Lecture 7: Model Identification, Calibration and Validation

Rafiqul Gani
(Plus material from Ian Cameron)

PSE for SPEED
Skyttemosen 6, DK-3450 Allerod, Denmark
rgani2018@gmail.com

www.pseforspeed.com

Overview

- Grey box models and model calibration
- Model parameter and structure estimation:
  - linear-nonlinear
  - static-dynamic
- Model validation

Grey-box Models

Process models

- developed from first engineering principles (white box part)
- part of their parameters and/or structure unknown (black box part)

are called grey-box models

Model Calibration
Conceptual Problem Statement

Given:

- grey-box model
- calibration data (measured data)
- measure of fit (loss function)

Estimate:

- the parameter values and/or structural elements

Model calibration: estimate the model parameters by matching the measured data

Model calibration: conceptual steps of solution

- Analysis of model specification
  - Scenario of simulation versus experiment for measured data
- Sampling of continuous time dynamic models
  - steady state optimization versus dynamic optimization
- Data analysis and preprocessing
  - Uncertainty and consistency issues
- Model parameter and structure estimation
  - Model fitting versus model discrimination
- Evaluation of the quality of the estimate

Model parameter and structure estimation

- Conceptual problem statement
- Least Squares parameter estimation*
  - estimation procedure
  - properties of the estimate
  - linear and nonlinear models
- Parameter estimation for static models
- Parameter estimation for dynamic models

Note: Ignores uncertainty of measured data (assumes correct and consistent data)

Problem statement of model parameter estimation – Example 1

Given:

- System model: $y^{(M)} = \mathrm{M}(x, p^{(M)})$ — *Log P = A - [B/(C + T)]*
- Measured data: $D[1,k] = \{x(i), y(i) \mid i = 1,\ldots,k\}$ — *T_j versus P^*_j*
- Loss function: $L(p) = \left\| y - y^{(M)} \right\|$ — *$L = \mathrm{Sum}_j\ (P_j - P^*_j)$*

Compute: an estimate $\hat{p}$ of $p^{(M)}$ (A, B, C) such that

$$L(p) = \| y - y^{(M)} \| \rightarrow \min$$ — *Min L subject to A, B, C*

Problem statement of model parameter estimation – Example 2

Given:

- System model: $y^{(M)} = \mathrm{M}(x, p^{(M)})$ $K_0 \exp(-E/RT)\, VC_A/F = C_A - C^0_A$
- Measured data: $D[1,k] = \{x(i), y(i) \mid i = 1,\ldots,k\}$ T_j versus C^*_{jA}
- Loss function: $L(p) = \left\| y - y^{(M)} \right\|$ $L = \mathrm{Sum}_j\,(C_{jA} - C^*_{jA})$

Compute: an estimate $\hat{p}$ of $p^{(M)}$ (K_0, E) such that

$$L(p) = \| y - y^{(M)} \| \rightarrow \min$$ *Min L subject to K_0, E*

Note: See also slide 18

Least Squares (LS) parameter estimation

Given:

- System model:

  linear in p, single $y^{(M)}$

$$y^{(M)} = x^T p = \sum_{i=1}^{n} x_i p_i$$

$$f(y) = w \sum_i N_i C_i$$

C_i is the set of p

- Measured data:

$$d(i) = (y(i); x_1(i), \ldots, x_n(i)), i = 1, \ldots, m$$

- Loss function:

$$L_{WSOS}(p) = \sum_{i=1}^{m} \sum_{j=1}^{m} (y(i) - y^{(M)}(i)) W_{ij} (y(i) - y^{(M)}(i))$$

Compute: an estimate $\hat{p}$ of $p^{(M)}$ such that

$$L(p) \rightarrow \min$$

Problem statement of model parameter estimation: Incidence matrix

Exercise: Make the incidence matrix and develop the calculation procedure

- Derive the equations for the optimal solution
- Make the incidence matrix (IC)
- Order the IC to a lower tridiagonal form
- Determine the calculation order

Properties of LS Parameter Estimation

Estimation: $(X^T W X)\hat{p} = X^T W y \quad \text{or} \quad \hat{p} = (X^T W X)^{-1} X^T y$

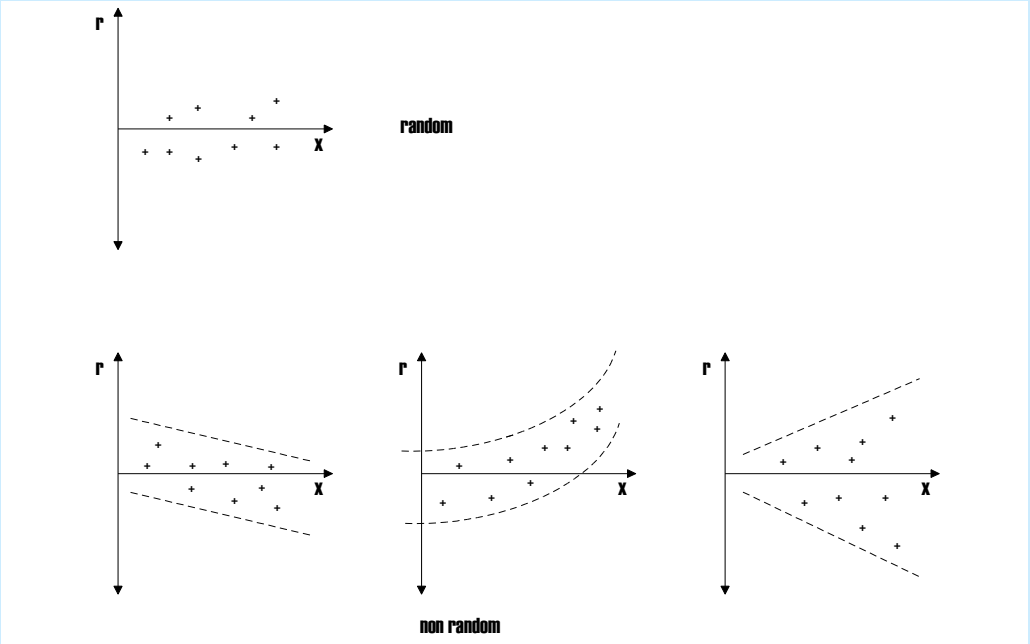
$$X = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_n(1) \\ x_1(2) & x_2(2) & \dots & x_n(2) \\ \dots & & & \dots \\ x_1(m) & x_2(m) & \dots & x_n(m) \end{bmatrix}, \quad y = \begin{bmatrix} y(1) \\ y(2) \\ \dots \\ y(m) \end{bmatrix}$$

with Gaussian measurement errors: $\hat{p} \sim \mathrm{N}(p, \mathbf{COV}\{\hat{p}\})$

- unbiased: $E\{\hat{p}\} = p$
- covariance matrix: $\mathbf{COV}\{\hat{p}\} = (X^T W X)^{-1} \Delta_\varepsilon$

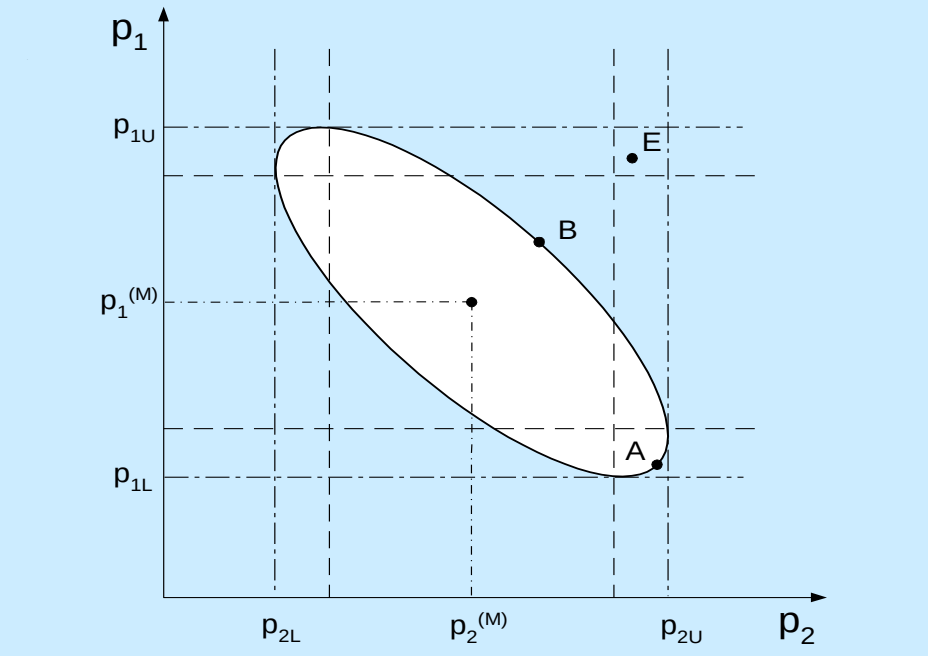
Assessing the Fit

Residuals are independent and $r \sim N(0, \Delta_\varepsilon)$



- residual tests
- correlation coefficient measures
- *See also plots of y (calculated) versus y* (experimental)*

Confidence Regions and Intervals



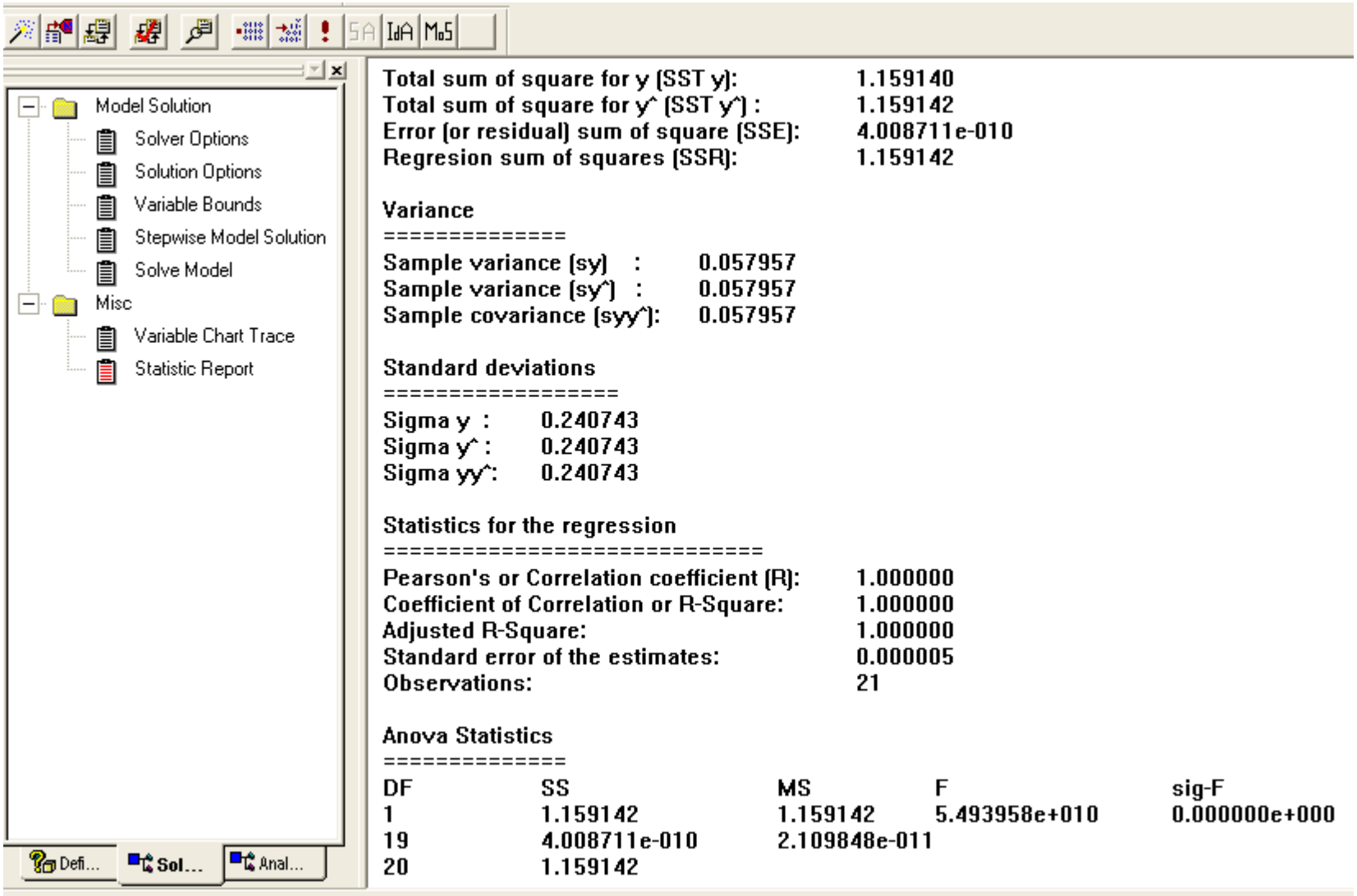
Individual confidence intervals:

$$\hat{p}_i \pm t(m,1-\frac{1}{2}\alpha)s_{p_i} \quad , \quad s_{p_i} = \sqrt{[\mathbf{COV}\{\hat{p}\}]_{ii}}$$

Parameter estimation features in MoT - 1

- Set-up parameter estimation problem
- Solve the problem
- See the ”statistics report”

Parameter estimation features in MoT - 2



LS Parameter Estimation for Nonlinear Models

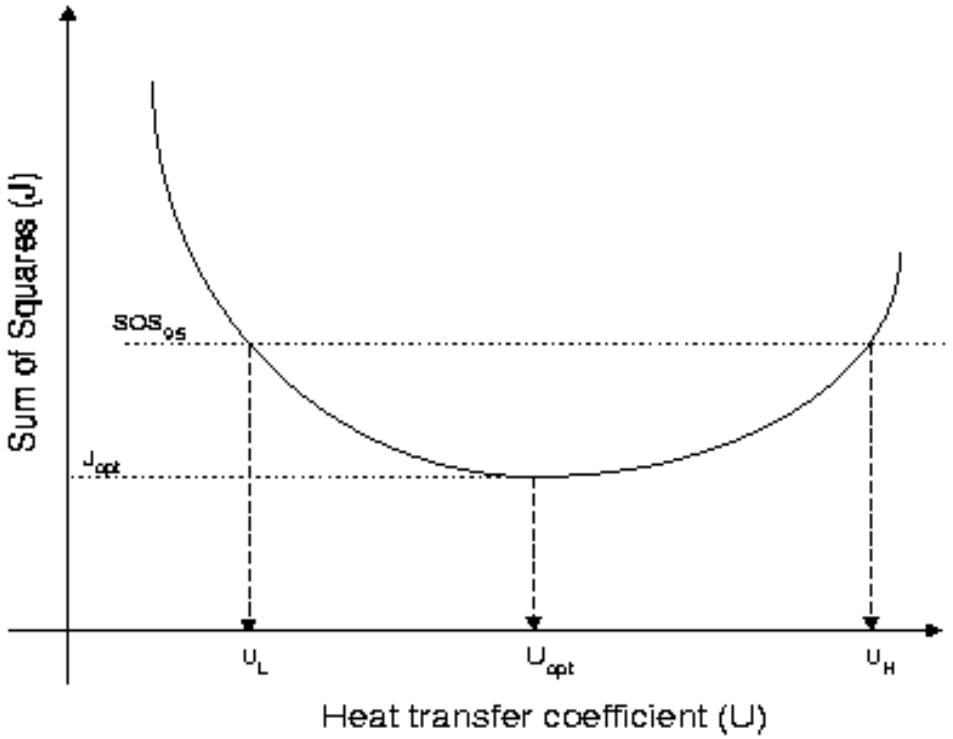
Solution

- Transformation into linear form
- Solution by (numerical) optimization
- Properties

$\hat{p}$ has lost its nice properties

- non-normally distributed
- confidence region and confidence intervals are not symmetric
- unbiased

Confidence Interval for Nonlinear Parameter Estimation



Sum-of-squares as function of a parameter

$$L_{SOS(p)} = \sum_{i=1}^{k} \left\| y(i) - y^{(M)}(i) \right\|_2^2$$

Static Models Linear in Parameters

General form

$$y^{(M)} = x^T p^{(M)} = \sum_{i=1}^{n} x_i p_i^{(M)}$$

Examples

$$y^{(M)} = \sum_{i=0}^{\nu} \xi^i p_i^{(M)} \quad , \quad x = [1 \ \xi \ \xi^2 \ldots \ \xi^{\nu}]^{\mathrm{T}}$$

CSTR

$$\frac{dC_A}{dt} = FC_{Ai} - FC_A - k_A C_A V$$

$$C_{Ai} - C_A = k_A \frac{C_A V}{F}$$

Steady state

$$y = k_A x$$

Linear model

Static models linear in parameters: problem set-up

$$\frac{dC_A}{dt} = FC_{Ai} - FC_A - k_A C_A V$$

$$C_{Ai} - C_A = k_A \frac{C_A V}{F}$$

$$y = k_A x$$

Given: a CSTR-model; data of C_{Ai} versus C_A (with fixed F and V)

Here: $\underline{y} = \underline{C_A}$; $\underline{x} = \underline{C_{Ai}}$; $p = k_A$

Residual function:

$$L(p) = \sum [(y_j - y_j^{(M)}) / ND]^2$$

Table of measured data:

F	V	C_{Ai}	C_A
…	…	…	…
		…	…
		…	…

Solution steps:

1.Assume value for p
2. Calculate $\underline{y} = \underline{C_A}$
3. Calculate $L(p)$
4. If $L(p) < \varepsilon$, stop. Otherwise, repeat from step 1

Identification: Model Parameter and Structure Estimation of Dynamic Models

Properties of the estimation problem

variables (y and x) are time dependent

$$D[1,k] = \{y(\tau), x(\tau) \mid \tau = 1,..., k\}$$

x : present and past inputs and outputs ordered

$$x(\tau) = \{u(\tau); y(\tau-1), u(\tau-1),..., y(\tau-k), u(\tau-k)\}$$

measurement errors on both y and x

Steps **1. sampling continuous time models**

2. estimation

Parameter Estimation of Dynamic Models Linear in Parameters

General form of the input-output model

$$y^{(M)}(\tau) = d^T(\tau)\, p^{(M)} = \sum_{i=1}^{n} d_i(\tau)\, p_i^{(M)}$$

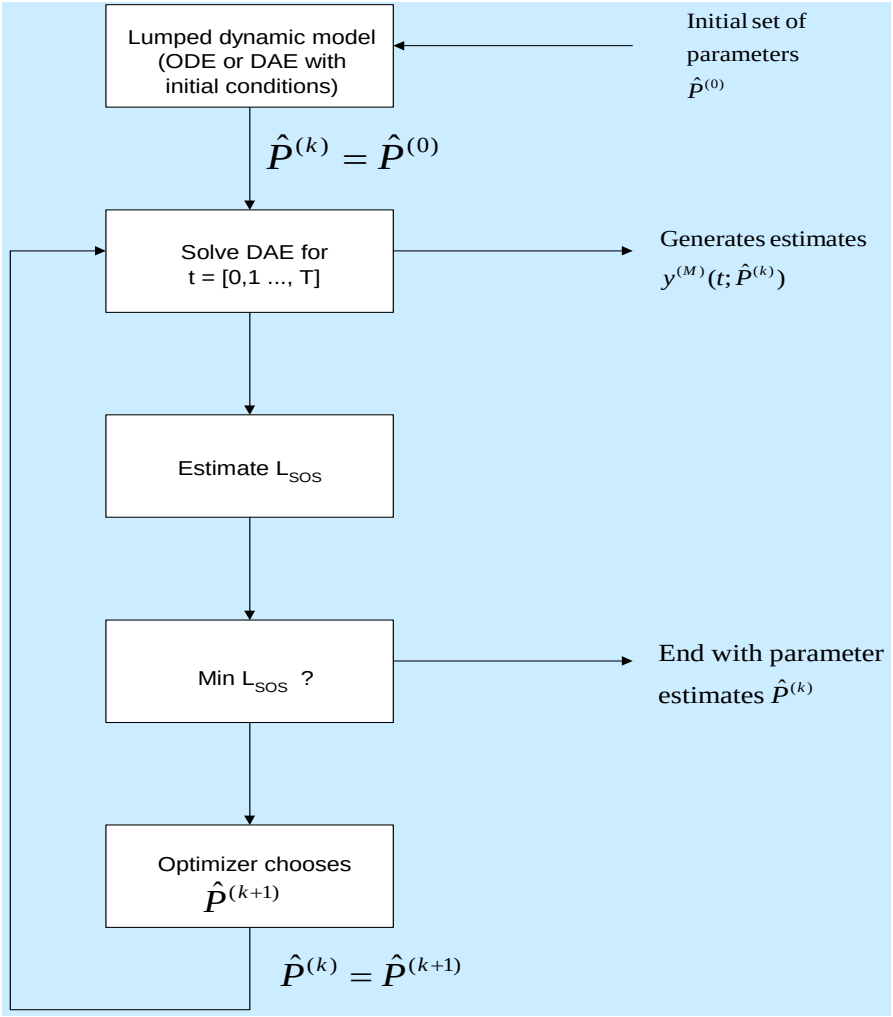
$$d_i(\tau) = \{ y_j(\tau - \theta), u_k(\tau - \vartheta) \mid j = 1,..,m;\ k = 1,..,r;\ 0 \le \theta, \vartheta \le k \}$$

LS parameter estimation with

$$L_{SOS} = \sum_{i=1}^{N} (y(i) - y^{(M)}(i))^2$$

Parameter estimation of nonlinear dynamic models

- General nonlinear programming problem
- Requires optimization code as outer loop
- Requires integration code as inner loop
- Solution tolerances are critical



Nonlinear parameter estimation

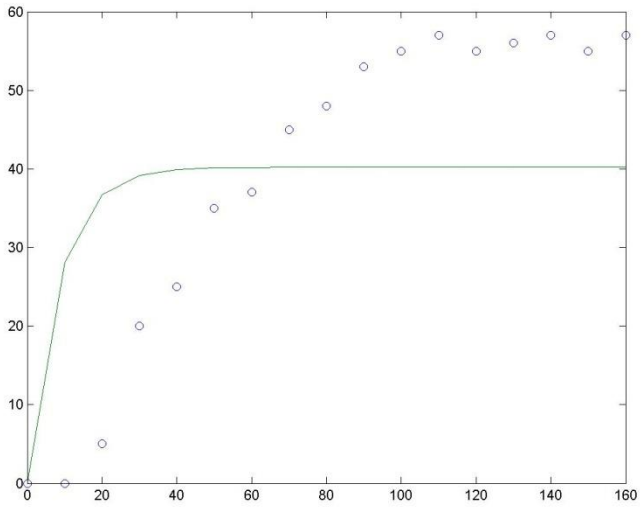
- **Process model:**

$$\frac{dx_1}{dt} = x_2$$

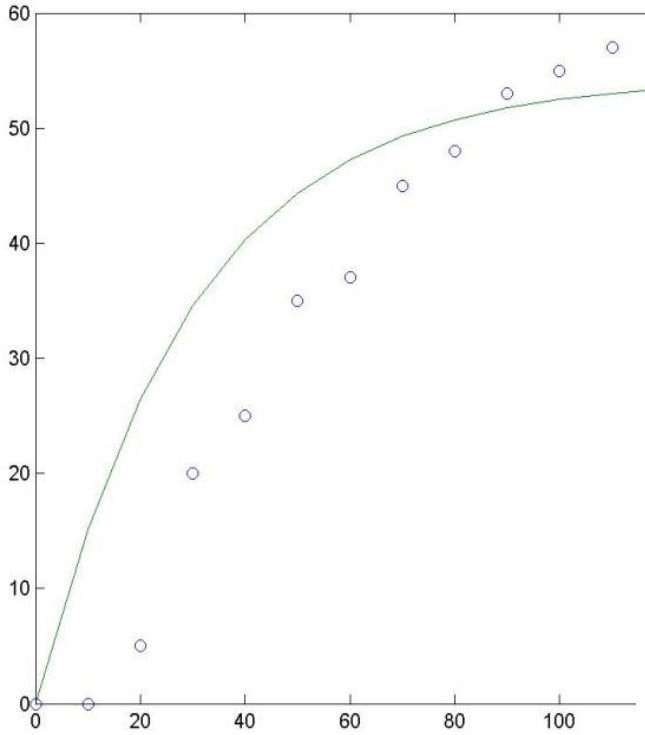
$$\frac{dx_2}{dt} = (p_1 - x_1 - 2p_1p_3x_2) / p_2^2$$

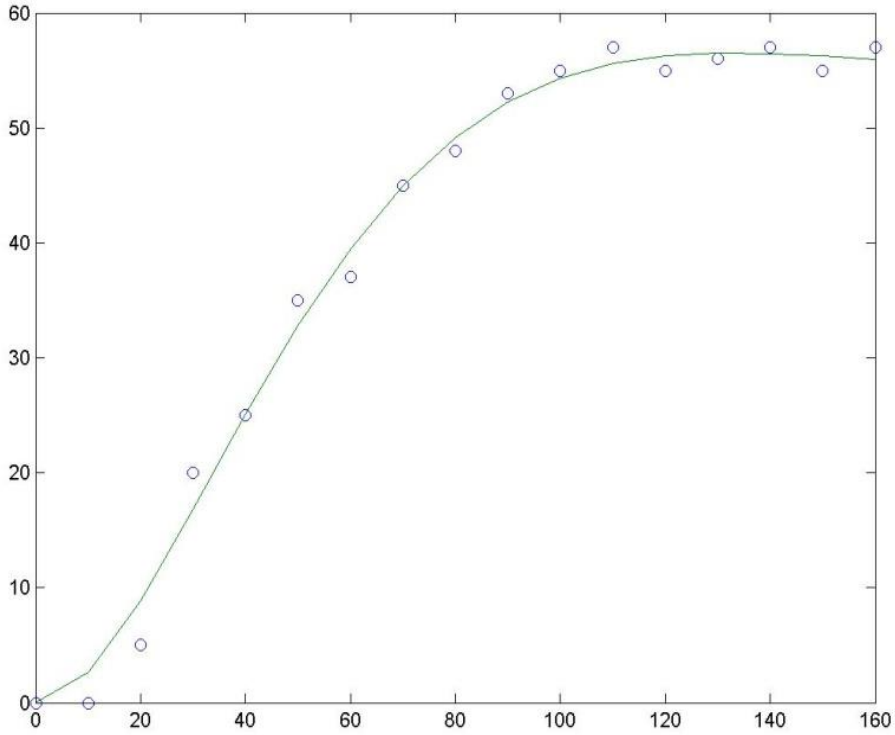
- **Process data:**

(+ initial estimates)



Nonlinear parameter estimation





X-axis: time;

Y-axis: concentration

Sensitivity analysis for parameter selection

Model: M = f(x, y, p); *for example*, **M = f (T, P, A, B, C)**

Calculate derivatives: dM/dp; *for example*,
df (T, P, A, B, C)/dA; **df (T, P, A, B, C)/dB; df (T, P, A, B, C)/dC**

If the absolute value of any derivative (dM/dp) **is close to zero, ignore the parameter**

Plot abs (y – y*) on Y-axis and ± x on X-axis

Follow tutorial in MoT

Sensitivity analysis for parameter selection

Sequential parameter estimation

- Decompose parameter set $\underline{p}$ into sub-sets
- Estimate each sub-set sequentially according to a determined hierarchy

❖ Useful in developing property models (group-contribution models)

❖ Complex kinetic models

Problem Statement of Model Structure Estimation

Given:

- System model: (not parametrized) $y^{(M)} = \mathrm{M}(x)$
- Set of measured data: $D[1,k] = \{x(i), y(i) \mid i = 1,\ldots,k\}$
- Loss function: $L(p) = \| y - y^{(M)} \|$

Compute: an estimate $\hat{\mathrm{M}}$ of M such that $L(p) \rightarrow \min$

+ "candidate structures" in DOM_{M}

Statistical Model Validation via Parameter Estimation:

Conceptual Problem Statement

Given:

- a calibrated model
- validation data (measured data)
- measure of fit (loss function)

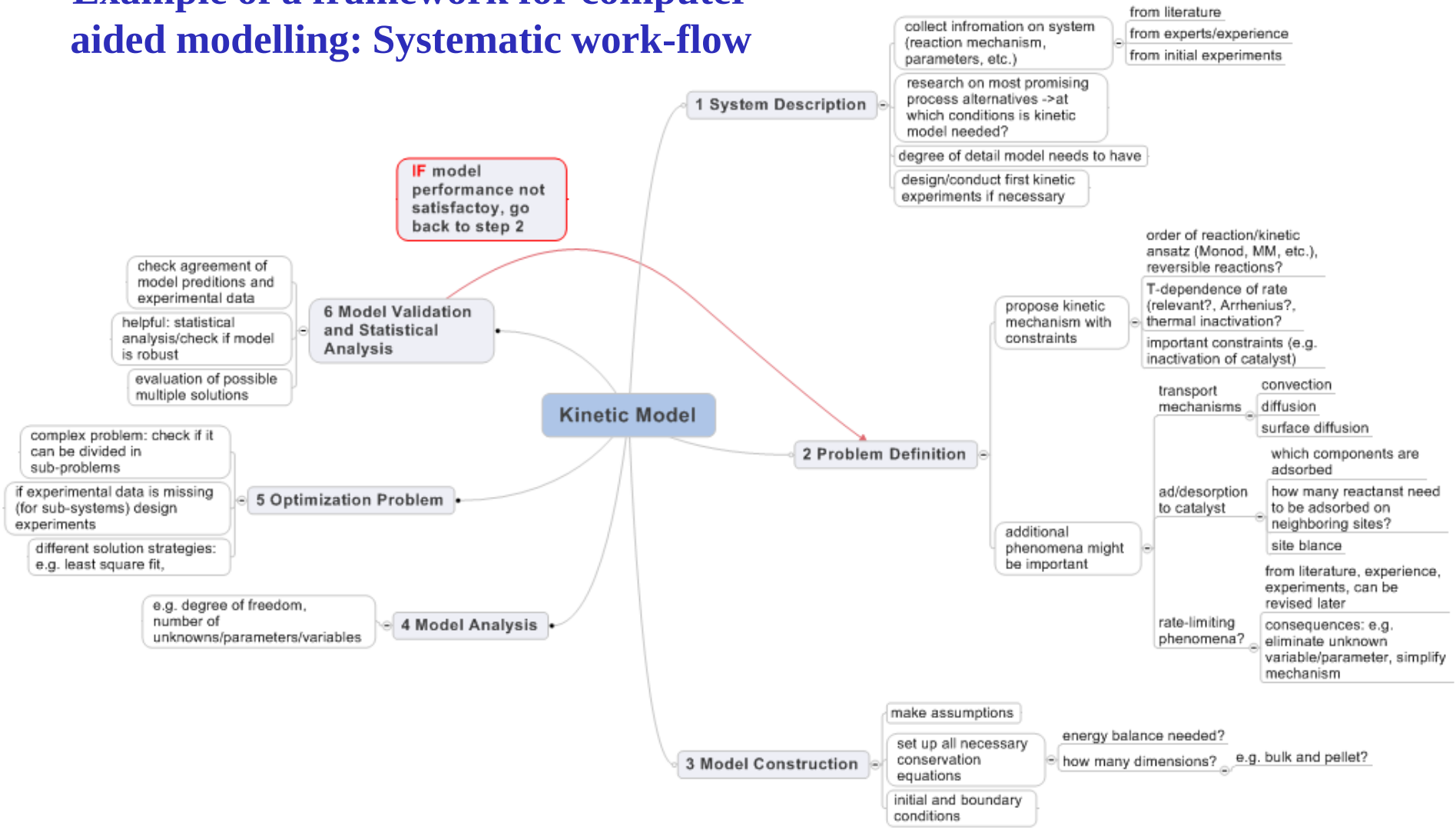
Question:

Is the calibrated model “good enough” for the purpose?

(Does it reproduce the data well?)

Kinetic modelling workflow

Example of a framework for computer aided modelling: Systematic work-flow



Lecture 8: Model discrimination

See tutorial document on model discrimination:

Problem – based on a given data-set with unknown uncertainty, perform parameter estimation and select the best model

Method used: Maximum liklihood principle

Data and Databases

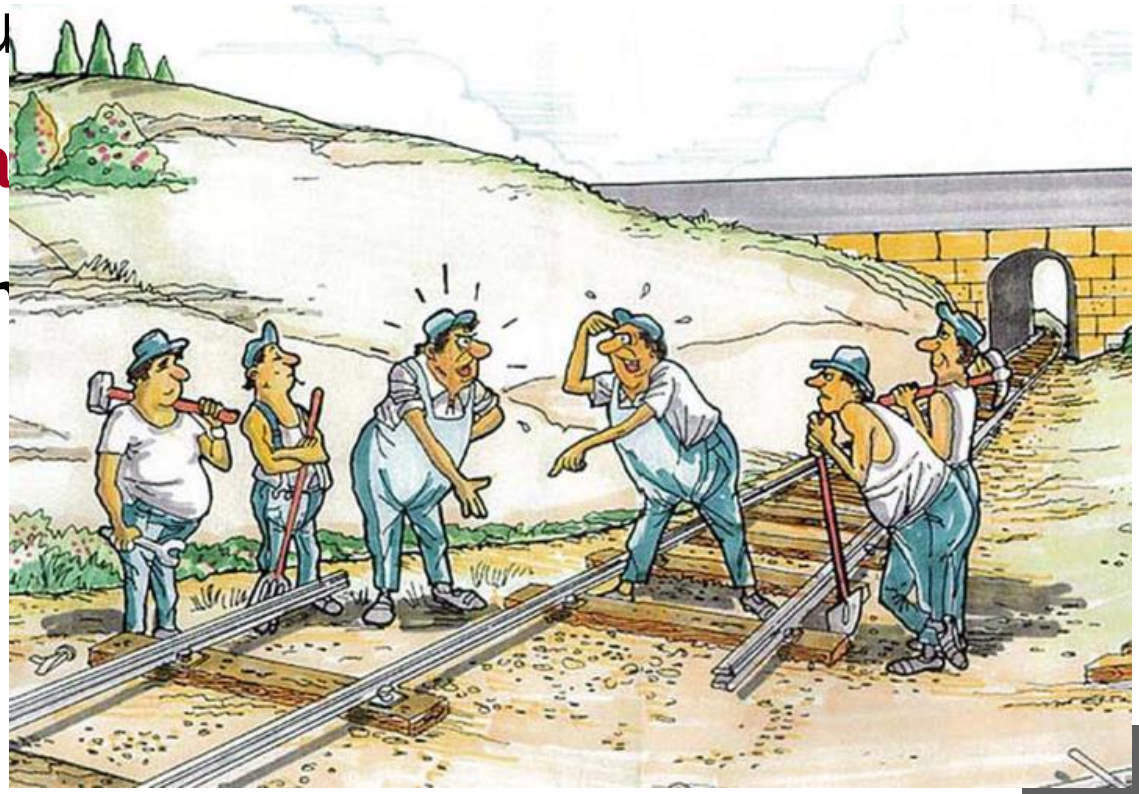
Issue: Data availability versus usability

Drinking from the water-hose problem V. Venkatasubramanian, 2011

Need to manage enormous amounts of data, information & knowledge – see work of NIST

Model rich reliability uncertain

- **Predictive power versus application range**
- **Data used for parameter regression**
- **Extrapolation featu**
- **Uncertainty estima**
- **Model improvemer**



Model Development

- Select function
- Collect data
- Perform regression
- Evaluate performance

Hierarchy of property models

Use a more predictive model to predict the missing parameter

Correlations — Pi = Ai + [Bi/(Ci + T)]

Molecular — Zc = (Pc*Vc)/(83.14*Tc)

CH3-; -CH2-; -OH; …..

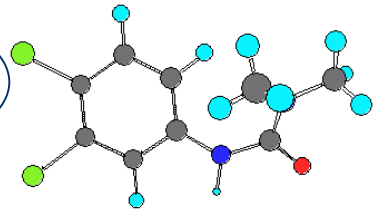
Groups — Tb = 222.543*log(Sum.Groups.I + Sum.Groups.II + Sum.Groups.III)

C, H, O, N, S, ….

Atoms — P=∑niPi + b(vχ0) + 2c(vχ1)

Use the smaller scale (atoms) model to predict the missing parameters of the larger scale (GC)

Micro



Accuracy (verification)

Predictive power (design)

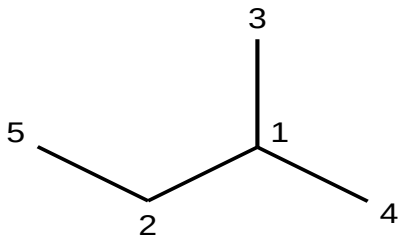
Property Model Development – GC concept

Organic Chemical & Polymer Properties

1st-order: CH_3 , CH_2, CH, OH, COOH, CH3CO, ….

2nd-order: $(CH_3)_2CH$, ..

3rd-order: HOOC-(CHn)m-COOH (m>2, n in 0..2)



$$ {}^{v}\chi^{0} \qquad {}^{v}\chi^{1} \qquad {}^{v}\chi^{2} $$

groups **connectivity index**

REPRESENTING DATA SET AS GROUPS & ATOMS

EXPERIMENTAL DATA SET COLLECTION

Property Model Development

Organic Chemical & Polymer Properties

DETERMINING ATOM or GROUP CONTRIBUTIONS → **MINIMIZATION OF SUM OF SQUARED RESIDUALS (experimental – estimated)**

MARRERO / GANI GC-method (2001)

$$f(y) = \sum n_i C_i + w\sum m_j D_j + z\sum o_k E_k + F$$

Atom-CI Model (2006)

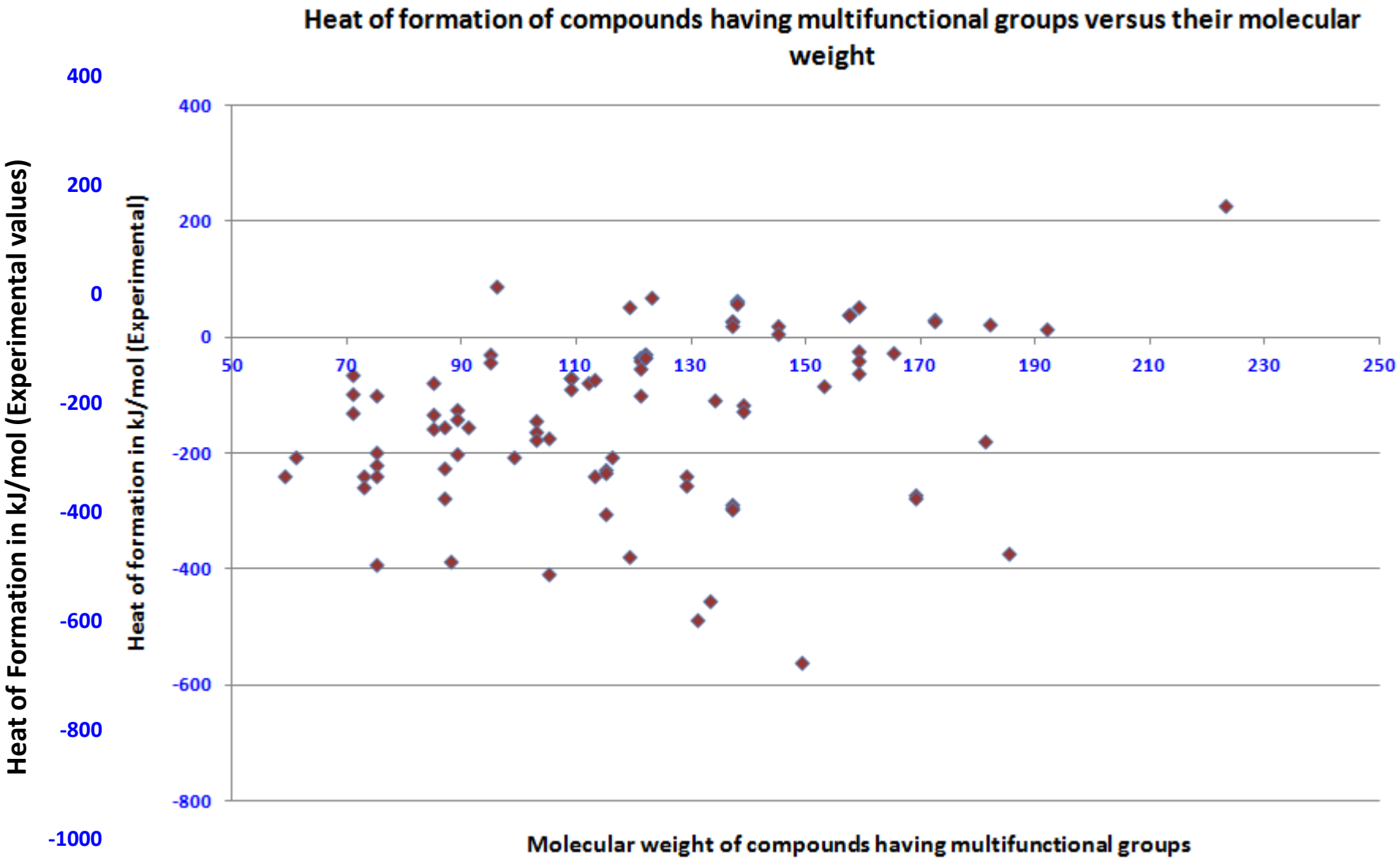
$$f(y)=\sum a_i A_i + b(^{v}\chi^{0}) + 2c(^{v}\chi^{1}) + 2c(^{v}\chi^{2}) + D$$

Note: f(y) is function of property x; example $f(y) = Exp(T_b / t_{bo})$

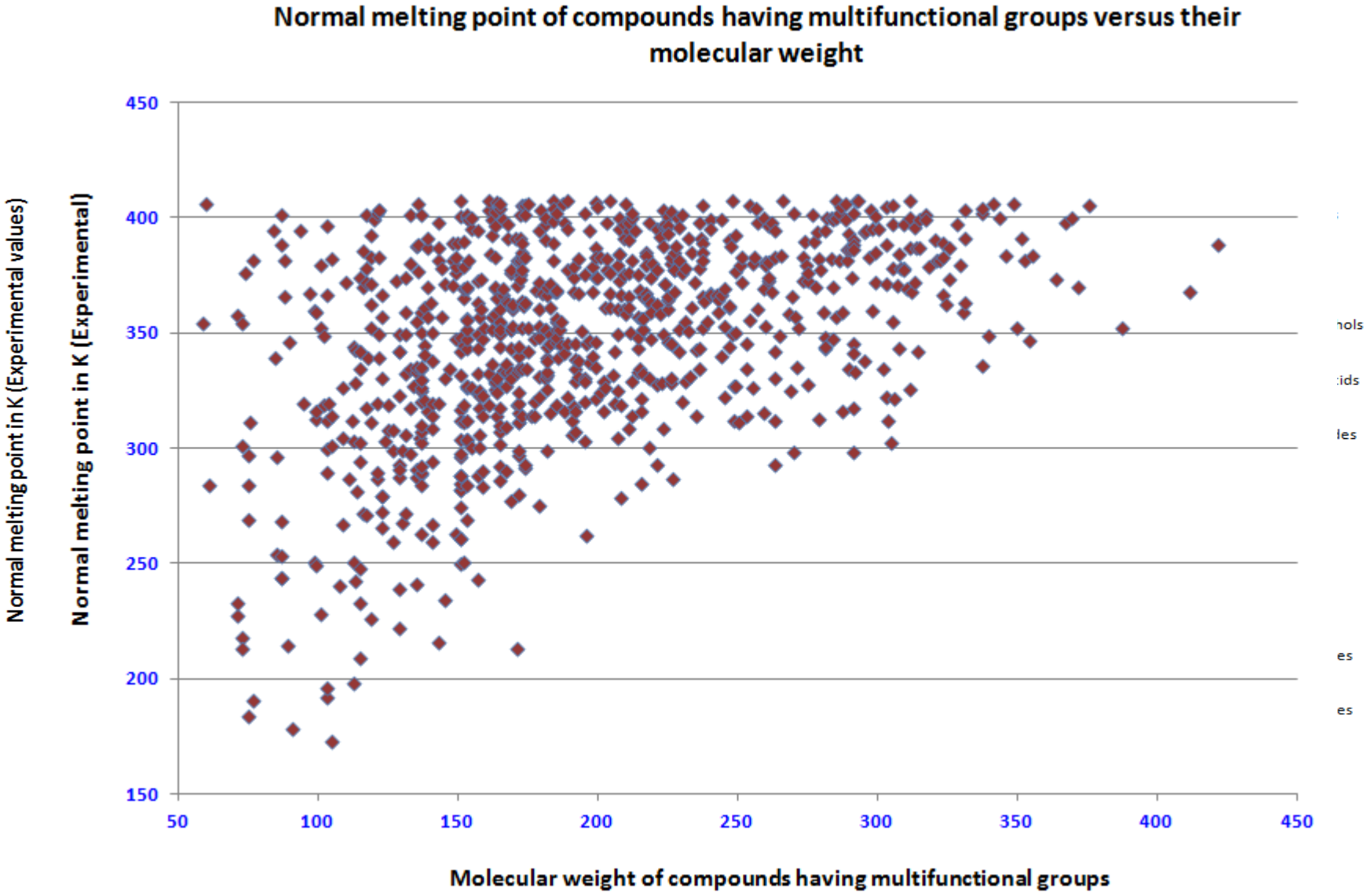
Knowledge representation & data ordering

Class of pure components	T_b	T_c	P_c	V_c	T_m	G_f	H_f	H_{fus}	$LogK_{ow}$	F_p	δ_D	δ_P	δ_H	H_v	H_{vb}	S_{vb}	δ	Ait	ω	V_m
Hydrocarbons	662	277	288	281	492	279	272	270	233	168	73	57	64	189	112	112	386	191	430	324
Oxygenated	1187	297	314	280	1493	266	245	237	1500	229	443	444	444	229	185	185	498	243	659	352
Nitrogenated	369	91	86	76	374	72	149	73	785	42	76	76	75	91	67	67	125	45	140	76
Chlorinated	202	38	30	29	149	29	28	18	271	21	75	74	74	34	30	30	62	28	65	49
Fluorinated	46	21	11	8	41	6	5	10	27	-	18	15	12	15	17	17	26	3	41	29
Brominated	99	8	8	8	89	8	10	11	41	5	29	29	29	23	12	12	20	6	19	15
Iodinated	30	4	5	5	27	5	5	2	10	-	9	9	9	9	7	7	8	-	8	8
P containing	2	2	-	1	2	-	1	1	2	-	-	-	-	-	-	-	1	-	1	1
Sulfonated	109	34	33	32	83	31	29	36	65	2	23	23	23	51	39	39	35	2	73	45
Si containing	14	2	2	-	4	-	-	-	6	2	-	-	-	-	-	-	4	4	2	2
Multifunctional	790	84	75	77	2429	53	138	103	9253	43	291	290	286	64	43	43	219	48	285	155
Total no. of components	**3510**	**858**	**852**	**797**	**5183**	**749**	**882**	**761**	**12193**	**512**	**1037**	**1017**	**1016**	**705**	**512**	**512**	**1384**	**570**	**1723**	**1056**

Knowledge representation & data analysis



Knowledge representation & data analysis



Methodology for parameter estimation and uncertainty analysis – Maximum-likelihood estimation theory

- The minimization of a cost function, $S(\boldsymbol{P}^*)$ i.e. sum of the squares of the difference between the experimental value $X^{\exp}$ and estimated value X^{pred},

1 $$S\left(\boldsymbol{P}^*\right)=\min\sum_{j=1}^{N}\left(X_j^{\exp}-X_j^{\text{pred}}\left(\boldsymbol{P}^*\right)\right)^2$$ yields the values of model parameters $\boldsymbol{P}^*$.

- The covariance matrix $COV\,(\boldsymbol{P}^*)$ for the estimated parameters is given by,

2 $$COV\left(\boldsymbol{P}^*\right)=\frac{SSE}{dof}\left(J\left(\boldsymbol{P}^*\right)^T J\left(\boldsymbol{P}^*\right)\right)^{-1}$$

- The Jacobian matrix $J(\boldsymbol{P}^*)$ is given by,

3 $$J\left(\boldsymbol{P}^*\right)=\frac{\partial X^{\,pred}\left(\boldsymbol{P}^*\right)}{\partial \boldsymbol{P}^*}$$

- The covariance matrix for estimated property value is then obtained by,

4 $$COV\left(X^{\text{pred}}\right)=\left(J\left(\boldsymbol{P}^*\right)\;COV\left(\boldsymbol{P}^*\right)\;J\left(\boldsymbol{P}^*\right)^T\right)$$

References: Seber and Wild, (1989). *Nonlinear Regression*, New York: Wiley.
Sin et al., Computers and Chemical Engineering, (2010), 34, 1385-1392.

Continued...

- The confidence interval of parameters, $\boldsymbol{P}^*$, at α_t significance level (usually a value of 0.05) is given as,

5 $$\boldsymbol{P}^*_{1-\alpha_t} = \boldsymbol{P}^* \pm \sqrt{\text{diag}\left(COV\left(\boldsymbol{P}^*\right)\right)} \,.\, t\left(dof,\ \alpha_t/2\right)$$

- The confidence interval of the predicted property value, X^{pred} , at α_t significance level is given as,

6 $$X^{pred}_{1-\alpha_t} = X^{pred} \pm \sqrt{\text{diag}\left(J\left(\boldsymbol{P}^*\right)COV\left(\boldsymbol{P}^*\right)J\left(\boldsymbol{P}^*\right)^T\right)} \,.\, t\left(dof, \alpha_t/2\right)$$

Following 21 pure component properties were considered for the analysis:

1. Normal melting point, 2. Normal boiling point, 3. Critical temperature, 4. Critical pressure, 5. Critical volume, 6. Standard enthalpy of formation, 7. Standard enthalpy of vaporization (298 K), 8. Standard enthalpy of vaporization (Tb), 9. Standard Gibbs energy, 10. Standard enthalpy of fusion, 11. Entropy of vaporization (Tb), 12. Liquid surface tension (298 K), 13. Liquid viscosity (300 K), 14. Flash point, 15. Auto ignition temperature, 16. Hansen solubility parameters, 17. Hildebrand solubility parameter, 18. Aqueous solubility, 19. Octanol/water partition coefficient, 20. Acentric factor, and 21. Liquid molar volume (298K).

Modelling Issues: Well conditioned data

Consistency, Uncertainty, Predictive Power

$$x^{pred} = f(y) = \sum n_i C_i + w\sum m_j D_j + z\sum o_k E_k + F$$

$$J\left(P^*\right) = \frac{\partial X^{pred}\left(P^*\right)}{\partial P^*}$$

Where *P is the set of model parameters**

The ”condition number” of *J* indicates if the selected model and data-set are identifiable. If the condition number is high, the data-set may not be compatible

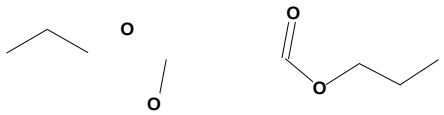
Results: 4.1 Application example
(Estimation of normal boiling point and the 95% confidence interval)

Butanedioic acid, dipropyl ester

CAS No: 925-15-5

Molecular formula: $C_{10}H_{18}O_4$

Molecular structure



First-order groups	Occurrences	Contribution
CH_3	2	0.9218
CH_2	4	0.5780
CH_2COO	2	2.1182

Second-order groups	Occurrences	Contribution
OOC-CHm-CHm-COO (n, m in 1..2)	1	0.2610

Third-order groups	Occurrences	Contribution
No third-order groups are involved		

$$T_b^{pred} = T_{bo}\, ln\left(\sum_i N_i C_i + w\sum_j M_j D_j + z\sum_k E_k O_k\right) = 528.23\ \text{K} \quad \text{Absolute error} = 523.95\ \text{K} - 528.23\ \text{K} = 4.28\ \text{K}$$

Marrero and Gani (2001) method = 535.72 K, Abs. Err. = 11.77 K ; Joback and Reid (1987) method = 544.64 K, Abs. Err. = 20.69 K

Constantinou and Gani (1994) method = 512.18 K, Abs. Err. = 11.77 K

Results: 4.2 Application example
(Calculation of 95% confidence interval of predicted property value)

Covariance matrix $COV(P^*)$ with dimensions (5 × 5) for the groups listed

2

	T_{bo}	CH_3	CH_2	CH_2COO	OOC-CHm-CHm-COO (n, m in 1..2)
T_{bo}	0.0781				
CH_3	6.50 E-04	6.28 E-05			
CH_2	-0.0008	-1.32 E-05	1.17 E-05		
CH_2COO	-0.00079	-1.4E-05	-8.1E-05	0.00021	
OOC-CHm-CHm-COO (n, m in 1..2)	-0.00189	-7.82 E-05	2.09 E-05	4.62 E-04	0.0082

Local sensitivity $J(P^*)$ with dimensions (5 × 5) of T_b model

3

$\delta T_b / \delta T_{b0}$	$\delta T_b / \delta CH_3$	$\delta T_b / \delta CH_2$	$\delta T_b / \delta CH_2COO$	$\delta T_b /\delta$ OOC-CHm-CHm-COO (n, m in 1..2)
2.157898	56.57944	113.1589	56.57945	28.28972

The 95% confidence interval of predicted normal boiling point is calculated by,

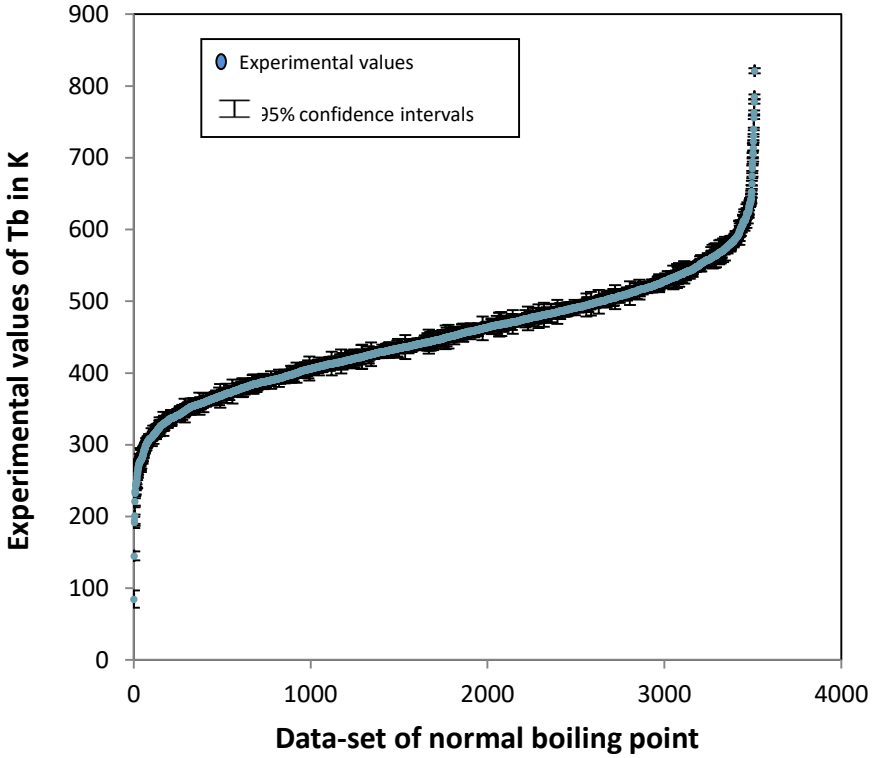
6

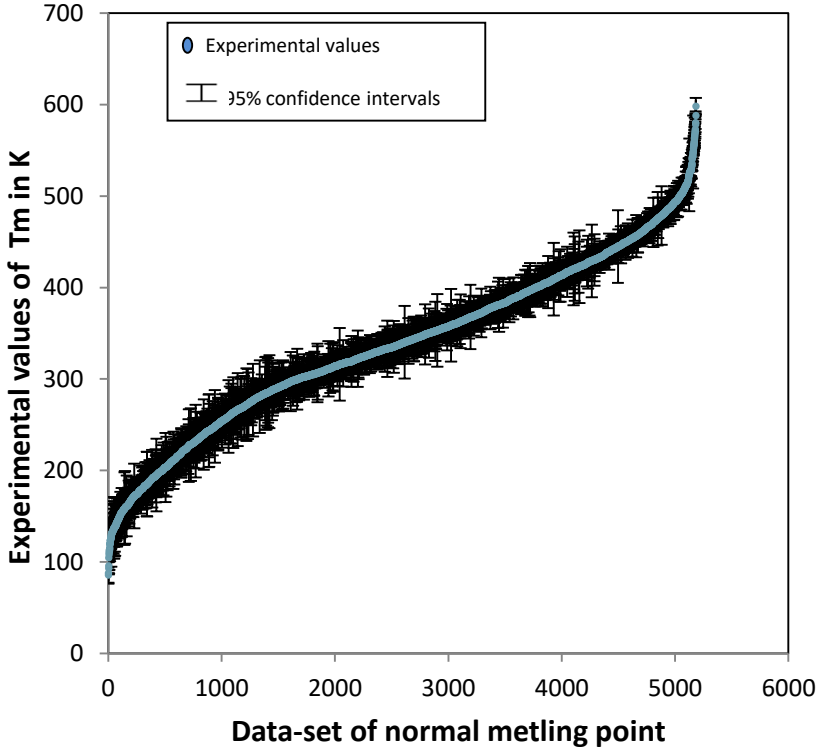
$$T_{b\,(1-0.05)}^{pred} = \underbrace{T_b^{pred}}_{528.23} \pm \underbrace{\sqrt{\mathrm{diag}\left(J(P^*)\,COV(P^*)\,J(P^*)^T\right)}}_{2.445} \cdot \underbrace{t(dof, \alpha/2)}_{1.9607} = 528.23\ \mathrm{K} \pm 4.79\ \mathrm{K}$$

It can be observed that experimental value of the normal boiling point (523.95 K) lies within predicted confidence interval indicating reliability of developed methodology.

Results: 4.3 Reliability of the developed methodology of property estimation and uncertainty analysis

Experimental values together with calculated 95% confidence intervals





- The most of the experimental values falls within calculated 95% confidence intervals.
- This analysis supports linear error propagation method for quantifying model prediction error.

Model improvement: Data analysis