file with $E(V)$ data.
strain2volume.m — Convert strain values into volumes.
strainbootstrap.m — Bootstrap resampling applied to a strain polynomial of fixed degree or to an average of strain polynomials.
strainevalE.m — Evaluate the energy for a strain polynomial.
straineval.m — Evaluate a collection of properties, including the energy, for a strain polynomial.
strainfit.m — Fit to a strain polynomial of fixed degree. This is used within several other routines, like **avgstrainfit.m** or **strainbootstrap.m**.
strainjumpfit.m — Iterative fitting to a strain polynomial and to a step function. The **guessjumps3.m** does a better and simpler job.
strainmin.m — Get the position of the minimum of a strain polynomial.
strainspinodal.m — Get the spinodal point, if present, of a strain polynomial.
volume2strain.m — Convert volume values into several kinds of strain.

The **allfits.m**, **asturfit.m**, and **checknoise.m** have been designed to produce a standalone run on a single datafile. The rest of the routines have been designed to be used interactively or within an OCTAVE script file.

We provide also several example data files containing $E(V)$ values:

mgo-y1.dat — Data from the calculation on MgO described in 2. This can be complemented with the data in **mgo-y1-compress.dat** and **mgo-y1-expand.dat**, for smaller and larger cell volumes, respectively.
w2k-lda-li.dat — Data from the WIEN2K FPLAPW calculation on bcc Li using the LDA functional. The data for Na, K, and Rb is also included in the corresponding files.

Finally, some script tests (**test01.m**, ...) are included, together with their corresponding output, to check the good behavior of everything. Run any example as:

```
   octave -q test01.m > test01.mylog
   diff test01.mylog test01.log
```

and look for non-trivial differences. Make sure that the OCTAVE version includes the OCTAVE-FORGE routines if you want to run the **nlf.m** routine.

7 Conclusions and outcome

A number of results can be extracted from the statistical analysis of theoretical $E(V)$ data described in this article:

(1) The conventional procedure of fitting nonlinearly an analytical EOS (BM3, Vinet, ...) can introduce a significant error in the equilibrium properties and in the derivatives calculated from $E(V)$ theoretical data.

(2) Birch-Murnaghan, Poirier-Tarantola, and several other EOS families come from a polynomial expansion in some expression of the strain. Whereas the traditional treatment produces and fits an EOS expression that is non linear in the set of $\{E_0, V_0, B_0, B'_0, B''_0, ...\}$ parameters, the result is exactly equivalent to the one produced by the linear fitting to the original polynomial.

(3) The fitting to the strain polynomials is strictly linear and it can be carried efficiently and robustly up to very high polynomial orders.

(4) There is an unlimited number of strain functions, each giving rise to a polynomial strain family of EOS.

(5) We have introduced an scheme of weights for averaging strain polynomials of any family based on: the square sum of residuals relative to the best fitting achieved, the number of data fitted, and the degree of the polynomial. The weights are used both to average the polynomial coefficients and the equilibrium properties of the several fits. This provides a full statistics, including error bars in the form of standard deviations, of the properties of the solid.

(6) The predictions of the different average strain EOS families do agree within the estimated error bars.

(7) The error bars obtained by the polynomial average method are more conservative (i.e. larger) than those obtained by bootstrap resampling.

(8) The polynomial average method, combined with a sufficiently large bootstrap resampling, is able to produce a robust fitting to theoretical data sets containing a few outlier points, and detect efficiently the outliers.

(9) We have developed a method for detecting, and eventually removing jumps in the $E(V)$ input data.

(10) This research has been introduced in the new version, soon to be publicly available, of the GIBBS code, far more robust than the previous version. A collection of OCTAVE routines are provided with this article for easy testing and further improvement of the described methods.

8 Acknowledgements

The development of this code has been done under financial support from the Spanish Ministerio de Ciencia y Tecnología and the ERDF of the European Union (project no. CTQ2006-02976). AOR is indebted to the Spanish Ministerio de Educación for a FPU grant. The authors belong to the MALTA group (CSD2007-0045 project, MEC Consolider Ingenio 2010 program).

References

[1] M. A. Blanco, E. Francisco, V. Luaña, GIBBS: isothermal-isobaric thermodynamics of solids from energy curves using a quasi-harmonic Debye model, Comput. Phys. Commun. 158 (2004) 57–72, source code distributed by the CPC program library: `http://cpc.cs.qub.ac.uk/summaries/ADSY`.

[2] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. Dal Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, R. M. Wentzcovitch, QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials, J. Phys.-Condens. Matter 21 (39), preprint: `http://arxiv.org/abs/0906.2569`.

[3] V. N. Zharkov, V. A. Kalinin, Equations of state for solids at high pressure and temperatures, Consultants Bureau, New York, 1971.

[4] F. D. Stacey, B. J. Brennan, R. D. Irvine, Finite strain theories and comparisons with seismological data, Surveys in Geophysics 4 (1981) 189–232.

[5] S. Eliezer, R. A. Ricci (Eds.), High-pressure Equations of State: Theory and Applications, North-Holland, New York, 1991.

[6] O. L. Anderson, Equations of State for Solids in Geophysics and Ceramic Science, Oxford UP, Oxford, UK, 1995.

[7] W. B. Holzapfel, Physics of solids under strong compression, Rep. Prog. Phys. 59 (1996) 29–90.

[8] J.-P. Poirier, Introduction to the Physics of the Earth's Interior, 2nd Edition, Cambridge University Press, Cambridge, UK, 2000.

[9] W. B. Holzapfel, Equations of state for solids under strong compression, Z. Kristall. 216 (2001) 473–488.

[10] S. Eliezer, A. K. Ghatak, H. Hora, Fundamentals of Equations of State, World Sci, Singapore, 2002.

[11] F. D. Stacey, High pressure equations of state and planetary interiors, Rep. Prog. Phys. 68 (2005) 341–383.

[12] S. M. Peiris, J. C. Gump, Equations of state and high-pressure phases of explosives, in: S. M. Peiris, G. J. Piermarini (Eds.), Static Compression of Energetic Materials, Springer, Berlin, Germany, 2008, pp. 99–126.

[13] F. D. Murnaghan, The compressibility of media under extreme pressures, Proc. Natl. Acad. Sci. USA 30 (1944) 244–247.

[14] F. Birch, Finite elastic strain of cubic crystals, Phys. Rev. 71 (1947) 809–824.

[15] F. Birch, Finite strain isotherm and velocities for single-crystal and polycrystalline NaCl at high pressures and 300 K, J. Geophys. Res. 83 (1978) 1257–1268.

[16] Maxima, a computer algebra system, version 5.20.1.
URL http://maxima.sourceforge.net/

[17] J.-P. Poirier, A. Tarantola, A logarithmic equation of state, Phys. Earth Planet. Int. 109 (1998) 1–8.

[18] G. Mie, Zurkinetischen Theorie der einatomigen Köper, Ann. d. Phys. 11 (1903) 657–697.

[19] P. Vinet, J. Ferrante, J. R. Smith, J. H. Rose, An universal equation of state for solids, J. Phys. C: Solid State Phys. 19 (1986) L467–L473.

[20] P. Vinet, J. H. Rose, J. Ferrante, J. R. Smith, Universal features of the equation of state of solids, J. Phys: Condens. Matter 1 (1989) 1941–1963.

[21] R. E. Cohen, O. Gülseren, R. J. Hemley, Accuracy of equation-of-state formulations, Amer. Mineralogist 85 (2) (2000) 338–344.

[22] W. B. Holzapfel, Equation of state for solids under strong compression, High Press. Res. 16 (1998) 81–126.

[23] V. García Baonza, M. Cáceres, J. Nuñez, Universal compressibility behavior of dense phases, Phys. Rev. B 51 (1995) 28–37.

[24] V. García Baonza, M. Taravillo, M. Cáceres, J. Nuñez, Universal features of the equation of state of solids from a pseudospinodal hypothesis, Phys. Rev. B 53 (1996) 5252–5258.

[25] M. Taravillo, V. García Baonza, J. Nuñez, M. Cáceres, Simple equation of state for solids under compression, Phys. Rev. B 54 (1996) 7034–7045.

[26] M. Álvarez Blanco, Métodos cuánticos locales para la simulación de materiales iónicos. Fundamentos, algoritmos y aplicaciones, Tesis doctoral, Universidad de Oviedo (Julio 1997).

[27] H. Anton, P. C. Schmidt, Theoretical investigations of the elastic constants in Laves phases, Intermetallics 5 (1997) 449–465.

[28] B. Mayer, H. Anton, E. Bott, M. Methfessel, J. Sticht, J. Harris, P. C. Schmidt, Ab-initio calculation of the elastic constants and thermal expansion coefficients of Laves phases, Intermetallics 11 (2003) 23–32.

[29] Y. Bard, Nonlinear parameter estimation, Academic Press, New York, 1974.

[30] N. R. Draper, H. Smith, Applied Regression Analysis, 3rd Edition, Wiley, New York, USA, 1998.

[31] R. R. Wilcox, Fundamentals of Modern Statistical Methods, 2nd Edition, Springer, New York, USA, 2010.

[32] J. J. Moré, D. C. Sorensen, K. E. Hillstrom, B. S. Garbow, The MINPACK project, in: W. J. Cowell (Ed.), Sources and Development of Mathematical Software, Prentice-Hall, xxx, 1984, pp. 88–111.
URL `http://www.netlib.org/minpack/`

[33] L. N. Trefethen, D. Bau, III, Numerical Linear Algebra, SIAM, Philadelphia, PA, 1997.

[34] L. Thomson, On the fourth order anharmonic equation of state of solids, J. Phys. Chem. Solids 31 (1970) 2003–2016.

[35] J. Bardeen, Compressibilities of the alkali metals, J. Chem. Phys. 6 (1938) 372–378.

[36] H. Akaike, A new look at the statistical model identification, IEEE Transac. Automatic Control 19 (1974) 716–723.

[37] R. A. Maronna, D. R. Martin, V. J. Yohai, Robust Statistics: Theory and Methods, Wiley, Chichester, England, 2006.

[38] A. Otero-de-la Roza, V. Luaña, Runwien: a text-based interface for the wien package, Comput. Phys. Commun. 180 (2009) 800–812, source code distributed by the CPC program library: `http://cpc.cs.qub.ac.uk/summaries/AECM_v1_0.html`.

[39] K. Schwarz, P. Blaha, G. K. H. Madsen, Electronic structure calculations of solids using the WIEN2k package for material sciences, Comput. Phys. Commun. 147 (2002) 71–76.

[40] K. Schwarz, P. Blaha, Solid state calculations using WIEN2k, Comput. Mater. Sci. 28 (2003) 259–273.

[41] H. Cynn, J. E. Clepeis, C.-S. Yao, D. A. Young, Osmium has the lowest experimentally determined compressibility, Phys. Rev. Lett. 88 (2002) 135701.

[42] T. Kenichi, Bulk modulus of osmium: High-pressure powder x-ray diffraction experiments under quasihydrostatic conditions, Phys. Rev. B 70 (2004) 012101.

[43] M. Hebbache, M. Zemzemi, Ab initio study of high-pressure behavior of a low compressibility metal and a hard material: osmium and diamond, Phys. Rev. B 70 (2004) 224107.