Improved Parameter Estimation in Kinetic Models: Selection and Tuning of Regularization Methods

Attila Gábor and Julio R. Banga

BioProcess Engineering Group, IIM-CSIC, Vigo, Spain
{attila.gabor,julio}@iim.csic.es

Abstract. Kinetic models are being increasingly used as a systematic framework to understand function in biological systems. Calibration of these nonlinear dynamic models remains challenging due to the nonconvexity and ill-conditioning of the associated inverse problems. Nonconvexity can be dealt with suitable global optimization. Here, we focus on simultaneously dealing with ill-conditioning by making use of proper regularization methods. Regularized calibrations ensure the best trade-offs between bias and variance, thus reducing over-fitting. We present a critical comparison of several methods, and guidelines for properly tuning them. The performance of this procedure and its advantages are illustrated with a well known benchmark problem considering several scenarios of data availability and measurement noise.

Keywords: Dynamic models, parameter estimation, Tikhonov regularization, regularization tuning.

1 Introduction

Dynamic mathematical models (i.e. kinetic models) are central in systems biology as a way to understand the function of biological systems [16], to generate new hypotheses, and to identify possible ways of intervention, especially in metabolic engineering [1]. Recent efforts are focused on the development and exploitation of large-scale kinetic models [28].

Parameter estimation aims to find the unknown parameters of the model which give the best fit to a set of experimental data. Parameter estimation belongs to the class of so called inverse problems, where it is important to include both, a priori (i.e. structural) and a posteriori (i.e. practical) parameter identifiability studies. In this way, parameters which cannot be measured directly will be determined in order to ensure the best fit of the model with the experimental results. This will be done by globally minimizing an objective function which measures the quality of the fit.

Global optimization methods must be used in order to avoid convergence to local solutions [2,3]. However, we also need to deal with the typical ill-conditioning of these problems [14], arising from (i) models with large number of parameters,

P. Mendes et al. (Eds.): CMSB 2014, LNBI 8859, pp. 45–60, 2014.

(ii) experimental data scarcity and (iii) significant measurement errors. As a consequence, we often obtain over-fitting of such kinetic models, i.e. calibrated models with reasonable fits to the available data but poor capability for generalization (low predictive value).

Regularization methods have a rather long history in inverse problems [9] as a way to surmount ill-posedness and ill-conditioning. The regularization process introduces additional information, usually by penalizing model complexity and/or wild behaviour. It also has links with Bayesian estimation in the sense that it can be regarded as a way of introducing prior knowledge about the parameters. It has been mainly used in fields dealing with estimation in distributed parameter systems, such as tomography (with applications in geophysics and medicine) and other image reconstruction techniques. Recently, it has enjoyed success in machine learning, gaining attention from the systems identification area [17].

However, the use of regularization in systems biology has been marginal [8], especially regarding kinetic models. Bansal et. al [4] compared Tikhonov and truncated singular value decomposition regularization for the linear regression model of green fluorescent protein reporter systems to recover transcription signals from noisy intensity measurements. Wang and Wang [31] presented a two stage Bregman regularization method for parameter estimations in metabolic networks. A clear conclusion from these studies is that for nonlinear inverse problems, there is no general recipe for the selection of regularization method and its tuning. Further, it is known that even for linear systems, choosing a method from the plethora of existing techniques is nontrivial [6].

Here we present a critical comparison of a wide range of regularization methods applicable to nonlinear kinetic models. Further, we detail a procedure with guidelines for regularization method selection and tuning. Finally, we use numerical experiments with a challenging benchmark problem to illustrate the usage and benefits of regularization.

2 Parameter Estimation in Dynamic Models

We consider kinetic models given by arbitrary nonlinear ordinary differential equations (ODEs) formulated as

$$\begin{aligned} &\frac{dx(t,\theta)}{dt} = f(u(t), x(t,\theta), \theta), \quad y(x,\theta) = g(x(t,\theta),\theta), \\ &x(t_0) = x_0(\theta), \quad t \in [t_0,\, t_f] \ , \end{aligned} \tag{1}$$

where the dynamics of the states $x \in \mathcal{R}_+^{n_x}$ are determined by the vectorfield $f(\cdot)$, $\theta \in \mathcal{P} \subset \mathcal{R}^{n_\theta}$ is the vector of model parameters (e.g. Hill-coefficients, reaction rate coefficients, Michaelis-Menten parameters, etc.), $u(t) \in \mathcal{R}^{n_u}$ denotes the time dependent stimuli, and $\tilde{y}(t)$ are measured values of the observed variables $y(x(t),\theta) \in \mathcal{R}^{n_y}$. The latter are related with the dynamic states via the observation function $g(x,\theta)$.

The parameter estimation problem is usually formulated as the maximization of the likelihood function. The measurement of the j-th observed quantity, taken at time t_i in the k-th experiment is assumed to be contaminated by random measurement error distributed according to the normal distribution, i.e. $\tilde{y}_{ijk} = y_{ijk}(x(t_i), \theta) + \epsilon_{ijk}$ and the error term $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2_{ijk})$, where σ^2_{ijk} is the error variance. Then, the maximization of the likelihood function leads to the minimization of the weighted least squares cost function [30].

$$Q_{\mathrm{LS}}(\theta) = \frac{1}{2} \sum_{k=1}^{N_{\mathrm{e}}} \sum_{j=1}^{N_{y,k}} \sum_{i=1}^{N_{t,k,j}} \left(\frac{y_{ijk}(x(t_i, \theta), \theta) - \tilde{y}_{ijk}}{\sigma_{ijk}} \right)^2 = \frac{1}{2} R(\theta)^T R(\theta) \ , \qquad (2)$$

where N_{e} is the number of experiments, $N_{y,k}$ is the number of observed compounds in the k-th experiment, $N_{t,k,j}$ is the number of measurement time points of the j-th observed quantity in the k-th experiment, and $R(\theta)$ is the normalized residual vector.

2.1 Optimization Method

It is well known that the cost function (2) can be highly nonlinear and nonconvex in the model parameters, so one should use global optimization in order to avoid local optima. However, the current state of the art in global optimization for this class of problems is still somewhat unsatisfactory. Deterministic global optimization methods [22,18] can guarantee global optimality but their computationally cost increases exponentially with the number of estimated parameters. Thus, stochastic methods [19], or meta-heuristic approaches [26] are better alternatives, given adequate solutions in reasonable time, although at the price of no guarantees.

Here, we have used a global-local hybrid metaheuristic which combines scatter search [25] with the very efficient adaptive nonlinear least squares algorithm NL2SOL [7]. In order to further increase the convergence rate of NL2SOL, the Jacobian of the normalised residual vector is computed based on the solution of the forward sensitivity equations corresponding to (1) via the SUNDIALS CVODES [11] software package.

3 Regularization Methods

Here we consider general family of penalty type regularization methods, which incorporate a term $\Gamma(\theta)$ in the optimization cost function

$$Q_{\mathrm{Reg}}(\theta) = Q_{\mathrm{LS}}(\theta) + \alpha \Gamma(\theta) \ . \qquad (3)$$

Specific methods differ in the form of the penalty; e.g. for Tikhonov regularization $\Gamma_{\mathrm{T}}(\theta) = ||W\theta||^2$, where $W \in \mathcal{R}^{n_\theta \times n_\theta}$ is a weighting matrix; for Bregman regularization [31] $\Gamma_{\mathrm{B}}(\theta) = \sum_{i=1}^{n_\theta} \theta_i \log(\theta_i)$; for LASSO regularization [29] $\Gamma_{\mathrm{L}}(\theta) = \sum_{i=1}^{n_\theta} |\theta_i|$; and the so-called elastic net [32] combines the Tikhonov and the LASSO regularization.

The Tikhonov regularized optimization problem can be formulated as

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \frac{1}{2} R(\theta)^T R(\theta) + \alpha (W\theta)^T (W\theta) \\ \text{subject to} \quad & \theta \in \mathcal{P}, \text{ Eqs. (1).} \end{aligned} \tag{4}$$

Since (4) is still a *nonlinear least squares* problem, the above mentioned optimization procedure, with NL2SOL as local method, is still fully applicable. Optimization methods for the LASSO regularization has been reviewed in [27].

3.1 Regularization Parameter (Tuning Methods)

One of the crucial step in the regularization of ill-posed problems is the choice of the regularization parameter α, which balances the model fit and the regularization penalty. Recent studies [6,24] have compared more than twenty parameter choice methods for *linear* inverse problems. In our study, we consider the problem of regularization parameter selection for the nonlinear dynamic problem (4) with the Tikhonov scheme above. However, it should be noted that the methods below are general and applicable for other penalty types, and can also be used in iterative regularization procedures [12]. Note that α is a continuous variable, but below we consider the selection among the set of discrete regularization parameters $\alpha_i = \alpha_{\max} \cdot q^i$, for $0 < q < 1$, and $i = 0, 1, 2 \ldots I$.

Optimal regularization (OR). The optimal regularization minimizes the distance between the estimated parameters and the unknown model parameters, i.e. the estimation error. The expected error in the estimated parameters can be decomposed [6] as

$$\mathbb{E}||\hat{\theta}_\alpha^\epsilon - \theta||^2 = ||\hat{\theta}_\alpha^0 - \theta||^2 + \mathbb{E}||\hat{\theta}_\alpha^0 - \hat{\theta}_\alpha^\epsilon||^2, \tag{5}$$

where $\hat{\theta}_\alpha^\epsilon$ is the estimated parameter vector using noisy measurement data, α is the regularization parameter, θ is the (in general unknown) nominal parameter vector and $\hat{\theta}_\alpha^0$ is the estimated parameters from noise-free data. The first term in the right hand side is the *regularization error*, which accounts for the regularization bias and is a monotonically increasing function of α. The second, variance term is the data noise *propagated error*, which monotonically decreases with increasing α. Therefore, a minimum of the estimation error is expected for a certain α, denoted by α_{opt}. In the discretized framework, if the resolution is fine enough, the problem of finding the optimal regularization parameter is reduced to the selection of the best candidate in the set $\{\alpha_i\}$. It should be noted that OR can only be computed for synthetic problems where the true parameters are known. In other words, the direct computation of (5) is impossible in real problems, since θ and the noise-free data are unknown in practice. The OR results presented below, for the sake of comparison, could be computed because the problems considered are synthetic.

Parameter choice methods. Since in general we do not know the true parameters (as this is obviously the objective of the estimation problem), several parameter choice methods have been developed to find the optimal regularization parameter in an indirect way. Most of the methods have been developed for linear inverse problems (see [6,9,21] and the references therein) or for nonlinear problems in combination with a local (Newton-type) optimization method [12], which cannot handle the nonconvexity of the objective function. In the following part of this section we shortly summarize a selection of existing regularization parameter choice methods that can be used for nonlinear problems in combination with our global metaheuristic optimization approach. In our implementation, the regularized estimates are first obtained on the whole set of α_i for illustrative purposes. However note that most methods can be used in an iterative way, thus reducing the number of regularized solutions required. Several methods require a maximum index I_m of the regularization parameter, such that the optimal index $i_{\text{opt}} \leq I_m$. Details on how the maximum index is computed for those methods are given in Appendix B.

Discrepancy principle (DP)[20]. The discrepancy principle chooses the regularization parameter such that the observed discrepancy between the data and the model prediction is explained by the measurement error, i.e. $||y(\hat{\theta}) - \tilde{y}|| \approx ||\epsilon||$. Since the residuals are normalised (2) and thus each element of the residuals contributes equally to the cost function, the principle chooses the index $n_{\text{DP}} = i$ for which

$$Q_{LS}(\hat{\theta}_{\alpha_i}) \leq \tau N_{\text{data}} \leq Q_{LS}(\hat{\theta}_{\alpha_{i-1}}),$$

where N_{data} is the total number of data and τ is a small tuning parameter of this method. We used $\tau = 1.5$ according to [6] and also 2.0 [9], but did not find significant differences. The results below correspond to $\tau = 1.5$.

Balancing principle (BP1, BP2) [15,6]. The balancing principle chooses the regularization parameter that balances the propagated error bound $||\hat{\sigma}_R(\hat{\theta}_{\alpha_k})||$ and the regularization error. Following [6], the balancing functional is defined as $b(i) = \max\limits_{i<k\leq I_m} \frac{||\hat{\theta}_{\alpha_i} - \hat{\theta}_{\alpha_k}||}{4||\hat{\sigma}_R(\hat{\theta}_{\alpha_k})||}$, where I_m is the maximum regularization index. We considered two submethods: in BP1 the term $||\hat{\sigma}_R(\hat{\theta}_{\alpha_k})||$ was approximated by a local, sensitivity based analysis (A.1) as shown in Appendix B; in case BP2 we used parameter estimates from 4 independent datasets to approximate the standard deviation of the parameters. Then, the smooth balancing functional was computed as $B(n) = \max\limits_{n\leq k\leq I_m} b(k)$. The optimal index ($n_{B1}$ and n_{B2} for the two cases respectively) according to the balancing principle is the first index i such that $B(i) \leq \kappa$, where κ is a tuning parameter. For our test problems $\kappa = 1$.

Hardened-balancing (HB) [5]. This method is a tuning parameter free version of the balancing principle. The smooth balancing functional $B(i)$ is defined as above, but the optimal index is chosen based on the minimisation as $n_{\text{HB}} = \arg\min\limits_{0\leq k\leq I_m} B(k)\sqrt{||\hat{\sigma}_R(\hat{\theta}_{\alpha_k})||}$, where $||\hat{\sigma}_R(\hat{\theta}_{\alpha_k})||$ is computed as (A.1).

Quasi-optimality criterion (QO) [9]. As the regularization parameter decreases, the corresponding estimated model parameters change. When the regularization parameter is large, the estimated parameters are heavily influenced by the regularization term, while for mildly regularized cases, the fit measure prevails. Quasi-optimality is achieved, when the variability of the estimated parameters is minimized, i.e. the optimal regularization index is defined as $n_{\mathrm{QO}} = \arg\min_{0\leq k\leq I_m} ||\hat{\theta}_{\alpha_k} - \hat{\theta}_{\alpha_{k+1}}||$. This method showed high sensitivity to the maximum regularization parameter index I_m.

L-curve method (LC1,LC2)[10]. When $||\hat{\theta}_{\alpha_i}||$ is plotted against $Q_{\mathrm{LS}}(\hat{\theta}_{\alpha_i})$ for $i = 1, 2 \ldots I$, an L-shaped curve is obtained (see Figure 1). The L-curve method chooses the corner point of the curve balancing the propagated error and the regularization error. We considered two variants: method LC1, which identifies the corner by finding the point of the L-curve that has the highest curvature. The corresponding regularization index is n_{LC1}. Method LC2 [23] finds the corner where the tangent of curve is -1, equivalently $n_{\mathrm{LC2}} = \arg\min_{0\leq k\leq I_m} Q_{\mathrm{LS}}(\hat{\theta}_{\alpha_k})||\hat{\theta}_{\alpha_k}||$.

Cross validation (CV_{χ^2},CV_{RSS})[17,24]. When further data is at hand, one can evaluate the performance of the calibrated models with parameters $\hat{\theta}_{\alpha_i}$, $i = 1, 2, \ldots I$ on a second dataset that was not used for the calibration. The performance of the models is measured either by residual sum of squares $\mathrm{RSS}_{\mathrm{CV}}$ or by the χ^2_{CV} defined in Appendix C. The optimal regularization parameter index is chosen as the index of the estimated parameter vector that performed the best in cross validation, i.e. the index selected by the method $\mathrm{CV}_{\mathrm{RSS}}$ is $n^{\mathrm{CV}}_{\mathrm{RSS}} = \arg\min_{0\leq k\leq I_m} \mathrm{RSS}_{\mathrm{CV}}(\hat{\theta}_{\alpha_k})$ and by the method CV_{χ^2} is $n^{\mathrm{CV}}_{\chi^2} = \arg\min_{0\leq k\leq I} \chi^2_{\mathrm{CV}}(\hat{\theta}_{\alpha_k})$.

4 Numerical Experiments

4.1 Test Problems

We have constructed 45 parameter estimation problems as test cases using the three-step metabolic pathway model [13], described in Appendix A. For a given stimuli, the model was simulated using the parameters in Table A.1 and the computed trajectories were sampled. These parameters and sampled trajectories are called the *nominal parameters* and the *nominal model predictions*, respectively. Then, random noise was added to the samples that generated an experimental dataset. We considered parameter estimation problems with:

- 3 levels of experimental data (8, 12 and 16 experiments),
- 3 noise levels per experiment (1, 5 and 10% additive Gaussian noise),
- 5 realizations of each scenario.

Therefore, the total number of scenarios is 45. For the cross validation based method 8 further sets of data were generated by the same procedure. The corresponding stimuli is indicated by "CV" in Table A.1. This data contains 5% error.

Each test problem is solved for a set of regularization parameters: $I = 24$ regularization parameters were chosen a-priory ranging from 10^3 to 10^{-8} equidistantly on logarithmic scale. Equivalently, $\alpha_i = 10^3 \cdot q^i$, for $i = 1, \dots I$ and $q = 0.3325$. Altogether, this results $24 \cdot 45 = 1080$ nonlinear, nonconvex estimation problems to be solved. We have not applied any scaling in the regularization, i.e. W is the unity matrix. The set of regularization parameters was chosen this way to give a uniform base for each regularization tuning method and for illustrative purposes. As mentioned above, a careful implementation of each method could reduce the required points.

4.2 Comparison Criteria

Each tuning method selected a regularized parameter estimate $\hat{\theta}_{\alpha_m}$ by solving (4) for the whole set of α_l, $l = 1, \dots I$ and applying the above tuning procedures. Then, the methods are compared based on well known metrics, such as the residual sum of squares $\mathrm{RSS}(\hat{\theta}_{\alpha_m})$, $\chi^2(\hat{\theta}_{\alpha_m})$ and model prediction error $\mathrm{PE}(\hat{\theta}_{\alpha_m})$ (for details see Appendix C). The inefficiency IE of a tuning method measures the estimation error EE in the chosen regularized estimate compared to the optimal regularized estimate

$$\mathrm{IE}(\hat{\theta}_{\alpha_m}) = \frac{\mathrm{EE}(\hat{\theta}_{\alpha_m})}{\mathrm{EE}(\hat{\theta}_{\alpha_{\mathrm{opt}}})} = \frac{||\theta - \hat{\theta}_{\alpha_m}||}{||\theta - \hat{\theta}_{\alpha_{\mathrm{opt}}}||} \ , \tag{6}$$

where $\hat{\theta}_{\alpha_{\mathrm{opt}}}$ is the parameter estimate based on the optimal regularization parameter α_{opt}, for which the index is $n_{\mathrm{opt}} = \arg\min_{0 \le l \le I} ||\theta - \hat{\theta}_{\alpha_l}||$.

5 Results

Figure 1A) shows the trade-off (3) between model fit $Q_{\mathrm{LS}}(\hat{\theta}_\alpha)$ and regularization penalty $\Gamma_{\mathrm{T}}(\hat{\theta}_\alpha)$ for a typical estimation scenario. Large regularization biases the estimation and cause large discrepancy between the model and the measured data (large Q_{LS}). As the regularization parameter decreases, the discrepancy decreases towards a lower limit, but the variability of the estimated parameters (given by the parameter norm) increases drastically. Some methods, such as the L-curve method, try to come up with an optimal trade-off between the two effects by finding the so-called knee-point of the curve (in other words, they treat the problem as a bi-criteria optimization where the L-curve is a Pareto-optimal set).

5.1 Estimation Error and Optimal Regularization

The parameter estimation error (A.6) was calculated for each of the 1080 estimation problem. Figure 2 shows this magnitude for the 5 replicates of a selected scenario (estimation using 8 datasets containing 10% noise). The error curves can be divided into three regions, as the regularization index increases (i.e., the

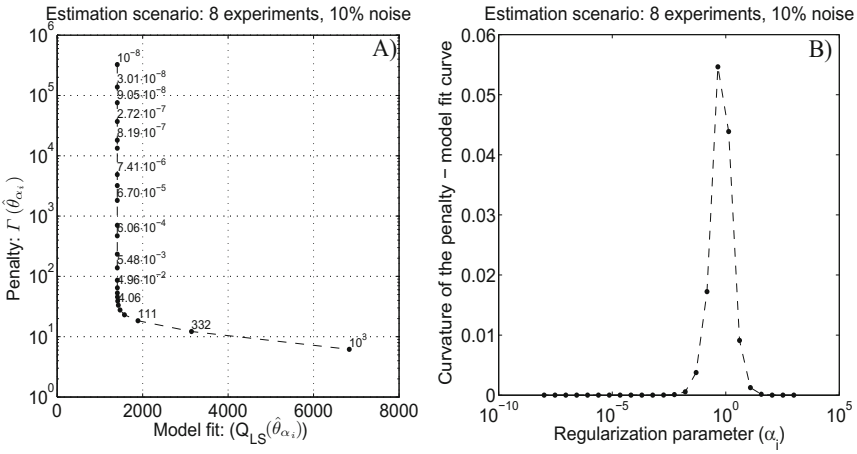

Fig. 1. Results of a regularized estimations. A) Trade-off between model fit and the penalty (3): for a given estimation scenario the regularization penalty $\Gamma_{\mathrm{T}}(\hat{\theta}_{\alpha_i})$ is plotted against the fit of the model $Q_{\mathrm{LS}}(\hat{\theta}_{\alpha_i})$ for each regularization parameter (denoted by the text next to each point), that results in an L-shape curve. Figure B) shows the curvature of the L-shape curve.

regularization parameter decreases). For large regularization parameters, the error in the estimated parameters is dominated by the regularization term (rE domain), while for small regularization parameter the noise propagated error is the main contributor (pE domain). In most cases we found that the propagated error levels off at some value $\mathrm{EE}_{\mathrm{lim}}$ as the regularization parameter reaches a certain limit α_{lim}, which varies with replicates. Below this limit, not only the error in the estimated parameters, but generally the estimated parameters themselves did not change, i.e. $\hat{\theta}_{\alpha_i} \approx \hat{\theta}_{\alpha_{i+1}}$ for all $\alpha_i < \alpha_{lim}$. The theory of inverse problems [9] shows that the regularization parameter must be larger than the smallest eigenvalue of the Hessian of the objective function (2), which can justify our results. Between the rE and pE regions one can find a domain qO, in which the estimation error is smaller than $\mathrm{EE}_{\mathrm{lim}}$. The minimum of the curve ($\mathrm{EE}_{\mathrm{min}}$) is taken at the optimal regularization parameter α_{opt}.

Similar trends and domains can be identified for all the estimation problems, in which the noise level is medium or high. Further, in these cases, the optimal regularization parameter index only slightly varies between 6 and 8. However, in the cases of small measurement noise (1%), the error in the estimated parameters due to the noise propagation is negligible, the pE region is flat and there is not a unique, optimal regularization parameter.

5.2 Performance of the Methods

The statistics described in Section 4.2 were calculated for each scenario and each value of the regularization parameter. The different regularization tuning

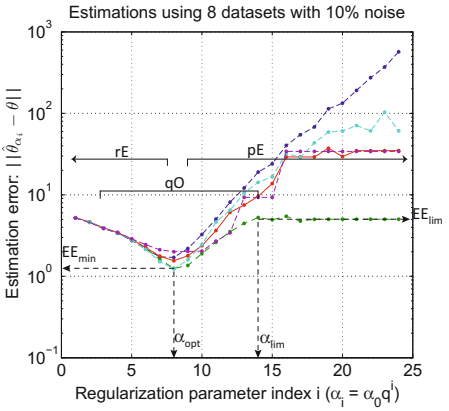

Fig. 2. Parameter estimation error. The estimation error is plotted against the regularization parameters for the 5 replicates of a selected scenario. The notations corresponding to the lowest, green curve: the regularization bias is dominating in domain rE; the noise propagated error is the main contributor to the estimation error in pE. Typically, the propagated error levels off at EE_{lim} as the regularization parameter reaches a certain limit α_{lim}. Any regularization parameter in the qO region gives lower estimation error, than without the regularization ($\alpha = 0$). The optimal regularization parameter α_{opt} corresponds to the minimum of the estimation error curve.

methods (Section 3.1) were used to find the regularization parameter for each scenario. To serve as a reference, we also computed the scores corresponding to the estimations *without* regularization (NR), i.e. $\alpha = 0$. Figure 3 shows the distribution of the inefficiencies given by (6), i.e. the relative parameter estimation error, computed for each regularization method. For the sake of clarity, only the estimation scenarios corresponding to the 10% noise are depicted in the figure. More detailed numerical results for all scenarios can be found in Table A.2 in Appendix D.

From Figure 3 we see that the estimation error grows rapidly as the number of experimental datasets decreases in the non-regularized estimations (NR), i.e. these estimations are greatly affected by ill-conditioning. The same trend can be observed based on the numerical results when the noise level of the data increases, leading to larger inefficiencies of the non-regularized estimations. These results also indicate that, using regularization, such estimation error is reduced up to 2 orders of magnitude. For the more ill-conditioned scenarios (more noise and less data), almost all regularization methods perform better than the non-regularized estimation. For the mildly ill-conditioned cases (more data with less noise), the discrepancy principle (DP) and the L-curve method based on the tangent condition (LC2) perform rather poorly due to over-regularization.

The cross validation based methods (CV_{χ^2}, CV_{RSS}) result in a generally low estimation error and perform the best for the cases when the calibration data is highly contaminated by noise. For the situations where there is no additional data set, the L-curve method based on the maximal curvature detection (LC1),

the quasi-optimality (QO) criteria and the balancing principle (BP1) are the best alternatives from estimation error point of view. All these methods performed similarly well for almost all cases. Among them, the LC1 performed also very well in the mildly ill-posed cases. Furthermore, the LC1 method outperformed the NR case in almost all scenarios from the prediction error (PE) point of view, too (see Table A.2).

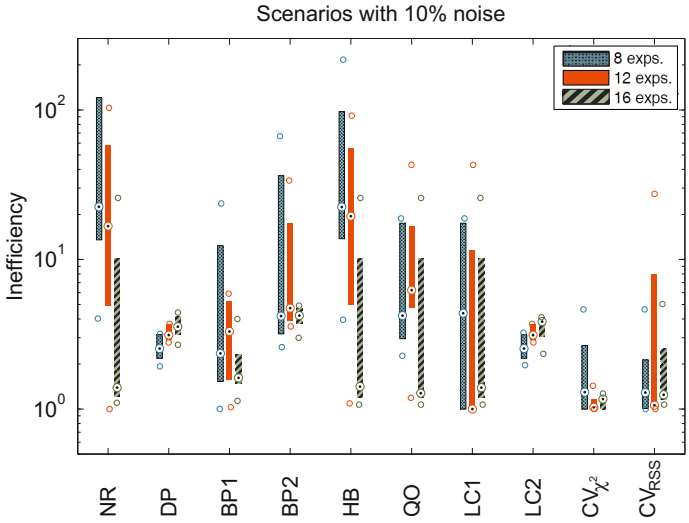

Fig. 3. Inefficiencies of tuning method for three selected scenarios. The inefficiency shows the parameter estimation error normalized by the estimation error with optimal regularization. The color indicates the number of datasets used in the estimations according to the legend. Each column represents the distribution of the 5 measurement error realizations: the circle with the black dot shows the median, the filled area spreads between the 25th and 75th percentiles of the points, the rest of the points are shown individually. NR: non-regularized solution, DP: discrepancy principle, BP: balancing principle, HB: hardened balancing, QO: quasi optimality criteria, LC: L-curve method, CV: cross validation based tuning methods.

6 Conclusions

In this study we considered regularization as a way to improve the calibration of (nonlinear) kinetic models in systems biology, reducing the typical ill-conditioning of these problems. We considered the Tikhonov regularization framework coupled with a global optimization solver. We focused on the specific question of regularization method selection and tuning. We compared several regularization parameter tuning methods, including the discrepancy principle, balancing principle, hardened balancing, quasi optimality criteria, L-curve method and cross validation based methods. The different methods were tuned and tested considering several scenarios of a challenging kinetic model.

Overall, the results obtained indicate that regularization can reduce the parameter estimation error very significantly (up to 2 orders of magnitude for the example considered). The results also indicate that, for the situations where a second data set is available, the cross validation (CV) χ^2 score based method gives the best tuning results. When no further data is available for cross-validation, the L-curve method based on the maximum curvature detection (LC2) is the most robust tuning algorithm.

Acknowledgement. This research was supported by the funding from EU FP7 ITN "NICHE", project no. 289384.

References

1. Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J., Jirstrand, M.: Kinetic models in industrial biotechnology - Improving cell factory performance. Metabolic Engineering 1–22 (April 2014)
2. Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J.A., Blom, J.G.: Systems biology: parameter estimation for biochemical models. FEBS Journal 276(4), 886–902 (2009)
3. Banga, J.R., Balsa-Canto, E.: Parameter estimation and optimal experimental design. Essays in Biochemistry 45, 195–210 (2008)
4. Bansal, L., Chu, Y., Laird, C., Hahn, J.: Regularization of Inverse Problems to Determine Transcription Factor Profiles from Fluorescent Reporter Systems. AIChE Journal 58(12), 3751–3762 (2012)
5. Bauer, F.: Some considerations concerning regularization and parameter choice algorithms. Inverse Problems 23(2), 837–858 (2007)
6. Bauer, F., Lukas, M.A.: Comparingparameter choice methods for regularization of ill-posed problems. Mathematics and Computers in Simulation 81(9), 1795–1841 (2011)
7. Dennis, J.E., Gay, D.M., Welsch, R.E.: An Adaptive Nonlinear Least-Squares Algorithm. ACM Transaction on Mathematical Software 7(3), 348–368 (1981)
8. Engl, H.W., Flamm, C., Kügler, P., Lu, J., Müller, S., Schuster, P., Philipp, K.: Inverse problems in systems biology. Inverse Problems 25(12), 123014 (2009)
9. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of Inverse Problems. Kluwer Academic Publishers (1996)
10. Hansen, P.C., O'Leary, D.P.: The use of the L-Curve in the regularization of discrete ill-posed problems. SIAM Journal on Scientific Computing 14(6), 1487–1503 (1993)
11. Hindmarsh, A.C., Brown, P.N., Grant, K.E., Lee, S.L., Serban, R., Shumaker, D.E., Woodward, C.S.: {SUNDIALS: Suite of Nonlinear and Differential / Algebraic Equation Solvers}. ACM Transaction on Mathematical Software 31(3), 363–396 (2005)
12. Kaltenbacher, B., Neubauer, A., Scherzer, O.: Iterative Regularization Methods for Nonlinear Ill-Posed Problems. Radon Series on Computational and Applied Mathematics. Walter de Gruyter, Berlin, New York (2008)
13. Kitano, H. (ed.): Foundations of Systems Biology. The MIT Press (2001)
14. Kravaris, C., Hahn, J., Chu, Y.: Advances and selected recent developments in state and parameter estimation. Computers & Chemical Engineering 51, 111–123 (2013), http://linkinghub.elsevier.com/retrieve/pii/S0098135412001779

15. Lepskii, O.: On a Problem of Adaptive Estimation in Gaussian White Noise. Theory of Probability & Its Applications 35(3), 454–466 (1991)
16. Link, H., Christodoulou, D., Sauer, U.: Advancing metabolic models with kinetic information. Current Opinion in Biotechnology 29, 8–14 (2014)
17. Ljung, L., Chen, T.: What can regularization offer for estimation of dynamical systems? In: 11th IFAC International Workshop on Adaptation and Learning in Control and Signal Processing. IFAC, vol. 5 (2013)
18. Miró, A., Pozo, C., Guillén-Gosálbez, G., Egea, J.A., Jiménez, L.: Deterministic global optimization algorithm based on outer approximation for the parameter estimation of nonlinear dynamic biological systems. BMC Bioinformatics 13(1), 90 (2012)
19. Moles, C.G., Mendes, P., Banga, J.R.: Parameter Estimation in Biochemical Pathways: A Comparison of Global Optimization Methods. Genome Research 13, 2467–2474 (2003)
20. Morozov, V.A.: Methods for Solving Incorrectly Posed Problems. Springer (1984)
21. Palm, R.: Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Ph.D. thesis, University of Tartu, Estonia
22. Papamichail, I., Adjiman, C.S.: Global optimization of dynamic systems. Comput. Chem. Eng. 28, 403–415 (2004)
23. Regiska, T.: A Regularization Parameter in Discrete Ill-Posed Problems. SIAM Journal on Scientific Computing 17(3), 740–749 (1996)
24. Reichel, L., Rodriguez, G.: Old and new parameter choice rules for discrete ill-posed problems. Numerical Algorithms 63(1), 65–87 (2012)
25. Rodriguez-Fernandez, M., Egea, J.A., Banga, J.R.: Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. BMC Bioinformatics 7, 483 (2006)
26. Rodriguez-Fernandez, M., Mendes, P., Banga, J.R.: A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Bio Systems 83(2-3), 248–265 (2006)
27. Schmidt, M., Fung, G., Rosaless, R.: Optimization Methods for l1-Regularization. Tech. rep. (2009), http://www.cs.ubc.ca/cgi-bin/tr/2009/TR-2009-19.pdf
28. Stanford, N.J., Lubitz, T., Smallbone, K., Klipp, E., Mendes, P., Liebermeister, W.: Systematic construction of kinetic models from genome-scale metabolic networks. PloS one 8(11), e79195 (2012)
29. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58(1), 267–288 (1996)
30. Walter, E., Prorizato, L.: Identification of Parametric Models from experimental data. Springer (1997)
31. Wang, H., Wang, X.C.: Parameter estimation for metabolic networks with two stage Bregman regularization homotopy inversion algorithm. Journal of Theoretical Biology 343, 199–207 (2014)
32. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67(2), 301–320 (2005), http://doi.wiley.com/10.1111/j.1467-9868.2005.00503.x

Authors Contribution

This paper describes one of the main tasks in the Ph.D. of Attila Gabor. The work presented here describes the methods that he has developed as well as the results that he has obtained. Julio R. Banga is his Ph.D. supervisor.

A Kinetic Model of a Three-Steps Metabolic Pathway – Details

The parameters, initial values and stimuli conditions corresponding to the estimation problems can be found in Table A.1. The ODEs read as:

$$\dot{G}_1 = \frac{V_1}{1 + (\frac{P}{Ki_1})^{ni_1} + (\frac{Ka_1}{S})^{na_1}} - k_1 G_1$$

$$\dot{G}_2 = \frac{V_2}{1 + (\frac{P}{Ki_2})^{ni_2} + (\frac{Ka_2}{M_1})^{na_2}} - k_2 G_2$$

$$\dot{G}_3 = \frac{V_3}{1 + (\frac{P}{Ki_3})^{ni_3} + (\frac{Ka_3}{M_2})^{na_3}} - k_3 G_3$$

$$\dot{E}_1 = \frac{V_4 G_1}{K_4 + G_1} - k_4 E_1$$

$$\dot{E}_2 = \frac{V_5 G_2}{K_5 + G_2} - k_5 E_2$$

$$\dot{E}_3 = \frac{V_6 G_3}{K_6 + G_3} - k_6 E_3$$

$$\dot{M}_1 = \frac{kcat_1 E_1 (\frac{1}{Km_1})(S - M_1)}{1 + \frac{S}{Km_1} + \frac{M1}{Km_2}} - \frac{kcat_2 E_2 \frac{1}{Km_3}(M_1 - M_2)}{1 + \frac{M_1}{Km_3} + \frac{M_2}{Km_4}}$$

$$\dot{M}_2 = \frac{kcat_2 E_2 \frac{1}{Km_3}(M_1 - M_2)}{1 + \frac{M_1}{Km_3} + \frac{M_2}{Km_4}} - \frac{kcat_3 E_3 \frac{1}{Km_5}(M_2 - P)}{1 + \frac{M_2}{Km_5} + \frac{P}{Km_6}}$$

B Finding the Maximal Regularization Index

We see from Figure 2, that the estimation error is levelling off for small regularization parameters, i.e. the regularization parameter does not influence the estimation problem any more. The goal of the maximal index is to find the minimum regularization parameter after which the estimation error levels off. However, the curve is not available in practice, since the nominal parameters is required to compute the estimation error. Alternatively, an estimate of the second term in (5) can be made based on the Hessian of the regularized cost function. Let $\tilde{R}(\theta, \alpha) = [R(\theta)^T \sqrt{\alpha}(W\theta)^T]^T$ be the augmented regularized residual vector (c.f. (4)) and define $F_R : \mathcal{R}^{n_p \times 1} \mapsto \mathcal{R}^{n_p \times n_p}$ as $F_R(\theta, \alpha) \doteq \frac{\partial \tilde{R}(\theta,\alpha)}{\partial \theta}^T \frac{\partial \tilde{R}(\theta,\alpha)}{\partial \theta}$. Note that,

Table A.1. Nominal parameter values, parameter estimation bounds, stimuli (input) values and initial conditions (I.C.) for the dynamic model of the 3-Steps Metabolic Pathway

Param.	**Value**	$[\mathbf{p_{lb}}, \mathbf{p_{ub}}]$	**Param.**	**Value**	$[\mathbf{p_{lb}}, \mathbf{p_{ub}}]$	**Param.**	**Value**	$[\mathbf{p_{lb}}, \mathbf{p_{ub}}]$
V1	1	$[10^{-12}, 10^{6}]$	V3	1	$[10^{-12}, 10^{6}]$	V6	0.1	$[10^{-12}, 10^{6}]$
Ki1	1	$[10^{-12}, 10^{6}]$	Ki3	1	$[10^{-12}, 10^{6}]$	K6	1	$[10^{-12}, 10^{6}]$
ni1	2	[0.1, 1]	ni3	2	[0.1, 1]	k_6	0.1	$[10^{-12}, 10^{6}]$
Ka1	1	$[10^{-12}, 10^{6}]$	Ka3	1	$[10^{-12}, 10^{6}]$	kcat1	1	$[10^{-12}, 10^{6}]$
na1	2	[0.1, 1]	na3	2	[0.1, 1]	Km1	1	$[10^{-12}, 10^{6}]$
k_1	1	$[10^{-12}, 10^{6}]$	k_3	1	$[10^{-12}, 10^{6}]$	Km2	1	$[10^{-12}, 10^{6}]$
V2	1	$[10^{-12}, 10^{6}]$	V4	0.1	$[10^{-12}, 10^{6}]$	kcat2	1	$[10^{-12}, 10^{6}]$
Ki2	1	$[10^{-12}, 10^{6}]$	K4	1	$[10^{-12}, 10^{6}]$	Km3	1	$[10^{-12}, 10^{6}]$
ni2	2	[0.1, 1]	k_4	0.1	$[10^{-12}, 10^{6}]$	Km4	1	$[10^{-12}, 10^{6}]$
Ka2	1	$[10^{-12}, 10^{6}]$	V5	0.1	$[10^{-12}, 10^{6}]$	kcat3	1	$[10^{-12}, 10^{6}]$
na2	2	[0.1, 1]	K5	1	$[10^{-12}, 10^{6}]$	Km5	1	$[10^{-12}, 10^{6}]$
k_2	1	$[10^{-12}, 10^{6}]$	k_5	0.1	$[10^{-12}, 10^{6}]$	Km6	1	$[10^{-12}, 10^{6}]$

Inputs:	[S]	[P]		[S]	[P]		[S]	[P]
exp. #1	0.1	0.050	exp. #7	10	0.368	exp. #13	10	0.050
exp. #2	0.1	1.0	exp. #8	10	1.0	exp. #14	0.1	0.368
exp. #3	0.464	0.136	exp. #9	0.1	0.136	exp. #15	0.464	0.050
exp. #4	0.464	1.0	exp. #10	0.464	0.368	exp. #16	2.15	0.136
exp. #5	2.15	0.05	exp. #11	2.15	1.000			
exp. #6	2.15	0.368	exp. #12	10	0.136			
CV. #1	1.0	0.02	CV. #4	4.0	0.02	CV. #7	8.0	0.02
CV. #2	1.0	0.2	CV. #5	4.0	0.2	CV. #8	8.0	0.2
CV. #3	1.0	0.8	CV. #6	4.0	0.8	CV. #9	8.0	0.8

Sampling time points: equidistantly 21 points on [0s 120s].

States	**I.C.**	**States**	**I.C.**	**States**	**I.C.**
G1	0.6667	E1	0.4	M1	1.419
G2	0.5725	E2	0.3641	M2	0.9346
G3	0.4176	E3	0.2946		

as $\alpha \to 0$, $F_R(\hat{\theta}_\alpha, \alpha)$ becomes the observed Fisher Information matrix (FIM). The inverse of the FIM (if exists) is the Cramer-Rao lower bound (CRLB) of the covariance matrix of the parameters [30]. Although, the FIM is practically non-invertible when the estimation is highly ill-posed, the inverse of F_R always exists for sufficiently large $\alpha > 0$ and invertible weighting matrix W. Thus, the α-dependent regularized CRLB is estimated by $\mathrm{CRLB}_R(\hat{\theta}_{\alpha_l}) = F_R(\hat{\theta}_{\alpha_l}, \alpha_l)^{-1}$. The regularized variance is therefore bounded by

$$\sigma_R^2(\hat{\theta}_{\alpha_l}) \geq \mathrm{diag}(\mathrm{CRLB}_R(\hat{\theta}_{\alpha_l})) = \hat{\sigma}_R^2(\hat{\theta}_{\alpha_l}). \tag{A.1}$$

The maximum regularization parameter corresponds to the index

$$I_m = \max_{1 \leq i \leq I} (i \mid ||\hat{\sigma}_R(\hat{\theta}_{\alpha_i})|| < 0.9 \max_{1 \leq k \leq I} (||\hat{\sigma}_R(\hat{\theta}_{\alpha_k})||)), \tag{A.2}$$

where 0.9 is a tuning parameter that tries to avoid the small numerical disturbances.

C Computational Details of Comparison Criteria

Some quantities, such as the residual sum of squares

$$\mathrm{RSS}(\hat{\theta}_{\alpha_l}) = \sum_{k=1}^{N_\mathrm{e}} \sum_{j=1}^{N_{y,k}} \sum_{i=1}^{N_{t,k,j}} \left(y_{ijk} \left(x(t_i, \hat{\theta}_{\alpha_l}), \hat{\theta}_{\alpha_l} \right) - \hat{y}_{ijk} \right)^2 \tag{A.3}$$

and

$$\chi^2(\hat{\theta}_{\alpha_l}) = \sum_{k=1}^{N_\mathrm{e}} \sum_{j=1}^{N_{y,k}} \sum_{i=1}^{N_{t,k,j}} \frac{\left(y_{ijk} \left(x(t_i, \hat{\theta}_{\alpha_l}), \hat{\theta}_{\alpha_l} \right) - \hat{y}_{ijk} \right)^2}{\sigma_{ijk}^2} \tag{A.4}$$

can be easily computed from the model prediction and the data. They measure the explanatory potential of the model with estimated parameter $\hat{\theta}_{\alpha_l}$.

Some quantities, as the *nominal parameters* θ and nominal model prediction (i.e. the measurement error free concentrations), are not known in practice. However, a synthetic framework let us compute these values and we can compare the regularization methods based on these quantities. The prediction error defined as

$$\mathrm{PE}(\hat{\theta}_{\alpha_l}) = \sum_{k=1}^{N_\mathrm{e}} \sum_{j=1}^{N_{y,k}} \sum_{i=1}^{N_{t,k,j}} (y_{ijk}(x(t_i, \hat{\theta}_{\alpha_l}), \hat{\theta}_{\alpha_l}) - y_{ijk}(x(t_i, \theta), \theta))^2 \tag{A.5}$$

measures the distance of the model prediction $y(x, \hat{\theta})$ and the noise-free underlying data y, that is unknown in practical applications. A model that tends to over-fit the data, i.e. fits also the noise in the data, likely to generate a good fit to the estimation data (small RSS value), but performs worst according to the PE.

The accuracy of the estimated parameters is measured by the estimation error:

$$\mathrm{EE}(\hat{\theta}_{\alpha_l}) = ||\hat{\theta}_{\alpha_l} - \theta|| \ , \tag{A.6}$$

which is the 2-norm measure of the deviation of the estimated parameters from the nominal parameters.

D Detailed Numerical Results

Table A.2 contains the averaged statistics corresponding to each tuning method.

Table A.2. Performance of the Parameter Choice Methods. Each statistics is obtained by taking the average of the five replicates. $\mathbf{N_{exps}}$ number of experiments used for the estimations, **N.**: amplitude of the noise in %. **NR**: non-regularized solution, **DP**: discrepancy principle, **B**: balancing principle, **HB**: hardened balancing, **QO**: quasi optimality criteria, **LC**: L-curve method, **CV**: cross validation based

Inefficiency averages: $\langle \mathrm{IE}(\hat{\theta}_{\alpha_m}) \rangle$

$\mathbf{N_e}$	N.	NR	DP	B1	B2	HB	QO	LC1	LC2	$\mathbf{CV_{\chi^2}}$	$\mathbf{CV_{RSS}}$
8	1	1.34	6.09	3.50	12.41	1.06	1.06	1.33	3.50	1.65	1.08
	5	1.34	2.98	1.46	4.12	1.25	1.23	1.25	2.43	1.06	1.12
	10	85.31	2.61	7.46	20.60	63.52	9.10	8.44	2.61	1.99	1.85
12	1	1.07	6.44	3.60	14.48	1.12	1.12	1.07	3.60	2.59	1.23
	5	1.37	3.55	1.66	4.92	1.29	1.31	1.10	3.55	1.23	1.23
	10	34.05	3.24	3.42	11.57	32.25	12.81	9.39	3.24	1.11	6.39
16	1	1.24	8.75	2.84	12.73	1.25	1.24	1.24	2.84	2.50	1.24
	5	1.59	4.17	1.90	6.20	1.56	1.60	1.60	4.17	1.16	1.33
	10	6.91	3.61	2.02	4.14	6.91	6.88	6.91	3.53	1.12	2.05

χ^2 averages: $\langle \chi^2(\hat{\theta}_{\alpha_m}) \rangle$

$\mathbf{N_e}$	N.	NR	DP1	B1	B2	HB	QO	LC1	LC2	$\mathbf{CV_{\chi^2}}$	$\mathbf{CV_{RSS}}$
8	1	1254.3	1635.4	1342.8	5597.5	1256.7	1256.7	1254.3	1342.8	1267.5	1256.2
	5	1305.5	1740.1	1324.2	6444.3	1305.6	1305.6	1305.6	1452.8	1307.1	1308.4
	10	1297.0	1767.5	1297.6	4543.6	1297.0	1486.9	1298.7	1767.5	1300.3	1301.5
12	1	1909.3	2215.6	1971.4	6474.4	1910.2	1910.2	1909.3	1971.4	1940.6	1911.8
	5	2041.2	2565.4	2064.4	6095.7	2041.4	2041.8	2041.6	2565.4	2042.5	2041.8
	10	1988.2	2477.8	1990.6	5137.9	1988.2	1988.3	1989.7	2477.8	1996.2	1994.7
16	1	2676.7	3845.5	2730.6	6997.3	2677.3	2677.3	2676.7	2730.6	2722.1	2676.6
	5	2593.2	3120.8	2622.4	6857.5	2593.4	2593.7	2593.2	3120.8	2599.1	2605.2
	10	2636.2	3514.7	2645.3	6374.5	2636.2	2636.2	2636.2	3344.3	2639.3	2637.4

RSS averages: $\langle \mathrm{RSS}(\hat{\theta}_{\alpha_m}) \rangle$

$\mathbf{N_e}$	N.	NR	DP1	B1	B2	HB	QO	LC1	LC2	$\mathbf{CV_{\chi^2}}$	$\mathbf{CV_{RSS}}$
8	1	0.15	0.21	0.17	0.78	0.15	0.15	0.15	0.17	0.16	0.15
	5	4.45	6.11	4.53	15.49	4.45	4.45	4.45	4.99	4.45	4.47
	10	15.11	20.45	15.14	33.08	15.11	17.66	15.16	20.45	15.17	15.20
12	1	0.24	0.28	0.25	0.80	0.24	0.24	0.24	0.25	0.24	0.24
	5	6.25	7.74	6.32	16.33	6.25	6.26	6.26	7.74	6.25	6.26
	10	22.39	28.15	22.47	41.76	22.39	22.39	22.41	28.15	22.47	22.46
16	1	0.31	0.47	0.32	0.86	0.31	0.31	0.31	0.32	0.32	0.31
	5	7.24	8.93	7.35	18.62	7.24	7.24	7.24	8.93	7.26	7.29
	10	31.32	41.37	31.51	58.06	31.32	31.33	31.32	40.03	31.42	31.37

PE averages: $\langle \mathrm{PE}(\hat{\theta}_{\alpha_m}) \rangle$

$\mathbf{N_e}$	N.	NR	DP1	B1	B2	HB	QO	LC1	LC2	$\mathbf{CV_{\chi^2}}$	$\mathbf{CV_{RSS}}$
8	1	0.069	0.244	0.126	0.799	0.069	0.069	0.069	0.126	0.080	0.070
	5	0.392	1.361	0.472	3.299	0.390	0.390	0.389	0.852	0.390	0.398
	10	0.729	2.383	0.718	3.520	0.731	1.446	0.703	2.383	0.719	0.716
12	1	0.069	0.215	0.115	0.760	0.071	0.071	0.069	0.115	0.091	0.074
	5	0.437	1.266	0.466	3.158	0.431	0.429	0.429	1.266	0.416	0.423
	10	0.696	2.353	0.723	3.765	0.698	0.698	0.645	2.353	0.624	0.657
16	1	0.080	0.401	0.111	0.742	0.082	0.081	0.080	0.111	0.103	0.080
	5	0.376	1.215	0.382	3.315	0.368	0.366	0.376	1.215	0.321	0.318
	10	0.770	3.041	0.852	4.762	0.770	0.765	0.771	2.809	0.788	0.775