Divergence-based spectral approximation with degree constraint as a concave optimization problem

Abstract

The Kullback-Leibler pseudo-distance, or divergence, can be used as a criterion for spectral approximation. Unfortunately this criterion is not convex over the most general classes of rational spectra. In this work it will be shown that divergence minimization is equivalent to a costrained entropy minimization problem, whose concave structure can be exploited in order to guarantee global convergence in the most general case.

D.1 Introduction

Divergence-based spectral approximation has deep connections with some of the most important estimators of ARMA spectral parameters. In fact it can be shown under mild conditions, that minima with respect to divergence tend to maximum-likelihood solutions as the data length grows to infinity. Furthermore prediction error and divergence are indeed equivalent criterions up to a normalization factor [3]. Although divergence and prediction error can be shown to be convex criterions over some restricted classes of spectra, in general they are not. Therefore convergence to a global minimum can not, in general, be guaranteed.

The purpose of this work is to address the global convergence problem when the minimization is performed over the most general classes of ARMA spectra. In order to do so it will be shown that divergence minimization is equivalent to a costrained, minimum entropy problem of special structure, namely concave minimization [1].

Concavity is not a property as desirable as convexity, in fact it can be shown that even simple concave problems are NP-hard, but it provides nonetheless useful properties. Many different algorithms exists in the literature that are tailored to solve concave problems. Ultimately it has been shown that these methods indeed converge to a global minimum [2].

The paper is outlined as follows. In section II the Spectral Approximation through Divergence is presented with particular focus on the parameterization of the chosen class of spectra to be considered. In section III an equivalent optimization problem with a concave structure will be introduced.

D.2 Spectral approximation

Given a spectral density Φ, spectral approximation amounts to finding a simpler spectral density Ψ that is close enough to the original one. Here the concepts of closeness and simplicity have to be given a precise meaning.

D.2.1 Parameterization of rational spectra of limited degree

One very important class of spectral densities to be considered is the one of rational spectra that corresponds to purely non deterministic processes (p.n.d). Within this class a common notion of complexity is the MacMillan degree. Therefore one may want to search for the approximation in the class of spectra of degree no greater than a selected n, a class that will be denoted by $\mathcal{F}_n$.

A way to better handle the set $\mathcal{F}_n$ is to parameterize it – i.e find a surjective map $h : X \to \mathcal{F}_n$, such that X is a convex set. In the following discussion two different such parameterizations will be introduced that have additional properties.

Let $\Psi \in \mathcal{F}_n$, and W be any of its minimum-phase spectral factors. Since Ψ is a spectrum of a p.n.d process, W is analytic outside the open unit disc and can be expressed as

$$W(z) = \frac{M(z)}{a(z)} = \frac{M_0 z^n + M_1 z^{n-1} + \dots M_n}{a_0 z^n + a_1 z^{n-1} + \dots a_n}$$

where $a_k \in \mathbb{R}$, $M_k \in \mathbb{R}^{m\times m}$ and $a(z)$ has no zeros in the unit circle $\mathbb{T}$.

Therefore, for any $\Psi \in \mathcal{F}_n$, there exist M_k and q_k such that

$$\Psi(z) = \frac{M(z)M(z^{-1})^T}{q(z)} \triangleq \check{\Psi}_z(\mathbf{M}, \mathbf{q}), \tag{D.1}$$

where

$$q(z) = a(z)a(z^{-1}) = \sum_{k=0}^{n} q_k \frac{z^k + z^{-k}}{2} \tag{D.2}$$

is strictly positive on the unit circle, $\mathbf{q} = [q_0, q_1 \dots q_n]^T$ and $\mathbf{M}$ is a compact notation for the collection of M_k, $k = 1 \dots n$.

Equation (D.1) defines a map $\check{\Psi} : \mathcal{S}_n \to \mathcal{F}_n$, where

$$\mathcal{S}_n \triangleq \{\mathbf{M}, \mathbf{q} : q(z) > 0, \forall z \in \mathbb{T}\},$$

sending $(\mathbf{M}, \mathbf{q})$ to the corresponding spectral density. Here $\mathcal{S}_n$ is a convex set while $\check{\Psi}$ is clearly surjective.

Proposition D.2.1 *For every fixed $z \in \mathbb{T}$ the map $\check{\Psi}_z$ defined in (D.1) is a function $\mathcal{S}_n \to \mathbb{S}^m$, where $\mathbb{S}^m$ is the set of symmetric matrices of size m, and is matrix convex on $\mathcal{S}_n$.*

Proof. The first statement follows directly from the fact (D.1) maps onto $\mathcal{F}_n$ and every spectrum therein will have no poles on the unit circle.

As can be seen in the appendix, $\check{\Psi}_z$ is matrix convex if and only if, for any $v \in \mathbb{R}^m$, $v^T \check{\Psi}_z v$ is convex. We have

$$\begin{aligned} v^T \check{\Psi}_z v &= \frac{\left[M(z)^\dagger v\right]^\dagger \left[M(z)^\dagger v\right]}{q(z)} = \sum_{j=0}^{m} \frac{\left|\left[M(z)^\dagger v\right]_j\right|^2}{q(z)} \\ &= \sum_{j=0}^{m} \left\{ \frac{\left\{Re\left[M(z)^\dagger v\right]_j\right\}^2}{q(z)} + \frac{\left\{Im\left[M(z)^\dagger v\right]_j\right\}^2}{q(z)} \right\} \end{aligned}$$

where $A^\dagger$ indicates the Hermitian of A. But

$$\left[M(z)^\dagger v\right]_j = \sum_{k=0}^{n} \left[M_k^T v\right]_j z^{-(n-k)}$$

so

$$Re\left[M(z)^\dagger v\right]_j = \sum_{k=0}^{n} \left[M_k^T v\right]_j Re[z^{-(n-k)}]$$

and

$$Im\left[M(z)^\dagger v\right]_j = \sum_{k=0}^{n} \left[M_k^T v\right]_j Im[z^{-(n-k)}],$$

in addition to $q(z)$, are linear function of the parameters.

It suffice now to note that $f(x, y) = \frac{x^2}{y}$ is convex in $\mathbb{R} \times \mathbb{R}_{++}$ as

$$\nabla^2 f(x, y) = \frac{2}{y^3} \begin{bmatrix} y \\ -x \end{bmatrix} \begin{bmatrix} y \\ -x \end{bmatrix}^T \geq 0$$

and, for every j

$$\frac{\left\{Re\left[M(z)^\dagger v\right]_j\right\}^2}{q(z)} = f \circ I_{Re}^j(\mathbf{M}, \mathbf{q}),$$

$$\frac{\left\{Im\left[M(z)^{\dagger}v\right]_j\right\}^2}{q(z)} = f \circ I^j_{Im}(\mathbf{M}, \mathbf{q})$$

where I^j_{Re} and I^j_{Im} are maps of obvious definition and, as noted above, also linear. Furthermore $q(z) > 0$ on the unit circle so both I^j_{Re} and I^j_{Im} map $\mathcal{S}_n$ onto $\mathbb{R} \times \mathbb{R}_{++}$. This assures each composite function is convex on $\mathcal{S}_n$ for every $z \in \mathbb{T}$, as long as their sum: $v^T \check{\Psi}_z v$. □

The importance of matrix convexity of a parameterization is that it may be used along with composition rules to eventually prove the convexity of an approximating criterion.

On the other hand it may happen that matrix concavity is a much more desirable property for the parameterization of $\mathcal{F}_n$.

Similarly to $\check{\Psi}$, let us introduce the map

$$\hat{\Psi}_z(c, \mathbf{B}, \mathbf{q}) \triangleq cI - \frac{B(z)B(z^{-1})^T}{q(z)}, \tag{D.3}$$

where $q(z)$ can be written as (D.2) and

$$B(z) = B_0 z^n + B_1 z^{n-1} + \dots B_n. \tag{D.4}$$

Similarly as above, $\mathbf{B}$ is a compact notation for the collection of $B_k \in \mathbb{R}^{m \times m}$. Consider its restriction to the set

$$\mathcal{H}_n \triangleq \{c, \mathbf{B}, \mathbf{q} : q(z) > 0, \Psi_z > 0, \forall z \in \mathbb{T}\} \tag{D.5}$$

that is therefore mapped by (D.3) onto $\mathcal{F}_n$.

Proposition D.2.2 *The set $\mathcal{H}_n$ defined in (D.5) is convex. Furthermore, for every fixed $z \in \mathbb{T}$, the map $\hat{\Psi}_z$ defined in (D.3) is a function $\mathcal{H}_n \to \mathbb{S}^n$ and it is matrix concave on $\mathcal{H}_n$.*

Proof. Consider

$$\hat{\Psi}_z(c, \mathbf{B}, \mathbf{q}) = cI - \check{\Psi}_z(\mathbf{B}, \mathbf{q})$$

as a function on the convex domain $\mathbb{R} \times \mathcal{S}_n$. From Proposition D.2.1 follows that $\hat{\Psi}_z$ takes values on $\mathbb{S}^m$ for every $z \in \mathbb{T}$ and is matrix concave on $\mathbb{R} \times \mathcal{S}_n$.

As $\mathcal{H}_n \subset \mathbb{R} \times \mathcal{S}_n$ the proposition follows. □

For (D.3) to be a parameterization it remains to show the following

Proposition D.2.3 *The map $\hat{\Psi} : \mathcal{H}_n \to \mathcal{F}_n$ defined in (D.3) is surjective.*

Proof. Any $\Psi \in \mathcal{F}_n$ can be factorized as

$$\Psi = \frac{M(z)M(z^{-1})^T}{a(z)a(z^{-1})}$$

where $M(z)$ and $a(z)$ are polynomials of order not greater than n. Furthermore $a(z)a(z^{-1})$ is positive on $\mathbb{T}$ for every spectrum in $\mathcal{F}_n$ so $\mathbf{q}$ can be chosen such that $q(z) = a(z)a(z^{-1})$. Additionally it is also a pseudopolynomial of order not greater than n. Furthermore, as $q(z) > 0$ on the unit circle, one can always choose a c big enough such that

$$cq(z)I - M(z)M(z^{-1})^T \geq 0 \tag{D.6}$$

on $\mathbb{T}$ and, as (D.6) is a pseudopolynomial matrix of order not greater than n, it can be factorized by a polynomial matrix $B(z)$ of the form (D.4) of the same order. So

$$cq(z)I - M(z)M(z^{-1})^T = B(z)B(z^{-1})^T$$

and it follows directly that with such $(c, \mathbf{B}, \mathbf{q}) \in \mathcal{H}_n$ we have

$$\Psi = \frac{M(z)M(z^{-1})^T}{a(z)a(z^{-1})} = cI - \frac{B(z)B(z^{-1})^T}{q(z)} = \hat{\Psi}_z(c, \mathbf{B}, \mathbf{q}).$$

□

D.2.2 Spectral approximation with divergence

The concept of closeness usually is defined by the choice of a suitable (pseudo) distance between spectra as it will be done in the following discussion.

Definition D.2.1 *Let x_1 and x_2 random variables on $\mathbb{R}^m$ with probability densities p_1 and p_2 respectively. The Kullback-Leibler divergence of x_2 from x_1 is*

$$\mathbf{D}(x_1||x_2) = \int_{\mathbb{R}^n} p_1(v) \log \frac{p_1(v)}{p_2(v)} dv \tag{D.7}$$

The Kullback-Leibler divergence can be seen as a pseudo-distance between random variables as $\mathbf{D}(x_1||x_2) \geq 0$ where the equality holds if and only if $p_1 = p_2$ almost everywhere. It is worth noting it is not a distance as it does not satisfy the triangle inequality, hence the use of the term *divergence*.

One can generalize the divergence of random variables to infer a corresponding concept for random processes as in the following.

Definition D.2.2 *Let x and y be discrete-time, jointly stationary processes on $\mathbb{R}^m$. The Kullback-Leibler divergence rate of y from x is*

$$\mathbf{D}(x||y) = \limsup_{N\to\infty} \frac{1}{N} \mathbf{D}(x_N||y_N) \tag{D.8}$$

where x_N and y_N are any windows of lenght N of x and y respectively.

The divergence rate is a widely used and accepted tool to infer how 'close' a random process is from another.

Furthermore, in the case of Gaussian processes, Stoorvogel and van Shuppen in [5] proved the following

Theorem D.2.1 *Let x and y be discrete-time, jointly stationary Gaussian processes with zero mean. Assume they admit spectral densities Φ and Ψ respectively, then*

$$\mathbf{D}(x||y) = \frac{1}{2}\int_{-\pi}^{\pi}\{tr[(\Phi-\Psi)\Psi^{-1}] - \log\det(\Phi\Psi^{-1})\}\frac{d\theta}{2\pi} \tag{D.9}$$

where, in the integral, Φ and Ψ are evaluated at $z = e^{i\theta}$.

Therefore the divergence rate can be used as a criterion for approximating a given spectrum Φ, that is assumed to be strictly positive, by a minimization with respect to Ψ over a suitable set of 'simple' spectra such as $\mathcal{F}_n$.

For compactness of notation it is better to consider the functional

$$J_\Phi(\Psi) = \int_{-\pi}^{\pi}\{tr(\Phi\Psi^{-1}) + \log\det\Psi\}\frac{d\theta}{2\pi} \tag{D.10}$$

and, as

$$\mathbf{D}(x||y) = \frac{1}{2}J_\Phi(\Psi) - \frac{1}{2}m - \frac{1}{2}\int_{-\pi}^{\pi}\log\det(\Phi)\frac{d\theta}{2\pi} \tag{D.11}$$

minimizing (D.10) is equivalent to minimizing (D.9).

It is worth noting that (D.10) can be reformulated as a convex functional. It suffice to consider it as a functional with respect to $Q = \Psi^{-1}$

$$\tilde{J}(Q) = J_\Phi(Q^{-1}) = \int_{-\pi}^{\pi}\{tr(\Phi Q) - \log\det Q\}\frac{d\theta}{2\pi}$$

and it is easy to see that it is the sum of two convex functionals. The two problems are equivalent as we are considering only spectra of p.n.d processes so that they are always invertible. This approach was followed in [4] and was combined with a linear parameterization that would lead to a convex problem. Unfortunately the model class $\mathcal{F}_n$ presented above can not be mapped completely by such linear parameterization, since the zeros are fixed by the choice of a finite number of basis functions.

To sum it up, convexity, the most desirable property, can not be achieved without sacrificing the generality of the model class.

D.3 Concave spectral approximation

The next step is to introduce an equivalent optimization problem that, while keeping the desired generality on the model class, will manifest an exploitable structure in combination with the parameterization (D.3).

D.3.1 An equivalent optimization problem

Let $\mathcal{X}$ be a class of spectral functions $\mathbb{C} \to \mathbb{S}^m$ that are strictly positive definite on the unit circle. Suppose also that $\mathcal{X}$ is closed with respect to the outer product with a positive real number – i.e if $\Psi \in \mathcal{X}$ then $\alpha\Psi \in \mathcal{X}$ for all $\alpha > 0$. The following holds

Proposition D.3.1 *Let $\Psi^* \in \mathcal{X}$ be any optimum of the problem*

$$\min_{\Psi \in \mathcal{X}} \int_{-\pi}^{\pi} tr\left[\Phi(e^{i\theta})\Psi^{-1}(e^{i\theta})\right] + \log\det\Psi(e^{i\theta})\frac{d\theta}{2\pi} \tag{D.12}$$

then

$$\int_{-\pi}^{\pi} tr\left\{\Phi(e^{i\theta})\left[\Psi^*(e^{i\theta})\right]^{-1}\right\}\frac{d\theta}{2\pi} = m. \tag{D.13}$$

Proof. Suppose the optimum Ψ^* does not satisfy the condition (D.13)

$$\int_{-\pi}^{\pi} tr[\Phi(\Psi^*)^{-1}]\frac{d\theta}{2\pi} \neq m.$$

Consider the objective function of (D.12) evaluated at $\alpha^{-1}\Psi^*$ for $\alpha > 0$

$$\begin{aligned} J(\alpha) &= J_\Phi(\alpha^{-1}\Psi^*) \\ &= \int_{-\pi}^{\pi} \{tr[\Phi(\alpha^{-1}\Psi^*)^{-1}] + \log\det(\alpha^{-1}\Psi^*)\}\frac{d\theta}{2\pi} \\ &= \alpha\int_{-\pi}^{\pi} tr[\Phi(\Psi^*)^{-1}]\frac{d\theta}{2\pi} - m\log\alpha + const \end{aligned} \tag{D.14}$$

The function (D.14) is strictly convex in $\mathbb{R}_{++}$ and as such, if it admits a stationary point α^*, α^* is its unique global minimum. In this case one obtains

$$0 = \left.\frac{dJ}{d\alpha}\right|_{\alpha=\alpha^*} = \int_{-\pi}^{\pi} tr[\Phi(\Psi^*)^{-1}]\frac{d\theta}{2\pi} - m(\alpha^*)^{-1}$$

so α^* such that

$$\alpha^*\int_{-\pi}^{\pi} tr[\Phi(\Psi^*)^{-1}]\frac{d\theta}{2\pi} = m$$

is the unique global minimum of (D.14).

It follows that $\alpha^* \neq 1$ and as such, by the strict convexity of (D.14), $J(\alpha^*) < J(1)$ i.e

$$J_\Phi((\alpha^*)^{-1}\Psi^*) < J_\Phi(\Psi^*).$$

Finally, by assumption on $\mathcal{X}$, $(\alpha^*)^{-1}\Psi^* \in \mathcal{X}$ hence Ψ^* can't be an optimum of (D.12).

□

A direct consequence of Proposition D.3.1 is that the problem (D.12) is equivalent to

$$\begin{aligned} &\min_{\Psi\in\mathcal{X}} \int_{-\pi}^{\pi} \log\det\Psi(e^{i\theta})\frac{d\theta}{2\pi} \\ &s.t \int_{-\pi}^{\pi} tr\left[\Phi(e^{i\theta})\Psi^{-1}(e^{i\theta})\right]\frac{d\theta}{2\pi} = m \end{aligned} \tag{D.15}$$

With similar arguments of proposition (D.3.1) one can prove the following

Proposition D.3.2 *Let $\Psi^* \in \mathcal{X}$ be any optimum of the problem*

$$\begin{aligned} &\min_{\Psi\in\mathcal{X}} \int_{-\pi}^{\pi} \log\det\Psi(e^{i\theta})\frac{d\theta}{2\pi} \\ &s.t \int_{-\pi}^{\pi} tr\left[\Phi(e^{i\theta})\Psi^{-1}(e^{i\theta})\right]\frac{d\theta}{2\pi} \leq m \end{aligned} \tag{D.16}$$

then

$$\int_{-\pi}^{\pi} tr\left\{\Phi(e^{i\theta})\left[\Psi^*(e^{i\theta})\right]^{-1}\right\}\frac{d\theta}{2\pi} = m. \tag{D.17}$$

Proof. Let us proceed analogously to the proof of (D.3.1) and suppose the optimum Ψ^* does not satisfy (D.17),in this case we have

$$\int_{-\pi}^{\pi} tr\left[\Phi(\Psi^*)^{-1}\right]\frac{d\theta}{2\pi} < m$$

Furthermore, let us consider the scalar optimization problem derived from (D.16) by the composition with $\Psi(\alpha) = \alpha^{-1}\Psi^*$ for $\alpha \in \mathbb{R}_{++}$:

$$\begin{aligned} &\min_{\mathbb{R}_{++}} \int_{-\pi}^{\pi} \log\det\alpha^{-1}\Psi^*\frac{d\theta}{2\pi} \\ &s.t \int_{-\pi}^{\pi} tr\left[\Phi(\alpha^{-1}\Psi^*)^{-1}\right]\frac{d\theta}{2\pi} \leq m \end{aligned}$$

that becomes

$$\begin{aligned} &\min_{\mathbb{R}_{++}} -m\log\alpha + const = J(\alpha) \\ &s.t \quad \alpha\int_{-\pi}^{\pi} tr\left[\Phi(\Psi^*)^{-1}\right]\frac{d\theta}{2\pi} \leq m \end{aligned} \tag{D.18}$$

As the objective function of (D.18) is strictly decreasing in α and

$$\int_{-\pi}^{\pi} tr\left[\Phi(\Psi^*)^{-1}\right]\frac{d\theta}{2\pi} > 0,$$

the optimum α^* will be such that

$$\alpha^* \int_{-\pi}^{\pi} \left[\Phi(\Psi^*)^{-1}\right] \frac{d\theta}{2\pi} = m.$$

Thus $\alpha^* > 1$, $J(\alpha^*) < J(1)$ and, as $(\alpha^*)^{-1}\Psi^* \in \mathcal{X}$, Ψ^* can't be an optimum of (D.16). □

Proposition D.3.1 and D.3.1 combined show that the problems (D.12) and (D.16) are equivalent i.e they have the same optima if any. In other words, divergence minimization is equivalent to entropy minimization under a normalization constraint.

Moreover one can find a lower bound for both problems. By the non-negativeness of the Divergence rate and from (D.11) follows

$$J_\Phi(\Psi) \geq m + \int_{-\pi}^{\pi} \log\det\Phi \frac{d\theta}{2\pi} \tag{D.19}$$

for every $\Psi \in \mathcal{X}$. The bound holds also for any optimum Ψ^* of (D.12) and using Proposition D.3.1 we obtain

$$\int_{-\pi}^{\pi} \log\det\Psi^* \frac{d\theta}{2\pi} \geq \int_{-\pi}^{\pi} \log\det\Phi \frac{d\theta}{2\pi}$$

but, by the equivalence of the two problems, Ψ^* is a global minimum for both. Therefore for every feasible Ψ of (D.16) the following lower bound also holds

$$\int_{-\pi}^{\pi} \log\det\Psi \frac{d\theta}{2\pi} \geq \int_{-\pi}^{\pi} \log\det\Phi \frac{d\theta}{2\pi}.$$

Thus the normalization constraint bounds the entropy of the solution being greater than the one of the target spectrum.

D.3.2 A concave formulation

Finally the parameterization (D.3) can be applied with the equivalent formulation (D.16) in order to show the concave nature of the problem at hand.

Proposition D.3.3 *The following problem, where $\hat{\Psi}$ and $\mathcal{H}_n$ are defined in (D.3) and (D.5) respectively*

$$\begin{aligned} &\min_{\mathcal{H}_n} \int_{-\pi}^{\pi} \log\det\hat{\Psi}_{e^{i\theta}}(c, \mathbf{B}, \mathbf{q}) \frac{d\theta}{2\pi} \\ &s.t \int_{-\pi}^{\pi} tr\left[\Phi(e^{i\theta})\hat{\Psi}_{e^{i\theta}}(c, \mathbf{B}, \mathbf{q})^{-1}\right] \frac{d\theta}{2\pi} \leq m \end{aligned} \tag{D.20}$$

is a concave optimization problem, i.e the problem of minimizing a concave function over a convex set.

Proof. The extended-value extensions of $g_1(X) = -\log\det(X)$ is matrix non-increasing (see Appendix) also, for any $X \in \mathbb{S}^m_{++}$ and $Y \in \mathbb{S}^m$,

$$\begin{aligned}\frac{\partial^2}{\partial t^2}\left\{-\log\det(X+tY)\right\}\Big|_{t=0} &= tr\left[YX^{-1}YX^{-1}\right] \\ &= tr\left[Z^2\right] \geq 0\end{aligned}$$

where $Z = YX^{-1}$ so g_1 is also convex on $\mathbb{S}^m_{++}$. By the composition rules in the Appendix and the Proposition D.2.2 the function $-\log\det\hat{\Psi}_z$ is convex for any $z \in \mathbb{T}$. The integration preserves convexity.

Similarly, the extended-value extensions of $g_2(A, X) = tr[AX^{-1}]$ for any $A \succeq 0$ is matrix non-increasing with respect to X and, for any $X \in \mathbb{S}^m_{++}$ and $Y \in \mathbb{S}^m$,

$$\begin{aligned}\frac{\partial^2}{\partial t^2} tr[A(X+tY)^{-1}]\Big|_{t=0} &= tr\left[AX^{-1}YX^{-1}YX^{-1}\right] \\ &= tr\left[AX^{-1}Z^2\right] \geq 0\end{aligned}$$

where Z is as above so g_2 is convex with respect to X on $\mathbb{S}^m_{++}$. Finally, by the composition rules in the Appendix and the Proposition D.2.2, the function $tr[\Phi(z)\hat{\Psi}_z^{-1}]$ is convex for any $z \in \mathbb{T}$ and the integral is still convex. □

The concave structure of the problem can be exploited in branch-and-bound techniques in defining efficient bounding procedures. An outline of such an algorithm is given in the Appendix. It must be stressed that such an algorithm is guaranteed to converge to a *global* optimum.

D.4 Closing remarks

In this paper a possible method for solving divergence-based spectral approximation problems in the most general model class was introduced. It is based on the solution of an equivalent, concave, minimum entropy problem. Furthermore several algorithms exists in the literature that are guaranteed to converge to a global optimum.

At this point the main contribution of this work is theoretical insight, namely understanding the structure of the problem. Unfortunately, to the author knowledge, there does not exist any concave minimization software available in the public domain. Thus the proposed method's performance remains to be verified in practice.

Appendix

Matrix monotonicity

Let us consider the set of symmetric matrices $\mathbb{S}^m \subset \mathbb{R}^{m\times m}$ and let $\mathbb{S}^m_+, \mathbb{S}^m_{++} \subset \mathbb{S}^m$ be the sets of positive semidefinite and positive definite matrices respectively.

The set $\mathbb{S}^m_+$ is a proper cone and let $\preceq$ be its associated generalized inequality - i.e for $X, Y \in \mathbb{S}^m$, $X \preceq Y$ iff $Y - X \in \mathbb{S}^m_+$.

A function $g : \mathbf{dom}(g) \to \mathbb{R}$ where $\mathbf{dom}(g) \subseteq \mathbb{S}^m$ is called *matrix non-decreasing* if for any $X, Y \in \mathbf{dom}(g)$

$$X \preceq Y \Longrightarrow g(X) \leq g(Y)$$

and *matrix increasing* if

$$X \preceq Y, X \neq Y \Longrightarrow g(X) < g(Y).$$

Similarly one can define *matrix non-increasing* and *matrix decreasing* functions.

Two examples will now be presented that are instrumental to the proof of Proposition D.3.2

Example D.5.1 *The function $tr(X^{-1})$ is matrix non-increasing on $\mathbb{S}^m_{++}$. In fact for any $X, Y \in \mathbb{S}^m_{++}$ such that $X \succeq Y$, they admit inverse and $Y^{-1} \succeq X^{-1}$ i.e there exists a $Z \in \mathbb{S}^m_+$ such that $X^{-1} = Y^{-1} - Z$ so*

$$tr(X^{-1}) = tr(Y^{-1} - Z) = tr(Y^{-1}) - tr(Z) \leq tr(Y^{-1}).$$

The same can be shown for the function $tr(AX^{-1})$ for any $A \succeq 0$.

□

Example D.5.2 *The function $\log\det X$ is matrix non-decreasing on $\mathbb{S}^m_{++}$. In fact for any $X, Y \in \mathbb{S}^m_{++}$ such that $X \succeq Y$, there exists a $Z \in \mathbb{S}^m_+$ such that $X = Y + Z$, and hence*

$$\begin{aligned}
\log\det X &= \log\det(Y + Z) \\
&= \log\det\left[Y^{1/2}\left(I + Y^{-1/2}ZY^{-1/2}\right)Y^{1/2}\right] \\
&= \log\det Y + \log\det\left(I + Y^{-1/2}ZY^{-1/2}\right) \\
&= \log\det Y + \sum \log(1 + \lambda_i),
\end{aligned}$$

where λ_i are the eigenvalues of $Y^{-1/2}ZY^{-1/2}$, and as this matrix is positive semidefinite, $\lambda_i \geq 0$ and

$$\log\det X \geq \log\det Y.$$

Of course the function $-\log\det X$ is matrix non-increasing on $\mathbb{S}^m_{++}$.

□

Matrix convexity

Let us consider the function $h : \mathcal{X} \to \mathbb{S}^m$ where $\mathcal{X} \subseteq \mathbb{R}^p$ is a convex set. h is said to be *matrix convex* if

$$h(\theta x + (1-\theta)y) \preceq \theta h(x) + (1-\theta)h(y)$$

for any $x, y \in \mathcal{X}$ and $\theta \in [0,1]$. Furthermore it is said to be *strictly matrix convex* if

$$h(\theta x + (1-\theta)y) \prec \theta h(x) + (1-\theta)h(y)$$

for any $x, y \in \mathcal{X}$, $x \neq y$ and $\theta \in (0,1)$. Similarly the function h is said to be *(strictly) matrix concave* if $-h$ is (strictly) matrix convex.

It can be shown that the function h as above is (strictly) matrix convex/concave iff for any non-zero $v \in \mathbb{R}^m$ the scalar function $v^T h(x) v$ is (strictly) convex/concave.

Composition rules

Let's define the *extended-value extension* $\tilde{g}$ of the real-valued function g as:

$$\tilde{g}(x) = \begin{cases} g(x) & if \quad x \in \mathbf{dom}(g) \\ +\infty & otherwise \end{cases}$$

that is defined in the whole $\mathbb{R}^p$.

It can be shown that $\tilde{g}$ is matrix non-increasing if g is and $\mathbf{dom}(g) = \mathbf{dom}(g) + \mathbb{S}^m_+$. In fact for any $X, Y \in \mathbb{S}^m$ such that $X \succeq Y$ if $X, Y \in \mathbf{dom}(g)$

$$\tilde{g}(X) = g(X) \leq g(Y) = \tilde{g}(Y).$$

On the other hand, either $Y \notin \mathbf{dom}(g)$ and $\tilde{g}(X) \leq +\infty = \tilde{g}(Y)$, or $X \notin \mathbf{dom}(g)$. In the latter case $X = Y + Z$ with $Z \succeq 0$ so $Y \notin \mathbf{dom}(g)$ otherwise $X \in \mathbf{dom}(g)$ too as $\mathbf{dom}(g) = \mathbf{dom}(g) + \mathbb{S}^m_+$.

Also the converse is true. Obviously if $\tilde{g}$ is matrix non-increasing so is g. Furthermore for any $Y \in \mathbf{dom}(g)$ and $Z \succeq 0$, $X = Y + Z \succeq Y$ thus $\tilde{g}(X) \leq \tilde{g}(Y)$. Here $X \in \mathbf{dom}(g)$, otherwise $\tilde{g}(X) = +\infty$ but this would imply $\tilde{g}(Y) = +\infty$ also and $Y \notin \mathbf{dom}(g)$. So $\mathbf{dom}(g) = \mathbf{dom}(g) + \mathbb{S}^m_+$. In a similar way one can show that $\tilde{g}$ is matrix non-decreasing iff g is and $\mathbf{dom}(g) = \mathbf{dom}(g) - \mathbb{S}^m_+$.

Example D.5.3 *Both* $-\log\det X$ *and* $tr(X^{-1})$ *are matrix non-increasing on their domain* $\mathbb{S}^m_{++}$. *Also* $\mathbb{S}^m_{++} + \mathbb{S}^m_+ = \mathbb{S}^m_{++}$ *so their extended-value extensions that are non-increasing. The same holds for* $tr(AX^{-1})$ *for every* $A \succeq 0$.

□

Given the functions $g : \mathbf{dom}(g) \to \mathbb{R}$ and $h : \mathcal{X} \to \mathbb{S}^m$ the composite function $f = g \circ h$ whose domain is defined as

$$\mathbf{dom}(f) = \{x \in \mathcal{X} : h(x) \in \mathbf{dom}(g)\}$$

has the following properties:

- if g is convex, $\tilde{g}$ is matrix non-decreasing and h is matrix convex $\Longrightarrow$ f is convex
- if g is convex, $\tilde{g}$ is matrix non-increasing and h is matrix concave $\Longrightarrow$ f is convex
- if g is concave, $\tilde{g}$ is matrix non-decreasing and h is matrix concave $\Longrightarrow$ f is concave
- if g is concave, $\tilde{g}$ is matrix non-increasing and h is matrix convex $\Longrightarrow$ f is concave

A short proof will be given to the second rule, the one used in Proposition D.3.2. The first step is to prove that $\mathbf{dom}(f)$ is convex, for doing so consider $x, y \in \mathbf{dom}(f)$ and $\theta \in [0,1]$. As $x, y \in \mathbf{dom}(h)$ and $h(x), h(y) \in \mathbf{dom}(g)$ then, by convexity, $\theta h(x) + (1-\theta)h(y) \in \mathbf{dom}(g)$. Also by matrix concavity of h

$$h(\theta x + (1-\theta)y) \succeq \theta h(x) + (1-\theta)h(y)$$

so, for some $Z \succeq 0$

$$h(\theta x + (1-\theta)y) = \theta h(x) + (1-\theta)h(y) + Z$$

thus $h(\theta x + (1-\theta)y) \in \mathbf{dom}(g)$ and therefore $\mathbf{dom}(f)$ is convex as stated.

Now, as $\tilde{g}$ is matrix non-increasing

$$g(h(\theta x + (1-\theta)y)) \leq g(\theta h(x) + (1-\theta)h(y))$$

and by convexity of g

$$g(\theta h(x) + (1-\theta)h(y)) \leq \theta g(h(x)) + (1-\theta)g(h(y)).$$

The last two inequalities combined show the convexity of $f = g \circ h$ on its domain.

Overview of a concave minimization algorithm

A concave minimization problem can be introduced in the following way

$$\inf_{x \in \mathcal{D}} f(x) \tag{D.21}$$

where $\mathcal{D} \subset \mathbb{R}^n$ is a closed, convex set with non-empty interior and $f(.)$ is a concave funcion over $\mathcal{D}$. It will be assumed that the origin belongs to the interior of $\mathcal{D}$ - i.e $\forall i : g_i(O) < 0$, this can always be guaranteed by an appropriate translation. An overview of the algorithm presented in [6] will now be presented.

At iteration k of the algorithm we assume to have at our disposal $\mathcal{M}_k$, a collection of convex polyhedral cones vertexed at O and having n edges whose union is guaranteed to contain an optimum. Each cone $M \in \mathcal{M}_k$ is generated by n vectors

$s_i^M \in \mathbb{R}^n$. In addition, we assume a lower bound $\mu(M) = \inf\{f(x) : x \in \mathcal{D} \cap M\}$ is known for each cone. Finally we have x_k the candidate optimal point and its corresponding objective value $\gamma_k = f(x_k)$.

At iteration $k = 0$ the algorithm is initialized as follows. Let $\mathcal{M}_0$ be any collection of cones whose union contains the feasible set $\mathcal{D}$. For each of these cones the lower bound is calculated by a *bounding rule* to be specified. The candidate point x_0 and its corresponding value γ_0 is chosen through the following minimization

$$\inf\left\{f(x) : x = \alpha s_i^M \in \mathcal{D}, \alpha \geq 0, M \in \mathcal{M}_0, i = 1, \dots n\right\}.$$

At iteration k the algorithm proceeds as follows. Let

$$\mathcal{R}_k = \{M \in \mathcal{M}_k : \mu(M) < \gamma_k\} \tag{D.22}$$

If $\mathcal{R}_k = \emptyset$ the algorithm ends as x_k is an optimal solution.

Otherwise select by a *selection rule* to be specified a cone $\bar{M} \in \mathcal{R}_k$ and split it by a *splitting rule* to be specified into two cones $\bar{M}_1$, $\bar{M}_2$. Additionally compute the lower bounds, $\mu(\bar{M}_1)$ and $\mu(\bar{M}_2)$ respectively, by the same bounding rule as above. Finally select the new candidate x_{k+1} point and its corresponding value by comparing the current to

$$\bar{x} = \arg\inf\left\{f(x) : x = \alpha s_i^{\bar{M}_j} \in \mathcal{D}, \alpha \geq 0, j = 1, 2\right\} \tag{D.23}$$

Set $\gamma_{k+1} = f(x_{k+1})$, $\mathcal{M}_{k+1} = (\mathcal{R}_k \setminus \{\bar{M}\}) \cup \{\bar{M}_1, \bar{M}_2\}$.

Compute next iteration $k + 1$.

The above described algorithm can either terminate after finitely many steps or continue indefinitely. In the latter case, convergence of the above algorithm depends on the choice of the individual rules, for selecting, splitting and bounding cones. In order to guarantee convergence we need to require certain properties from these rules, namely

Theorem D.5.1 *Whenever the algorithm generates an infinite number of steps and if the following conditions hold*

- *The selection rule is* ultimately complete*, i.e*

$$\inf\left\{f(x) : x \in \mathcal{D} \bigcap \left(\bigcup_{p=1}^{\infty} \bigcap_{k=p}^{\infty} \mathcal{R}_k\right)\right\} \geq \gamma^* \tag{D.24}$$

  as $\gamma^* = \lim_{k\to\infty} \gamma_k$.

- *The splitting rule is* exhaustive*, i.e for any infinite decreasing sequence of cones generated by the algorithm, its intersection is an half-line.*

- *The bounding rule is* consistent*, i.e*

$$\lim_{h\to\infty} [\mu(M_h) - \gamma_{k_h}] = 0. \tag{D.25}$$

for any infinite decreasing sequence of cones $M_h \in \mathcal{R}_{k_h}$ whose intersection is a half-line non entirely contained in $\mathcal{D}$.

then as $k \to \infty$

$$f(x_k) = \gamma_k \searrow \inf\{f(x) : x \in \mathcal{D}\}. \tag{D.26}$$

The proof of the theorem can be found in [6] along some simple choices of rules that satisfy (D.5.1).

D.6 Bibliography

[1] H.P. Benson. Concave minimization: Theory, applications and algorithms, in: Handbook of global optimization, 1995.

[2] R. Horst and H. Tuy. *Global Optimization: Deterministic Approaches*. Springer-Verlag, 1996.

[3] A. Lindquist and G. Picci. *Linear Stochastic Systems: A Geometric Approach to Modeling, Estimation and Identification*. yet to be published.

[4] Anders Lindquist. Prediction-error approximation by convex optimization, in: Modeling, estimation and control: Festschrift in honor of giorgio picci on the occasion of his sixty-fifth, 2007.

[5] A. A. Stoorvogel and J.H. van Schuppen. System identification with information theoretic criteria, in: Identification, adaptation, learning, 1996.

[6] H. Tuy, T.V Thieu, and Q. Thai NG. A conical algorithm for globally minimizing a concave function over a closed convex set. *Mathematics of Operations Research*, 10(3), 1985.