# Usage of Robust Regression for Approximation of Thermodynamic Data*

Alexey L. Voskov[0000−0002−9211−5563]

Department of Chemistry, Lomonosov Moscow State University, Moscow, Russia, 119991
alvoskov@gmail.com

**Abstract.** M-estimators based on Huber and Andrews sine loss functions were successfully used for approximation of heat capacities and heat contents of K-substituted natrolite and petalite by means of the weighted sum of Einstein functions. It automatically excluded outliers for petalite and narrow peak of lambda-transition for K-natrolite.

**Keywords:** heat capacity · heat content · K-natrolite petalite · robust regression · thermodynamic models

## 1 Introduction

Evaluation of parameters of thermodynamic models from experimental data is a very common problem of nonlinear optimization. It is usually based on the weighted non-linear least squares method. Selection of the statistical weights is a complex problem due to different accuracy of experimental data and possible presence of systematic errors and outliers. Different schemes of their automatic selection were suggested [8,9], but all of them are based on the least squares method that is not robust to outliers. However, outliers may be excluded by the robust regression, i.e. by replacement of the sum of squares by other objective functions, e.g. by so called M-estimators:

$$F(\beta) = \sum_{k=1}^{n} \rho\left(\omega_k \frac{y_k^{\text{calc}}(\beta) - y_k^{\text{exp}}}{\sigma}\right) = \sum_{k=1}^{n} \rho\left(\frac{\omega_k r_k}{\sigma}\right) \tag{1}$$

where $n$ is the number of points, $r_k$ are residuals, $y_k^{\text{calc}}$ and $y_k^{\text{exp}}$ are calculated and experimental values respectively, $\beta$ is the model parameters column vector, $\rho(t)$ is the loss function, $\sigma$ is the scaling factor, $\omega_k$ are statistical weights.

However, minimization of eq. 1 requires special algorithms and much more computational power than the least squares method.

The aim of this work is to demonstrate the applicability of M-estimators for approximation of heat capacities and heat contents of individual substances. Experimental data for petalite $\text{LiAlSi}_4\text{O}_{10}$ and K-substituted natrolite (K-natrolite) $\text{Na}_{0.01}\text{K}_{1.85}\text{Mg}_{0.01}\text{Ca}_{0.04}\text{Al}_{1.96}\text{Si}_{3.04}\text{O}_{10} \cdot 2.72\text{H}_2\text{O}$ were be used as examples.

* Supported by the Russian Foundation for Basic Research for providing financial support (Grant No. 20-03-00575) and by the "Chemical Thermodynamics and Theoretical Materials Science" program (No. 121031300039-1).

## 2 Used Algorithm

In this work the the iterative reweighted least squares (IRLS) algorithm combined with Levenberg-Marquardt type regularization technique was used for finding eq. 1 minimum. IRLS for robust regression was suggested by Mudrov, Kushko et al. [6]. This quasi-Newton method is based on the numerical solution of the system of equations that express necessary condition for eq. 1 minimum:

$$\frac{\partial F}{\partial \beta_i} = \sum_{k=1}^{n} \psi\left(\frac{r_k \omega_k}{\sigma}\right) \cdot \frac{\omega_k}{\sigma} \frac{\partial r_k}{\partial \beta_i} = 0;\ \psi(t) \equiv \dot{\rho}(t) \equiv \frac{\partial \rho(t)}{\partial t} \tag{2}$$

IRLS uses two simplifications to get rid of the second derivatives. The first one is linearisation of the deviations $r_k$ near the initial approximation $\beta^\circ$:

$$r_k(\beta) = r_k(\beta^\circ) + \sum_{j=1}^{m} \frac{\partial r_k}{\partial \beta_j} (\beta_j - \beta_j^\circ) \Rightarrow r = r^\circ + Jp;\ p = \beta - \beta^\circ \tag{3}$$

where $J$ is Jacobian ($n \times m$ matrix), $m$ is the number of parameters, $r$ is the residuals column vector, The second step is exclusion of the $\ddot{\rho}(t)$ function by introduction of so called weight function $w(t) = \psi(t)/t$ [4]. If we assume that $w(t) \approx w(t_0)$ then $\ddot{\rho}(t) \approx w(t_0)$ and eq. 2 simplifies to:

$$\sum_{k=1}^{n} w\left(\frac{\omega_k r_k^\circ}{\sigma}\right) \cdot \omega_k^2 J_{ki} r_k^\circ + \sum_{j=1}^{m} p_j \sum_{k=1}^{n} w\left(\frac{\omega_k r_k^\circ}{\sigma}\right) \cdot \omega_k^2 J_{ki} J_{kj} = 0 \tag{4}$$

It gives the next matrix formula for the iteration (step) $p$ and the covariance matrix $C$ for the model parameters:

$$\beta - \beta^\circ = p = -\left(JWJ\right)^{-1} J^\top W r^\circ;\ C = \frac{(r^\circ)^\top W r^\circ}{n - m} \left(JWJ\right)^{-1} \tag{5}$$

where $W$ is the $n \times n$ diagnonal matrix with the $W_{kk} = \omega_k^2 \cdot w(\omega_k r_k / \sigma)$ elements.

The IRLS algorithm was embedded into the CpFit program [11] designed for approximation of heat capacities and heat contents of substances. It was used as the replacement of the least squares method and included the next steps:

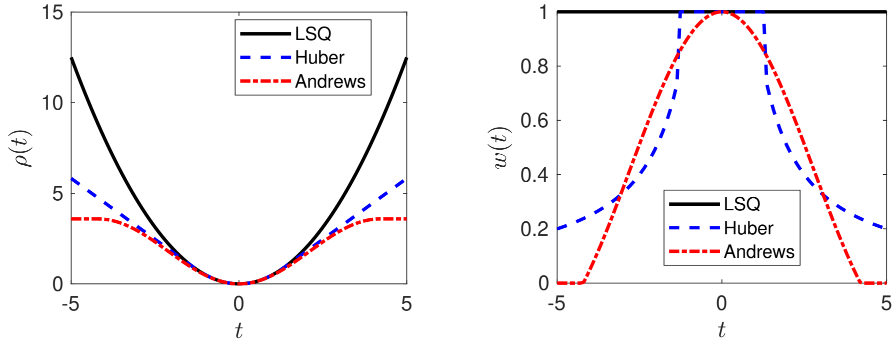
1. Find the initial approximation by the least squares method.
2. Estimate the scaling factor in eq. 1 using the robust estimation of standard error based on median [10]:

   $$\sigma = \Phi^{-1}(0.75) \cdot \text{median}\,|r| = 1.483 \cdot \text{median}\,|r| \tag{6}$$

   where $\Phi^{-1}(x)$ is the inverse cumulative distribution function for standard normal distribution.
3. Run the IRLS iterations combined with Levenberg-Marquardt type regularization technique using the given data and the loss function $\rho(t)$.

**Table 1.** Loss functions for M-estimators (see eq. 1) and their derivatives and weight functions used in this work; $a$ is the tuning constant for 95% asymptotic efficiency.

| Function | $\rho(t)$ | $\psi(t)$ | $\dot{\psi}(t)$ | $w(t)$ | $t$ ranges | $a$ |
|---|---|---|---|---|---|---|
| Square | $0.5t^2$ | $t$ | 1 | 1 | $t \in [-\infty; +\infty]$ | — |
| Huber | $\begin{cases} 0.5t^2 \\ a\,(\lvert t\rvert - 0.5a) \end{cases}$ | $t$<br>$a\,\mathrm{sign}\,t$ | 1<br>0 | 1<br>$\lvert t\rvert^{-1}$ | $\lvert t\rvert \leq a$<br>$\lvert t\rvert > a$ | 1.345 |
| Andrews | $\begin{cases} a^2\left(1 - \cos\frac{t}{a}\right) \\ 2a^2 \end{cases}$ | $a\sin\frac{t}{a}$<br>0 | $\cos\frac{t}{a}$<br>0 | $\frac{a}{t}\sin\frac{t}{a}$<br>0 | $\lvert t\rvert \leq a\pi$<br>$\lvert t\rvert > a\pi$ | 1.339 |



(a)

(b)

**Fig. 1.** (a) Loss functions $\rho(t)$ (b) the corresponding weight functions $w(t)$. Solid, dashed, dash-n-dotted lines — square, Huber and Andrews functions respectively.

The $\rho(t) = 0.5t^2$ (i.e. the least squares method), Huber and Andrews sine loss function were used in this work, their $\rho(t)$, $\psi(t)$ and $w(t)$ are given in Table 1 and at Figure 1. The tuning constants $a$ for 95% asymptotic efficiency in the case of normal distribution of errors were taken from Holland and Welsch [4].

Andrews sine and Huber functions are piecewise. They turn into $AT^2$ at smaller $t$ and to constant and linear functions respectively at larger $t$. This reduces values of their weight functions $w(t)$ at larger $t$ and influence of outliers. For Andrews sine function $w(t)$ reaches 0 for finite values of $t$. It causes exclusion of potential outliers from the optimization. In the case of Huber function $w(t)$ is always positive.

Huber function is convex and Andrews sine function is not (see Figure 1). The latter one belongs to redescending M-estimators that have non-convex $\rho(t)$, $\psi(t)$ with local extrema and $\lim_{t\to\infty}\psi(t) = 0$. They allow to totally exclude outliers from the optimization but may lead to non-convex objective function (see eq. 1) even in the case of linear regression. This increases the possibility of reaching local minimum instead of global and requires more careful selection of the initial approximation [1,5].

## 3 Experimental Data

Experimental data for K-substituted natrolite and petalite heat capacity and heat content were considered in this work. They are summarized in Table 2.

**Table 2.** Experimental data for K-natrolite and petalite; $N$ is the number of points; “ad.cal” — adiabatic calorimetry, DSC – differential scanning calorimetry, “up.lim.” – upper limits, ”std.dev” — standard deviation; $H$ was obtained by drop calorimetry.

| Compound | Data type | $N$ | $T$ / K | Uncertainty | Reference |
|---|---|---|---|---|---|
| K-natrolite | $C_p$ (ad.cal.) | 71 | 7.4–302.1 | 5%, 2%, 0.5% for $<$ 10, 10–20, $>$ 20 K (up.lim.) | Paukov et al. [7] |
| Petalite | $C_p$ (ad.cal.) | 83 | 5.6–381 | $<$ 10%, 0.3% for $<$ 20 K, $T \geq 20$ K (up.lim.) | Hemingway et al. [3] |
| | $C_p$ (DSC) | 17 | 340–500 | 1.0% (upper limits) | Hemingway et al. [3] |
| | $C_p$ (ad.cal.) | 41 | 10.7–302 | 2%, 0.5%, 0.2% for 10–20 K, 20–50 K, 50-300 K (std.dev.) | Bennington et al. [2] |
| | $H_T - H_{298.15}$ | 17 | 403–1194 | 0.4% (std.dev.) | Bennington et al. [2] |

These data were already approximated earlier by Voskov et al. [10,11] using the least squares method and the weighted sum of Einstein functions:

$$C_p(T) = \sum_{i=1}^{m} \alpha_i C_{\mathrm{E}}\left(\frac{\theta_i}{T}\right); \quad \frac{C_{\mathrm{E}}(x)}{R} = \frac{3x^2 e^x}{(e^x - 1)^2} \tag{7}$$

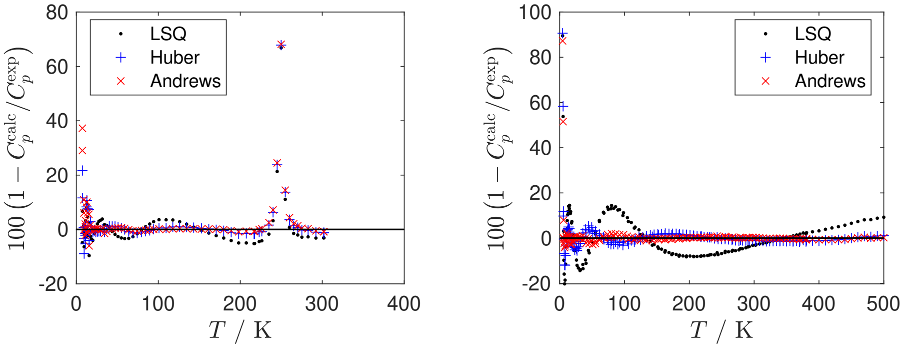
$$H_T - H_0 = \int_0^T C_p(T)\,dT = \sum_{i=1}^{m} \alpha_i H_{\mathrm{E}}\left(\frac{\theta_i}{T}\right); \quad \frac{H_{\mathrm{E}}(x)}{RT} = \frac{3x}{e^x - 1} \tag{8}$$

where $m$ is the number of terms, $R$ is the universal gas constant, $C_{\mathrm{E}}(x)$ is Einstein function, $\alpha_i$ and $\theta_i$ are model parameters that are found by the minimization of eq. 1. They may be considered as a crude approximation of phonon spectrum but due to anharmonism and possible Schottky anomalies they are closer to ad-hoc parameters. However the approximation based on the least squares method required manual exclusion of low-temperature outliers for petalite [10] (at $T =$ 4.57 and 5.27 K) and of heat capacity anomaly for K-natrolite [11] (at $T = 210 - 300$ K with a narrow peak at 250.32 K). The experimental data were approximated by eq. 7 without manual exclusion of the outliers and $C_p$ anomaly using the $\omega_{k,C} = 1/C_{p,k}^{\mathrm{exp}}$ and $\omega_{k,H} = 1/\Delta H_k^{\mathrm{exp}}$ statistical weights, i.e. relative deviations.

## 4 Results and their Discussion

The results of approximation for both substances are shown at Figure 2. Higher uncertainties at $T < 25$ K are due to less accurate experimental data, see Table 2.

The corresponding $\alpha_i$ and $\theta_i$ values are given in Tables 3 and 4. Extra digits are left intentionally: parameters confidence intervals because parameters are correlated to each other, i.e. $C$ from eq. 5 is not diagonal. This is typical for linear and nonlinear regression. Number of terms for petalite is not equivalent for different models because attempts to increase number of terms up to 5 in all models caused ill conditioned optimization problems.



(a)

(b)

**Fig. 2.** Relative error of $C_p$ approximation vs $T$ for models based on different M-estimators for (a) K-natrolite and (b) petalite.

**Table 3.** Model parameters for K-natrolite based on different M-estimators.

| Loss func. | Parameters |
|---|---|
| Square | $\vec{\alpha} = [16.6747 \pm 4.1; 6.80433 \pm 1.3; 1.61822 \pm 0.51; 0.139818 \pm 0.46]$<br>$\vec{\theta}/\mathrm{K} = [784.343 \pm 170; 216.143 \pm 41; 84.9998 \pm 26; 46.9228 \pm 30]$ |
| Huber | $\vec{\alpha} = [15.7714 \pm 8.0; 6.9136 \pm 1.7; 3.29675 \pm 1.2; 0.80118 \pm 0.14]$<br>$\vec{\theta}/\mathrm{K} = [948.901 \pm 460; 312.067 \pm 150; 130.076 \pm 38; 64.3238 \pm 4.5]$ |
| Andrews | $\vec{\alpha} = [15.7097 \pm 13; 7.02547 \pm 5.8; 4.14298 \pm 3.3; 1.05665 \pm 0.31]$<br>$\vec{\theta}/\mathrm{K} = [1042 \pm 810; 367.368 \pm 270; 158.74 \pm 46; 69.0627 \pm 4.2]$ |

The least squares method is sensitive to the $C_p$ anomaly and outliers: the obtained models are not accurate and undergo oscillations. M-estimators based on Huber and Andrews sine function are much less sensitive to them. Standard "baseline" (i.e. not taking into account $C_p$ anomalies) entropies $S^{\circ,\mathrm{BL}}_{298.15}$ were calculated for all models, see Table 5. For K-natrolite uncertainties are 1.1%, 0.5% and 0.3% for quadratic, Huber and Andrews loss functions; the reference value $S^{\circ,\mathrm{BL}}_{298.15} = 437.7\ \mathrm{J} \cdot (\mathrm{mol} \cdot \mathrm{K})^{-1}$ was taken from [11]. For petalite the uncertainties are 2.2%, 0.2% and $< 0.1\%$; the reference value $S^{\circ,\mathrm{BL}}_{298.15} = 232.7\ \mathrm{J} \cdot (\mathrm{mol} \cdot \mathrm{K})^{-1}$ was taken from [10]. Andrews sine loss function leads to more accurate values

**Table 4.** Model parameters for petalite based on different M-estimators.

| Loss func. | Parameters |
|---|---|
| Square | $\vec{\alpha} = [10.8013 \pm 0.48; 1.98900 \pm 0.19; 0.123494 \pm 0.026]$<br>$\vec{\theta}$ / K $= [564.984 \pm 29; 122.377 \pm 6.6; 43.0569 \pm 2.4]$ |
| Huber | $\vec{\alpha} = [8.55870 \pm 1.1; 5.40312 \pm 1.2; 1.36527 \pm 0.22; 0.119787 \pm 0.030]$<br>$\vec{\theta}$ / K $= [994.092 \pm 160; 350.399 \pm 46; 107.811 \pm 8.1; 43.7175 \pm 2.6]$ |
| Andrews | $\vec{\alpha} = [6.60117 \pm 2.2; 6.83701 \pm 2.0; 2.12692 \pm 1.1; 0.834509 \pm 0.35;$<br>$0.0661169 \pm 0.035]; \vec{\theta} \ / \ K = [1339.72 \pm 470; 517.876 \pm 140; 201.653 \pm 64;$<br>$88.2320 \pm 13; 38.3133 \pm 4.7]$ |

**Table 5.** Relative standard errors of approximation, $S^{\circ,\mathrm{BL}}_{298.15}$ values and results of 5-fold cross-validation.

| Compound | Loss func. | $10^2 s_{C_p}$ | $10^2 s^{\mathrm{test}}_{C_p}$ | $10^2 s^{\mathrm{train}}_{C_p}$ | $10^2 s_{\Delta H}$ | $10^2 s^{\mathrm{test}}_{\Delta H}$ | $10^2 s^{\mathrm{train}}_{\Delta H}$ | $\frac{S^{\circ,\mathrm{BL}}_{298.15}}{\mathrm{J\cdot(mol\cdot K)^{-1}}}$ |
|---|---|---|---|---|---|---|---|---|
| K-natrolite | LSQ | 4.4 | 3.17 | 3.45 | — | — | — | 445.1 |
| | Huber | 1.5 | 3.28 | 1.76 | — | — | — | 440.0 |
| | Andrews | 0.87 | 2.60 | 1.49 | — | — | — | 438.9 |
| Petalite | LSQ | 9.5 | 8.3 | 7.7 | 15 | 14 | 12 | 237.4 |
| | Huber | 2.0 | 2.3 | 2.2 | 2.4 | 3.0 | 2.9 | 232.2 |
| | Andrews | 0.51 | 0.85 | 0.45 | 0.29 | 0.47 | 0.21 | 232.6 |

of entropies, but during the optimization it sometimes manual tuning of initial approximation for petalite. Further research is required for automatic selection of initial approximation in the stepwise regression.

Although parameters confidence intervals were controlled to avoid overfitting, $k$-fold cross-validation with $k = 5$ was made for all models. The results are present in Table 5, all standard errors were estimated by means of eq. 6. $s$ were calculated for parameters from Tables 3 and 4 before cross-validation. $s^{\mathrm{test}}$ and $s^{\mathrm{train}}$ were evaluated as mean standard errors for test and training sets respectively; $s^{\mathrm{test}}$ and $s^{\mathrm{train}}$ were estimated by means of eq. 6.

For Andrews sine functions obtained $s_{C_p}$ and $s_{\Delta H}$ are close to the standard errors of existing models: for K-natrolite $s_{C_p} = 0.68\%$ (restored from model parameters from [11]) and for petalite $s_{C_p} = 0.46\%$, $s_{\Delta H} = 0.091\%$ [10]. In the case of K-natrolite $s^{\mathrm{test}}$ is about 2–3 times higher than $s$ for Huber and Andrews sine M-estimators. For petalite both $s^{\mathrm{test}}$ and $s^{\mathrm{train}}$ are close to $s$. Such differences may be connected with presence of $\lambda$-transition of K-natrolite and random fluctuations during sampling procedure may have stronger influence. The cross-validation results show that the models are not overfitted.

## 5 Conclusion

Robust regression based on M-estimators and the IRLS algorithm were successfully applied for approximation of isothermal heat capacity and heat content of K-substituted natrolite and petalite. This approach allowed to automatically exclude outliers. However, it is not designed for estimation of random and systematic errors of different data series, and may be combined with other schemes of statistical weights assignment if required. It also can't replace critical data evaluation of available experimental data but may help to find anomalies and outliers.

## 6 Data Availability

CpFit program is available at the site of Laboratory of Chemical Thermodynamics (http://td.chem.msu.ru). Data files for K-natrolite and petalite are published as Mendeley data set (http://dx.doi.org/10.17632/gbgnkr3f2x.1).

## References

1. Baselga, S., Klein, I., Suraci, S.S., de Oliveira, L.C., Matsuoka, M.T., Rofatto, V.F.: Global optimization of redescending robust estimators. Mathematical Problems in Engineering **2021**, 9929892 (2021), https://doi.org/10.1155/2021/9929892
2. Bennington, K.O., Stuve, J.M., Ferrante, M.J.: Thermodynamic properties of petalite ($Li_2Al_2Si_8O_{20}$). U.S. Bureau of Mines, Report of investigations 8451 (1979), https://hdl.handle.net/2027/mdp.39015006379187
3. Hemingway, B.S., Robie, R.A., Kittrick, J.A., Grew, E.S., Nelen, J.A., London, D.: The heat capacities of osumilite from 298.15 to 1000 K, the thermodynamic properties of two natural chlorites to 500 K, and the thermodynamic properties of petalite to 1800 K. Am. Mineral. **69**(7–8), 701–710 (1984)
4. Holland, P.W., Welsch, R.E.: Robust regression using iteratively reweighted least-squares. Comm. Statist. Theory Methods **6**(9), 813–827 (1977), https://doi.org/10.1080/03610927708827533
5. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust Statitics. Theory and methods. John Wiley & Sons, Ltd (2006)
6. Mudrov, V.I., Kushko, V.L., Mikhailov, V.I., Osovitskii, E.M.: Experiments on usage of the least absolute deviations method for orbital information processing problems [in Russian]. Kosmicheskie issledovania (Cosmic Research) **6**, 502–514 (1968)
7. Paukov, I.E., Kovalevskaya, Y.A., Seretkin, Y.V., Belitskii, I.A.: The thermodynamic properties and structure of potassium-substituted natrolite in the phase transition region. Russ. J. Phys. Chem. A **76**(9), 1406–1410 (2002)
8. Paulson, N.H., Zomorodpoosh, S., Roslyakova, I., Stan, M.: Comparison of statistically-based methods for automated weighting of experimental data in CALPHAD-type assessment. Calphad **68**, 101728 (2020), https://doi.org/10.1016/j.calphad.2019.101728
9. Rudnyi, E.B.: Statistical model of systematic errors: An assessment of the Ba–Cu and Cu–Y phase diagram. Chemom. Intell. Lab. Systems **36**(2), 213–227 (1997), https://doi.org/10.1016/S0169-7439(96)00069-X
10. Voskov, A.L.: Description of thermodynamic functions of aluminosilicates with the zeolite-like composition by sums of Einstein-Planck functions. Russ. J. Inorg. Chem **65**, 765–772 (2020), https://doi.org/10.1134/S0036023620050265
11. Voskov, A.L., Kutsenok, I.B., Voronin, G.F.: CpFit program for approximation of heat capacities and enthalpies by Einstein-Planck functions sum. Calphad **61**, 50–61 (2018), https://doi.org/10.1016/j.calphad.2018.02.001