ON GRADIENT-BASED SEARCH FOR MULTIVARIABLE SYSTEM ESTIMATES

Adrian Wills * **Brett Ninness** * **Stuart Gibson** *

* *School of Electrical Engineering and Computer Science, University of Newcastle, Australia. Corresponding author email:* Adrian.Wills@newcastle.edu.au *Phone:+61 2 49215204 Fax:+61 249216993*

Abstract: This paper addresses the design of gradient based search algorithms for multivariable system estimation. In particular, the work here considers so-called 'full parametrization' approaches, and establishes that the recently developed 'Data Driven Local Coordinate' (DDLC) methods can be seen as a special case within a broader class of techniques that are designed to deal with rank-deficient Jacobians. This informs the design of a new algorithm that, via a strategy of dynamic Jacobian rank determination, is illustrated to offer enhanced performance.

Keywords: System identification, parameter estimation, gradient based search

1. INTRODUCTION

In the field of dynamic system identification, the so-called Maximum Likelihood (ML) principle and its relations, such as Prediction Error (PE) techniques, play a key role. Despite the advantages of ML/PE methods, their practical deployment is not always straightforward. This is largely due to the non-convex optimisation problems that are often implied. Typically, these are solved via a gradient-based search strategy such as a Newton type method or one of its derivatives (Ljung, 1999; T.Söderström and P.Stoica, 1989; Dennis and Schnabel, 1983).

The success of this sort of approach depends on the chosen system parametrization. Selecting the latter can be difficult, particularly in the multivariable case where the cost contours resulting from natural canonical state-space parametrizations imply poor numerical conditioning during gradient-based search (Deistler, 2000; McKelvey, 1998).

Indeed, the possibility of avoiding these parametrization-based difficulties is one of the key reasons for the recent intense interest in State Space Subspace based System Identification (4SID) methods (van Overschee and Moor, 1996; Larimore, 1990; Verhaegen, 1994). With these techniques, every element of every matrix in the state space model is estimated, which this paper terms a '*fully parametrized*' model structure.

More recently, the use of these fully parametrized structures has been investigated in the context of gradient based search for ML/PE estimates (McKelvey *et al.*, 2004; McKelvey and Helmersson, 1999; Bergboer *et al.*, 2002; Verdult *et al.*, 2002; Lee and Poolla, 1999). An essential point of this work is to recognise that a full parametrization is an over-parametrization, but that a minimal parametrization which is (locally) linearly related to the full parametrization can be simply derived.

These local representations have been dubbed '*Data Driven Local Co-ordinates*' (DDLC). Their employment reduces computational requirements, and also (usually) avoids rank deficiency of the prediction error Jacobian. This simplifies the computation of search directions. These features render DDLC-based gradient search as a very effective means for finding ML/PE estimates of multivariable systems, and indeed it has become the default method for multi-variable system estimation implemented in the widely used Matlab System Identification Toolbox (Ljung, 2004).

This paper is also directed at the development of gradient based search methods for ML/PE estimation of multivariable systems. Full parametrizations are also employed here, but instead of employing a DDLC-based local re-parametrization to avoid Jacobian rank-deficiencies, alternative "robust" strategies are proposed for computing search directions.

These robust methods for computing search directions effectively involve discarding elements of the Jacobian matrix that lie in its own kernel. A first main result of this paper is that if this discarding is implemented whereby a certain fixed dimensional subspace of the Jacobian is eliminated in the computation of a search direction, then this ensuing direction is *identical* to that obtained via DDLC methods.

While this may be of independent interest in terms of providing insight into the search mechanism inherent to DDLC techniques, its main significance in relation to this paper is that it establishes that DDLC methods are simply a particular choice within a range of search direction alternatives that are designed to be robust to Jacobian rank.

As such, this paper proposes and then examines a strategy of employing *dynamic* Jacobian rank allocation as part of search direction computation.

2. PROBLEM SETTING

This paper considers the problem of system estimation using the following innovations-form state-space model structure:

$$\left[\begin{array}{c} x_{t+1} \\ y_t \end{array}\right] = \left[\begin{array}{cc} A & B \\ C & D \end{array}\right] \left[\begin{array}{c} x_t \\ u_t \end{array}\right] + \left[\begin{array}{c} K \\ I \end{array}\right] v_t. \tag{1}$$

Here $u_t \in \mathbf{R}^m$ is the observed input, $y_t \in \mathbf{R}^p$ is the observed output, $v_t \in \mathbf{R}^p$ is a zero mean i.i.d. stochastic process that models measurement corruptions, and the state $x_t \in \mathbf{R}^n$.

In order to compute estimates, this paper will employ a *full parametrization* of the system matrices in (1). Specifically, this work will address the estimation of a parameter vector $\theta \in \mathbf{R}^{n_\theta}$ given as

$$\theta^T \triangleq \left[\mathrm{vec}\,\{A\}^T, \mathrm{vec}\,\{B\}^T, \mathrm{vec}\,\{C\}^T, \mathrm{vec}\,\{D\}^T, \mathrm{vec}\,\{K\}^T\right]. \tag{2}$$

Here, the $\mathrm{vec}\,\{\cdot\}$ operator is one which forms a vector from a matrix by stacking its columns on top of one another. Via this, and the innovations form of the structure (1), the steady-state mean square optimal one-step-ahead predictor $\widehat{y}_{t|t-1}(\theta)$ associated with a model parametrized by θ can be simply expressed as (Ljung, 1999)

$$\begin{aligned} \widehat{x}_{t+1|t} &= (A - KC)\widehat{x}_{t|t-1} + (B - KD)u_t + Ky_t, \\ \widehat{y}_{t|t-1}(\theta) &= C\widehat{x}_{t|t-1} + Du_t. \end{aligned} \tag{3}$$

Therefore, with the assumption that $\{v_t\}$ is Gaussian distributed as $v_t \sim \mathcal{N}(0, \sigma^2 I_p)$, $\sigma^2 \in \mathbf{R}^+$, and neglecting constant terms which are immaterial to the estimation process, the associated log likelihood function for the data is given as

$$L(\theta) = -\frac{Np}{2}\log\sigma^2 - \frac{1}{\sigma^2}\|E(\theta)\|^2. \tag{4}$$

Here the prediction error vector $E(\theta)$ is defined as

$$E(\theta) \triangleq \left[y_1^T - \widehat{y}_{1|0}^T(\theta), y_2^T - \widehat{y}_{2|1}^T(\theta), \cdots, y_N^T - \widehat{y}_{N|N-1}^T(\theta)\right]^T \tag{5}$$

and the norm used in (4) is the Euclidean one. Notice that, according to (4) there is an essential decoupling between the estimation of σ^2 and the elements of the parameter vector θ defined in (2). Namely, under the model structure (1), the ML estimate $\widehat{\theta}$ is given as an element satisfying

$$\widehat{\theta} \in \{\theta \in \mathbf{R}^{n_\theta} : \|E(\theta)\| \leq \|E(\overline{\theta})\|, \quad \forall \overline{\theta} \in \mathbf{R}^{n_\theta}\}. \tag{6}$$

If the aforementioned Gaussian assumption is violated, then the criterion (6) will no longer yield a Maximum-Likelihood solution. However, it will still specify a minimum prediction error norm estimate that will, asymptotically in observed data length, possess statistical properties that are closely related to those of a Maximum-Likelihood solution.

Balancing these attractive features, $\widehat{\theta}$ defined by (6) cannot be specified in closed form due to the nonlinear dependence of $E(\theta)$ on θ. In recognition of this, the previous work (McKelvey *et al.*, 2004; McKelvey and Helmersson, 1999; Bergboer *et al.*, 2002; Verdult *et al.*, 2002; Lee and Poolla, 1999) has focused on this problem of finding minima of $\|E(\theta)\|$, and has explored a gradient search approach. This paper is also directed at studying these methods, and seeks to propose, analyse and empirically substantiate effective variants of them.

3. GRADIENT-SEARCH BASED METHODS

Gradient search strategies have long been employed in a wide variety of system identification applications (Ljung, 1999). These methods are commonly motivated (Dennis and Schnabel, 1983, Chap. 10)) by an argument that the quadratic nature of $\|E(\theta)\|^2$ suggests the use of a linear approximation of $E(\theta)$ about a current guess θ_k of a minimiser according to

$$E(\theta) \approx E(\theta_k) + E'(\theta_k)(\theta - \theta_k), \quad E'(\theta_k) \triangleq \left.\frac{\partial E(\theta)}{\partial \theta}\right|_{\theta=\theta_k} \tag{7}$$

in which case the ensuing approximation

$$\min_\theta \|E(\theta)\| \approx \min_\theta \|E(\theta_k) + E'(\theta_k)(\theta - \theta_k)\| \tag{8}$$

by virtue of being affine in θ, does have a solution which can be found in closed form. However, this solution is only unique if the Jacobian $E'(\theta_k)$ is of full column rank.

The Levenberg–Marquardt and Gauss–Newton approaches ((Dennis and Schnabel, 1983, Chap. 10), (Nocedal and Wright, 1999, Chap. 10) and (Fletcher, 1987, Chap. 6)), are undoubtedly the most famous gradient based iterative search methods used in solving problem (6), and both use the linear approximation given in (7). Indeed, both methods involve the calculation of a search direction p and also (eg. in the 'damped' Gauss-Newton case) a step length α, such that the $k+1$'st iterate θ_{k+1} in the search for $\widehat{\theta}$ is found from the previous iterate θ_k by

$$\theta_{k+1} = \theta_k + \alpha p. \tag{9}$$

In particular, the search direction p for both methods is obtained from

$$\begin{aligned} p \in \{d \in \Delta \subseteq \mathbf{R}^{n_\theta} : \|E(\theta_k) - E'(\theta_k)d\| \\ \leq \|E(\theta_k) - E'(\theta_k)\bar{d}\|, \ \forall \bar{d} \in \mathbf{R}^{n_\theta}\}, \end{aligned} \tag{10}$$

In the Levenberg-Marquardt method, Δ (assumed here to be a sphere) is chosen in an adaptive manner according to how well the local approximation predicts the actual algorithm performance.

When a 'damped' Gauss-Newton technique is used instead, the search region is taken as $\Delta = \mathbf{R}^{n_\theta}$, and a second stage is introduced to compute a step length $\alpha > 0$ such that $\|E(\theta_k + \alpha p)\| < \|E(\theta_k)\|$.

In both cases the search direction defined by (10) is any vector p which satisfies

$$\left[E'(\theta_k)^T E'(\theta_k) + \lambda I\right] p = -E'(\theta_k)^T E(\theta_k), \quad \lambda \geq 0 \quad (11)$$

where $\lambda \in \mathbf{R}$ is taken as zero in the Gauss–Newton method, and is any positive value that ensures that $p \in \Delta$ in the Levenberg–Marquardt case.

4. DATA DRIVEN LOCAL CO-ORDINATES (DDLC)

The full parametrization (2) is an over-parametrization in that the set of systems represented by (1) is a manifold of dimension $n_\theta - n^2 = n(m + 2p) + mp$. This implies, since the Jacobian $E'(\theta_k)$ has n_θ columns, that the Jacobian is rank deficient, with a kernel of dimension of (at least) n^2. Therefore, the search direction (11) in the damped Gauss–Newton case of $\lambda = 0$ is not uniquely defined.

In reaction to this, several authors have proposed the use of a certain $n_\theta - n^2$ dimensional minimal parametrization that is related to the full parametrization (2) via an affine transformation, and which has been dubbed 'Data Driven Local Co-ordinates' (DDLC) (McKelvey *et al.*, 2004; McKelvey and Helmersson, 1999; Bergboer *et al.*, 2002; Verdult *et al.*, 2002; Lee and Poolla, 1999).

In this work, the key idea has been to identify the set of systems parametrized by $\theta \in \mathbf{R}^{n_\theta}$ that are input-output equivalent. This can be conveniently described by a mapping $S_\theta(T) : \mathbf{R}^{n_\theta} \to \mathbf{R}^{n_\theta}$ that depends on an arbitrary invertible matrix $T \in \mathbf{R}^{n \times n}$ according to

$$S_\theta(T) = \begin{bmatrix} \boldsymbol{T^{-1}AT} \\ \boldsymbol{T^{-1}B} \\ \boldsymbol{CT} \\ \boldsymbol{D} \\ \boldsymbol{T^{-1}K} \end{bmatrix}. \quad (12)$$

This mapping is clearly nonlinear with respect to T. However, since $S_\theta(T)$ is differentiable on $M_n \triangleq \{T \in \mathbf{R}^{n \times n} : \det(T) > 0\}$ (Lee and Poolla, 1999), a linear approximation applying locally for a perturbation ΔT around $T = I_n$ may be derived as

$$S_\theta(I_n + \Delta T) \approx S_\theta(I_n) + S'_\theta(I_n) \operatorname{vec}\{\Delta T\} = \theta + Q \operatorname{vec}\{\Delta T\}$$

where

$$Q \triangleq S'_\theta(I_n) = \left.\frac{\partial S_\theta(T)}{\partial \operatorname{vec}\{T\}}\right|_{T=I_n} = \begin{bmatrix} A^T \otimes I_n - I_n \otimes A \\ B^T \otimes I_n \\ -I_n \otimes C \\ \varnothing_{mp \times n^2} \\ K^T \otimes I_n \end{bmatrix}. \quad (13)$$

This implies that a parameter space update in the search direction $p \triangleq Q \operatorname{vec}\{\Delta T\}$ for any ΔT will locally yield a system with equivalent input-output properties, and hence, an unchanged value of $\|E(\theta)\|$. Therefore, it seems reasonable to restrict search directions to be orthogonal to the columns of Q.

In recognition of this, the works (McKelvey and Helmersson, 1997; McKelvey and Helmersson, 1999; Lee and Poolla, 1999; Bergboer *et al.*, 2002; Verdult *et al.*, 2002) have suggested the use of a local co-ordinate structure, termed DDLC, that is minimal in the sense that distinct points in parameter space correspond to non input-output equivalent systems. More specifically, a vector $\beta \in \mathbf{R}^{n_\theta - n^2}$ is used to parametrize this local co-ordinate system according to

$$\theta(\beta) = \theta + P\beta \quad (14)$$

where the columns of P are chosen by (for example) a singular value decomposition of Q, and satisfy the requirements

$$P^T P = I, \qquad P^T Q = 0, \qquad \mathscr{R}(P) \oplus \mathscr{R}(Q) = \mathbf{R}^{n_\theta}. \quad (15)$$

Here $\mathscr{R}(Q) \triangleq \{x : x = Qy \text{ for some } y\}$ is the column space of Q and similarly for P.

Thus, according to the local parameterisation (14), θ can only move in directions that are a linear combination of the columns of P, i.e. in directions $P\beta$. Hence, we may treat θ as a function of β and obtain the following problem related to (8)

$$\min_\beta \|E(\theta_k) + E'(\theta_k)P\beta\|. \quad (16)$$

As explored by (McKelvey and Helmersson, 1997; McKelvey and Helmersson, 1999; Bergboer *et al.*, 2002), the benefit of solving (16) is that β typically has dimension n^2 less than that of θ and it seems reasonable to expect that the computational load is therefore diminished. In order to understand the properties of this DDLC approach, note that according to the local parametrization (14), the prediction error vector $E(\cdot)$ can be restated as a function of β by defining a new function $E_\theta : \mathbb{R}^{n_\theta - n^2} \to \mathbb{R}^{Np}$ according to

$$E_\theta(\beta) \triangleq E(\theta(\beta)) = E(\theta + P\beta), \quad (17)$$

where the subscript denotes that θ, and consequently P, are fixed. Furthermore, the Jacobian of E_θ is given (via application of the chain-rule) as

$$E'_\theta(0) \triangleq \left.\frac{\partial E_\theta(\beta)}{\partial \beta}\right|_{\beta=0} = E'(\theta)P. \quad (18)$$

where we use the identity that $E_\theta(0) = E(\theta)$. Therefore, using this relationship, a Levenberg-Marquardt or Gauss-Newton method, in accordance with the previous discussion, may be used to solve (6) by computing a search direction q as

$$\begin{aligned} q \in \{d \in \Delta \subseteq \mathbf{R}^{n_\beta} : \|E(\theta_k) - E'_{\theta_k}(0)d\| \\ \leq \|E(\theta_k) - E'_{\theta_k}(0)\bar{d}\|,\ \forall \bar{d} \in \mathbf{R}^{n_\beta}\}, \end{aligned} \quad (19)$$

where $n_\beta \triangleq \operatorname{rank}(P)$. This in turn is satisfied by any q which solves

$$\left[E'_{\theta_k}(0)^T E'_{\theta_k}(0) + \lambda I\right] q = -E'_{\theta_k}(0)^T E(\theta_k). \quad (20)$$

With this in mind, DDLC based estimation methods proceed, at iteration k, as follows: 1. Compute the matrix P from (15); 2. Solve (20) for q; 3. Use this solution to update θ_k according to

$$\theta_{k+1} = \theta_k + \alpha P q. \quad (21)$$

5. RAPPROCHEMENT BETWEEN FULL AND DDLC PARAMETRIZED SEARCH

Computing a search direction p via the solution of (11) or an update direction q via solution of (20) is (relatively) straightforward in the Levenberg–Marquardt

situation, since $\lambda > 0$ ensures positive definiteness of the left hand side co-efficient matrix in (11) and (20), hence the uniqueness of either p or q.

However, when a Gauss–Newton search strategy is employed in which $\lambda = 0$, then this same coefficient matrix may well be rank deficient in the case of poor input excitation. It will certainly be rank deficient in the situation where the search direction p pertaining to a fully parametrized search direction is sought.

The need to deal with this sort of rank deficiency is well recognised in the general theory of gradient based optimisation. In particular, it is routinely handled by the employment of a pseudo-inverse (Golub and Loan, 1989, Section 5.5.3) of the possibly rank deficient co-efficient matrix. The ensuing scheme is denoted as a robust Gauss–Newton strategy to signify that rank deficient and full-rank cases are handled simultaneously (Nocedal and Wright, 1999, Chap. 10).

The most common implementation of robust Gauss-Newton methods employ a singular-value-decomposition (SVD) of the Jacobian matrix since this allows a straightforward and computationally robust means to compute the pseudo-inverse. To provide further detail on this point, define the SVD of $E'(\theta)$ as

$$E'(\theta) = USV^T = [U_1, U_2] \begin{bmatrix} S_1 & \varnothing \\ \varnothing & \varnothing \end{bmatrix} \begin{bmatrix} V_1^T \\ V_2^T \end{bmatrix} = U_1 S_1 V_1^T. \tag{22}$$

Concentrating for a moment of the full parametrization approach, then using this SVD we can obtain a solution to (11) for any value of $\lambda \geq 0$ (i.e. both Levenberg-Marquardt and Gauss-Newton methods) according to

$$p = -V_1(S_1^2 + \lambda I)^{-1} S_1 U_1^T E(\theta). \tag{23}$$

This follows since, from equations (11) and (22), p is required to satisfy (recall that $V^T V = VV^T = I$)

$$V(S^2 + \lambda I)V^T p = -V_1 S_1 U_1^T E(\theta). \tag{24}$$

Hence, the search direction p given in (23) can be validated by direct substitution into (24).

Moving to the DDLC case, we will require the following Lemma that establishes a connection between P and V_1.

Lemma 5.1. Let Q be given by (13) and a corresponding matrix P satisfy the equations in (15). Let $E'(\theta)$ be expressed by its SVD as in (22) and let $r_p \triangleq \text{rank}(P)$ and $r_v \triangleq \text{rank}(V_1)$. Then $r_v \leq r_p$, and $V_1 = PR$ for some matrix $R \in \mathbf{R}^{r_p \times r_v}$ with $R^T R = I$ and $V_1 = PP^T V_1$.

Proof: Since $S_\theta(T)$ parametrizes a system that is input-output equivalent to that parametrized by θ then

$$E(S_\theta(T)) = E(\theta). \tag{25}$$

Therefore, differentiating with respect to $\boldsymbol{T}$ at the point $T = I_n$ provides (recall the definition (13))

$$E'(\theta)S'_\theta(I_n) = E'(\theta)Q = 0 \tag{26}$$

and hence, via the singular value decomposition (22) and the fact that $U_1 S_1$ is full-rank

$$U_1 S_1 V_1^T Q = 0 \quad \Rightarrow \quad V_1^T Q = 0. \tag{27}$$

Hence, for any $z \in \mathscr{R}(V_1)$ it follows that $z \in \mathscr{N}(Q^T) \triangleq \{x : Q^T x = 0\}$, which from equation (15) implies that $z \in \mathscr{R}(P)$. Since z was arbitrary then $\mathscr{R}(V_1) \subseteq \mathscr{R}(P)$ which implies that $r_v \leq r_p$.

Using the above argument, any column of V_1 can be expressed as a linear combination of the columns of P, hence the expression $V_1 = PR$. Furthermore, since $P^T P = I$ and $V_1^T V_1 = I$ then $I = V_1^T V_1 = R^T P^T PR = R^T R$. Moreover, since $P^T V_1 = P^T PR = R$ then $V_1 = PR = PP^T V_1$. □

This result allows us to express the DDLC based search direction in terms of the SVD (22) as follows.

Lemma 5.2. The search direction q given by

$$q = -P^T V_1(S_1^2 + \lambda I)^{-1} S_1 U_1^T E(\theta). \tag{28}$$

satisfies equation (20) for all $\lambda \geq 0$.

Proof: According to equations (18) and (22), and using the properties that $P^T P = I$ and $VV^T = I$ we can express (20) as

$$P^T V(S^2 + \lambda I)V^T Pq = -P^T V_1 S_1 U_1^T E(\theta). \tag{29}$$

Substituting for q using (28) and exploiting the identities that $R^T R = I$, $R = P^T PR = P^T V_1$ and $V_1 = PR = PP^T V_1$ we get

$$\begin{aligned} P^T V(S^2 + \lambda I)V^T Pq &= -P^T V(S^2 + \lambda I)V^T P \\ &\quad \times P^T V_1(S_1^2 + \lambda I)^{-1} S_1 U_1^T E(\theta), \\ &= -P^T V_1 S_1 U_1^T E(\theta). \end{aligned}$$

□

These results now deliver the main technical result of this paper.

Corollary 5.1. Let Q be given by (13) and a corresponding matrix P satisfy the equations in (15). Let $E'(\theta)$ be expressed by its SVD as in (22) and let p and q be given by (23) and (28) respectively. Then the full parameterisation and DDLC parameterisation search directions coincide. That is $p = Pq$.

Proof: From Lemma 5.1 we know that $V_1 = PR = PP^T V_1$. Therefore, Pq may be expressed as

$$\begin{aligned} Pq &= -PP^T V_1(S_1^2 + \lambda I)^{-1} S_1 U_1^T E(\theta), \\ &= -V_1(S_1^2 + \lambda I)^{-1} S_1 U_1^T E(\theta) = p. \end{aligned}$$

□

The significance of this result is that, since it establishes that $p = Pq$, the search update (9) of $\theta_{k+1} = \theta_k + \alpha p$ using a fully parametrized model, and the search update (21) implied by a (locally minimal) DDLC parametrization of $\theta_{k+1} = \theta_k + \alpha Pq$ are identical.

As a consequence, gradient search employing a DDLC parametrization can be viewed as being a special case of gradient search using a full parametrization where any ensuing rank deficiency in the Jacobian E' is accommodated via a pseudo-inverse.

6. AN EXTENDED GAUSS–NEWTON APPROACH

The arguments of the previous section depended on the singular value decomposition (22) of the prediction error vector Jacobian E' where it is assumed that S_1 is a diagonal matrix with strictly positive entries.

The preceding sections have considered the DDLC approach where, according to Theorem 5.1, it is recognised that S can have no more than $n_\theta - n^2$ non-zero entries. However, if the input excitation is poor, for example, then even in case of employing DDLC, S may have less than $n_\theta - n^2$ non-zero entries, and hence some sort of on-line determination of the effective non-zero singular values in S is necessary.

This will involve a thresholding procedure, and in determining how this should be decided, it is important to recognise that if columns of V_1 are retained which correspond to singular values which are positive, but very small, then this entails a consideration of search directions which may well have negligible effect on the cost function.

Motivated by this, and the results of the preceding section which have established that DDLC-based gradient search corresponds to a particular (fixed) choice of singular value truncation, the remainder of this paper proposes and profiles an extended approach whereby the truncation point is chosen on-line and adaptively.

In particular, with the notation that the diagonal entries of S_1 are denoted as a sequence $\{\mu_1, \cdots, \mu_k\}$, this paper proposes to adaptively truncate them by restricting the singular value spread μ_r/μ_1 (for some $r \leq k$) to some small value γ; more precisely, given some small value γ then choose r such that $\mu_r < \mu_1\gamma$.

Moreover, this paper proposes that γ be changed on-line according to the size of the previous step length α according to the following reasoning. If $\alpha = 1$ on the previous iteration, then the algorithm is likely to be close to a local minima so the singular value spread γ is decreased. Vice-versa, if $\alpha < 0.5^5$ (corresponding to five bisections of the step length) then γ is increased.

The intuition underlying this approach is that it is worth focusing attention on directions in which $\|E(\theta)\|$ is sufficiently sensitive to changes in θ, and it is worth ignoring overly flat "valley" directions. The precise details of how this paper proposes these ideas be implemented are encapsulated in the following algorithm definition.

Algorithm 1. **Robust Gauss–Newton based search:** Given an initial guess θ_0, initialise $\alpha_{\min} = 0.5^5$, $\gamma = 10^{-4}$, and iterate the following steps starting with $k = 0$.

(1) Determine the prediction error vector Jacobian $E'(\theta_k)$;
(2) Compute the singular value decomposition

$$E'(\theta_k) = USV^T = U_1 S_1 V_1^T \tag{30}$$

(3) Find the index r of the smallest singular value that satisfies $S_1(r) > \gamma S_1(1)$.
(4) Let U_r, V_r be the first r columns of U_1 and V_1 respectively and let S_r be a diagonal matrix formed from the the first r entries of $\mathrm{diag}(S_1)$. Compute a search direction p as

$$p = -V_r S_r^{-1} U_r^T E(\theta_k). \tag{31}$$

(5) Initialise the step length $\alpha = 1$ and perform the following
  (a) If $V_N(\theta_k + \alpha p) < V_N(\theta_k)$ then goto step 6;
  (b) Otherwise, update $\alpha \leftarrow 0.5\alpha$ and goto (a);
(6) If $\alpha = 1$ then update $\gamma \leftarrow \min\{10^{-7}, 0.25\gamma\}$;
(7) If $\alpha \leq \alpha_{\min}$ then update $\gamma \leftarrow \max\{S_1(1), 2\gamma\}$;
(8) Set $\theta_{k+1} = \theta_k + \alpha p$ and update $k \leftarrow k + 1$;
(9) Check termination conditions (for example, $\|\mathrm{d}E(\theta_k)/\mathrm{d}\theta\| \leq$ tolerance) and stop if satisfied. Otherwise return to step 1 and repeat.

Empirical study of the performance of this algorithm relative to existing approaches together with an analysis of computational requirements now form the remainder of this paper.

7. EMPIRICAL STUDY

Although this paper is primarily concerned with providing multivariable estimates, we begin with a SISO example to emphasise that Algorithm 1 provides a dividend even in simple situations where it might otherwise be thought unnecessary.

More specifically, we begin by considering a scenario in which data is generated by simulation of the following SISO third order system

$$y_t = G(q)u_t + v_t, \quad G(q) = \frac{1.6q^3 + 3.5q^2 + 2q + 0.003}{q^3 + 1.1q^2 + 0.7q - 0.05} \tag{32}$$

The performance of Algorithm 1 with regard to estimating this system is first evaluated via Monte–Carlo analysis which involves 500 runs over different data and noise realisations. In each run $N = 500$ samples of the input signal and measurement noise were generated according to $u_t \sim \mathcal{N}(0, 1)$, $v_t \sim \mathcal{N}(0, 0.01)$, and then 500 samples of y_t were found according to (32). Four algorithms for finding estimates of (32) via the state-space model structure (1) were then implemented:

(1) Algorithm 1 as discussed above (denoted by rGN);
(2) Robust Gauss-Newton back-stepping algorithm (denoted $\mathrm{rGN}_{10^{-4}}$) with a fixed singular-value tolerance of $\gamma = 10^{-4}$ (this is equivalent to Algorithm 1 but with Steps 6 and 7 removed);
(3) Robust Gauss-Newton back-stepping algorithm (denoted $\mathrm{rGN}_{10^{-7}}$) with a fixed singular-value tolerance of $\gamma = 10^{-7}$;
(4) DDLC based gradient search (denoted PEM) as implemented via Matlab's System Identification Toolbox Version 6.0 (SIT6) `pem.m` routine.

Within this set of simulations signified by S1, the initialisation of θ_0 was performed both by initial deployment of an N4SID subspace estimation algorithm (S1a) and by simply using a random value (S1b). This latter case is included in order to study robustness to initial value and robustness against capture in local minima. All methods were run for one hundred iterations, unless they terminated earlier due to $\|\mathrm{d}E(\theta_k)/\mathrm{d}\theta\| \leq 10^{-4}$.

Table 1 then profiles the performance of the various algorithms mentioned above by showing the number of failures for each of them, where a failure is defined to be a situation in which

$$\|E(\theta)\|^2 > 1.3 \sum_{t=1}^{N} \|v_t\|^2. \tag{33}$$

Clearly, for the case of simulation S1b, Algorithm 1 is significantly more robust than the DDLC method where Jacobian subspace dimension is fixed, or when it is made adaptive in a fixed manner whereby a constant small tolerance is used to determine when singular values are essentially equal to zero.

In order to further examine these issues, but on a broader class of problems, the above Monte–Carlo scenario was repeated, but this time with a different randomly chosen 3'rd order SISO system on each run. This trial is labelled as S2 (a and b to denote N4SID-based and random initialisations). Failures were again judged according to the criterion (33) and are presented in Table 1.

This again illustrates, now for a much wider range of SISO systems, that with good initialisation, Algorithm 1 and DDLC based gradient search offer equivalent performance, while for poor initialisation, Algorithm 1 offers enhanced performance.

Progressing now to the multivariable case of input, output and state dimensions increased to $m = 2$, $p = 2$ and $n = 8$, and all other parameters as in the previous SISO case, the results, denoted as S3, with epithets a and b according to N4SID and random initialisation are presented in Table 1. This provides clear evidence that the conclusions, in terms of enhanced robustness of Algorithm 1, that arose from the previous SISO study, apply (it seems with greater emphasis) in the multivariable case which has been the main impetus for this paper.

Given this evidence, it seemed worthwhile to pursue it further by increasing input, output and state dimensions to $m = 3$, $p = 3$ and $n = 18$ and denoting the results as S4 (a and b for N4SID and random initialisation). These results comprise the final entries of Table 1.

In particular, we note that consideration of the row labelled S4a in Table 1 indicates that, for a higher dimensional system (18 state, 3 inputs/outputs), and even for a "good" N4SID initialisation, use of Algorithm 1 may reduce a 12.4% (subsequent) failure rate for a DDLC-based method to zero.

As to the overall utility of this algorithm, the question of convergence rate deserves consideration. To this end, a further simulation study, denoted by S5, was conducted with input, output and state dimensions given by $m = 2$, $p = 2$ and $n = 8$ respectively (as in simulation S3). A fixed system and a fixed initialisation point were chosen once, and randomly, and used for every simulation run. Only the input excitation was chosen randomly for each simulation, as described in previous simulations.

The number of failures, presented in Table 1 corresponding to row S5, is seen to be relatively low for three of the four algorithms, and we conclude that this system is not inherently difficult to estimate from the given initialisation point. Figure 1 presents the average prediction error cost, excluding failures, at each iteration and we conclude that, for this particular simulation, not only is Algorithm 1 more robust, but has the fastest convergence rate.

	rGN	$rGN_{10^{-4}}$	$rGN_{10^{-7}}$	PEM
S1a	0	0	0	0
S1b	34	43	169	435
S2a	0	0	0	2
S2b	8	17	156	69
S3a	0	0	0	4
S3b	3	12	241	162
S4a	0	0	2	62
S4b	12	12	331	414
S5	0	8	197	13

Table 1. *Number of failures for different algorithms (columns) under different conditions (rows).*

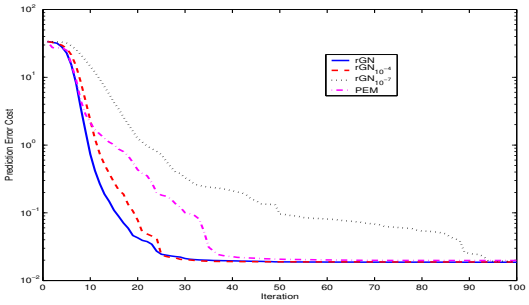

Fig. 1. *Average prediction error cost for rGN, $rGN_{10^{-4}}$, $rGN_{10^{-7}}$ and PEM algorithms for simulation S5.*

8. CONCLUSION

In this paper it is shown, under mild conditions, that when using a full-parameterisation method, the strategy of data-driven-local-coordinates (DDLC) is identical to a more widely known strategy in the optimisation literature of using Jacobian pseudo-inverses. The utility of this observation is that it informs the development of a new algorithm developed here that uses a strategy of dynamic Jacobian rank determination, which, via empirical analysis, is illustrated to offer enhanced performance.

REFERENCES

Bergboer, Niek H., Vincent Verdult and Michel H.G. Verhaegen (2002). An efficient implementation of Maximum Likelihood identification of LTI state-space models by local gradient search. In: *Proceedings of the 41st IEEE CDC, Las Vegas, USA*.

Deistler, Manfred (2000). *Model Identification and Adaptive Control*. Chap. System Identification - General Aspects and Structure. Springer-Verlag.

Dennis, J.E. and Robert B. Schnabel (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice Hall.

Fletcher, R. (1987). *Practical Methods of Optimization*. John Wiley & Sons Ltd.. Chichester.

Golub, Gene and Charles Van Loan (1989). *Matrix Computations*. Johns Hopkins University Press.

Larimore, W. (1990). Canonical variate analysis in identification, filtering and adaptive control. In: *Proceedings of the 29th IEEE Conference on Decision and Control, Hawaii*. pp. 596–604.

Lee, L. H. and K. Poolla (1999). Identification of linear parameter-varying systems using nonlinear programming. *Journal of Dynamic Systems, Management, and Control* **121**, 71–78.

Ljung, Lennart (1999). *System Identification: Theory for the User, (2nd edition)*. Prentice-Hall, Inc.. New Jersey.

Ljung, Lennart (2004). *MATLAB System Identification Toolbox Users Guide, Version 6*. The Mathworks.

McKelvey, T., A. Helmersson and T. Ribarits (2004). Data driven local coordinates for multivariable linear systems and their application to system identification. *Automatica* **40**, 1629–1635.

McKelvey, T. and A. Helmersson (1997). System identification using an over-parametrized model class - improving the optimization algorithm. In: *Proc. 36th IEEE Conference on Decision and Control*. San Diego, California, USA. pp. 2984–2989.

McKelvey, Tomas (1998). Discussion: 'on the use of minimal parametrizations in multivariable ARMAX identification' by R.P. Guidorzi. *European Journal of Control* **4**, 93–98.

McKelvey, Tomas and Anders Helmersson (1999). A dynamical minimal parametrization of multivariable linear systems and its application to optimization and system identification. In: *Proc. of the 14th World Congress of IFAC*. Vol. H. Beijing, P. R. China. pp. 7–12.

Nocedal, J. and S. J. Wright (1999). *Numerical Optimization*. Springer-Verlag. New York.

T.Söderström and P.Stoica (1989). *System Identification*. Prentice Hall. New York.

van Overschee, Peter and Bart De Moor (1996). *Subspace Identification for Linear Systems*. Kluwer Academic Publishers.

Verdult, Vincent, Niek Bergboer and Michel Verghaegen (2002). Maximum Likelihood identification of multivariable bilinear state-space systems by projected gradient search. In: *Proceedings of the 41st IEEE CDC, Las Vegas, USA*.

Verhaegen, M. (1994). Identification of the deterministic part of MIMO state space models in innovations form from input-output data. *Automatica* **30**(1), 61–74.