Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 90 (2008) 116–122
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Estimation and testing for the common intersection point

Andrew L. Rukhin

Department of Mathematics and Statistics, University of Maryland at Baltimore County, Baltimore, MD 21250 USA
Statistical Engineering Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 USA

Received 30 October 2006; received in revised form 12 August 2007; accepted 13 August 2007
Available online 19 August 2007

Abstract

The problems of estimating a common intersection point from a collection of straight lines and of testing the hypothesis that such a point exists are considered. The relationship to the error-in-variables regression and to the intersection–union principle is explored. Robust rank-based procedures for this problem are suggested, and the results of Monte Carlo simulation are reported. An example of isokinetic relationship in hexachlorobiphenyl is reviewed.

Keywords: Linear regression; Isokinetic relationship; Intersection–union principle; Error-in-variables regression; Rank-based inference; Robustness

1. Introduction: determination of the iso-equilibrium point

The goal of this paper is to develop statistically rigorous procedures for confidence estimation of a common intersection point from a collection of straight lines and for testing the hypothesis that such a point exists. These problems appear in several chemistry applications, in particular, in the study of the so-called isokinetic relationship [1–3]. The author encountered them when consulting on statistical analysis of an empirical dependence study between enthalpies and entropies in a series of related reactions among polychlorinated biphenyls. In this study it was important to establish inferential techniques for the compensation and the iso-equilibrium effects. In statistical terms, the issues are to confirm a linear relationship between the two variables and to test the existence of a common intersection point for several linear models (in which case the isokinetic hypothesis holds.).

An Excel program to test the isokinetic hypothesis has been presented in [4]. We discuss properties of this likelihood ratio test in Section 2 which also contains the form of confidence regions for the intersection point. Section 3 is dedicated to robust rank-based procedures in this problem. Simulation results and an example are given in Section 4. Section 5 discusses an example of isokinetic relationship in hexachlorobiphenyls.

2. Least squares model

A number, say, ℓ, $\ell \geq 3$, of linear statistical models is postulated. According to these the data y_{ik} has the form

$$y_{ik} = m_i x_{ik} + b_i + \epsilon_{ik},\ k = 1, \ldots, n_i,\ i = 1, \ldots, \ell. \tag{1}$$

Here m_i, b_i denote the unknown slope and intercept of the ith model. In chemistry applications x's correspond to reciprocals of temperatures, y's are logarithms of rate or equilibrium constants. In this section ϵ_{ik} are supposed to be independent normal errors with mean 0 and the unknown but common variance σ^2. The unbiased estimates of the slope and of the intercept are provided by the classical least squares estimators,

$$\hat{m}_i = \frac{\sum_k (x_{ik} - \overline{x}_i)(y_{ik} - \overline{y}_i)}{S_i^2}, \tag{2}$$

$$\hat{b}_i^{\star} = \overline{y}_i - \hat{m}_i \overline{x}_i. \tag{3}$$

Here $\overline{x}_i = n_i^{-1} \sum_k x_{ik}$, $\overline{y}_i = n_i^{-1} \sum_k y_{ik}$, $S_i^2 = \sum_k (x_{ik} - \overline{x}_i)^2$.

According to the null hypothesis H_0 all lines $y = m_i x + b_i$ have a common intersection point, i.e. for some x_0 and y_0,

$$m_1 x_0 + b_1 = m_2 x_0 + b_2 = \ldots = m_\ell x_0 + b_\ell = y_0, \tag{4}$$

or

$$b_i = y_0 - m_i x_0,\ i = 1, \ldots, \ell. \tag{5}$$

E-mail address: adrew.rukhin@nist.gov (A.L.Rukhin).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.08.004

In other words, ℓ straight lines having a common intersection point is equivalent to the linear relationship among ℓ pairs (b_i, m_i) with y_0 for the intercept, and $-x_0$ for the slope.

Since $\hat{m}_i$ and $\hat{b}_i^{\star}$ are dependent, it is convenient to put

$$\hat{b}_i = \hat{b}_i^{\star} + \hat{m}_i \overline{x}_i = \overline{y}_i. \tag{6}$$

Then $\hat{m}_i$, $\hat{b}_i$ are uncorrelated normal with means m_i, $m_i\overline{x}_i + b_i$, and variances $\mathrm{Var}(\hat{m}_i)=\sigma^2 S_i^{-2}$, $\mathrm{Var}(\hat{b}_i)=\sigma^2/n_i$. We also put $S_P^2=\sum_{ik}[y_{ik}-\overline{y}_i-\hat{m}_i(x_{ik}-\overline{x}_i)]^2$, which has the distribution of the form $\sigma^2\chi^2(\nu)$, where $\nu=\sum(n_i-2)=N-2\ell$.

To find the maximum likelihood estimators of x_0 and y_0, one can use only (complete) sufficient statistics $\hat{m}_i$, $\overline{y}_i$, and S_P^2. The negative log-likelihood function for these statistics has the form (up to a constant term not involving unknown m_i, x_0, y_0 and σ^2),

$$L(\{m_i\}, x_0, y_0, \sigma) = N \log \sigma + \frac{1}{2\sigma^2}\left[\sum_i S_i^2(m_i - \hat{m}_i)^2 + \sum_i n_i(m_i(\overline{x}_i - x_0) - \overline{y}_i + y_0)^2 + S_P^2\right]. \tag{7}$$

If the minimizer $\hat{x}_0$ were found, the value of m_i corresponding to the largest likelihood is

$$\hat{\hat{m}}_i = \frac{S_i^2\hat{m}_i + n_i(\overline{x}_i - \hat{x}_0)(\overline{y}_i - \hat{y}_0)}{S_i^2 + n_i(\overline{x}_i - \hat{x}_0)^2} = \frac{\sum_k (x_{ik} - \hat{x}_0)(y_{ik} - \hat{y}_0)}{\sum_k (x_{ik} - \hat{x}_0)^2}, \tag{8}$$

as

$$S_i^2 + n_i(\overline{x}_i - x_0)^2 = \sum_k (x_{ik} - x_0)^2. \tag{9}$$

Thus, $\hat{\hat{m}}_i$ is a convex combination of two slopes, $\hat{m}_i$ and $(\overline{y}_i - \hat{y}_0)/(\overline{x}_i - \hat{x}_0)$.

Algebra shows that the weighted sum of "residuals" $\overline{y}_i - \hat{m}_i(\overline{x}_i - \hat{x}_0)$ estimates y_0,

$$\hat{y}_0 = \frac{\sum_i \frac{n_i S_i^2(\overline{y}_i - \hat{m}_i(\overline{x}_i - \hat{x}_0))}{S_i^2 + n_i(\overline{x}_i - \hat{x}_0)^2}}{\sum_i \frac{n_i S_i^2}{S_i^2 + n_i(\overline{x}_i - \hat{x}_0)^2}}. \tag{10}$$

However, determination of $\hat{x}_0$ demands numerical minimization of the function of x_0,

$$G(x_0) = \sum_i \frac{n_i S_i^2(\overline{y}_i - \hat{m}_i(\overline{x}_i - \hat{x}_0))^2}{S_i^2 + n_i(\overline{x}_i - x_0)^2} - \frac{\left[\sum_i \frac{n_i S_i^2(\overline{y}_i - \hat{m}_i(\overline{x}_i - x_0))}{S_i^2 + n_i(\overline{x}_i - x_0)^2}\right]^2}{\sum_i \frac{n_i S_i^2}{S_i^2 + n_i(\overline{x}_i - x_0)^2}}$$

$$= \sum_i \frac{n_i S_i^2}{S_i^2 + n_i(\overline{x}_i - x_0)^2}\left[\overline{y}_i - \hat{m}_i(\overline{x}_i - x_0) - \frac{\sum_j \frac{n_j S_j^2(\overline{y}_j - \hat{m}_j(\overline{x}_j - x_0))}{S_j^2 + n_j(\overline{x}_j - x_0)^2}}{\sum_j \frac{n_j S_j^2}{S_j^2 + n_j(\overline{x}_j - x_0)^2}}\right]^2. \tag{11}$$

To use iterative optimization methods, one needs an initial approximation for $\hat{x}_0$, which is discussed below. The two-dimensional likelihood optimization in x_0, y_0 is considered in [5].

After $\hat{x}_0$ is found, the (biased) maximum likelihood estimator of σ^2 is determined from the formula,

$$\hat{\sigma}^2 = \frac{G(\hat{x}_0) + S_P^2}{N}, \tag{12}$$

and

$$\max L(\{m_i\}, x_0, y_0, \sigma) = N + \frac{N}{2}\log\ \hat{\sigma}^2. \tag{13}$$

It is easy to see that in the unconstrained model

$$\max L(\{m_i\}, \{b_i\}, \sigma) = N + \frac{N}{2}\log\ \frac{S_P^2}{N}. \tag{14}$$

These formulas give the form of the likelihood ratio test statistic, based on $NG(\hat{x}_0)\ /\ {S_P}^2$, which has approximate χ^2-distribution with $\ell - 2$ degrees of freedom.

The test statistic in [4],

$$F = \frac{\nu G(\hat{x}_0)}{(\ell - 2)S_P^2}, \tag{15}$$

is suggested to reject the null hypothesis: when $F > F_\alpha(\ell - 2, \nu)$, the α-critical point of F-distribution with $\ell - 2$ and ν degrees of freedom. Indeed for large N, one can replace N by ν and the approximate χ^2-distribution with $\ell - 2$ degrees of freedom by the F-distribution with $\ell - 2$ and ν degrees of freedom.

The statistic (15) can be motivated by the *intersection–union principle*. Indeed, the likelihood ratio test of the null hypothesis (4) for a *known* x_0 in our notation has the form

$$F_{x_0} = \frac{\nu G(x_0)}{(\ell - 1)S_P^2}. \tag{16}$$

Thus, $F=(\ell-1)(\ell-2)^{-1}\min_{x0} F_{x0}$ with F_{x0} having the (exact) F-distribution with $\ell - 1$ and ν degrees of freedom [6]. If our null hypothesis is rejected when $\min_{x0} F_{x0}$ exceeds $F_\alpha(\ell - 1, \nu)$, then the level of the resulting test will be bounded above by α [7], i.e. the corresponding test will be conservative. But since $F_\alpha(\ell - 2, \nu) < (\ell - 1)(\ell - 2)^{-1} F_\alpha(\ell - 1, \nu)$, the definition (15) attempts to compensate for conservativeness of the intersection–union test by using smaller degrees of freedom, $\ell - 2$, not $\ell - 1$.

The known duality between hypothesis testing and confidence sets gives the $(1-\alpha)$-confidence region for x_0,

$$R_1 = \left\{x_0 : G(x_0) \le F_\alpha(\ell - 1,\ \nu)\frac{(\ell - 1)S_P^2}{\nu}\right\}. \tag{17}$$

A necessary condition for this region to be an interval is that

$$\lim_{|x_0|\to\infty} G(x_0) = \sum_i S_i^2(\hat{m}_i - \overline{m})^2 > F_\alpha(\ell - 1,\ \nu)\frac{(\ell - 1)S_P^2}{\nu}. \tag{18}$$

Here $\overline{m}=\ell^{-1}\sum_i \hat{m}_i$.

When x_0 and y_0 are both known, the likelihood ratio test of the null hypothesis (4) has a similar form

$$F_{x_0 y_0} = \frac{\nu G(x_0, y_0)}{\ell S_P^2}, \tag{19}$$

where

$$G(x_0, y_0) = \sum_i \frac{n_i S_i^2(\overline{y}_i - \hat{m}_i(\overline{x}_i - x_0) - y_0)^2}{S_i^2 + n_i(\overline{x}_i - x_0)^2}. \tag{20}$$

Now $F_{x_0y_0}$ has the F-distribution with ℓ and ν degrees of freedom. As above, one obtains the $(1-\alpha)$-confidence region for (x_0, y_0),

$$R = \left\{(x_0, y_0) : G(x_0, y_0) \le F_\alpha(\ell, \nu)\frac{\ell S_P^2}{\nu}\right\}. \tag{21}$$

Under condition (18) this region typically has an elliptical shape. Since $(n_i S_i^2) / [S_i^2 + n_i(\overline{x}_i - x_0)^2] \le n_i$, R always contains the ellipsoid,

$$\sum_i n_i(\overline{y}_i - \hat{m}_i(\overline{x}_i - x_0) - y_0)^2 \le F_\alpha(\ell, \nu)\frac{\ell S_P^2}{\nu}. \tag{22}$$

To find an initial approximation to optimize $G(x_0)$, one can use the fact that under the null hypothesis the least squares estimators $\hat{m}_i$ and $\overline{y}_i$ satisfy a linear statistical relationship,

$$\hat{m}_i = m_i + \delta_i, \tag{23}$$

$$\overline{y}_i = y_0 + m_i(\overline{x}_i - x_0) + \epsilon_i. \tag{24}$$

Provided that $\mathrm{Var}(\hat{m}_i) \equiv \sigma_m^2$, $\mathrm{Var}(\overline{y}_i) \equiv \sigma_b^2$, and $\overline{x}_i \equiv \overline{x}$ (which implies balancedness of the model), the results of the error-in-variables regression theory are applicable assuming that the ratio $\sigma_b^2/\sigma_m^2 = \lambda$, $\lambda > 0$, is known. In practice, for moderately unbalanced models, one can take n and $\overline{x}$ to be average values of n_i and $\overline{x}_i$, and estimate λ as $(S_1^2/n_1 + \ldots + S_\ell^2 / n_\ell) / \ell$. A classical solution of error-in-variables regression parameter estimation [8] leads to the following estimators of x_0 and y_0,

$$\hat{x}_0 = \overline{x} - \frac{s_{bb} - \lambda s_{mm} + \sqrt{(s_{bb} - \lambda s_{mm})^2 + 4\lambda s_{mb}^2}}{2s_{mb}}, \tag{25}$$

and

$$\hat{y}_0 = \overline{m}(\hat{x}_0 - \overline{x}) + \overline{y}. \tag{26}$$

Here

$$s_{mm} = \sum_i (\hat{m}_i - \overline{m})^2, \tag{27}$$

$$s_{bb} = \sum_i (\overline{y}_i - \overline{\overline{y}})^2, \tag{28}$$

$$s_{mb} = \sum_i (\hat{m}_i - \overline{m})(\overline{y}_i - \overline{\overline{y}}) \tag{29}$$

with $\overline{\overline{y}} = \ell^{-1}\sum_i \overline{y}_i$. An experienced chemometrist might have a reasonable idea about the relative order of uncertainties in estimators $\hat{m}_i$ and $\hat{b}_i$, which can be used to choose λ. For example, if $\sigma_m^2 \ll \sigma_b^2$, i.e. if $S_i^2 \gg n_i$, then one can employ the classical linear regression formulas,

$$\hat{x}_0 = \overline{x} - \frac{s_{mb}}{s_{mm}}. \tag{30}$$

In the perfectly balanced case when $n_i \equiv n$, $S_i^2 \equiv S^2$, and $\overline{x}_i = \overline{x}$, all weights in (10) are equal, (26) holds, and $\hat{x}_0$ is the minimizer of the ratio of two quadratic polynomials in x_0,

$$G(x_0) = \frac{S^2\sum_i \left[\overline{y}_i - \overline{\overline{y}} - (\hat{m}_i - \overline{m})(\overline{x} - x_0)\right]^2}{S^2/n + (\overline{x} - x_0)^2}. \tag{31}$$

Thus (25) is met with $\lambda = n^{-1}S^2$.

An alternative initial approximation can be obtained by averaging the coordinates of intersection points for pairs of lines. Now we sum up the results of this Section.

Theorem 2.1. *The likelihood ratio test of H_0 when the errors are normally distributed is given by Eq.* (15)*, where $\hat{x}_0$ is the minimizer of the function $G(x_0)$, with Eq.* (25) *furnishing an initial approximation. The confidence region R for (x_0, y_0) is indicated in Eq.* (21)*, and the confidence region for x_0 is provided by Eq.* (17).

3. Least-absolute-deviations regression

The normality assumption in the linear regression model (1) may not hold as in many applications the errors have a heavier tail distribution than normal. For this reason procedures insensitive to outliers are desirable.

In this section we accept a setting when errors ϵ's have a density f with zero median, $f(0) > 0$. One of the commonly used robust estimators of the slope m_i is

$$\widetilde{m}_i = \mathrm{wmed}\left\{\frac{y_{ik} - y_{ij}}{x_{ik} - x_{ij}}\right\},\ k = 1, \ldots, n_i,\ i = 1, \ldots, \ell, \tag{32}$$

where wmed means the weighted median of data points in the ith regression model taken over all values $k < j$, such that $x_{ik} \ne x_{ij}$, with the weights $|x_{ik} - x_{ij}|$. In other words, the $n_i(n_i - 1)/2$ ratios, $(y_{ik} - y_{ij}) / (x_{ik} - x_{ij})$, are sorted from smallest to largest, and the observation in this series corresponding to the cumulative sum of similarly ordered weights closest to $\sum_{k<j} |x_{ik} - x_{ij}|/2$ is the weighted median, e.g. [9]. This estimator is known to minimize the sum of absolute deviations $\sum_k |y_{ik} - m_i x_{ik}|$ (or $\sum_k |y_{ik} - y_0 - m_i(x_{ik} - x_0)|$), as well as the Wilcoxon-type dispersion measure,

$$D(m_i) = \frac{\sqrt{12}}{n_i + 1}\sum_k \left(r_k(m_i) - \frac{n_i + 1}{2}\right)(y_{ik} - m_i x_{ik}). \tag{33}$$

Here $r_k(m_i)$ is the rank of $y_{ik} - m_i x_{ik}$ among their ordered values. Thus $D(m_i)$ provides a measure of goodness-of-fit for the ith model.

Another popular estimator is

$$\widetilde{m}_i = \mathrm{med}\left\{\frac{y_{ik} - y_{ij}}{x_{ik} - x_{ij}}\right\},\quad i = 1, \ldots, \ell, \tag{34}$$

the median of empirical slopes. This estimator attempts to make the residuals and x's uncorrelated after the Kendall tau coefficient,

$$\frac{\sum_{k,j} \mathrm{sign}\ \left(r_k(m_i) - r_j(m_i)\right)\left(q_{ik} - q_{ij}\right)}{n_i(n_i - 1)}. \tag{35}$$

Here for fixed i, q_{ik} denotes the rank of x_{ik}. For both of these statistics the corresponding estimator of the intercept is the median of residuals, $\tilde{e}_{ik} = y_{ik} - \tilde{m}_i x_{ik}$,

$$\widetilde{b}_i = \mathrm{med}(\widetilde{e}_{ik}). \tag{36}$$

In the symmetric case, $f(-x) = f(x)$, the recommended intercept estimator is $\mathrm{med}\{(\tilde{e}_{ik} + \tilde{e}_{ij}) / 2\}$, but it will not be the maximum likelihood estimator in our setting.

Different estimators of the slope, $\check{m}_i$, and of the intercept, $\check{b}_i$, are found as minimizers of the sum of absolute deviations $\sum_k |y_{ik}-m_i x_{ik}-b_i|$ for $i=1,\ldots,\ell$. By using these statistics, one can construct the likelihood ratio test assuming that the errors obey the Laplace (double exponential) distribution with a density,

$$f(u)=\frac{1}{2\sigma}\ \exp\ \left\{-\frac{|u|}{\sigma}\right\}. \tag{37}$$

In this case the (negative) log-likelihood function has the form

$$L(\{m_i\},\{b_i\},\sigma)=N\ \log\ \sigma+\frac{1}{\sigma}\sum_{i,k}|y_{ik}-m_i x_{ik}-b_i|, \tag{38}$$

so that $\check{m}_i$ and $\check{b}_i$ are maximum likelihood estimators which can be evaluated after the algorithms in [10]. Also

$$L(\{m_i\},x_0,y_0,\sigma)=N\ \log\ \sigma+\frac{1}{\sigma}\sum_{i,k}|y_{ik}-y_0-m_i(x_{ik}-x_0)|, \tag{39}$$

and $\widetilde{m}_i$ in (32) is the maximum likelihood estimator of m_i for any fixed x_0, y_0 so that it is the maximum likelihood estimator under the null hypothesis.

As in Section 2, the likelihood ratio test statistic is

$$\widetilde{F}=\frac{2(N-\ell)\left[\min_{x0}\ \sum_{i,k}|y_{ik}-\widetilde{y}_0-\widetilde{m}_i(x_{ik}-x_0)|-\sum_{i,k}|y_{ik}-\check{m}_i x_{ik}-\check{b}_i|\right]}{(\ell-2)\sum_{i,k}|y_{ik}-\check{m}_i x_{ik}-\check{b}_i|}, \tag{40}$$

where

$$\widetilde{y}_0=\widetilde{y}_0(x_0)=\mathrm{med}\{y_{ik}-\widetilde{m}_i(x_{ik}-x_0)\}. \tag{41}$$

The distribution of $\tilde{F}$ under H_0 can be approximated by the F-distribution with $\ell-2$ and $2(N-\ell)$ degrees of freedom. Indeed, $2\sum_{ik}|y_{ik}-\check{m}_i x_{ik}-\check{b}_i|\ /\ \sigma$ has approximate χ^2-distribution with $2\sum(n_i-1)=2(N-\ell)$ degrees of freedom, and the numerator in Eq. (40) is a multiple of χ^2-random variable with $\ell-2$ degrees of freedom.

Table 1
The observed significance level of the normal-based test (15) and the rank-based test (40) for $\alpha=0.05$ and several sample numbers ℓ

		$\ell=3$	$\ell=4$	$\ell=5$	$\ell=10$	$\ell=20$
I	(15)	0.05	0.05	0.05	0.05	0.05
	(40)	0.04	0.04	0.04	0.04	0.04
II	(15)	0.08	0.07	0.06	0.05	0.05
	(40)	0.02	0.02	0.01	0.01	0.01
III	(15)	0.04	0.04	0.05	0.05	0.05
	(40)	0.03	0.03	0.03	0.02	0.02
IV	(15)	0.02	0.02	0.02	0.02	0.04
	(40)	0.01	0.01	0.006	0.005	0.004
V	(15)	0.05	0.05	0.05	0.05	0.04
	(40)	0.02	0.02	0.02	0.01	0.01

Table 2
The rescaled by σ^2 mean squared errors of estimators of x_0, y_0 based on normal theory $\hat{x}_0$ and $\hat{y}_0$, and rank-based $\tilde{x}_0$ and $\tilde{y}_0$ for several sample numbers ℓ

		$\ell=3$	$\ell=4$	$\ell=5$	$\ell=10$	$\ell=20$
I	(15)	0.077	0.044	0.031	0.012	0.006
	(40)	0.113	0.068	0.047	0.021	0.008
II	(15)	0.140	0.094	0.063	0.024	0.012
	(40)	0.082	0.053	0.033	0.015	0.008
III	(15)	0.077	0.066	0.043	0.022	0.011
	(40)	0.013	0.012	0.011	0.005	0.002
IV	(15)	$4.57e^6$	$2.33e^6$	$7.24e^5$	$8.6e^4$	$2.0e^4$
	(40)	0.223	0.131	0.087	0.037	0.016
V	(15)	0.186	0.116	0.098	0.033	0.014
	(40)	0.152	0.093	0.077	0.021	0.010

It is known [11] that the approximate joint distribution of $\check{m}_i$, $\check{b}_i$, as well as of $\tilde{b}_i$, $\tilde{m}_i$, is normal with the mean $(b_i,\ m_i)$ and the covariance matrix

$$\frac{1}{4f^2(0)n_i S_i^2}\begin{pmatrix} S_i^2+n_i\,\overline{x}_i^2 & -n_i\,\overline{x}_i \\ -n_i\,\overline{x}_i & n_i \end{pmatrix}. \tag{42}$$

By using this fact one can apply the results of the error-in-variables robust regression theory [12] and estimate $(x_0,\ y_0)$ under H_0 as the minimizer of

$$2f(0)\sum_i\frac{|\widetilde{b}_i+\widetilde{m}_i x_0-y_0|}{\sqrt{\mathrm{Var}\left(\widetilde{b}_i+\widetilde{m}_i x_0-y_0\right)}}=\sum_i\frac{S_i|\widetilde{b}_i+\widetilde{m}_i x_0-y_0|}{\sqrt{S_i^2/n_i+(\overline{x}_i-x_0)^2}}. \tag{43}$$

In the balanced case

$$\widetilde{y}_0=\mathrm{med}\{\widetilde{b}_i+\widetilde{m}_i x_0\}, \tag{44}$$

while $\tilde{x}_0$ is the minimizer in x_0 of the ratio,

$$\frac{\sum_i|\widetilde{b}_i+\widetilde{m}_i x_0-\widetilde{y}_0|}{\sqrt{S^2/n+(\overline{x}-x_0)^2}}. \tag{45}$$

This point is a good initial approximation to x_0 in Eq. (40).

Theorem 3.1. *The likelihood ratio test of H_0, when the errors have a Laplace distribution, is given by $\tilde{F}$ in Eq.* (40) *with $\tilde{m}_i$ defined by Eq.* (32) *and $\tilde{y}_0$ as in Eq.* (41).

According to our simulations some of which are reported in Section 4, the test based on Eq. (40) is conservative. Upon the whole the procedure (40) turns out to be quite robust against y-outliers, which are of most interest in chemical applications. Also it has much higher power against many alternatives for which the regression lines do not intersect.

The corresponding $(1-\alpha)$-confidence set for $(x_0,\ y_0)$ has the form

$$\left\{(x_0,y_0):\sum_{i,k}|y_{ik}-y_0-\widetilde{m}_i(x_{ik}-x_0)|\right. \tag{46}$$
$$\left.\le\left[\frac{\ell F_\alpha(\ell,2(N-\ell))}{2(N-\ell)}+1\right]\sum_{i,k}|y_{ik}-\check{m}_i x_{ik}-\check{b}_i|\right\},$$

Table 3
The power of the normal-based test (15) and of the rank-based test (40) against the parallel lines alternative for $\alpha=0.05$ and several sample numbers ℓ

		$\ell=3$	$\ell=4$	$\ell=5$	$\ell=10$	$\ell=20$
I	(15)	0.04	0.05	0.06	0.06	0.06
	(40)	0.99	0.99	0.99	0.99	0.99
II	(15)	0.06	0.06	0.06	0.06	0.07
	(40)	1.00	1.00	1.00	1.00	1.00
III	(15)	0.05	0.05	0.05	0.05	0.06
	(40)	0.99	0.99	0.99	1.00	1.00
IV	(15)	0.05	0.06	0.08	0.12	0.17
	(40)	0.98	0.99	0.99	0.99	0.99
V	(15)	0.01	0.02	0.04	0.04	0.04
	(40)	0.99	0.99	0.99	1.00	1.00

and this region is always convex. Similarly, with $S(x_o)=\sum_{i,k}|y_{ik}-\tilde{y}_0-\tilde{m}_i(x_{ik}-x_0)|$, the $(1-\alpha)$-confidence interval for x_0 is

$$\left\{x_0 : S(x_0)\le\left[\frac{(\ell-1)F_\alpha(\ell-1,2(N-\ell))}{2(N-\ell)}+1\right]\sum_{i,k}|y_{ik}-\check{m}_i x_{ik}-\check{b}_i|\right\}. \quad (47)$$

4. Monte Carlo simulation results

Monte Carlo simulation was used to evaluate comparative performance of the normal-based method given in Eq. (15) and with the robust rank-based test from Eq. (40). We used the following error distributions: (I) the normal with $\sigma=0.1$, (II) the Laplace distribution with $\sigma=0.196/\log 10=0.8512\ldots$, (III) the logisitic distribution, with $\sigma=0.196/\log 19$, (IV) the Cauchy distribution with $\sigma=0.196/\tan(0.45\pi)$, (V) the contaminated normal distribution with the distribution function $0.8\Phi(x/\sigma)+0.2\Phi(x/(3\sigma))$, $\sigma=0.196/3.47$. The choice of σ was motivated by matching 95%th percentiles of all these distributions. The simulations were performed for $n_i\equiv 40$ with x_{ik}, $k=1,\ldots, n$ being uniformly distributed on the interval $(-0.5, 0.5)$, $b_i\equiv 0$, $m_i=i$, $i=1,\ldots, \ell$. Different x designs, namely, those corresponding to the order statistics from normal or double exponential distributions and different, smaller values of n_i were also employed. These results are not reported here as the conclusions were similar. In these simulations H_0 holds with $x_0=y_0=0$.

Table 4
The confidence coefficients (upper lines) and the standard errors (lower lines) of the normal-based confidence interval (17) and of the rank-based interval (47) for several sample numbers ℓ

		$\ell=3$	$\ell=4$	$\ell=5$	$\ell=10$	$\ell=20$
I	(17)	0.96	0.95	0.95	0.95	0.95
		0.051	0.039	0.029	0.015	0.010
	(47)	0.92	0.89	0.87	0.82	0.80
		0.053	0.045	0.039	0.021	0.017
II	(17)	0.94	0.94	0.93	0.92	0.91
		0.077	0.065	0.051	0.047	0.045
	(47)	0.95	0.96	0.96	0.95	0.93
		0.043	0.027	0.012	0.010	0.007
III	(17)	0.95	0.95	0.95	0.96	0.96
		0.033	0.025	0.016	0.005	0.003
	(47)	0.94	0.93	0.92	0.87	0.84
		0.022	0.016	0.011	0.004	0.002
IV	(17)	0.60	0.66	0.73	0.82	0.87
		0.343	0.348	0.203	0.068	0.005
	(47)	0.98	0.98	0.97	0.94	0.89
		0.025	0.019	0.014	0.006	0.003
V	(17)	0.95	0.95	0.95	0.95	0.95
		0.024	0.017	0.012	0.005	0.004
	(47)	0.95	0.93	0.88	0.80	0.79
		0.015	0.010	0.008	0.003	0.002

Table 5
The data in the HEXA example

x	y_1	k_1	y_2	k_2	y_3	k_3	y_4	k_4
3.28947	−3.8014	8	−3.6950	9	−3.5981	8	−3.3868	6
3.35570	−4.0126	13	−3.8342	13	−3.7163	13	−3.3131	13
3.43643	−5.0970	6	−4.6106	6	−4.3681	5	−4.0172	5
3.52113	−6.3649	8	−5.5038	9	−4.9789	10	−4.3025	10
3.61011	−8.2089	8	−7.0854	7	−6.1159	8	−4.9982	8

The variable x denotes the values of inverse temperature, y_ℓ is the averaged logarithms of rate, k_ℓ is the number of repeats, $\ell=1,\ldots, 4$.

Table 1 contains the observed values of the significance level of tests (15) and (40) when $\alpha=0.05$. It turns out that the rank-based test (40) is conservative while (15) maintains its significance level rather well (except for the Cauchy distribution.) Table 2 reports the observed values of the mean squared errors of estimators of x_0, y_0 for the normal theory based $\hat{x}_0$, $\hat{y}_0$, and of the Laplace distribution motivated procedure $\tilde{x}_0$, $\tilde{y}_0$ from (40). These errors are rescaled by the corresponding values of σ^2. Not unexpectedly, the estimator $\tilde{x}_0$, $\tilde{y}_0$ while exhibiting in the case I almost the same performance as the normal theory based procedure, outperforms that rule in all other cases. For the Cauchy distribution the contrast is most dramatic as the ratio of the sum of the mean squared errors of $\hat{x}_0$ and $\hat{y}_0$ and of $\tilde{x}_0$ and $\tilde{y}_0$ exceeds 10^7. Also for non-normal distributions the latter estimators typically have much smaller bias than $\hat{x}_0$ and $\hat{y}_0$.

The empirical power of these tests is reported in Table 3 in the case when $m_i\equiv 0$, $b_i=i$, $i=1\ldots, \ell$, i.e., the regression lines are parallel. Even for the normal distribution the power of the

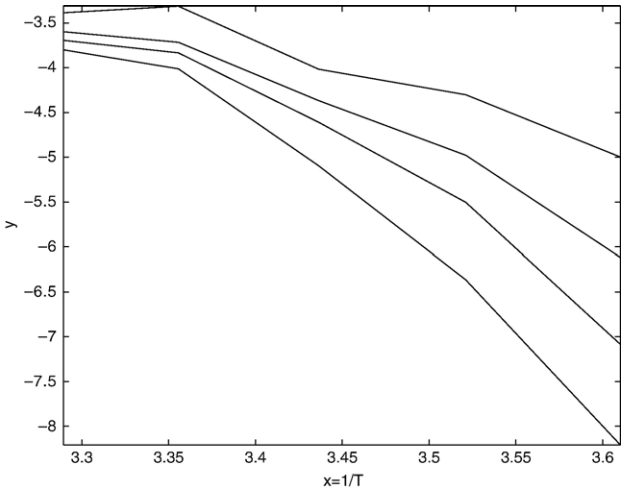

Fig. 1. The data in the HEXA example. The variable x stands for inverse temperature, y's are the averaged logarithms of rate constants.

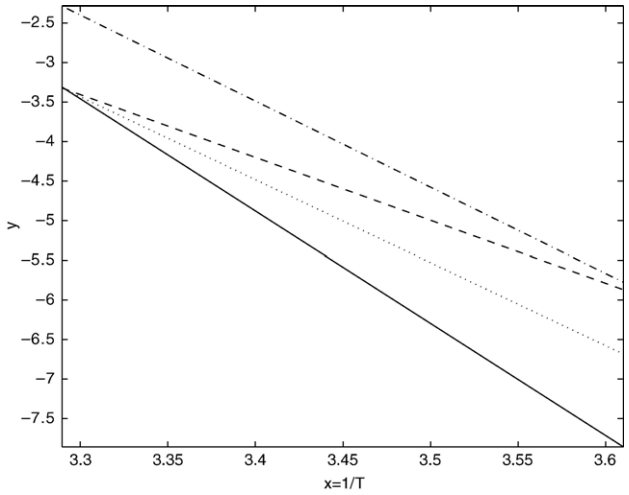

Fig. 2. Four fitted straight lines in the HEXA example. The first sample line is solid, the second dotted (:), the third is dashed (–), and the fourth is marked by dashdots (-.).

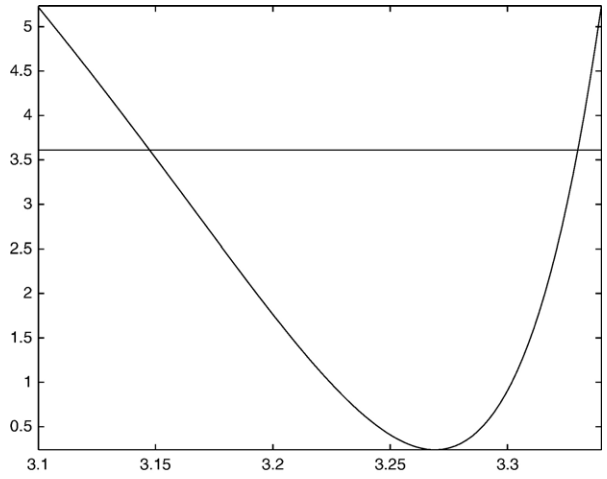

Fig. 4. The graph of function $G(x_0)$ for the HEXA set. The straight line corresponds to $\ell F_\alpha(\ell, \nu)S_P^2/\nu$.

test (15), which is based on this distribution, is just slightly above its nominal level $\alpha=0.05$. This power does not exceed 0.17. However, Eq. (40) has almost zero probability of Type II error. In the Table 4 the standard error (half-width) and confidence coefficient of the confidence intervals (17) and (47) of both procedures are given.

For all considered non-normal distributions the rank-based methods outperform the normal theory based inferential procedures. In this sense the double exponential distribution seems to provide a better model than the normal distribution. Note that one can choose one of these models by using the likelihood ratio test of the most powerful invariant test [13] applied to the residuals.

5. HEXA example

Dr. H. Bamford (Chesapeake Biological Laboratory, University of Maryland) has kindly provided the author with a data set of enthalpies and entropies for hexachlorobiphenyls. In this set summarized in Table 5, with the number of different samples $\ell=4$, x's correspond to reciprocals of temperatures (enthalpies), and y's reported in this table are the averaged logarithms of rate or equilibrium constants (entropies).

In the experiment for each fixed value of the temperature (five distinct values are given in the first column of the Table 5), several repeated measurements were taken. Thus for the first value $T=1/3.28947$ the first sample had $k_1=8$ repeats, the second had $k_2=9$, etc.. The total sample sizes are: $n_1=8+13+6+8+8=43$, $n_2=9+13+6+9+7=44$, $n_3=8+13+5+10+8=44$, $n_4=6+13+5+10+8=42$. The full data set is reproduced in [14]. Fig. 1 shows the scatter-plot of these samples.

The slope estimators are $\hat{m}_1=-14.1876$, $\hat{m}_2=-10.4815$, $\hat{m}_3=-7.9460$, $\hat{m}_4=-10.8954$, and $\bar{y}_1=-5.3429$, $\bar{y}_2=-4.7703$, $\bar{y}_3=-4.4921$, $\bar{y}_4=-3.9640$; $\bar{x}_1=3.4328$, $\bar{x}_2=3.4275$, $\bar{x}_3=3.4367$, $\bar{x}_4=3.4437$; $S_1^2=0.5556$, $S_2^2=0.5513$, $S_3^2=0.5706$, $S_4^2=0.5251$, $S_P^2=0.0895$. The null hypothesis consists in

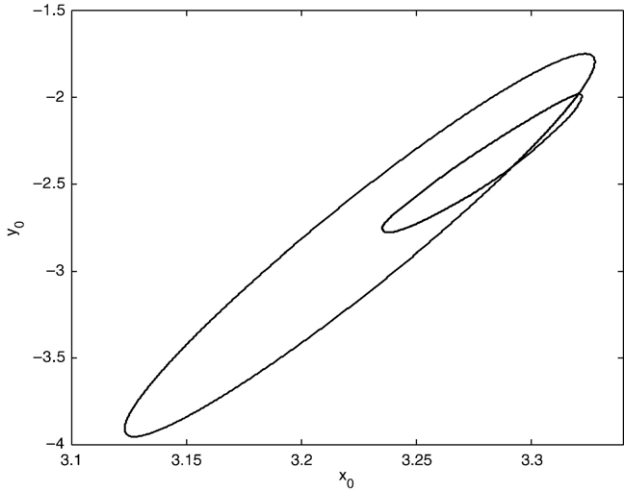

Fig. 3. Two confidence regions for the intersection point (x_0, y_0) in the HEXA example.

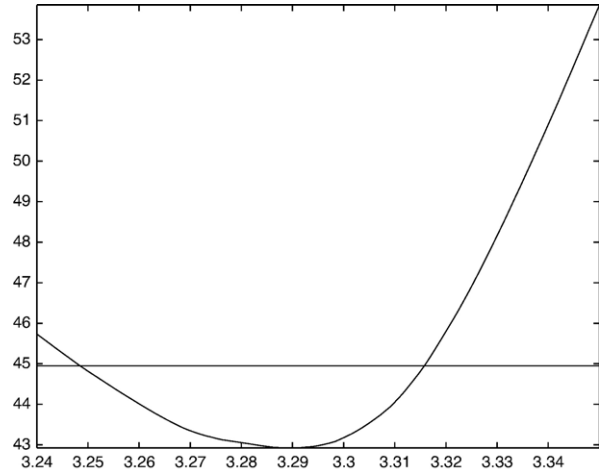

Fig. 5. The graph of function $S(x_0)$ for the HEXA set. The straight line corresponds to $[0.5(\ell-1)F_\alpha(\ell-1, 2(N-\ell))/(N-\ell)+1]\sum|y_{ik}-\tilde{m}_i x_{ik}-\tilde{b}_i|$.

existence of a common intersection point of the fitted lines portrayed in Fig. 2.

The normal theory procedure (15) gave the following answers: $\hat{x}_0=3.279$ and $\hat{y}_0=-3.155$; $F=2.2076$ degrees of freedom 2 and 165, so that the P-value is about 0.126. The corresponding values obtained from rank-based method (47) are $\tilde{x}_0=3.289$, $\tilde{y}_0=-3.185$ with larger P-value 0.877. The approximate normal error-in-variables estimators (25) and (26) were 3.269 and −3.064.

The two 95%-confidence regions for x_0, y_0 are portrayed in Fig. 3. The normal-based confidence region from Eq. (21) (centered at 3.269, −3.075) is much wider than the rank-based region obtained from the corresponding test Eq. (46) (centered at 3.291, −3.190.) This phenomenon has been confirmed by many simulations of non-normal distributions. The normal theory gives 95%-confidence region for x_0 based on Eq. (15) as (3.148, 3.292), while that obtained from rank-based (47) is much shorter: (3.2674, 3.3159) (Figs. 4 and 5).

6. Conclusions and acknowledgement

The suggested robust testing procedure (40) is a good alternative to the normal distribution based test (15). The nature of the latter method is elucidated. An advantage of the rank-based method is that commonly it leads to smaller confidence intervals (regions) for the intersection point, whereas the normal theory based procedure may result in very wide or even infinite intervals especially when outliers are present. Typically the rank-based procedure has larger power function, much smaller bias and smaller mean squared error for distributions whose tails are heavier than normal.

This research was supported by NSA grant #H98230-06-1-0068. The author is grateful to Stefan Leigh for his interest in this work, and for NIST summer students Van Molino and Chiu Yeung for help with calculations.

References

[1] F. Kita, W. Adam, P. Jordan, W.M. Nau, J. Wirz, J. Am. Chem. Soc. 121 (1999) 9265–9275.
[2] L. Liu, Q-X. Guo, Chem. Rev. 101 (2001) 673–695.
[3] H.A. Bamford, D.L. Poster, R.E. Huie, J.A. Baker, Environ. Sci. Technol. 36 (2002) 4395–4402.
[4] C. Ouvrard, M. Berthelot, T. Lamer, O. Exner, J. Chem. Inf. Comput. Sci. 41 (2001) 1141–1144.
[5] H.C. Chen, S.P. Yuan, Commun. Stat., Theory Methods 11 (1982) 395–409.
[6] F.A. Graybill, Theory and Application of the Linear Model, 1976, pp. 285–290, Duxbury, North Scituate, MA.
[7] G. Casella, R. Berger, Statistical Inference, 2nd edition, 2002, p. 395, Duxbury, Pacific Grove, CA.
[8] W.A. Fuller, Measurement Error Models, John Wiley, New York, 1987, p. 31.
[9] D. Draper, Stat. Sci. 3 (1988) 239–271.
[10] P. Bloomfield, W.L. Steiger, Least Absolute Deviations: Theory, Applications and Algorithms, Birkhauser, Boston, 1983, pp 12–73 and pp 181–212.
[11] T.P. Hettmansperger, J.W. McKean, Robust Nonparametric Statistical Methods, Arnold, London, 1998, p. 168.
[12] C.-L. Cheng, J. Van Ness, Statistical Regression with Measurement Error, Arnold, London, 1999, p. 198.
[13] V.A. Uthoff, Ann. Stat. 1 (1973) 170–174.
[14] A.L. Rukhin, Technical Report. Department of Mathematics and Statistics, University of Maryland Baltimore County, TR 2007-2 http://www.math.umbc.edu/misc/technicalpapers/index2007.html.