On Global Optimization of the Likelihood Function for Linear Systems

Bernard Hanzon [1]

joint work with

Wolfgang Scherrer[2]

[1]Department of Mathematics, University College Cork, Ireland
[2]Department of Econometrics, Operations Research and System Theory (EOS), TU Vienna, Austria.

ERNSI, Lyon, 27 September 2017

Linear Dynamical Gaussian State-Space Systems

- Discrete time linear state-space system:

$$x_{t+1} = Ax_t + B\epsilon_t$$

$$y_t = Cx_t + \epsilon_t$$

where $\{\epsilon_t\}$ is Gaussian i.i.d. with zero mean and variance equal to a matrix(parameter) Ω; $t = 1, 2, \ldots, T$, initial state x_1, state dimension n, and output dimension m.

Linear Dynamical Gaussian State-Space Systems

- Note: This system is in innovations form. This means that the ϵ_t coincide with the prediction error if one predicts the next observation by the conditional expectation of the output variable given the previous observations of the output variable. Note however that time-invariance of the matrices A, B, C, Ω has been assumed here, while stationarity of the time series has NOT been assumed.

Augmented System

- Augmented system:

$$x_{t+1} = Ax_t + \bar{B}\bar{\epsilon}_t$$

$$\bar{y}_t = \bar{C}x_t + \bar{\epsilon}_t$$

$t = 0, 1, \ldots, T$, initial state $x_0 = 0$, and

$$\bar{B} = [B, x_1], \bar{C} = \begin{bmatrix} C \\ 0 \end{bmatrix}, \bar{\epsilon}_t = \begin{bmatrix} \epsilon_t \\ \delta_t \end{bmatrix}, \bar{y}_t = \begin{bmatrix} y_t \\ \delta_t \end{bmatrix},$$

$$y_t = \epsilon_t = 0 \text{ for } t \leq 0; \delta_0 = 1, \ \delta_t = 0 \text{ for } t > 0.$$

Inverse System

- Inverse System:

$$x_{t+1} = (A - \bar{B}\bar{C})x_t + \bar{B}\bar{y}_t,$$

$$\bar{\epsilon}_t = -\bar{C}x_t + \bar{y}_t.$$

Let $\bar{A} := A - \bar{B}\bar{C} = A - BC$. We have

$$x_t = \sum_{j \geq 0} \bar{A}^j \bar{B} \bar{y}_{t-j-1}, \text{ for } t \leq T + 1.$$

Inverse System

- Let $X := [x_T, x_{T-1}, \ldots, x_T]$, $R := [\bar{B}, \bar{A}\bar{B}, \ldots, \bar{A}^{T-1}\bar{B}]$ and

$$\bar{H}(y) = \begin{bmatrix} \bar{y}_{T-1} & \bar{y}_{T-2} & \cdots & \bar{y}_1 & \bar{y}_0 \\ \bar{y}_{T-2} & \bar{y}_{T-3} & \cdots & \bar{y}_0 & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ \bar{y}_1 & \bar{y}_0 & & 0 & 0 \\ \bar{y}_0 & 0 & \ldots & 0 & 0 \end{bmatrix}$$

Then

$$X = R.\bar{H}(y).$$

Furthermore, let $Y = [\bar{y}_T, \bar{y}_{T-1}, \ldots, \bar{y}_1]$ and $E = [\bar{\epsilon}_T, \bar{\epsilon}_{T-1}, \ldots, \bar{\epsilon}_1]$, then

$$Y = CX + E.$$

Likelihood

- Likelihood function (negative log likelihood really) for this model: $I(\bar{A}, \bar{B}, C, \Omega) =$

$$\log\det\Omega + \frac{1}{T}\sum_{t=1}^{T}\epsilon_t'\Omega^{-1}\epsilon_t = \log\det\Omega + \mathrm{tr}\left(\left(\frac{1}{T}EE'\right)\Omega^{-1}\right)$$

- For given $(\bar{A}, \bar{B}, C)$ (i.e. given E) the optimizer for Ω is $\hat{\Omega} = \left(\frac{1}{T}EE'\right)$. The "concentrated" likelihood, up to an additive constant, is

$$I(\bar{A}, \bar{B}, C, \hat{\Omega}) = \log\det\left(\frac{1}{T}EE'\right).$$

Concentrated Likelihood

- For given $(\bar{A}, \bar{B})$ (i.e. given $X = R.\bar{H}(y)$) the optimizer for C is obtained by projection (regression) of Y onto the row space of X :

$$Y = \hat{C}X + \hat{E}, \text{ where } X\hat{E}' = X(Y - \hat{C}X)' = 0$$

If X has full rank n then
$\hat{C} = YX'(XX')^{-1}, \hat{E} = Y - YX'(XX')^{-1}X,$
$\hat{E}\hat{E}' = YY' - YX'(XX')^{-1}XY'.$

- If the state matrix X_0 of a pair $(\hat{A}_0, \hat{B}_0)$ has full rank then the "concentrated" likelihood function $l_3(\bar{A}, \bar{B})$ is continuous in a neighbourhood of this pair (and hence in a neighbourhood of the corresponding parameter vector θ_0).

Rank deficiency of X

- It can be shown that rk $X < n$ implies perfect prediction(s) for certain components or linear combinations of the outputs up to the time point $T - \text{corank}(X)$. This can be shown to imply that the infimum of the negative log-likehood function is minus infinity. It is possible to detect this possibility by inspecting whether the $nm \times (T - n)$ data matrix

$$H_n^T(y) := \begin{pmatrix} y_{T-1} & \cdots & y_n \\ y_{T-2} & \cdots & y_{n-1} \\ \vdots & & \vdots \\ y_{T-n} & \cdots & y_1 \end{pmatrix}$$

has a non-trivial left kernel. If not then the rank of X will be n for all pairs $(\bar{A}, \bar{B})$ for which R has rank n. From now on we will assume that this is the case, without loss of generality.

Dependence on the row space of R

- Note that the concentrated likelihood function is invariant under a linear state space transformation given by a non-singular matrix S, say: gives new reachable pair $(S\bar{A}S^{-1}, S\bar{B})$, corresponding reachability matrix $S.R$, hence corresponding state matrix $S.X$, hence corresponding matrix $\hat{E}\hat{E}' = YY' - YX'S'(SXX'S')^{-1}SXY' = YY' - YX'(XX')^{-1}XY'$.
- Therefore the likelihood only depends on the *row space* row(R) of R.

Geometry of the row spaces

- The row spaces all have the same dimension n and hence form a subset of the collection of all $n-$dimensional subspaces of $(m+1)T-$dimensional Euclidean space. In geometry this last collection is called a Grassmannian, which is known to be compact differential manifold. It follows that the closure of our family of row spaces of R as a subspace of the Grassmann manifold must also be compact.
- One can ask whether it will also be a differentiable manifold. This can be answered in the positive by the explicit construction of an atlas of charts (each chart being an open subset of Euclidean space of dimension $n(m+1)$.).

Backward reachability matrix

- First note that if $\bar{A}$ is invertible, then the row space of R is the same as that of the matrix $(\bar{A}^{-(T-1)}\bar{B}, \bar{A}^{-(T-2)}\bar{B}, \ldots, \bar{B})$ which is a "backward" reachability matrix of the pair $(\bar{A}_b, \bar{B}_b) := (\bar{A}^{-1}, \bar{B})$.
- In the "backward" reachability matrix one can allow the matrix $\bar{A}_b$ to become singular. This corresponds to one or more of the eigenvalues of the matrix $\bar{A}$ going to infinity in modulus.
- To allow for the possibility that $\bar{A}$ has eigenvalues (going to) zero as well as eigenvalues "going to infinity in modulus", we can *additively decompose* the reachable pair $(\bar{A}, \bar{B})$ into a reachable pair $(\bar{A}_f, \bar{B}_f)$, with n_f *small* eigenvalues and a reachable pair $(\bar{A}_b^{-1}, \bar{B}_b)$ with n_b *large* eigenvalues, where $n_f + n_b = n$, such that there is a *gap* between the spectral radius of $\bar{A}_f$ and the inverse of the spectral radius of $\bar{A}_b$, where this inverse could possibly be infinity.

Forward-backward reachability matrix

- The corresponding reachability matrix has row space equal to

$$\text{row}(R) = \text{row}\begin{pmatrix} \bar{B}_f & \bar{A}_f\bar{B}_f & \dots & \bar{A}_f^{T-2}\bar{B}_f & \bar{A}_f^{T-1}\bar{B}_f \\ \bar{A}_b^{T-1}\bar{B}_b & \bar{A}_b^{T-2}\bar{B}_b & \dots & \bar{A}_b\bar{B}_b & \bar{B}_b \end{pmatrix}$$

- Note that the matrix appearing in the right-hand side of this equation is well-defined even if $\bar{A}_b$ is *singular*.
- Question: Can we construct a *finite* set of parametrizations that cover all cases?
  Answer: Yes; this can be done in two steps.

Finite set of covering parametrizations

- Step 1. Consider $0 < a_0 < b_0 < a_1 < b_1 < \ldots < a_n < b_n < \infty$. Then the closed intervals $[a_0, b_0], [a_1, b_1], \ldots, [a_n, b_n]$ each have positive length and are pairwise disjoint. With each $j = 0, 1, 2, \ldots, n$ we associate the set S_{j,n_f,n_b} of all pairs $(\bar{A}, \bar{B})$ such that $\mathrm{rk}(R(\bar{A}, \bar{B})) = n$ and for which n_f of the eigenvalues (multiplicities included) of $\bar{A}$ have modulus less than a_j and $n_b = n - n_f$ of the eigenvalues (multiplicities included) have modulus larger than b_j.
- We claim that each pair $(\bar{A}, \bar{B})$ with $\mathrm{rk}(R(\bar{A}, \bar{B})) = n$ is element of S_{j,n_f,n_b} for at least one value of $j \in \{0, 1, \ldots, n\}$ and $n_f \in \{0, 1, \ldots, n\}$ while $n_b = n - n_f$. Reason is that the n moduli of eigenvalues of $\bar{A}$ can be member of at most n of the $n + 1$ intervals, so one of the intervals can serve as a *gap* in the collection of moduli of the spectrum.

Finite set of covering parametrizations

- Step 2. For each S_{j,n_f,n_b} we can construct a finite atlas of local parametrizations. This can be done by existing methods: Take $\rho_j \in (a_j, b_j),\ j = 0, 1, 2, \ldots, n$. Then $(\rho_j^{-1}\bar{A}_f, \bar{B}_f)$ as well as $(\rho_j \bar{A}_b, \bar{B}_b)$ form a stable pair in the sense that the spectral radius of the first matrix of the pair is less than one. Such stable pairs can be parametrized using input-normal SDPS (subdiagonal pivot structure) forms, for which a finite atlas exists (see H-Olivi-Peeters, SYSID 2009). Transformations between such input-normal forms is performed by *orthogonal* state-space transformation matrices.

Finite set of covering parametrizations

- Note that we obtain a *finite* set of local parametrizations covering all cases. Each chart can be taken to be a bounded nonempty open set in $n(m+1)-$dimensional Euclidean space. It follows that if we can extend the criterion function to the closure of these charts and *if* the extension is continuous, then from topology we know that on the extended space, which is *compact*, the criterion function will attain an optimum.

Regularization

- However on the extended space the matrix X cannot be guaranteed to be of full rank n and hence continuity of the likelihood function cannot be guaranteed. We propose to use a *regularization* by replacing $(XX')^{-1}$ in the formula for C by a well-defined matrix, by replacing each of the singular values of XX' less than δ, where $\delta > 0$ is a chosen treshold value, by δ. Note that we do not change X elsewhere in the formula(s). We can show that the resulting regularized likelihood function is continuous on the extended space. Therefore the minimum of this regularized likelihood function is attained on the extended space.

Lipschitz continuity

- To find the minimum of the regularized likehood function, one can apply the following technique:
- Subdivide the space in a *finite* number of small pieces, each one contained in a small ball (this is possible due to the compactness of the space).
- Calculate the value of the criterion function at one point in each piece. If the variation of the function values within each piece is small then taking the minimum of the computed function values gives a good approximation of the true minimum value over all the points. If we know an upper bound to the variation within each piece then we also know a lower bound of the true minimum. We can take the minimum of the computed function values as the upper bound to the true minimum.

Lipschitz continuity

- We can also decide in which pieces the minimum could possibly be attained and in which pieces this is not the case.
- By taking the pieces sufficiently small (e.g. by subdividing them further) one can make the results arbitrarily precise.
- What is required for such a procedure to work is *Lipschitz continuity* of the criterion function.
- A function $f(x)$ is Lipschitz continuous if there exists a number L, called a Lipschitz constant, such that for all pairs of points x_1, x_2 we have

$$|f(x_1) - f(x_2)| \leq L d(x_1, x_2)$$

where $d(x_1, x_2)$ denotes the distance between the points x_1, x_2.

Lipschitz continuity

- Using a nice result of Wihler(2009) on Hölder continuity of matrix functions, we can show that the regularized criterion function is Lipschitz continuous. A Lipschitz constant can be obtained constructively.

Recent and astonishing findings

- We discovered an astonishing fact. Most easily explained by considering the scalar order one case: The reachability matrix of a backward pair $(\bar{A}_b, \bar{B}_b), \bar{A}_b \neq 0$ and $\bar{B}_b = (b_1, b_2)$ is

$$R = [\bar{A}_b^{T-1}\bar{B}_b, \ldots, \bar{A}_b\bar{B}_b, \bar{B}_b]$$

and the corresponding states are given by

$$x_t = \sum_{j=1}^{t} \bar{A}_b^{T-j}\bar{B}_b\bar{y}_{t-j} = b_1 \sum_{j=1}^{t-1} \bar{A}_b^{T-j} y_{t-j} + b_2\bar{A}_b^{T-t}, t = 1, 2, \ldots, T.$$

Therefore we obtain a simple backward recursion

$$x_T = b_1 \sum_{j=1}^{T-1} \bar{A}_b^{T-j} y_{T-j} + b_2$$

$$x_t = \bar{A}_b x_{t+1} - b_1\bar{A}_b^T y_t \text{ for } t = T-1, T-2, \ldots, 1.$$

Recent and astonishing findings

Now set $b_1 = -\bar{A}_b^{-T}$ and $b_2 = y_T - b_1 \sum_{j=1}^{T-1} \bar{A}_b^{T-j} y_{T-j}$. Then the above recursion gives:

$$x_T = y_T$$

$$x_t = \bar{A}_b x_{t+1} + y_t = y_t + \bar{A}_b y_{t+1} + \ldots + \bar{A}_b^{T-t} y_T, \;\; t = T-1, \ldots, 1$$

and in the limit $\bar{A}_b \to 0$ one gets a *perfect prediction*, since $x_t \to y_t$. Hence the negative log likelihood goes to minus infinity! It follows that the maximum likelihood estimator *does not exist* in the original model (i.e. in the space without the additional points), nor in the extended space (i.e. with the additional points) and this is true whatever the data sequence observed is.

Recent and astonishing findings

- To approach the infimum (=minus infinity) one approximately encodes in b_2 *all the data observed*! However the corresponding sequence of Markov parameters $\bar{C}\bar{A}^{j-1}\bar{B},\ j = 1, \ldots, T$ gets extremely large. And the corresponding model is (of course) completely useless in practice, as the slightest perturbation in the data will give extremely large prediction errors.
- Using our regularization this infimum will not be reached, as the relaxation forces the Markov parameters to be bounded!
- However the choice of the regularization parameter at times can have a large effect on the outcome which is unsatisfactory.

Recent and astonishing findings

- Theoretically one should perhaps penalize the amount of information encoded in the parameters. Usually we measure that by the dimension of the parameter space. Here one however encodes an approximation of the whole data sequence in one scalar parameter so the usual measure of information appears to break down here!

Concluding remarks and further research

- The method described to find the global optimum can be combined with gradient search. In this way one may speed up the procedure.
- Practical implementation of the methods suggested may benefit from parallel computing.
- It is likely that efficiency gains can be made by working further on the practical implementation of such algorithms.
- Having a good heuristic estimate can also speed up the procedure, as it may allow for quick elimination of a lot of points in such an extensive search algorithm (namely if their criterion function values are bad compared to the one found with the heuristic method)

Concluding remarks and further research

- The global optimization algorithms are likely to be computationally intensive and may become (too) slow for larger problems.
- Global optimization algorithms can be used to check whether a given heuristic method produces the global optimum or not in smaller problems. This may help in fine-tuning such a heuristic method.

Concluding remarks and further research

- The method has brought to light a weakness in the concept of maximum likelihood for the class of models considered, as the negative log likelihood always has infimum minus infinity. How to handle this will require further thought. Using the regularization avoids the problem but may at times make the answer very dependent on the precise choice of the regularization parameter which is an unsatisfactory situation.

Thank You

Thank you![1]

[1]With thanks to TUWien and INRIA for additional financial support for this project

References1

[1] Thomas P. Wihler (2009) On the Hölder Continuity for Matrix Functions for Normal Matrices, Journal of Inequalities in Pure and Applied Mathematics (JIPAM), vol. 10(4):5, 2009.

[2] Ralf Peeters, Martine Olivi, Bernard Hanzon (2009) Balanced realization of lossless systems: Schur parameters, canonical forms and applications 15th IFAC Symposium on System Identification Saint-Malo , pp.273-283

[3] Bernard Hanzon, Martine Olivi, Ralf Peeters (2009) Subdiagonal pivot structures and associated canonical forms under state isometries 15th IFAC Symposium on System Identification, St Malo, pp.1620-1625

References2

[1] Th. Ribarits, M. Deistler and B. Hanzon, An analysis of separable least squares data driven local co-ordinates for maximum likelihood estimation of linear systems (2005), Automatica, Special Issue on Data-Based Modeling and System Identification, 41(3).

[2] M. Deistler, B. Pötscher, The Behaviour of the Likelihood Function for ARMA Models, Advances in Applied Probability, vol. 16, no. 43 (Dec. 1984), pp. 843–866.

[3] E.A. Galperin, The Cubic Algorithm for Optimization and Control, NP Research Publ., Montreal, Canada, 1990.

References3

[1] B. Hanzon, M. Olivi and R.L.M. Peeters, Balanced Realizations of Discrete-Time Stable All-Pass Systems and the Tangential Schur Algorithm, Linear Algebra and Its Applications, vol. 418, 793-820, 2006.

[2] Bernard Hanzon and Ralf L.M. Peeters, Balanced Parametrizations of Stable SISO All-Pass Systems in Discrete Time, Math. Control Signals Systems, vol.13, 2000, pp.240-276.