Bulletin of Mathematical Biology (2010) 72: 697–718
DOI 10.1007/s11538-009-9466-y

ORIGINAL ARTICLE

Properties of the Proximate Parameter Tuning Regularization Algorithm

Martin Brown[a], Fei He[a,*], Stephen J. Wilkinson[b]

[a] *Control Systems Centre, School of Electrical and Electronic Engineering, The University of Manchester, Manchester M60 1QD, UK*
[b] *Department of Chemical and Process Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK*

Received: 9 December 2008 / Accepted: 2 October 2009 / Published online: 5 January 2010

Abstract An important aspect of systems biology research is the so-called "reverse engineering" of cellular metabolic dynamics from measured input-output data. This allows researchers to estimate and validate both the pathway's structure as well as the kinetic constants. In this paper, the recently published 'Proximate Parameter Tuning' (PPT) method for the identification of biochemical networks is analysed. In particular, it is shown that the described PPT algorithm is essentially equivalent to a sequential linear programming implementation of a constrained optimization problem. The corresponding objective function consists of two parts, the first emphasises the data fitting where a residual 1-norm is used, and the second emphasises the proximity of the calculated parameters to the specified nominal values, using an ∞-norm. The optimality properties of PPT algorithm solution as well as its geometric interpretation are analyzed. The concept of optimal parameter locus is applied for the exploration of the entire family of optimal solutions. An efficient implementation of the parameter locus is also developed. Parallels are drawn with 1-norm parameter deviation regularization which attempt to fit the data with a minimal number of parameters. Finally, a small example is used to illustrate all of these properties.

Keywords Systems biology · Parameter estimation · Inverse modelling · Regularization · Proximate parameter tuning

1. Introduction

In molecular systems biology research, the dynamical properties of biochemical signalling pathways are usually represented as a set of nonlinear ordinary differential equations (ODEs) using the "well-stirred" (Conrad and Tyson, 2006) assumption. When the pathway structure is known, the identification of uncertain kinetic parameters becomes

[*] Corresponding author.
E-mail address: hefei_m@hotmail.com (Fei He).

the central task. However, this is still a limiting step in many systems biology studies (Voit, 2000) due to (1) very limited time series data involving only a few of the state variables; (2) the complexity and nonlinearity of the pathways which contain a large number of reaction species and kinetic parameters; (3) some parameter estimates are highly correlated and the model's outputs are insensitive to certain parameters. From a mathematical viewpoint, this leads to a highly ill-conditioned system identification problem with sparse, noisy exemplar data which produces a large uncertainty in the estimated parameters. To try and overcome this problem, some methods attempt to (Okino and Mavrovouniotis, 1998; Liebermeister et al., 2005; Jin et al., 2007) reduce the model's complexity, which results in the number of parameters being (much) less than the number of experimental data. Other approaches (Banga et al., 2003; Rodriguez et al., 2006a, 2006b) improve the global searching ability of the optimization algorithm using hybrid methods which combine stochastic and deterministic optimization techniques. However, since the biochemical pathway structures are often known, the former strategy can be less appealing, and from a computational viewpoint the later approaches are usually computationally costly and problem dependent.

Therefore, not all of the parameters can be reliably estimated in practice due to poor parameter identifiability (Yue et al., 2006; Gutenkunst et al., 2007). In a strict sense, parameter identifiability asks whether, for a given model structure, there exists a (hypothetically noise free) data set from which the parameters can be uniquely determined. In a systems biology context, there are limits on the amount of data and the degree of excitation that can be used to probe the system, as well as a large amount of experimental noise that corrupts the measurements, all of which significantly affect the degree of identifiability. The limited information content of the exemplar data when compared with the model's complexity results in a parameter estimation process that is singular or near singular. As such the optimization problem is ill-posed (Muller et al., 2008) since there are many competing parameter values which will give the same, or similar, performance when measured against the exemplar data.

In this case, a minimal model may be used to explain the main variations in the data and some form of model selection process is often employed to determine the most significant components. In this paper, an alternative strategy is considered, which retains the original model structure but seeks to make a trade-off between data fitting accuracy and the closeness of the estimated parameters to given nominal values. This can be represented using either a Bayesian prior or as regularization constraints (Lei and Jorgensen, 2001; Papadopoulos and Brown, 2007; Brown et al., 2008; He et al., 2008). This is also the approach adopted by the PPT algorithm (Wilkinson et al., 2008). By using regularization techniques, the parameter estimation process is transformed such that the parameter updates are not independent; rather they lie in small but important subspaces. The regularization constraints penalise the movement of the parameters from their prior/nominal values. The variances of the parameter estimates are therefore reduced at the expense of introducing model bias (Johansen, 1997). However, for kinetic pathway models, it is usually impossible to obtain a perfect fit between dynamic experimental data and the model. Thus, the model bias introduced by regularization is negligible provided that the weight on the regularization term is chosen appropriately (Hansen, 1997).

In this paper, the original PPT algorithm is interpreted as a regularization algorithm and the convergence and optimality properties of PPT are analyzed in detail. In addition, the parameter sensitivity and efficient calculation of the optimal parameter locus are also analyzed and developed for both linear static and nonlinear dynamic models. The regularization interpretation, noise assumptions, and convergence analysis of PPT algorithm are analyzed in Section 2. The dual problem and geometric interpretation of optimal solution properties are given in Section 3.1, with discussion regarding the 1-norm and ∞-norm on the parametric constraint term. The properties and efficient implementation of the optimal parameter locus are provided in Section 3.2. Finally, an illustrative pathway example is studied in Section 4 which validates the theoretical analyses.

2. Regularization interpretation of the PPT algorithm

The proximate parameter tuning algorithm (PPT) (Wilkinson et al., 2008), is a sequential method for estimating the model's parameters. The aim is to balance the model's data fitting ability with a measure of deviation from the given parameters' nominal values. The nominal parameter values represent prior knowledge and, as such, are used to constrain the final parameter estimates when information from the training data is either weak or non-existent. However, in the original paper, only the local sequential algorithm for iteratively updating the parameter values was described. It was difficult to analyse the algorithm's convergence and optimality properties as well as the similarities with other techniques. This section addresses these gaps by introducing a constrained primal optimization problem and showing that the core PPT algorithm corresponds to a sequential, locally linearized sub-problem. Differences with the full PPT algorithm (e.g. weighting factors and additional criteria) are then described. Local parameter convergence is then discussed in terms of the aforementioned primal problem and links are drawn with other regularization approaches.

2.1. *Constrained primal problem*

Consider a constrained optimization approach to parameter estimation where the aim is to minimize a 1-norm of the (logarithmic) residuals while simultaneously keeping the parameters close to the supplied nominal values. In this context, "close" is measured in terms of the maximum absolute (logarithmic) deviation. This can be represented as a constrained (primal) optimization problem of the form:

$$\begin{aligned} \min \quad & f(\mathbf{k}) = \left\| \log \mathbf{y} - \log \hat{\mathbf{y}}(\mathbf{k}) \right\|_1 \\ \text{s.t.} \quad & \left\| \log \mathbf{k}^{\text{nom}} - \log \mathbf{k} \right\|_\infty \le \tilde{k} \end{aligned} \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^N$ is the vector of measured exemplars and $\hat{\mathbf{y}}$ is the corresponding vector of predictions, $\mathbf{k} \in \mathbb{R}^m$ is the model's parameter vector and $\mathbf{k}^{\text{nom}}$ is the vector of the nominal values supplied by the user from existing experience and $\tilde{k}$ is the user defined maximum parameter deviation. A 1-norm is used to penalise the (log transformed) residuals, which results in a median estimator, rather than the more usual mean estimator associated with the least squares problem. This is important when the measurements are unreliable and outliers exist. The ∞-norm constraint on the parameter values means that the maximum deviation from the nominal values must be less than the given value $\tilde{k}$. The logarithmic transformations of both the outputs and the parameters reflect the fact that the relative (rather than absolute) change is being penalised, as discussed further in Section 2.4.

Obviously, when $\tilde{k}$ is large, the optimal parameter vector is the median estimator which simply minimizes the 1-norm of the residuals. When $\tilde{k} = 0$, the only feasible solution is $\mathbf{k} = \mathbf{k}^{\text{nom}}$ and as $\tilde{k}$ increases in size, the optimal solution is a compromise between minimizing the residual 1-norm whilst constraining the parameter values to lie within a hypercube around the nominal values. The corresponding dual problem and optimality properties are discussed further in Section 3.1.

2.2. *Relation to the PPT algorithm*

The original PPT algorithm was specified in terms of solving a sequence of linear programming (LP) algorithms. In this section, the primal representation of the sequential linear programming problem associated with (1) is derived and comparisons are drawn with the PPT algorithm.

2.2.1. *Sequential linear programming*

Consider the constrained optimization problem given in (1). Introducing the non-negative slack variable $\tilde{\mathbf{y}}$ which represents the absolute value of the residuals in log space:

$$\tilde{\mathbf{y}} = \left|\log(\mathbf{y}) - \log\left(\hat{\mathbf{y}}(\mathbf{k})\right)\right| \tag{2}$$

then (1) can be re-formulated as

$$\begin{aligned} \min \quad & f(\mathbf{k}, \tilde{\mathbf{y}}) = \mathbf{1}^T \tilde{\mathbf{y}} \\ \text{s.t.} \quad & \tilde{\mathbf{y}} \geq \log \mathbf{y} - \log \hat{\mathbf{y}}(\mathbf{k}) \\ & \tilde{\mathbf{y}} \geq -\left(\log \mathbf{y} - \log \hat{\mathbf{y}}(\mathbf{k})\right) \\ & \mathbf{1}\tilde{k} \geq \log \mathbf{k}^{\text{nom}} - \log \mathbf{k} \\ & \mathbf{1}\tilde{k} \geq -\left(\log \mathbf{k}^{\text{nom}} - \log \mathbf{k}\right) \end{aligned} \tag{3}$$

where $\mathbf{1} \in \mathbb{R}^m$ is a column vector of ones. Here, the objective function is linear and the only non-linearity in (3) occurs in the constraints which represent the model's output dependence on its parameters. Now consider linearizing the optimization problem (3) around the current parameter value $[\mathbf{k}^0, \tilde{\mathbf{y}}^0]$ where the incremental parameter updates are defined by $[\Delta\mathbf{k}, \Delta\tilde{\mathbf{y}}] = [\log(\mathbf{k}) - \log(\mathbf{k}^0), \log(\tilde{\mathbf{y}}) - \log(\tilde{\mathbf{y}}^0)]$. This produces the sequential local, LP sub-problem:

$$\begin{aligned} \min \quad & f(\Delta\mathbf{k}, \tilde{\mathbf{y}}) = \mathbf{1}^T \tilde{\mathbf{y}} \\ \text{s.t.} \quad & \tilde{\mathbf{y}} \geq \log \mathbf{y} - \log \hat{\mathbf{y}}\left(\mathbf{k}^0\right) - \mathbf{S}\left(\mathbf{k}^0\right)\Delta\mathbf{k} \\ & \tilde{\mathbf{y}} \geq -\left(\log \mathbf{y} - \log \hat{\mathbf{y}}\left(\mathbf{k}^0\right) - \mathbf{S}\left(\mathbf{k}^0\right)\Delta\mathbf{k}\right) \\ & \mathbf{1}\tilde{k} \geq \log \mathbf{k}^{\text{nom}} - \log \mathbf{k} \\ & \mathbf{1}\tilde{k} \geq -\left(\log \mathbf{k}^{\text{nom}} - \log \mathbf{k}\right) \end{aligned} \tag{4}$$

where the sensitivity matrix $\mathbf{S}(\mathbf{k}^0)$ is the local derivative of the model's outputs with respect to the parameters, and defined as $\mathbf{S}(\mathbf{k}^0) = d\log(\hat{\mathbf{y}}(\mathbf{k}^0))/d\log(\mathbf{k})$. When considering both the 1 and ∞-norms on both the residuals and parameters as in the original PPT algorithm, the corresponding LP-sequential algorithm can be derived in a similar way. As the objective and constraints are linear, the sequential PPT algorithm is computationally efficient. This has been exploited in a recent analysis of eukaryotic protein translation (Dimelow and Wilkinson, 2009), where the PPT algorithm was utilised to generate repeated solutions using nominal values that were sampled from the uncertain parameter space. It is now possible to draw similarities and differences with the PPT algorithm.

2.2.2. *Relation to the PPT algorithm*

In Appendix A, the original sequential PPT algorithm is given by (A.13) and the equivalent primal optimization problem is given by (A.1). The two optimization problems are very similar and the splitting of the parameter and prediction updates into positive and negative components in (A.13) does not alter the original optimization problem. There exist a number of differences between the original sequential PPT algorithm (A.1) and the simplified primal optimization problem (1):

1. The original PPT objective function (A.1) contains both 1 and ∞-norms on both the residuals and the parameters.
2. The original PPT objective function (A.1) is a weighted combination of the different terms, rather than the constrained approach in (1).
3. There are a number of weighting parameters in the original PPT algorithm (A.1), rather than the single maximum deviation parameter in (1).

Each of these differences will now be discussed.

2.2.3. *Residual and parameter objective norms*

The reason for employing 1 and ∞-norms in the original PPT algorithm was to simplify the transformation of the original optimization problem to an equivalent sequential LP problem. This is because both of these norms can be expressed as linear objectives or constraints on the parameters and residuals by introducing slack variables. Also, the reason for considering a linear combination of a 1-norm and ∞-norm (of both the residuals and the parameters) was to try and mimic a more general p-norm, where differences in the weighing parameters would allow a gradual transition between the 1-norm and ∞-norm optimal solutions. However, this requires the weights on the 1 and ∞-norms to be functions of the optimal parameter estimates, which are obviously unknown prior to the optimization calculation. Therefore, the original motivation for including weighted 1 and ∞-norms on both the residuals and parameter deviations does not apply and despite the original PPT optimization problem involving more criteria, the core of the approach is described by (1).

2.2.4. *Constrained optimization and weighted objectives*

Another obvious difference between (1) and (A.1) is that in the former, the parameter deviations are specified as a constraint, whereas in the latter, they are included in the objective function as a weighted criterion. Whilst these are similar for convex, differentiable problems, there is an important difference in this case.

For (1), let's assume that the objective function is convex with a single (global) minimum. Then the constrained solution will lie on the boundary of the constraining hypercube, assuming that $\tilde{k}$ is sufficiently small. Then the estimated parameters, $\hat{\mathbf{k}}(\tilde{k})$, will vary smoothly as $\tilde{k}$ is changed. Now consider the (simplified) PPT weighted primal objective:

$$\min f(\mathbf{k}) = \left\| \log \mathbf{y} - \log\bigl(\hat{\mathbf{y}}(\mathbf{k})\bigr) \right\|_1 + \alpha \left\| \log \mathbf{k}^{\text{nom}} - \log \mathbf{k} \right\|_\infty \tag{5}$$

where α is the non-negative weighing parameter. This is a piecewise continuous function, where derivative discontinuities occur whenever a residual is zero or when the absolute value of pairs (or more) of parameter deviations are equal to each other. As the weighting parameter, α, is slowly varied the optimal point will remain unchanged until one of one sided gradients at that point change sign and the optimal parameter values perform a discrete jump to the next optimal point. Therefore, the set of solutions from the original PPT algorithm forms a discrete subset of the continuous span of solutions given by (1) and a small change in the PPT weighting parameter would either have no effect or would produce a jump change.

2.2.5. *Weighting parameters*

In the original PPT algorithm (A.1), there are number of weighting parameters involving both the prediction and the parameter errors, whereas in (1), only a single maximum parameter deviation is considered. Obviously, separate non-negative weighing parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be introduced in (1) so that each data point is weighted as

$$\left\| \log \mathbf{y} - \log \hat{\mathbf{y}}(\mathbf{k}) \right\|_{1,\boldsymbol{\alpha}} = \sum_i \alpha_i \left| \log y_i - \log \hat{y}_i(\mathbf{k}) \right| \tag{6}$$

and the parameter deviations are constrained by

$$\begin{aligned} \boldsymbol{\beta}^{\max} \tilde{k} &\geq \log \mathbf{k}^{\text{nom}} - \log \mathbf{k} \\ \boldsymbol{\beta}^{\min} \tilde{k} &\geq -\bigl(\log \mathbf{k}^{\text{nom}} - \log \mathbf{k}\bigr) \end{aligned} \tag{7}$$

Without loss of generality, all of this can be incorporated into (1), and the simplified form (1) will be considered for the rest of the paper.

2.3. *Convergence analysis*

Pathway parameter estimation using the PPT algorithm involves solving a sequence of LP sub-problems. For computational efficiency, the LP-PPT algorithm, only uses first order gradient/sensitivity information (no second order approximation of Hessian is considered). As the PPT algorithm relies on the first order, local sensitivities, convergence to the global minimum of the objective function (2) cannot be guaranteed. However, if both objective function and constraints in (4) are smooth and locally convex, and good prior (nominal) parameters $\mathbf{k}^{\text{nom}}$ are provided within the global minimum's basin of attraction, the algorithm will find the global optimal solutions. The use of only first order information is a distinctive feature of the PPT algorithm and numerical experience on a number of examples in the original paper indicates reasonable convergence properties. Here, the emphasis is on the computational efficiency of a single iteration. However, the overall performance of the algorithm is a balance between this and the total number of iterations required for convergence which might well be improved by use of the second order

Hessian, although the computational cost of computing this exactly is large. A comparison of the relative performance of the second approach relative to the first order approach is beyond the scope of this paper.

It is well known that using a sequential LP optimization algorithm will converge slowly when the problem is badly conditioned. Regularization techniques like ridge regression which is based on measuring the 2-norm of the parameter deviation, have been shown to implicitly improve the condition of the original ill-posed optimization problem by implicitly adding a small positive definite matrix onto the local Hessian. However, rather than altering the form of the Hessian, the 1 and ∞-norm measures considered in this paper constrain the parameter update direction to lie in a sub-space aligned with the negative gradient. The step length for all examples in the original PPT paper (Wilkinson et al., 2008) is considered a unity full step, although an adaptive strategy is more general and robust.

2.4. *Analysis of the cost function*

In this section, the structures of both the residual objective function and the parameter deviation constraint in (1) are analysed. The role of the log transformations in each term is also discussed.

2.4.1. *Noise assumptions*

In (1), the performance function is the 1-norm of the residual between the log measurements and log predictions. Usually, a (squared) 2-norm measure is employed (Hansen, 1997; Golub et al., 1999) so that the residual variance is minimized which corresponds to an additive Gaussian, independent and identically distributed (i.i.d.) noise assumption. Using a 1-norm on the untransformed residuals (Conrad and Tyson, 2006) assumes that the noise model is additive Laplacian, which has heavier tails. Practically, it corresponds to a median value estimator which means that large residuals would have less influence or leverage on the solution, and thus is more robust to outliers. The log measurement/prediction transformations assume that the noise process is multiplicative Laplacian:

$$y = y \times r \tag{8}$$

where r is a Laplacian distribution with zero mean as given in Fig. 1 (upper). Residuals that correspond to doubling or halving the prediction have the same effect, and it should be noted that the logarithmic/exponential and multiplicative transformations ensure that the measurements remain non-negative. There will be a tendency to under predict the actual target value and as such the transformed residual signal will, in general, have a positive mean value and non-negative (Fig. 1 (nether)), unlike the more normal zero mean assumption (Kozubowski and Podgorski, 2003).

2.4.2. *Parameter norm measures*

The ∞-norm on the logarithmic parameter deviations in (1) is used to make the problem well posed by introducing some bias in order to minimize the parameter variance when the estimation variance is small. The practical effect is for the estimated parameters to remain at their nominal values and only disturb them when there is evidence in the data.

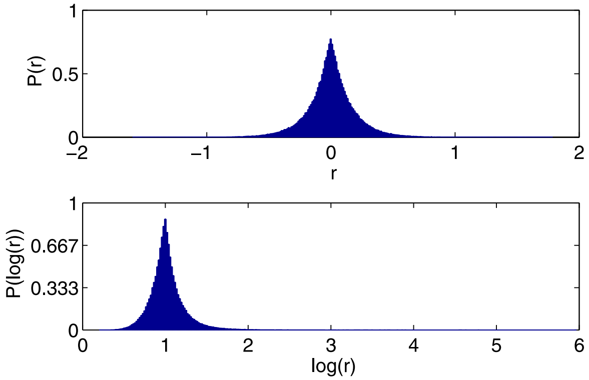

Fig. 1 Laplacian noise (upper) with a zero mean and standard deviation of 0.2 and the transformed log-Laplacian noise (nether).

The ∞-norm parameter deviation measure is equivalent to

$$\max_j \log\left|\frac{k_j^{\text{nom}}}{k_j}\right| \tag{9}$$

with all the parameters are being constrained at the same rate when the normalised updates are changing equally. Clearly, the nominal values bias the final parameter values as inferred from the data. This is often desirable for biochemical problems in which data is scarce and expected prior values can be estimated independently by comparison with similar parameters, *de novo* calculations or *in vitro* measurements.

3. Properties of the PPT algorithm

In Section 2, it was shown that the iterative PPT algorithm minimizes a regularization function which is composed of a 1-norm of the log residuals and an ∞-norm of the log parameter deviations. In this section, the optimality properties are discussed and, in particular, a technique for calculating the complete parameter locus is presented which illustrates the solution's sensitivity to the empirical data. For simplicity, the theoretical analysis is mainly based on linear static models, $\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{k}}$. In addition, without loss of generality, the log transformation on the outputs and parameters are ignored. The optimization problem can then be stated as

$$\begin{aligned} &\min_{\mathbf{k}} && \|\mathbf{y} - \mathbf{X}\mathbf{k}\|_1 \\ &\text{s.t.} && \left\|\mathbf{k}^{\text{nom}} - \mathbf{k}\right\|_\infty \leq \tilde{k} \end{aligned} \tag{10}$$

For non-linear dynamic pathway models, this linear static model representation could be obtained by either transforming the dynamic ODE model to a static model (Papadopoulos and Brown, 2007; Brown et al., 2008) using the "collocation" method, or by considering the local sensitivity linearisation at a parameter point (He et al., 2008) as given in (4). Since the former approach would generally introduce bias in the estimates (Brown et al., 2008), the latter approach is the focus of this paper. Accordingly, the analysis of the PPT algorithm's optimality properties for linear static models can also be applied to non-linear dynamic models as discussed in Section 3.2.2.

3.1. Properties and graphical interpretation

The dual PPT optimization problem is now analyzed as it provides an important theoretical basis for understanding and constructing the parameter sensitivity locus.

3.1.1. Dual problem

At optimality, the linearized primal PPT problem (1) can be analysed as the active sets of parameters and data are the same are the same as the original non-linear problem. The dual of the linear programming primal problem terms can be expressed in terms of the corresponding Lagrange multipliers, $[\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2, \boldsymbol{\lambda}^3, \boldsymbol{\lambda}^4]^T \geq \mathbf{0}$, where $\boldsymbol{\lambda}^i$ is associated with the ith set of inequality constraints in (4). Then according to Karush–Kuhn–Tucker (KKT) condition give

$$\begin{aligned} &\boldsymbol{\lambda}^3 - \boldsymbol{\lambda}^4 = -\mathbf{X}^T\left(\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2\right) \\ &\boldsymbol{\lambda}^1 + \boldsymbol{\lambda}^2 = \mathbf{1} \end{aligned} \tag{11}$$

Using these conditions, a number of simple, but extremely useful insights into the Lagrange multipliers can be easily proven. In the following, the active parameters are those parameters k_i for which $|k_i^{\mathrm{nom}} - k_i| = \tilde{k}$ and the active residuals r_i are those exemplars for which $y_i - \mathbf{x}_i^T\mathbf{k} = 0$. These definitions follow directly from the KKT conditions.

Theorem 1. *The Lagrange multipliers corresponding to the inactive residuals are binary and satisfy*:

$$\boldsymbol{\lambda}^1 - \boldsymbol{\lambda}^2 = \operatorname{sgn}(\mathbf{r}) \tag{12}$$

where vector $\mathbf{r}$ *denotes the set of inactive residuals that* $r_i = y_i - \mathbf{x}_i^T\mathbf{k} \neq 0$.

Proof: From (11), the Lagrange multipliers associated with the residuals, $\{\boldsymbol{\lambda}^1, \boldsymbol{\lambda}^2\}$, are non-negative and sum to unity. When a residual is non-zero, one of the corresponding Lagrange multipliers must therefore be 0 and the other must be 1. Therefore, the Lagrange multipliers for the non-zero residuals are binary. In addition, when the residuals are positive $\lambda_i^1 = 1, \lambda_i^2 = 0$ and when the residuals are negative $\lambda_i^1 = 0, \lambda_i^2 = 1$. Therefore, (12) holds. Alternatively, when a residual is active (zero), both constraints are active and the corresponding Lagrange multipliers are both non-zero $0 < \lambda_i^1, \lambda_i^2 < 1$. □

Theorem 2. *In the non-zero residual sub-space, the Lagrange multipliers on the parameters correspond to the gradient of the* 1*-norm errors*:

$$\boldsymbol{\lambda}^3 - \boldsymbol{\lambda}^4 = \nabla_{\mathbf{k}} p(\mathbf{k}) \tag{13}$$

where $p(\mathbf{k})$ *is the* 1*-norm residual objective function in* (10).

Proof: From (11) and (12), it follows that:

$$\boldsymbol{\lambda}^3 - \boldsymbol{\lambda}^4 = -\mathbf{X}^T \operatorname{sgn}(\mathbf{r}) \tag{14}$$

when the residuals are inactive. By definition, the 1-norm objective function can be written as $p(\mathbf{k}) = \mathbf{r}(\mathbf{k})^T \operatorname{sgn}(\mathbf{r}(\mathbf{k}))$ and differentiating with respect to the parameters gives $\nabla_{\mathbf{k}} p(\mathbf{k}) = -\mathbf{X}^T \operatorname{sgn}(\mathbf{r})$ where the residual sign vector is obviously constant in local regions. Therefore, (13) holds. □

This demonstrates a clear relationship between the (dual) Lagrange multipliers and the geometry of optimization problem. There is also a direct link between a residual becoming zero and a parameter becoming inactive.

Theorem 3. *In general, the number of inactive parameters is equal to the number of zero residuals.*

Proof: When a parameter is inactive, the corresponding Lagrange multipliers are zero, i.e. $\lambda_i^3 = \lambda_i^4 = 0$. However, (11) demonstrates that for inactive residuals $\boldsymbol{\lambda}^3 - \boldsymbol{\lambda}^4 = -\mathbf{X}^T \operatorname{sgn}(\mathbf{r})$, which will, in general, be a non-zero vector (note that $-\mathbf{X}^T \operatorname{sgn}(\mathbf{r})$ may be zero for some components when the gradient of the objective function is zero for a parameter, but in general, this is unlikely to happen). When a residual r_i is zero, both the corresponding Lagrange multipliers $(\lambda_i^1, \lambda_i^2)$ are non-zero as the corresponding constraints are both active. Hence,

$$\boldsymbol{\lambda}_I^3 - \boldsymbol{\lambda}_I^4 = -\left(\mathbf{X}_{I,I}^T \operatorname{sgn}(\mathbf{r}_I) + \mathbf{X}_{I,A}^T\left(\boldsymbol{\lambda}_A^1 - \boldsymbol{\lambda}_A^2\right)\right) = 0 \tag{15}$$

where the subscript in $\boldsymbol{\lambda}^3$, $\boldsymbol{\lambda}^4$ and the first subscript in $\mathbf{X}^T$ denote the inactive parameter set I; the second subscript in $\mathbf{X}^T$, i.e. I or A, refer to the sets of non-zero and zero residuals (denoted inactive and active residuals), respectively. For the ith element of the left-hand side to be zero (corresponding to the ith inactive parameter), there must exist at least one corresponding zero residual associated with one data point. The residual Lagrange multipliers $\boldsymbol{\lambda}_A^r$, which is defined as $\boldsymbol{\lambda}_A^r = \boldsymbol{\lambda}_A^1 - \boldsymbol{\lambda}_A^2$ is further determined by

$$\boldsymbol{\lambda}_A^r = -\left(\mathbf{X}_{I,A}^T\right)^{-1}\left(\mathbf{X}_{I,I}^T \operatorname{sgn}(\mathbf{r}_I)\right) \tag{16}$$

where $\mathbf{X}_{I,A}^T$ is required to be an invertible matrix. This is only true when each row of the matrix $\mathbf{X}_{I,A}^T$ denoting the corresponding data is linearly independent from each other. Otherwise, linearly dependent data should be removed to ensure the invertibility (full rank) of this matrix. □

3.1.2. Geometrical analysis of the constrained solutions

These theorems can be interpreted geometrically as the constraint parameter $\tilde{k}$ varies. When $\tilde{k}$ is small, a much greater weight is placed on keeping the parameters close to their nominal values. Assuming the data fitting problem is fairly ill-conditioned and that the nominal parameter values are not chosen to lie in the small volume in parameter space where the sensitivity derivatives are small, then the normalized parameter updates will lie at the corner of the hypercube of size $\tilde{k}$ which first intersects with objective function's contour, as shown in Fig. 2 (left).

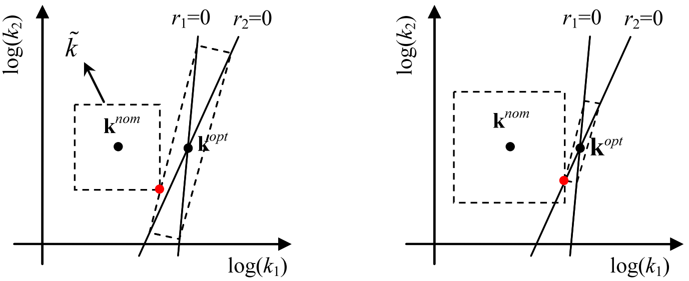

Fig. 2 An illustration of the optimal solution (red dot) for the PPT algorithm when close to the nominal parameter values (left) or close to the maximum likelihood solution (right). For simplicity, it has been assumed that there is a linear relationship between log(y) and log(k_i) and that there exists two exemplars.

When $\tilde{k}$ is larger, a greater emphasis is placed on fitting the data and in general, the optimal parameter estimate will lie at a corner (derivative discontinuity) of the objective function contour and along a side the parameter deviation hypercube, as illustrated in Fig. 2 (right). In this case, it corresponds to one or more of the residuals being zero and the effect of using a 1-norm performance function is to ensure one or more residuals are zero (distribution median). It should be noted that the number of non-active parameter deviations is bounded above by the number of exemplars used in the objective function.

3.1.3. Minimal parameter disturbance analysis

For comparison with ∞-norm constraint, a 1-norm to parameter deviation constraint corresponds to a continuous, relaxed version of the discrete optimization problem which attempts to reduce the residuals whilst minimizing the number of parameter deviations. This is expressed as

$$\begin{aligned} \min \quad & f(\mathbf{k}) = \left\| \log \mathbf{y} - \log \hat{\mathbf{y}}(\mathbf{k}) \right\|_1 \\ \text{s.t.} \quad & \left\| \log \mathbf{k}^{\text{nom}} - \log \mathbf{k} \right\|_1 \leq \tilde{k} \end{aligned} \tag{17}$$

and is illustrated in Fig. 3. Initially, one parameter deviates from its nominal value and as $\tilde{k}$ increases an increasing number of parameters are gradually perturbed from their nominal values. This performs a form of "soft feature selection" (Papadopoulos and Brown, 2007; He et al., 2008).

3.2. Optimal parameter locus of PPT algorithm

Model sensitivity is an important subject in systems biology as it can be used to identify important parameters and perform model selection. However, the solution sensitivity to important design parameters is often neglected and in the PPT algorithm, the most critical parameter is maximum parameter deviation $\tilde{k}$. The optimal parameter estimates can therefore be viewed as a parameter loci $\mathbf{k}(\tilde{k})$ (Papadopoulos and Brown, 2007) and the sensitivity of the parameter estimates to selected values of $\tilde{k}$ can then be studied. This idea have been studied for 1-norm parameter deviation constraints (Papadopoulos and Brown, 2007; He et al., 2008), and the following sections adapt this work for ∞-norm parameter deviation constraints, as used in the PPT algorithm.

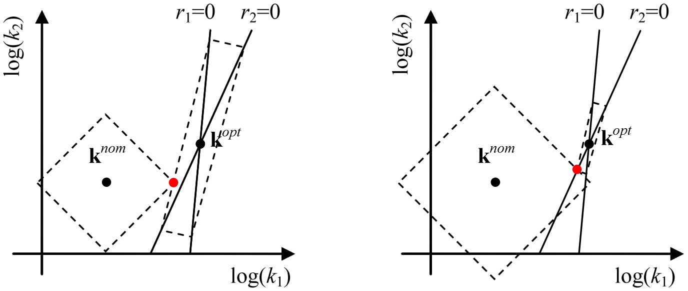

Fig. 3 Using a 1-norm to measure parameter deviation means that only a minimal number of parameters are updated when the deviations are small, i.e. only one parameter (k_1) is active when optimal parameters left from the nominal values (left), whereas, both two parameters are active when optimal solutions are close the maximum likelihood solutions (right).

3.2.1. Parameter locus for linear models

The properties and construction of an optimal parameter locus for linearized PPT algorithm will now be developed. This is based on the simple geometric insights illustrated in Fig. 2.

3.2.1.1. Properties of parameter locus Some basic results about the optimal parameter locus can be easily established.

1. $\mathbf{k}(0) = \mathbf{k}^{\text{nom}}$ which is the starting point for the parameter locus.
2. $\mathbf{k}(\tilde{k}) = \mathbf{k}^*$ where $\mathbf{k}^* = \arg\min_{\mathbf{k}} \|\mathbf{y} - \mathbf{X}\mathbf{k}\|_1$, for all $\tilde{k} \geq \|\mathbf{k}^{\text{nom}} - \mathbf{k}^*\|_\infty$ which is the end point (median estimator) for the parameter locus.
3. The optimal parameter locus, $\mathbf{k}(\tilde{k})$, is piecewise linear. The locus is linear within each region where the active data/parameter sets do not change.
4. The parameter space is partitioned into regions $\mathbf{R}^i$ where $\|\mathbf{y} - \mathbf{X}\mathbf{k}\|_1$ has constant gradients. The boundaries of the regions occur when $r_i = y_i - \mathbf{x}_i^T\mathbf{k} = 0$, as illustrated in Fig. 2. Within each region, the active parameter and data sets do not alter and the parameter locus is linear.
5. Assuming the active parameter space is non-empty, the optimal solution occurs at a vertex of the constraining hypercube in active parameter (sub)space and in the inactive parameter (sub)space, the optimal solution occurs in the interior of the relevant side of the hypercube.

Two important properties about the optimal value of the active and inactive parameter sets can also be easily established and geometrically illustrated in Fig. 4.

Theorem 4 (Active Parameters). *The relevant vertex of the active parameter space corresponds to label* $-\text{sgn}(\nabla_{\mathbf{k}_A} p(\mathbf{k}))$ *which is constant in the relevant region in parameter space.*

Proof: This can be easily established by considering the gradient of the linear performance function and can be considered to be over the non-zero residual space only. Firstly, since it is an ∞-norm constraint on the parameter deviations as in (10), in the non-zero residual parameter space, the active parameter will evolve along one of the vertexes of the

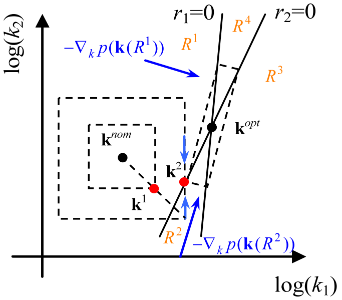

Fig. 4 An illustration of active and inactive parameters for PPT algorithm. For parameter estimate $\mathbf{k}^1$ in region $\mathbf{R}^1$, both parameters $\mathbf{k}_1$ and $\mathbf{k}_2$ are active; for parameter estimate $\mathbf{k}^2$ that is in the boundary of region $\mathbf{R}^1$ and $\mathbf{R}^2$, $\mathbf{k}_1$ is active but $\mathbf{k}_2$ is inactive.

∞-norm hypercube. Secondly, according to Theorem 2, the direction of the corresponding vertex must be the one that optimally minimize the 1-norm performance function $p(\mathbf{k})$. Therefore, the direction of the relevant vertex is $-\text{sgn}(\nabla_{\mathbf{k}_A} p(\mathbf{k}))$. □

This is illustrated in Fig. 4, for parameter estimate $\mathbf{k}^1$ in region $\mathbf{R}^1$, both parameters (k_1, k_2) are active, and they lie on the vertex of hypercube, and the direction of corresponding parameter updates in this region is labelled as $-\text{sgn}(\nabla_{\mathbf{k}} p(\mathbf{k})), \forall \mathbf{k} \in \mathbf{R}^1$.

Theorem 5 (Inactive Parameters). *In the inactive parameter space, the optimal value lies on the hypercube's (inactive) face which passes through the active vertex. The value occurs where the interior face intersects with set of zero residuals.*

Proof: One parameter is defined to be inactive, only if it is no longer updated along the mapping of the constraint hypercube's vertex (e.g. $\mathbf{k}^2$ in Fig. 4). However, since the rest active parameters would still evolve along the mapping of hypercube's vertex in the corresponding coordinates, thus this inactive parameter must be positioned on the mapping of hypercube's (inactive) face which passes through the corresponding "active vertex". On the other hand, since the mapping of the negative gradients $-\nabla_{\mathbf{k}_A} p(\mathbf{k}), \forall \mathbf{k} \in \mathbf{R}^i$ of corresponding two neighbouring regions $\mathbf{R}^i$ to the inactive parameter's coordinate would be just opposite in direction, the value of corresponding inactive parameter could only be where the hypercube's inactive face intersects with the zero residual (as shown in Fig. 4). □

The necessary but not sufficient condition, for an active parameter becoming inactive is further discussed in the next sub-section.

3.2.1.2. Calculating the parameter locus Given these properties, it is possible to calculate the optimal parameter locus in an iterative and computationally efficient manner:

1. The starting point is $\mathbf{k} = \mathbf{k}^{\text{nom}} \in \mathbf{R}^0$ and all the parameters can be assumed to be active. The corresponding region is labelled $\mathbf{R}^i$ where $i = 1$.

2. Within $\mathbf{R}^i$ the active parameters are updated according to

$$\begin{aligned}&\mathbf{k}_A(\tilde{k}) = \mathbf{k}_A\left(\mathbf{R}^{i-1}\right) + \tilde{k}\Delta\mathbf{k}_A\\&\Delta\mathbf{k}_A \propto \operatorname{sgn}\left(-\nabla_{\mathbf{k}_A} p(\mathbf{k})\right), \quad \mathbf{k} \in \mathbf{R}^i\end{aligned} \tag{18}$$

The inactive parameters are the solution to the problem:

$$\begin{cases} r_1 = y_1 - \mathbf{x}_1^T[\mathbf{k}_{I1};\ldots\mathbf{k}_{In};\mathbf{k}_{A1};\ldots\mathbf{k}_{A(m-n)}] = 0\\ \qquad\vdots\\ r_n = y_n - \mathbf{x}_n^T[\mathbf{k}_{I1};\ldots\mathbf{k}_{In};\mathbf{k}_{A1};\ldots\mathbf{k}_{A(m-n)}] = 0\end{cases} \tag{19}$$

Therefore, when there exist n inactive parameters $\mathbf{k}_I = [\mathbf{k}_{I1};\ldots\mathbf{k}_{In}]$, there are at least n zero data residuals, i.e. $r_i = y_i - \mathbf{x}_i^T\mathbf{k} = 0$ and the n inactive parameters can therefore be uniquely determined by solving the n equations in (19), in which it is assumed that the $m-n$ active parameters $\mathbf{k}_A = [\mathbf{k}_{A1};\ldots\mathbf{k}_{A(m-n)}]$ are known. The inactive parameters can be updated as

$$\begin{aligned}&\mathbf{k}_I(\tilde{k}) = \mathbf{k}_I\left(\mathbf{R}^{i-1}\right) + \left(\tilde{k} - \tilde{k}^{i-1}\right)\Delta\mathbf{k}_I\\&\Delta\mathbf{k}_I \propto \left(\mathbf{X}_{I,A}^T\right)^{-1}\mathbf{y}_A\end{aligned} \tag{20}$$

where $\tilde{k}^{i-1}$ denotes the value of maximum parametric deviation of the last region $\mathbf{R}^{i-1}$ and $\mathbf{y}_A$ is the "active" output data set associated with zero residuals.
3. When the parameter locus enters a new region $\mathbf{R}^{i+1}$, the active and inactive sets change.
   (i) Active parameter becomes inactive. The necessary condition for an active parameter becoming inactive is when a residual becomes zero and there is a corresponding sign change in $\nabla_k p()$ for that active parameter.
   (ii) Inactive parameter becomes active. This occurs when the optimal solution for an inactive parameter lies outside the hypercube's face boundary.
4. The algorithm will terminate at $\mathbf{k}^*$.

This process is illustrated in Fig. 5.

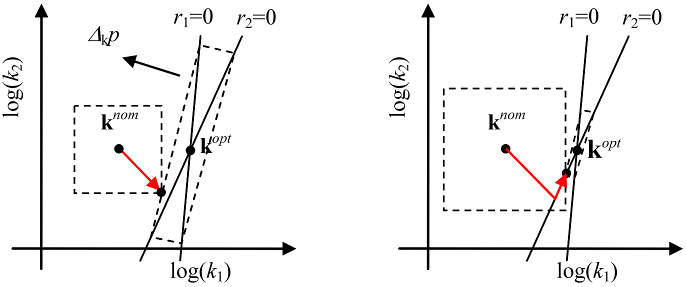

Fig. 5 A simple illustration of how the optimal parameter locus (red) evolves for PPT algorithm.

3.2.2. *PPT parameter locus for dynamic models*

For dynamic pathway models, the linear static assumption made in Section 3.2.1 is only locally valid. This means that the previously described piecewise linear parameter locus is an approximation of the true nonlinear parameter locus, in particular, the true parameter locus $\mathbf{k}(\tilde{k})$ will be a piecewise continuous nonlinear curves and no analytic solution exists. However, it is possible to approximate the true parameter locus because the starting point of the parameter locus and the active/inactive constraints are equivalent to the linear, static case. The key difference is that the parameter locus gradient $\nabla_{\tilde{k}}\mathbf{k}(\tilde{k})$ is locally calculated and then numerically integrated. This is used in the example in the following section and a more detailed description of this process is given in a related paper (Brown et al., 2009).

4. Example

The first illustrative pathway example in the original PPT paper (Wilkinson et al., 2008) is revisited here. It contains two reactions in series that convert species 1 to species 3 via an intermediate species 2 as shown in Fig. 6. The initial as well as the nominal value of reaction parameters are set to be $k_1^{\text{nom}} = 0.15\ \text{s}^{-1}$ and $k_2^{\text{nom}} = 0.015\ \text{s}^{-1}$, and the initial concentration of species 1 is 1 μM whereas species 2 and 3 have zero initial concentration. When setting model parameters the same as the optimal values in Conrad and Tyson (2006) ($k_1^* = 0.0103\ \text{s}^{-1}$, $k_2^* = 0.0152\ \text{s}^{-1}$), the generated time series data of species 2 is used as training data with additive zero mean Gaussian i.i.d. noise ($\sigma = 0.01$).

The logarithmic PPT optimization problem (1) is used for parameter estimation with a 1-norm on the residuals and a ∞-norm on the parameter deviations. The parameter deviations are constrained by $\tilde{k}$ which lies in the range 0 to ∞, generating the parameter loci. By sampling $\tilde{k}$ from 0.0001 to 40, the estimated parameter values and optimal parameter locus are given in Table 1 and Fig. 7.

Table 1 shows that as the parameter deviation constraint bound $\tilde{k}$ increases from a small value, the estimated parameter values evolve gradually from the initial nominal values to the median estimates. The optimal parameter locus that evolves as a function of $\tilde{k}$ is shown in parameter space in Fig. 7. When $\tilde{k}$ is small, optimal parameter values are

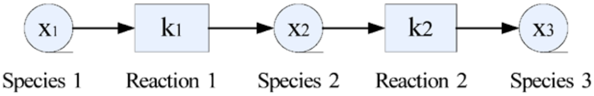

Fig. 6 Pathway structure of the illustrative irreversible reaction example.

Table 1 Optimal estimated parameters under different regularization parameter constrain values

	$\tilde{k} = 0.0001$	0.001	0.005	0.01	0.05	0.1	0.5	1	5	10	20	40
k_1	0.1500	0.1497	0.1483	0.1466	0.1337	0.1191	0.0474	0.0150	0.0102	0.0102	0.0102	0.0102
k_2	0.0150	0.0150	0.0148	0.0147	0.0134	0.0119	0.0100	0.0138	0.0152	0.0152	0.0152	0.0152
$\log(k_1)$	−0.8240	−0.8249	−0.8289	−0.8339	−0.8739	−0.9239	−1.3239	−1.8239	−1.9923	−1.9922	−1.9922	−1.9922
$\log(k_2)$	−1.8240	−1.8249	−1.8289	−1.8339	−1.8739	−1.9239	−2.0012	−1.8612	−1.8195	−1.8195	−1.8195	−1.8195

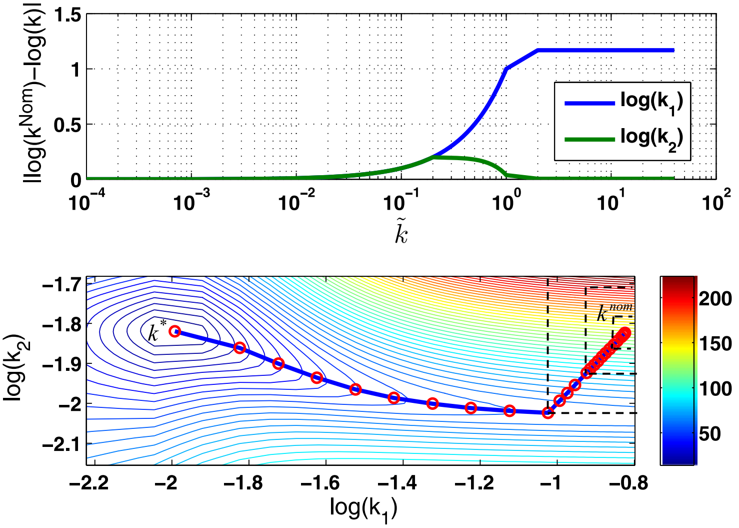

Fig. 7 Optimal parameter locus of PPT algorithm for the illustrative exemplar pathway. The logarithmic estimated parameter deviations from the nominal values plot against the parameter deviation constrains $\tilde{k}$ (upper). Two dimensional optimal parameter locus plot in logarithmic parameter space (nether). The contours are the 1-norm performance function, and the dashed black lines represent the ∞-norm parametric constraints.

close to nominal values k^{nom}, and both parameters are active so they evolve along the vertex of the ∞-norm box constraints. When the box constraint vertex hits one of the zero residual curves, parameter k_2 become inactive but k_1 still remains active. At this point, the parameter locus deviates from the -45 degree vertex direction and evolves along the zero residual trajectory until it reaches the maximum likelihood solution k^*. It should be noted that this zero residual curve is non-linear rather than piecewise linear for linear static cases. This matches the theoretical analysis of optimal parameter locus in Section 3.2. However, it is worth noting that in the parameter space where the vertex of the ∞-norm box constraint intersects with the zero residual curve the cost function is badly conditioned. This explains why the parameter locus evolves along the zero residual trajectory that lies in the valley of objective function contours, since these two are fairly close to each other. This is further illustrated by the zero residual trajectory plots in Fig. 8.

The dashed dot lines in Fig. 8 are (a subset of) the zero residual trajectories, in the parameter space, each of which is determined by $r_i = y_i - \hat{y}_i(\mathbf{k}) = 0$ corresponding to the ith data point. In this example, 300 measurement data points are generated, with only 10 representative data points (y_i) are shown here. It is obvious that most of the zero residual trajectories lie along the valley of the cost function, since the 1-norm optimization cost function is quite badly conditioned; and as can be expected, all the zero residual curves cross at the maximum likelihood solution. The red dashed dot line denotes the “optimal” zero residual trajectory determined by the 1-norm cost function, and the optimal parameter locus (as in Fig. 7) only varies its direction when the vertex of ∞-norm box constraint hits this “optimal” zero residual trajectory.

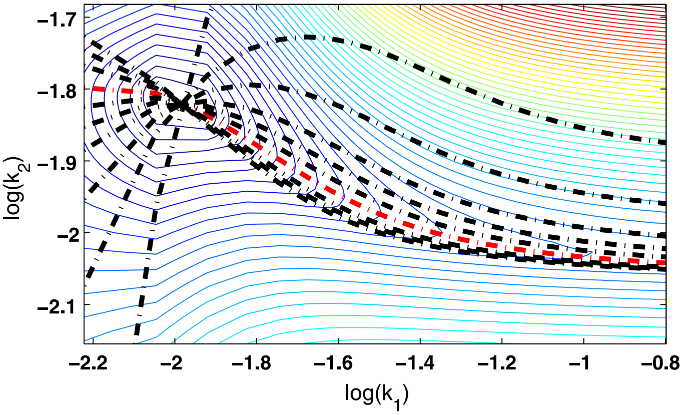

Fig. 8 Selected zero residual curves plot (dashed dot line) in the parameter space.

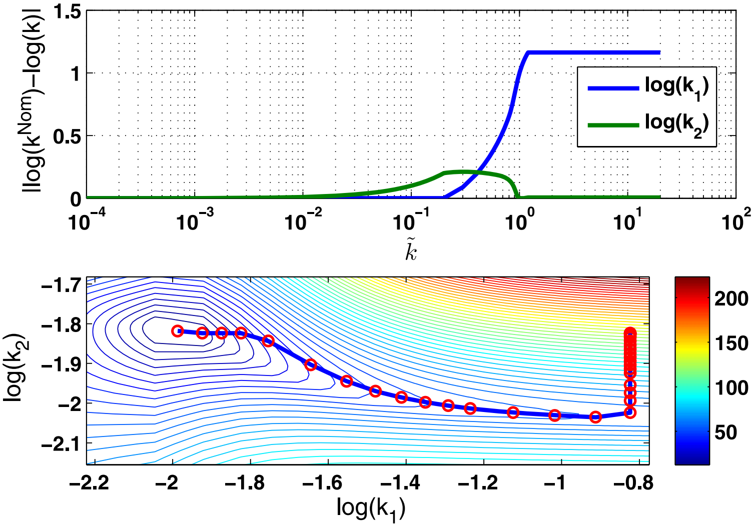

Fig. 9 Optimal parameter locus of using 1-norm measure on the parameter deviation for illustrative exemplar pathway.

For comparison with the ∞-norm constraint, the parameter locus using the 1-norm parameter deviation is also given in Fig. 9.

Figure 9 verifies that using the parameter deviation 1-nom corresponds to a minimal parameter perturbation strategy as described in Section 3.1.3. As the parameter deviation constraint bound, $\tilde{k}$, initially increases, only one parameter, k_2, becomes active. When the 1-norm constraint intersects with the zero residual trajectory, both parameters become active, and the optimal parameter locus evolves along this curve. Since the nominal value of parameter k_2^{nom} is very close to its optimal value k_2^* in this example, the parameter k_2

becomes inactive (deviation to the nominal value equal to zero) when the optimal locus is close to the maximum likelihood (median) solution.

5. Conclusions

This paper has provided a detailed analysis of the properties of the PPT algorithm, a recently published method for inverse modelling in systems biology. This has been done from a regularization perspective where the balance between the information content of measured data and prior knowledge has been explored. In computational systems biology, this is important since there is often limited information available, both from experimental data and prior knowledge. It is therefore necessary to combine all sources of information into model identification algorithms. The PPT algorithm uses prior knowledge of the network structure and also best guesses for the individual parameter values.

The choice of weighting between the prior parameter values and the exemplar data is important for the inverse modeller. In this paper, the concept of the optimal parameter locus used to explore the entire family of optimal solutions, and hence analyse the sensitivity of the model. Another key aspect of the PPT algorithm is its computational efficiency as it uses a relatively cheap LP sub-problem at each iteration. A rational derivation of this LP problem for nonlinear dynamic pathway system has been provided. It is also shown that the use of the 1-norm on the logarithmically transformed residuals essentially implies an assumption of log Laplacian noise. The dual form of the primal PPT problem is presented to prove the equivalence of the number of inactive parameters and the number of zero residuals at optimality. Using this and other properties we demonstrate how the parameter locus can be calculated efficiently and we provide a graphical interpretation of how it evolves.

In computational systems biology, we are getting increasingly comfortable with analysing underdetermined models in which the parameters are not individually identifiable. This is because our immediate aim is to characterise our model in terms of a restricted range of outputs and these outputs may be tightly constrained by the available data even though many individual parameter values might not be. In this context, efficient algorithms for inverse modelling such as the PPT algorithm can be used to extract maximum value from what data is currently available and help prioritise the experiments needed to reduce the uncertainty that remains. Since the PPT regularization algorithm is biased toward prior parameter values, a thorough understanding of its properties as presented in this paper is essential for its successful application. Such algorithms will drive the iterative improvement between modelling and experimentation to deliver truly predictive computational models for systems biology.

Appendix A: PPT algorithm

The basic PPT algorithm is focussed on the case when the information in the exemplar data is of a lower dimension than the model, so estimating the parameters is an ill-posed problem. As an extreme case, a situation is described in Wilkinson et al. (2008) where the model has 83 parameters and only two features are available: the time to peak value and the actual peak value. The PPT algorithm is formulated as an iterative sequence where, at each step, a linear programming (LP) problem is solved. This LP sub-problem is a linearisation of the following unconstrained minimisation problem:

$$\begin{aligned}\min f(\mathbf{k}) = {} & \left\|\boldsymbol{\alpha}\left(\log \mathbf{k}^{\mathrm{nom}} - \log \mathbf{k}\right)\right\|_1 + \bar{\alpha}\left\|\log \mathbf{k}^{\mathrm{nom}} - \log \mathbf{k}\right\|_\infty \\ & + \left\|\boldsymbol{\beta}\left(\log \mathbf{y} - \log\left(\hat{\mathbf{y}}(\mathbf{k})\right)\right)\right\|_1 + \bar{\beta}\left\|\log \mathbf{y} - \log\left(\hat{\mathbf{y}}(\mathbf{k})\right)\right\|_\infty \end{aligned} \tag{A.1}$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are diagonal matrices of positive weighting coefficients and $\bar{\alpha}$, $\bar{\beta}$ are positive scalar weighting coefficients.

This objective function given by (A.1) is a linear combination of the (weighted) 1-norm and ∞-norm of both the logarithmically transformed parameter deviations from their nominal values and the logarithmically transformed residuals.

In order to transform (A.1) into an LP problem, we introduce two complementary non-negative vectors $\boldsymbol{\Delta}\tilde{\mathbf{y}}^+$ and $\boldsymbol{\Delta}\tilde{\mathbf{y}}^-$ which are defined by

$$\boldsymbol{\Delta}\tilde{\mathbf{y}}^+ - \boldsymbol{\Delta}\tilde{\mathbf{y}}^- = \log \mathbf{y} - \log\left(\hat{\mathbf{y}}(\mathbf{k})\right) \tag{A.2}$$

Then it can be seen that sum of these two new vectors must be greater than or equal to the absolute value of the right-hand side of (A.2):

$$\boldsymbol{\Delta}\tilde{\mathbf{y}}^+ + \boldsymbol{\Delta}\tilde{\mathbf{y}}^- \geq \left|\log \mathbf{y} - \log\left(\hat{\mathbf{y}}(\mathbf{k})\right)\right| \tag{A.3}$$

and consequently we can also relate them to the 1-norm term appearing in the objective (A.1) using the following inequality:

$$\hat{\boldsymbol{\beta}}^T\left(\boldsymbol{\Delta}\tilde{\mathbf{y}}^+ + \boldsymbol{\Delta}\tilde{\mathbf{y}}^-\right) \geq \left\|\boldsymbol{\beta}\left(\log \mathbf{y} - \log\left(\hat{\mathbf{y}}(\mathbf{k})\right)\right)\right\|_1 \tag{A.4}$$

where $\hat{\boldsymbol{\beta}}$ is the vector of diagonal elements of the diagonal matrix $\boldsymbol{\beta}$. We also introduce an additional slack variable $\tilde{\mathbf{y}}$ to which we apply the following constraints:

$$\begin{aligned} \mathbf{1}\tilde{y} &\geq \boldsymbol{\Delta}\tilde{\mathbf{y}}^+ - \boldsymbol{\Delta}\tilde{\mathbf{y}}^- \\ \mathbf{1}\tilde{y} &\geq -\left(\boldsymbol{\Delta}\tilde{\mathbf{y}}^+ - \boldsymbol{\Delta}\tilde{\mathbf{y}}^-\right) \end{aligned} \tag{A.5}$$

which automatically imply the following relationship with respect to the ∞-norm:

$$\tilde{y} \geq \left\|\log \mathbf{y} - \log\left(\hat{\mathbf{y}}(\mathbf{k})\right)\right\|_\infty \tag{A.6}$$

In an entirely analogous manner, we introduce similar non-negative vectors $\boldsymbol{\Delta}\tilde{\mathbf{k}}^+$ and $\boldsymbol{\Delta}\tilde{\mathbf{k}}^-$ and a slack variable $\tilde{k}$ for the parameter deviations:

$$\boldsymbol{\Delta}\tilde{\mathbf{k}}^+ - \boldsymbol{\Delta}\tilde{\mathbf{k}}^- = \log \mathbf{k}^{\mathrm{nom}} - \log \mathbf{k} \tag{A.7}$$

$$\begin{aligned} \mathbf{1}\tilde{k} &\geq \boldsymbol{\Delta}\tilde{\mathbf{k}}^+ - \boldsymbol{\Delta}\tilde{\mathbf{k}}^- \\ \mathbf{1}\tilde{k} &\geq -\left(\boldsymbol{\Delta}\tilde{\mathbf{k}}^+ - \boldsymbol{\Delta}\tilde{\mathbf{k}}^-\right) \end{aligned} \tag{A.8}$$

which similarly imply the following:

$$\hat{\boldsymbol{\alpha}}^T\left(\boldsymbol{\Delta}\tilde{\mathbf{k}}^+ - \boldsymbol{\Delta}\tilde{\mathbf{k}}^-\right) \geq \left\|\boldsymbol{\alpha}\left(\log \mathbf{k}^{\mathrm{nom}} - \log \mathbf{k}\right)\right\|_1 \tag{A.9}$$

$$\tilde{k} \geq \left\|\log \mathbf{k}^{\mathrm{nom}} - \log \mathbf{k}\right\|_\infty \tag{A.10}$$

where $\hat{\boldsymbol{\alpha}}$ is the vector of diagonal elements of the diagonal matrix $\boldsymbol{\alpha}$.

Finally, the local relationship between the new residual and parameter deviation vectors is established by the linearisation of the function $\log(\hat{\mathbf{y}}(\mathbf{k}))$ about the current point in parameter space which we shall denote by $\mathbf{k}^0$:

$$\begin{aligned}\log\big(\hat{\mathbf{y}}(\mathbf{k})\big) &= \log\big(\hat{\mathbf{y}}\big(\mathbf{k}^0\big)\big) + \mathbf{S}\big(\mathbf{k}^0\big)\big(\log\mathbf{k} - \log\mathbf{k}^0\big)\\ &= \log\big(\hat{\mathbf{y}}\big(\mathbf{k}^0\big)\big) + S\big(\mathbf{k}^0\big)\big(\log\mathbf{k} - \log\mathbf{k}^0\big)\\ \log\big(\hat{\mathbf{y}}(\mathbf{k})\big) &= \log\big(\hat{\mathbf{y}}\big(\mathbf{k}^0\big)\big) + \mathbf{S}\big(\mathbf{k}^0\big)\big(\big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+0} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-0}\big) - \big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-}\big)\big)\end{aligned}$$

Substituting this linearisation into (A.2) yields:

$$\mathbf{\Delta}\tilde{\mathbf{y}}^{+} - \mathbf{\Delta}\tilde{\mathbf{y}}^{-} = \log\mathbf{y} - \log\big(\hat{\mathbf{y}}\big(\mathbf{k}^0\big)\big) - \mathbf{S}\big(\mathbf{k}^0\big)\big(\big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+0} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-0}\big) - \big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-}\big)\big) \tag{A.11}$$

The above equality provides a step to take in parameter space in order to reduce the residual vector from its current value: $\log\mathbf{y} - \log(\hat{\mathbf{y}}(\mathbf{k}^0))$. When considering that we are using a succession of linear approximations to the true dependence of outputs on parameters, we may wish to be conservative in moving in some directions at a given iteration of the algorithm. This can be achieved by the introduction of a diagonal step length matrix $\boldsymbol{\gamma}$ whose diagonal elements are between zero and unity:

$$\mathbf{\Delta}\tilde{\mathbf{y}}^{+} - \mathbf{\Delta}\tilde{\mathbf{y}}^{-} = \boldsymbol{\gamma}\big(\log\mathbf{y} - \log\big(\hat{\mathbf{y}}\big(\mathbf{k}^0\big)\big)\big) - \mathbf{S}\big(\mathbf{k}^0\big)\big(\big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+0} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-0}\big) - \big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-}\big)\big) \tag{A.12}$$

Note that the inequalities (A.4), (A.6), (A.9) and (A.10) relating the new variables to the respective 1-norm and ∞-norm terms in the unconstrained problem (A.1) become strict equalities under the pressure of minimisation. We can therefore use these relations, together with constraints (A.5), (A.8), and (A.12), to formulate an equivalent constrained LP problem:

$$\begin{aligned}\min\quad & f\big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+}, \mathbf{\Delta}\tilde{\mathbf{k}}^{-}, \tilde{k}, \mathbf{\Delta}\tilde{\mathbf{y}}^{+}, \mathbf{\Delta}\tilde{\mathbf{y}}^{-}, \tilde{y}\big)\\ & \quad = \hat{\boldsymbol{\alpha}}^{T}\big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+} + \mathbf{\Delta}\tilde{\mathbf{k}}^{-}\big) + \bar{\alpha}\tilde{k} + \hat{\boldsymbol{\beta}}^{T}\big(\mathbf{\Delta}\tilde{\mathbf{y}}^{+} + \mathbf{\Delta}\tilde{\mathbf{y}}^{-}\big) + \bar{\beta}\tilde{y}\\ \text{s.t.}\quad & \mathbf{1}\tilde{k} \geq \mathbf{\Delta}\tilde{\mathbf{k}}^{+} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-}\\ & \mathbf{1}\tilde{k} \geq -\big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-}\big)\\ & \mathbf{1}\tilde{y} \geq \mathbf{\Delta}\tilde{\mathbf{y}}^{+} - \mathbf{\Delta}\tilde{\mathbf{y}}^{-}\\ & \mathbf{1}\tilde{y} \geq -\big(\mathbf{\Delta}\tilde{\mathbf{y}}^{+} - \mathbf{\Delta}\tilde{\mathbf{y}}^{-}\big)\\ & \mathbf{\Delta}\tilde{y}^{+} - \mathbf{\Delta}\tilde{\mathbf{y}}^{-} = \boldsymbol{\gamma}\big(\log\mathbf{y} - \log\big(\hat{\mathbf{y}}\big(\mathbf{k}^0\big)\big)\big)\\ & \qquad\qquad - \mathbf{S}\big(\mathbf{k}^0\big)\big(\big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+0} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-0}\big) - \big(\mathbf{\Delta}\tilde{\mathbf{k}}^{+} - \mathbf{\Delta}\tilde{\mathbf{k}}^{-}\big)\big)\end{aligned} \tag{A.13}$$

The problem above is the LP sub-problem that is solved at each iteration of the PPT algorithm. When comparing this to the sequential LP problem (4) in the main text, the following differences should be noted:

1. The formulation above is more general in terms of its objective, the weightings and the step lengths, etc. and also the symmetry between the residuals and parameter deviations (both norms considered for each). However, (4) in the main text is easier to analyse.
2. The $\mathbf{\Delta\tilde{k}}^{+}$ and $\mathbf{\Delta\tilde{k}}^{-}$ variables above relate to the parameter deviations from their nominal values rather than the change from the previous (linearized) point in parameter space which is denoted as $\mathbf{\Delta k}$ in (4). This is a point of terminology only.
3. In principle, the formulation above allows positive and negative parameter deviations and residuals to be weighted differentially so as to represent non-symmetric probability distributions.

References

Banga, J.R., Moles, C.G., Alonso, A.A., 2003. Global optimization of bioprocesses using stochastic and hybrid methods. In: Frontiers in Global Optimization, Nonconvex Optimization and Its Applications, vol. 74, pp. 45–70. Kluwer Academic, Dordrecht.

Brown, M., He, F., Zhan, C., Yeung, L.F., 2008. Nonparametric collocation ODE parameter estimation: application in biochemical pathway modelling. In: UKACC International Conference on Control, September 2008, Manchester, UK.

Brown, M., He, F., Papadopoulos, G., 2009. Dynamic basis pursuit regularization for complex biochemical pathway identification. Accepted for IEEE Conference on Decision and Control, Shanghai, P.R. China, December.

Conrad, E.D., Tyson, J.J., 2006. Modeling molecular interaction networks with nonlinear ordinary differential equations. In: Szallasi, Z., Stelling, J., Periwal, V. (Eds.), Systems Modelling in Cellular Biology: From Concept to Nuts and Bolts. MIT Press, Cambridge.

Dimelow, R.J., Wilkinson, S.J., 2009. Control of translation initiation: a model-based analysis from limited experimental data. J. R. Soc. Interface 6, 51–62.

Golub, G.H., Hansen, P.C., Leary, D.P.O., 1999. Tikhonov regularization and total least squares. SIAM J. Matrix. Anal. Appl. 21, 185–194.

Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., Sethna, J.P., 2007. Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol 3, 1871–1878.

Hansen, P.C., 1997. Rank-deficient and Discrete Ill-posed Problems: Numerical Aspects of Linear Inversion. SIAM Monogr. Math. Model. Comput., vol. 4. SIAM, Philadelphia.

He, F., Brown, M., Yeung, L.F., 2008. On the complexity-sensitivity trade-off for the NF-κB pathway modelling. In: IEEE World Congress on Computational Intelligence, Hong Kong, pp. 3932–3939.

Jin, Y., Yue, H., Liang, Y., Kell, D.B., 2007. Improving data fitting of a signal transduction model by global sensitivity analysis. In: American Control Conference, New York City, USA, pp. 2708–2713.

Johansen, T.A., 1997. On Tikhonov regularization, bias and variance in nonlinear system identification. Automatica 33(3), 441–446.

Kozubowski, T.J., Podgorski, K., 2003. Log-Laplace distributions. Int. Math. J. 3, 467–495.

Lei, F., Jorgensen, S.B., 2001. Estimation of kinetic parameters in a structured yeast model using regularization. J. Biotechnol. 88, 223–237.

Liebermeister, W., Baur, U., Klipp, E., 2005. Biochemical network models simplified by balanced truncation. FEBS J. 272, 4034–4043.

Muller, S., Lu, J., Kugler, P., Engl, H.W., 2008. Parameter identification in systems biology: solving ill-posed inverse problem using regularization. Report of Johann Radon Institute for Computational and Applied Mathematics (RICAM), 2008-25. http://www.ricam.oeaw.ac.at/publications/reports/08/rep08-25.pdf.

Okino, M.S., Mavrovouniotis, M.L., 1998. Simplification of mathematical models of chemical reaction systems. Chem. Rev. 98, 391–408.

Papadopoulos, G., Brown, M., 2007. Feature sensitivity of biochemical signalling pathways. In: IEEE Symp. Comp. Intell. Bioinf. Comput. Biol., April 2007, Hawaii, USA, pp. 373–380.

Rodriguez, M., Mendes, P., Banga, J.R., 2006a. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Biosystems 83, 248–265.

Rodriguez, M., Egea, J.A., Banga, J.R., 2006b. Novel metaheuristic for parameter estimation in nonlinear dynamic biology systems. BMC Bioinformatics 7, 483–501.

Voit, E.O., 2000. Computational Analysis of Biochemical Systems. Cambridge University Press, Cambridge.

Wilkinson, S.J., Benson, N., Kell, D.B., 2008. Proximate Parameter Tuning for biochemical networks with uncertain kinetic parameters. Mol. Biosyst. 4, 74–97.

Yue, H., Brown, M., Kell, D.B., Knowles, J., Wang, H., Broomhead, D., 2006. Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis: a case study of an NF-κB signalling pathway. Mol. Biosyst. 2(12), 640–649.