WEIGHTED LOW RANK APPROXIMATION AND REDUCED RANK LINEAR REGRESSION

Karl Werner and Magnus Jansson

Department of Signals, Sensors and Systems
Royal Institute of Technology (KTH), SE-100 44 Stockholm, Sweden
Email: karl.werner@s3.kth.se

ABSTRACT

The weighted low-rank approximation (WLRA) problem is considered in this paper. The problem is that of approximating one matrix with another matrix of lower rank, such that the weighted norm of the difference is minimized. The problem is fundamental in a new method for reduced rank linear regression that is outlined here, as well as in areas such as two-dimensional filter design and data mining.

The WLRA problem has no known closed form solution in the general case, but iterative methods have previously been suggested. Non-iterative methods that are asymptotically optimal for the linear regression and related problems are developed in this paper. Computer simulations, where the new methods are compared to one step of the well-known alternating projections algorithm, show significantly improved performance.

1. INTRODUCTION

The *weighted low-rank approximation* (WLRA) problem for real matrices is

$$\min_{C:\text{rank}\{C\}=r} V(C),$$
$$V(C) = \text{vec}^T(\Psi - C)Q\text{vec}(\Psi - C) = \|\Psi - C\|_Q \quad (1)$$

This means finding a rank r matrix $C \in \Re^{m\times n}$ such that $\Psi - C$ is minimized under the weighting defined by the positive definite (p.d.) matrix $Q \in \Re^{mn\times mn}$.

Problems of the form (1) appear in for example filter design. In that context the main motivation is to find two one-dimensional filters (vertical and horizontal) that as closely as possible approximate a given two-dimensional filter response. The introduction of a weighting matrix Q allows for a variable relative sensitivity in different areas of the filter's frequency response [1], [2]. Note, however, that the weighting considered in that context is slightly less general than the one considered here (corresponding to a diagonal Q). Diagonal weighting also appears in a data-mining context in factor analysis of tabulated data [3].

In this paper it is shown how a WLRA can be used as a step in a method for parameter estimation in reduced rank linear regressions, allowing for a noise model with both temporal and spatial correlation. The treated regression problem has applications in signal processing, econometrics and statistics.

There is no known closed form solution to (1) in the general case. However, for certain classes of weighting matrices a globally optimal solution can be found. One such class is where $Q = (Q_1 \otimes Q_2)$ for some p.d. $Q_1 \in \Re^{n\times n}$ and $Q_2 \in \Re^{m\times m}$, for which the solution can be found via the singular value decomposition [4]. An important specialization of this is the *unweighted* case, $Q = I$.

For a general Q, iterative algorithms of Newton and steepest descent type are developed in [4]. These methods typically require tens of iterations for convergence to a (locally) minimal solution. Another well-known algorithm is the alternating projections algorithm, see e.g. [4].

In the following, non-iterative algorithms for the WLRA problem are developed under a small residual assumption. That is,

$$\Psi = C + X,$$
$$\text{rank}\{C\} = r \quad (2)$$

where the elements of the matrix X are assumed to have small magnitudes.

2. TWO ALGORITHMS FOR THE WLRA

Two one step Newton algorithms for the WLRA problem are derived in this section. The methods are based on two different parameterizations of the criterion function $V(C)$. For small enough X, the minimizer of $\|\Psi - C\|_Q$ under the rank constraint will be arbitrarily close to the minimizer of $\|\Psi - C\|_I$. The optimal solution to the unweighted problem ($Q = I$) could therefore be used as initial value. This means that higher order terms of a Taylor series expansion of the criterion function can be ignored, and a Newton algorithm will reach a stationary point in one step.

2.1. The one-step correction method

A standard way to capture the low-rank constraint is to introduce the (non-unique) parameterization $C = AB^T$, $A \in \Re^{m\times r}$, $B \in \Re^{n\times r}$, where both A and B are full rank matrices. A vectorized version of this parametrization is

$$\theta_{A,B} = [\text{vec}^T(A)\ \text{vec}^T(B^T)]^T = [a^T\ b^T]^T \quad (3)$$

The criterion function to be minimized becomes

$$V(C) = V(\theta_{A,B}) = \text{vec}^T(\Psi - AB^T)Q\text{vec}(\Psi - AB^T) \quad (4)$$

Also let $\hat{\theta}_{A,B}$ be a parametrization of an initial estimate $\hat{C} = \hat{A}\hat{B}^T$. In the following, a correction vector $\tilde{\theta}_{A,B}$ is sought in order

to minimize $V(\hat{\theta}_{A,B} + \tilde{\theta}_{A,B})$. Performing a series expansion of the gradient of the criterion function around $\hat{\theta}_{A,B}$ gives

$$\begin{aligned} 0 =& V'(\hat{\theta}_{A,B} + \tilde{\theta}_{A,B}) \\ =& V'(\hat{\theta}_{A,B}) + V''(\hat{\theta}_{A,B})\tilde{\theta}_{A,B} + \epsilon(\hat{\theta}_{A,B}, \tilde{\theta}_{A,B}), \end{aligned} \quad (5)$$

If (2) is valid then the term $\epsilon(\hat{\theta}_{A,B}, \tilde{\theta}_{A,B})$ is small. The Newton step (of minimum norm) is given by

$$\tilde{\theta}_{A,B} = -H^{+} V'(\hat{\theta}_{A,B}), \quad (6)$$

where H^{+} denotes the Moore-Penrose pseudo inverse of the rank deficient asymptotic Hessian H that is defined below. The Hessian and gradient are derived in Appendix A.1. They are

$$\begin{aligned} V''(\hat{\theta}_{A,B}) &\simeq H = 2(\hat{B} \otimes I_m \; I_n \otimes \hat{A})^T Q (\hat{B} \otimes I_m \; I_n \otimes \hat{A}) \\ V'(\hat{\theta}_{A,B}) &= 2(\hat{B} \otimes I_m \; I_n \otimes \hat{A})^T Q \mathrm{vec}(\hat{A}\hat{B}^T - \Psi) \end{aligned} \quad (7)$$

where $\simeq$ denotes equality in the dominating terms (based on the small residual assumption). Interestingly, the same update-equation (6) is obtained if a quadratic approximation of the criterion function (4) is considered. Let $\theta_{A,B} = \hat{\theta}_{A,B} + \tilde{\theta}_{A,B}$. Then with $\psi = \mathrm{vec}(\Psi)$:

$$\begin{aligned} V(A,B) &= \|Q^{1/2}\mathrm{vec}(\Psi - AB^T)\|_2^2 = \|Q^{1/2}\big(\psi - \mathrm{vec}(\hat{A}\hat{B}^T) \\ &\quad - \mathrm{vec}(\hat{A}\tilde{B}^T) - \mathrm{vec}(\tilde{A}\hat{B}^T) + \mathrm{vec}(\tilde{A}\tilde{B}^T)\big)\|_2^2 \\ &\simeq \|Q^{1/2}\big(\psi - \mathrm{vec}(\hat{A}\hat{B}^T) - \mathrm{vec}(\hat{A}\tilde{B}^T) \\ &\quad - \mathrm{vec}(\tilde{A}\hat{B}^T)\big)\|_2^2 \equiv \tilde{V}(\tilde{A}, \tilde{B}) \end{aligned} \quad (8)$$

The minimum norm minimizers of $\tilde{V}(\tilde{A}, \tilde{B})$ are given by (6).

2.2. Parameterizing the null-space of C

The main motivation for the next approach is to reduce the number of parameters to be updated, and thereby the computational complexity. The algorithm derived here differs from one iteration of the Newton algorithms presented in [4] in that the small residual assumption (2) is used to make simplifications that reduce the computational complexity drastically. The same underlying criterion-function is used, however.

It is clear that minimizing (4) with respect to (w.r.t.) A gives (for a fixed B)

$$\begin{aligned} a_m &\equiv \arg\min_{a=\mathrm{vec}(A)} V(A,B) \\ &= \arg\min_{a} \|Q^{1/2}\psi - Q^{1/2}(B \otimes I_m)a\|_2^2 \\ &= \Big((B^T \otimes I_m)Q(B \otimes I_m)\Big)^{-1}(B^T \otimes I_m)Q\psi \end{aligned} \quad (9)$$

Inserting (9) into (4) and defining $\bar{B} \equiv Q^{1/2}(B \otimes I_m)$ yield

$$V(B) = \|(I_{mn} - \bar{B}(\bar{B}^T\bar{B})^{-1}\bar{B}^T)Q^{1/2}\psi\|_2^2 \quad (10)$$

Now, define $N \in \Re^{n \times (n-r)}$ to be a full column-rank matrix such that $N^T B = 0$ ($\Rightarrow CN = AB^T N = 0$) and $\bar{N}^T = (N^T \otimes I_m)Q^{-1/2}$. Then $\bar{N}^T\bar{B} = 0$ and $\Pi_{\bar{B}}^{\perp} = \Pi_{\bar{N}}$[1]. Thus

$$\begin{aligned} V(B) &= \|\Pi_{\bar{B}}^{\perp} Q^{1/2}\psi\|_2^2 = \|\bar{N}(\bar{N}^T\bar{N})^{-1}\bar{N}^T Q^{1/2}\psi\|_2^2 \\ &= \psi^T Q^{1/2}\bar{N}(\bar{N}^T\bar{N})^{-1}\bar{N}^T Q^{1/2}\psi \equiv f(N) \end{aligned} \quad (11)$$

[1] $\Pi_X = X(X^TX)^{-1}X^T$ is the orthogonal projection matrix onto the range space of X and $\Pi_X^{\perp} = I - \Pi_X$ is the orthogonal projection matrix onto the null space of X^T.

Given an $N = \hat{N}$ minimizing $f(N)$, it is easy to find the $\hat{C}$ minimizing $V(C)$. To that end, reconsider (9) and note that

$$\begin{aligned} \mathrm{vec}(\hat{C}) &= Q^{-1/2}\hat{\bar{B}}a_m = Q^{-1/2}\hat{\bar{B}}(\hat{\bar{B}}^T\hat{\bar{B}})^{-1}\hat{\bar{B}}^T Q^{1/2}\psi \\ &= Q^{-1/2}\Pi_{\hat{\bar{B}}} Q^{1/2}\psi = Q^{-1/2}\Pi_{\hat{\bar{N}}}^{\perp} Q^{1/2}\psi \\ &= \psi - Q^{-1/2}\hat{\bar{N}}(\hat{\bar{N}}^T\hat{\bar{N}})^{-1}\hat{\bar{N}}^T Q^{1/2}\psi \end{aligned} \quad (12)$$

This means that minimizing $V(C)$ w.r.t. C is equivalent to minimizing $f(N)$ w.r.t. N. The main advantage of the second approach is that the number of parameters to update typically is reduced compared to the first approach.

By using the same reasoning as in the section above it can be concluded that one Newton step approximately will reach a stationary point.

The Hessian and gradient of the criterion function (11) are derived and simplified, using the small residual assumption (2), in Appendix A.2

$$\begin{aligned} f''(\hat{N}) &\simeq H = 2(I_{n-r} \otimes \hat{C}^T)(\hat{\bar{N}}^T\hat{\bar{N}})^{-1}(I_{n-r} \otimes \hat{C}) \\ f'(\hat{N}) &\simeq 2(I_{n-r} \otimes \hat{C}^T)(\hat{\bar{N}}^T\hat{\bar{N}}^T)^{-1}\bar{N}Q^{1/2}\psi \end{aligned} \quad (13)$$

3. REDUCED RANK LINEAR REGRESSION

The problem considered is parameter estimation in the reduced rank linear regression

$$y(t) = Cx(t) + e(t), \quad (14)$$

where the matrix $C \in \Re^{m \times n}$ has the known rank r and $e(t)$ is noise with zero mean, possibly with both temporal and spatial correlation. The noise model employed here is therefore more general than that of other methods, e.g. [5]. The noise and the input signal $x(t)$ are uncorrelated. The aim of the algorithm is to estimate C based on the observed data $\{x(i),\ y(i)\}_{i=1}^{M}$.

The algorithm will only be outlined here, a more thorough discussion, including asymptotical analysis and a comparison to other methods such as those described in [5] and [6], is presented in [7].

A three step procedure, motivated by the *extended invariance principle* [8], is used. The first step is an unconstrained least squares fit of a matrix Ψ to the data:

$$\Psi = \left(\frac{1}{M}\sum_{t=1}^{M} y(t)x^T(t)\right)\left(\frac{1}{M}\sum_{t=1}^{M} x(t)x^T(t)\right)^{-1} \quad (15)$$

Insertion of (14) into this expression gives

$$\begin{aligned} \Psi &= \hat{R}_{yx}\hat{R}_{xx}^{-1} = C + \hat{R}_{ex}\hat{R}_{xx}^{-1} \\ &\lim_{M\to\infty} \hat{R}_{ex} = 0, \end{aligned} \quad (16)$$

where the last equality holds with probability one (w.p. 1) and in the mean-square sense. In the second step the covariance matrix $Q_\psi = \mathrm{E}[\mathrm{vec}(\hat{R}_{ex})\mathrm{vec}^T(\hat{R}_{ex})]$ of ψ is estimated. It is [7]

$$M^2 Q_\psi = \mathrm{E}[\sum_{t=1}^{M}\sum_{s=1}^{M}(\hat{R}_{xx}^{-1}x(t)x^T(s)\hat{R}_{xx}^{-1} \otimes e(t)e^T(s))] \quad (17)$$

As suggested in [9] and [10] a good approximation of Q_ψ, that is guaranteed to be positive semi definite is

$$\hat{Q}_\psi = \frac{1}{M^2}\sum_{\tau=-M}^{M}(M - |\tau|)(\hat{R}_{xx}^{-1}\hat{R}_{xx}(\tau)\hat{R}_{xx}^{-1} \otimes \hat{R}_{ee}(\tau)) \quad (18)$$

The noise autocorrelations can be estimated using the observed data and Ψ [7]. Finally, the estimate of C is the solution to the WLRA problem

$$\hat{C} = \underset{C:\text{rank}\{C\}=r}{\arg\min} \|\Psi - C\|_{\hat{Q}_\psi^{-1}} \tag{19}$$

Since the covariance of $\hat{R}_{ex}$ tends to zero with increasing M, the approximations made in the derivations of the Hessians will be valid. As M increases the probability density function of Ψ will be concentrated around C and the approximations made when omitting higher order terms in the Taylor series expansions will hold. The Newton methods presented in Section 2 are thus asymptotically optimal for the problem at hand.

4. SIMULATION STUDY

Monte Carlo-type simulations of the linear regression problem described above are presented in this section.

The noise was generated by a generalized MA process

$$e(t) = \sum_{k=0}^{T-1} M_k w(t-k) \tag{20}$$

where $M_k \in \Re^{m\times m}$ are randomly generated matrices, each element was drawn from a gaussian distribution with zero-mean and unit variance. The process $w(t)$ is gaussian white noise with covariance $\sigma_w^2 I$. The input signal, $x(t)$, was generated similarly using sequences that are realizations of white gaussian noise processes with covariance $\sigma_u^2 I$. Both the noise and the input signal were regenerated in each Monte Carlo experiment. The coefficient matrices M_k were kept constant throughout the experiments. The covariance length T was fixed to 70 for both the noise and the input signal. The SNR$\equiv \sigma_u^2/\sigma_w^2$, was fixed to 20 while the size of the data set M was varied. The rank one regression matrix was

$$C = \begin{pmatrix} -0.3634 & -0.6984 & 0.1761 & 0.0838 \\ -0.1058 & -0.2033 & 0.0513 & 0.0244 \\ -0.1415 & -0.2719 & 0.0686 & 0.0326 \\ -0.1539 & -0.2957 & 0.0746 & 0.0355 \end{pmatrix} \tag{21}$$

The results of the simulations are presented in Figure 1. The MSE for the unconstrained least squares fit ($\hat{C} = \Psi$) and for unweighted rank reduction, obtained by setting $\hat{Q}_\psi = I$ in (19), are also included. Clearly, the two Newton methods perform very similarly. The new methods perform better than one step of the alternating projections method that was included for comparison. The empirical estimate of the MSE[2] reaches the asymptotical value also for relatively small M which indicates that the new methods are very well suited for the linear regression application.

5. CONCLUSIONS

Two new Newton-step methods were derived in order to find approximate solutions to the WLRA problem. As a step in a new method for reduced rank linear regression these methods are asymptotically optimal. Computer simulations suggest that the new methods perform well also on practical sample set sizes. When compared to one step of the alternating projections algorithm (which is comparable in terms of computational complexity) the new methods give significantly improved performance.

[2]MSE=E[$\frac{1}{mn}\|C-\hat{C}\|_F^2$]

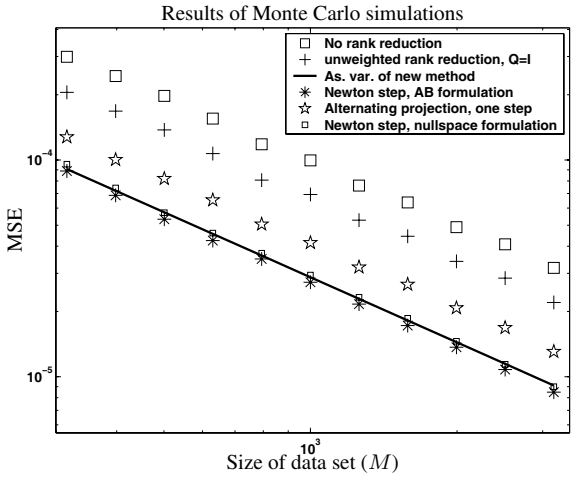

Fig. 1. Empirical MSE of the estimated linear regression matrices for sample set sizes M ranging from 316 to 3162. The theoretical asymptotical variance for the regressor parameter estimator is given by the solid line. Each point is averaged over 1500 independent realizations. All methods were applied to the same data.

A. DERIVATION OF THE LIMITING HESSIANS

A.1. The first parameterization

In this section the gradient and Hessian of the criterion function (4) are derived. The index A, B of $\theta_{A,B}$ will in the following be dropped. The criterion function can be written as

$$\begin{aligned} V(\theta) = \psi^T Q\psi - a^T(B^T \otimes I_m)Q\psi - \psi^T Q(B\otimes I_m)a \\ + a^T(B^T\otimes I_m)Q(B\otimes I_m)a \end{aligned} \tag{22}$$

This immediately gives

$$\frac{\partial V(\theta)}{\partial a} = 2(B^T\otimes I_m)Q\big((B\otimes I_m)a - \psi\big) \tag{23}$$

Performing parallel calculations for $b = \text{vec}(B^T)$ gives the sought gradient (7). The Hessian follows from the partial derivatives:

$$\frac{\partial^2 V(\theta)}{\partial a \partial a^T} = 2(B^T\otimes I_m)Q(B\otimes I_m) \tag{24}$$

$$\frac{\partial^2 V(\theta)}{\partial b \partial b^T} = 2(I_n\otimes A^T)Q(I_n\otimes A) \tag{25}$$

$$\begin{aligned} \frac{\partial^2 V(\theta)}{\partial b \partial a^T} = 2(\frac{\partial}{\partial b}(B^T\otimes I_m))Q\text{vec}(AB^T-\Psi) \\ + 2(B^T\otimes I_m)Q(I_n\otimes A) \end{aligned} \tag{26}$$

According to assumption (2) the quantity $\Psi - AB^T = X$ will be small close to the true value of A and B and (7) results.

A.2. The second parameterization

Here the gradient and Hessian of (11) will be derived. In order to differentiate $f(\bar{N})$, investigate the differential of $\bar{N}(\bar{N}^T\bar{N})^{-1}\bar{N}^T$:

$$\begin{aligned} d\bar{N}(\bar{N}^T\bar{N})^{-1}\bar{N}^T + \bar{N}(\bar{N}^T\bar{N})^{-1}d\bar{N}^T \\ - \bar{N}(\bar{N}^T\bar{N})^{-1}(d\bar{N}^T\bar{N} + \bar{N}^T d\bar{N})(\bar{N}^T\bar{N})^{-1}\bar{N}^T \end{aligned} \tag{27}$$

See [11] for the necessary relations. Then

$$df(\bar{N}, d\bar{N}) = \psi^T Q^{1/2}\big((I_{mn} - \bar{N}L^{-1}\bar{N}^T)d\bar{N}L^{-1}\bar{N}^T + \bar{N}L^{-1}d\bar{N}^T(I_{mn} - \bar{N}L^{-1}\bar{N}^T)\big)Q^{1/2}\psi \quad (28)$$

where $L \equiv (\bar{N}^T\bar{N})$. Since the differential is a scalar function it is clear that (28) can be written

$$df(\bar{N}, d\bar{N}) = 2\psi^T Q^{1/2} d\bar{N}L^{-1}\bar{N}^T Q^{1/2}\psi - 2\psi^T Q^{1/2}\bar{N}L^{-1}\bar{N}^T d\bar{N}L^{-1}\bar{N}^T Q^{1/2}\psi \quad (29)$$

In order to proceed define $\text{vec}(D) \equiv L^{-1}\bar{N}^T Q^{1/2}\psi$, $D \in \Re^{m\times(n-r)}$ and $\text{vec}(F) \equiv Q^{-1}\text{vec}(DN^T)$, $F \in \Re^{m\times n}$. Also note that $d\bar{N} = Q^{-1/2}(dN \otimes I_m)$. This gives

$$\begin{aligned} df(N; dN) =& 2\psi^T(dN \otimes I_m)\text{vec}(D) \\ & - 2\text{vec}^T(D)(N^T \otimes I_m)Q^{-1}(dN \otimes I_m)\text{vec}(D) \\ =& 2\left(\text{vec}^T(\Psi^T D) - \text{vec}^T(F^T D)\right)\text{vec}(dN) \end{aligned} \quad (30)$$

The identification theorem gives the gradient

$$f'(N) = 2\text{vec}((\Psi - F)^T D) = \nabla f \quad (31)$$

Proceeding to find the Hessian it is convenient to first calculate the differentials of D and F

$$\begin{aligned} d\text{vec}(D) =& L^{-1}d\bar{N}^T Q^{1/2}\psi - L^{-1}(d\bar{N}^T\bar{N} + \bar{N}^T d\bar{N})\text{vec}(D) \\ =& L^{-1}\big(\text{vec}(\Psi dN) - \text{vec}(F dN) \\ & - (N^T \otimes I_m)Q^{-1}\text{vec}(DdN^T)\big), \\ d\text{vec}(F) =& Q^{-1}\text{vec}(DdN^T) + Q^{-1}\text{vec}(dDN^T) \end{aligned} \quad (32)$$

The differential of the gradient (31) is then

$$\begin{aligned} \frac{1}{2}d\nabla f =& (I_{n-r} \otimes (\Psi^T - F^T))\text{vec}(dD) - (D^T \otimes I_n)\text{vec}(dF^T) \\ =& \big((I_{n-r} \otimes (\Psi^T - F^T)) \\ & - (D^T \otimes I_n)K_{m,n}Q^{-1}(N \otimes I_m)\big)\text{vec}(dD) \\ & - (D^T \otimes I_n)K_{m,n}Q^{-1}(I_n \otimes D)K_{n,(n-r)}\text{vec}(dN) \end{aligned} \quad (33)$$

where $K_{x,y}$ is the matrix satisfying $K_{x,y}\text{vec}(X) = \text{vec}(X^T)$ for any $X \in \Re^{x\times y}$. From (33) the Hessian can be found as

$$\begin{aligned} \frac{1}{2}f''(N) =& (I_{n-r} \otimes (\Psi^T - F^T))L^{-1}(I_{n-r} \otimes (\Psi - F)) \\ & + (D^T \otimes I_n)K_{m,n}Q^{-1}(N \otimes I_m)L^{-1} \\ & \quad \times (N^T \otimes I_m)Q^{-1}(I_n \otimes D)K_{n,(n-r)} \\ & - (I_{n-r} \otimes (\Psi^T - F^T))L^{-1} \\ & \quad \times (N^T \otimes I_m)Q^{-1}(I_n \otimes D)K_{n,(n-r)} \\ & - (D^T \otimes I_n)K_{m,n}Q^{-1} \\ & \quad \times (N \otimes I_m)L^{-1}(I_{n-r} \otimes (\Psi - F)) \\ & - (D^T \otimes I_n)K_{m,n}Q^{-1}(I_n \otimes D)K_{n,(n-r)} \end{aligned} \quad (34)$$

Using the small residual assumption (2) and (12) it can be noted that:

$$\text{vec}(\Psi - F) = \psi - \text{vec}(F) \simeq \text{vec}(C), \quad (35)$$

Also, since $C\hat{N} \approx 0$, it can be concluded that $D \approx 0$. Thus the asymptotic Hessian and gradient are given by (13).

B. REFERENCES

[1] D.J. Shpak, "A weighted-least-squares matrix decomposition method with application to the design of two-dimensional digital filters," in *Proceedings of the 33rd Midwest Symposium on Circuits and Systems*, Aug 1990, pp. 1070–1073.

[2] S.-C Pei W.-S. Lu and P.-H Wang, "Weighted low-rank approximation of general complex matrices and its application in the design of 2-d digital filters," *IEEE Transactions on Circuits and Systems*, vol. 44, no. 7, pp. 650–655, Jul. 1997.

[3] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proceedings of the Twentieth International Conference on Machine Learning.*, 2003.

[4] R. Mahony J.H. Manton and Y. Hua, "The geometry of weighted low-rank approximations," *IEEE Transactions on Signal Processing*, vol. 51, pp. 500–514, Feb. 2003.

[5] P. Stoica and M. Viberg, "Maximum likelihood parameter and rank estimation in reduced-rank linear regressions," *IEEE Transactions on Signal Processing*, vol. 44, pp. 3069–3078, Dec. 1996.

[6] T. Gustafsson and B. D. Rao, "Statistical analysis of subspace-based estimation of reduced-rank linear regressions," *IEEE Transactions on Signal Processing*, vol. 50, no. 1, pp. 151–159, Jan. 2002.

[7] K. Werner and M. Jansson, "Parameter estimation for reduced-rank multivariate linear regressions in the presence of correlated noise," in *Proceedings of Asilomar'03*, Nov. 2003.

[8] P. Stoica and T. Söderström, "On reparameterization of loss functions used in estimation and the invariance principle," *Signal Processing*, vol. 17, pp. 383–387, 1989.

[9] P. Stoica and M. Jansson, "Mimo system identification: State-space and subspace approximations versus transfer function and instrumental variables," *IEEE Transactions on Signal Processing*, vol. 48, no. 11, pp. 3087–3099, Nov. 2000.

[10] P. Stoica and M. Jansson, "Estimating optimal weights for instrumental variable methods," in *Digital Signal Processing*, vol. 11, pp. 252–268. 2001.

[11] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley Series in Probability and Mathematical Statistics. John Wiley and Sons, 1988.