

Matricial Wasserstein and Unsupervised Tracking

Lipeng Ning, Romeil Sandhu, Tryphon T. Georgiou, and Allen Tannenbaum

Abstract

The context of this work is spectral analysis of multivariable times-series as this may arise in processing signals originating in antenna and sensor arrays. The salient feature of these time signals is that they contain information about moving scatterers/targets which may not be known a priori. That is, neither the number nor the physical properties of scatterers may be known in advance, a fact which necessitates that analysis needs to be model free. Thus, what is important is to attain reliable and high resolution spectral estimates based on short-time observations due to the expected motion of objects within the scattering field.

Traditional spectral analysis methods such as spectrograms and maximum entropy, Capon, etc. techniques, are often severely constrained by the non-stationary nature of time-series, which necessitates very short observation records. Thus, our goal has been to develop natural regularization techniques that allow smooth interpolation of spectrograms in time, thereby improving resolution and robustness. Since power spectra are matrix-valued measure, we sought to develop geometric tools that are based on weak* continuous metrics, such as Wasserstein metrics, only for matrix-valued functions. The present work is largely based on [1] where such a theory was laid out.

Introduction

Traditional techniques in sensor arrays, detection, and estimation rely on periodogram-based methods, maximum-entropy techniques, or beamforming. While these are ubiquitous, they are severely limited when dealing with non-stationary time-series. Assimilation of data from disparate sources and dealing with systemic biases are quite challenging on their own, and they are even more so, when dealing with non-stationary time-series; the non-stationary nature of the signal content necessitates that spectral analysis is based on short observation records.

Our approach has been to seek natural ways to quantify distance between spectra, and to develop geometric tools based on that. More specifically, power spectra and probability distributions can be viewed as points on a suitable manifold. Then, slowly time-varying power spectra can be viewed as flows (geodesics) on this manifold, and interpolation/extrapolation of power spectral estimates as well as

L. Ning is with Brigham and Women's Hospital, Harvard Medical School, Psychiatry Neuroimaging Laboratory 1249 Boylston St. Boston, MA 02421; email: lning@bwh.harvard.edu

T.T. Georgiou is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, Minnesota MN 55455, USA; email: tryphon@umn.edu

R. Sandhu and A. Tannenbaum are with the Departments of Computer Science and Applied Mathematics/Statistics, Stony Brook University, Stony Brook, NY 11794. email: arobertan@gmail.com

The research was partially supported by the AFOSR under Grants FA9550-12-1-0319 and FA9550-15-1-0045.

uncertainty quantification can all be carried out within the associated metric topology. A desirable feature of metrics that would make them suitable for applications is weak* continuity; this is the property that small changes in a distribution affects measurements in a continuous manner. Interesting, L_2 , L_1 , Kullback-Leibler divergence and several many other popular notions of distance fail in this regard. To this end we sought to generalize the notion of *Wasserstein distance*, which is natural in this respect for scalar distributions.

Our interest in matrix valued measures stems from the fact that those represent power spectra of vector-valued time-series. In turn, vector-valued time-series may represent measurements of different modalities across a distributed array of sensors that reflect frequency/color, polarization, spatial characteristics, and other attributes that are thought to characterize target properties. Thus, we are interested in a “transport-based geometry” for such matrix-valued distributions as well as a “transport theory” that is flexible with regard to the preservation of mass/power across time. Advances on this front make it possible to tackle in a natural way smoothing and interpolation between inconsistent data sets or time-varying characteristics of a series as well as the computation of optical flow for object tracking in series of frames. Besides the relevance of geodesics as a tool for modeling, tracking, morphing, data assimilation/association, etc., for power spectra and images alike, we expect to advance concepts of resolution and the quantification of uncertainty in rigorous terms.

Background

Our topic, to define and explore Wasserstein-like metrics and the corresponding geometry for matrix-valued densities or measures, is an attempt to build on the classical subject of optimal mass transport (OMT). This subject has in recent years witnessed a fast developing phase with many applications in physics, probability theory, economics, etc. Standard references are [2], [3], [4] and some of the most significant recent developments that in particular have inspired our work are traced to [5], [6], [7].

Example

We wish to exemplify the spirit of our approach by a simple example, which would be quite challenging (unless, one uses hindsight and tailors a parametric method to the task). The task is to track slowly time-varying sinusoidal signals in noise that represent the echo from a pair of targets, whose position changes with regard to an antenna array at the same time. We wish to emphasize that we seek *non-parametric* techniques, and although our example will be treated parametrically (since a model is known), that in general would not be the case. Hence, our interest in non-parametric techniques.

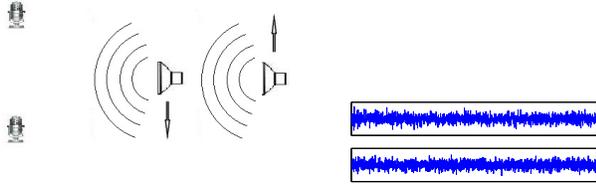


Fig. 1: Sensors and sources – correlated time-series

Consider two sources of sound (equivalently, scatterers in an antenna array’s field of view) as shown in Figure 1 moving relative to each other in opposite directions and, exchanging positions in the process, relative to the pair of stationary microphones. The emitted sounds are recorded by each microphone in considerable amount of ambient noise. Their respective frequencies and intensities vary due to Doppler shift and due to the change in their relative proximity to the two sensors. Computer simulated signals are also shown in Figure 1. The task is to distinguish the relative position of the two sources using frequency analysis of the recorded time-series.

The time-series is vector-valued (having two entries = number of sensors/channels). Thus, the power spectrum is matrix-valued (2×2 in this case). Short time maximum entropy reconstruction of the power spectrum is shown in Figure 2. Our convention is to show in the (1,1)-subplot the spectrogram of the first sensor, the (2,2)-subplot that of the second sensor, in the (1,2)-subplot the absolute value of the cross spectrum and in the (2,1)-subplot its phase. Next, in Figure 3 we display, following the same convention, the regularized spectrogram where we used optimal mass transport geometry and interpolated the time-distributions in Figure 2 by an OMT geodesic.

At each frequency and time, either plot represents the (color-coded) intensity of a Hermitian matrix, namely, the value of the power spectral matrix density. (The (1,1) and (2,2) entries are real, while the (1,2) and (2,1) are complex conjugate of each other and their real part and phase are displayed accordingly as indicated above.) Singular value decomposition reveals the directionality of the incoming energy. Thus, by identifying at each point in time the maxima of the power, we determine the corresponding singular vectors which reflect the relative portion of power in each sensor for the corresponding power source, thereby, pointing to its relative position (after suitable calibration). Figures 4 and 5 show the paths that the corresponding singular vectors traced over time, based on the maximum entropy spectrogram of Figure 2 and that of the OMT geodesic reconstruction in Figure 3, respectively. What is especially revealing is the dramatic improvement in resolution and consistency afforded by the use of OMT geodesics. To some degree, this is to be expected since the OMT geometry induces a natural (weakly continuous) metric where the approximation takes place and, automatically, the optimal path determined.

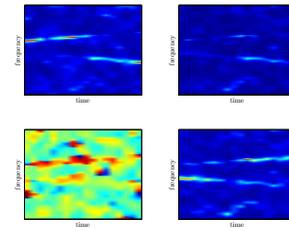


Fig. 2: Maximum entropy matrix-spectrogram

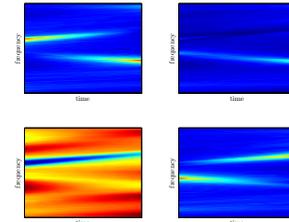


Fig. 3: OMT geodesic regularized matrix-spectrogram

In more detail, the schematics in Figures 6 and 7 explain how the paths of singular vectors corresponding to maxima of spectral power were obtained from the matricial spectrograms. Frequencies of maximal power are identified by computing maxima of the trace of the matricial power spectrum. We select the two frequencies where the power has local maxima and compute the corresponding singular vector of the (2×2 matrix) value of the power spectrum f . These are shown in Figures 6 and 7 by red and green color, respectively. Note that at each of the two frequencies there are two singular vectors (since f is 2×2). The one corresponding to the largest singular value is drawn in continuous line and the other in dashed, and scaled to reflect the size of the corresponding singular value. A “cleaner” view of the path that the “top” singular vectors traverse is given in Figures 4 and 5, respectively, as noted earlier.

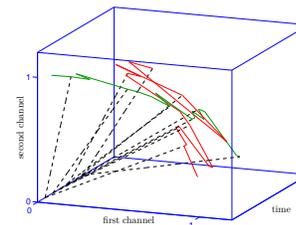


Fig. 4: Path of singular vectors of main harmonics (ME spectrogram), i.e., relative power in each sensor.

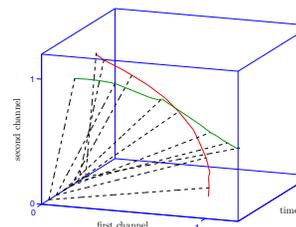


Fig. 5: Path of singular vectors of main harmonics (OMT spectrogram), i.e., relative power in each sensor.

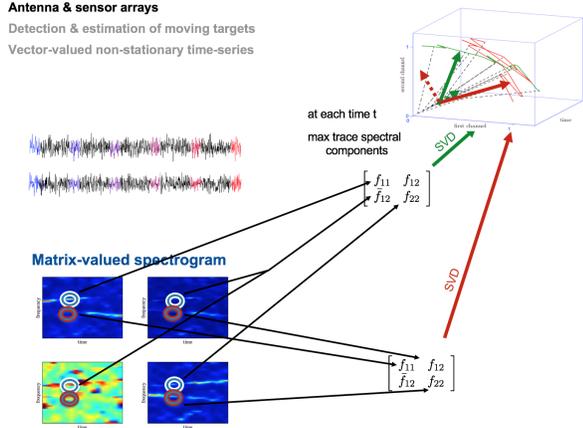


Fig. 6: Singular vectors corresponding to peak power vs. time based on ME-spectrogram

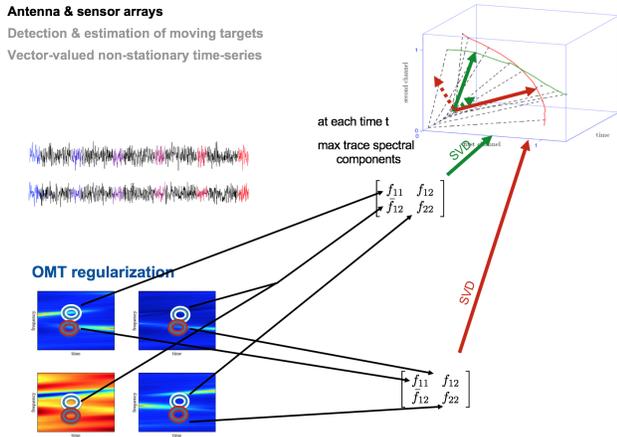


Fig. 7: Singular vectors corresponding to peak power vs. time based on OMT-spectrogram

We provide additional intuition on how mass corresponds between the two end-point marginals in an optimal mass transport framework by considering the matrix-valued power spectra as in Figure 8. The two end-point spectra are also shown in Figure 9. These represent spectra and cross spectra for measurement in two channels and, therefore, the power spectral density is a 2×2 matrix-valued function. The geodesic in Figure 8 produces a drift of the power between the two channels. Non-parametric techniques of interpolation typically suffer from a “push-pop” phenomenon instead of shifting gracefully the power between channels. It is instructive to note how the frequency of the dominant harmonic shifts from one channel to the other in Figure 8.

Rudiments of matrix-valued Wasserstein

The scalar OMT theory has been adapted in [8] to model slowly time-varying changes in power spectra of time series and has been used for statistical estimation, data assimilation, and morphing. While in scalar time-series, the power content may drift over time across frequencies (e.g., when

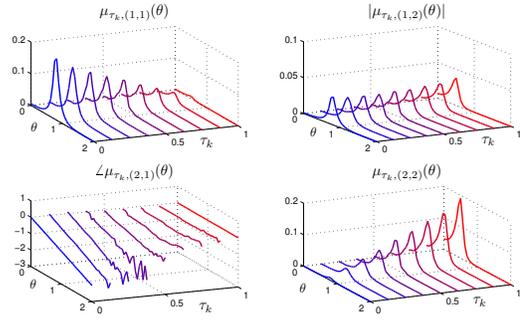


Fig. 8: Geodesic path “morphing” between two multivariable power spectra.

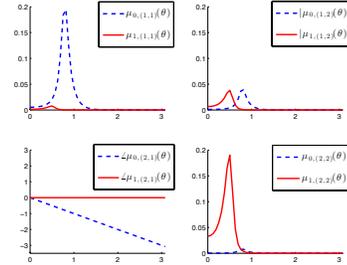


Fig. 9: End-point power spectra.

considering Doppler effects, echolocation of a moving target, etc.), in vector-valued time series the power spectral content may shift principle directions as well. In fact, such a rotation of the power-spectral content is typical in general antenna-arrays when a scatterer changes position with respect to array elements. Therefore, a concept of transport between matrix-valued densities requires that we take into account both, the cost of shifting power across frequencies as well as the cost of rotating the corresponding principle axes. We now briefly discuss such a “non-commutative” Monge-Kantorovich transportation and a corresponding metric.

Consider two probability density functions μ_0 and μ_1 supported on \mathbb{R} and let $\mathcal{M}(\mu_0, \mu_1)$ be the set of probability measures m on $\mathbb{R} \times \mathbb{R}$ with μ_0 and μ_1 as marginals, i.e.

$$\int_{\mathbb{R}} m(x, y) dy = \mu_0(x), \quad \int_{\mathbb{R}} m(x, y) dx = \mu_1(y), \quad m \geq 0.$$

Probability densities are thought of as distributions of mass and the optimal mass transport problem is to determine

$$\mathcal{T}_c(\mu_0, \mu_1) := \inf_{m \in \mathcal{M}(\mu_0, \mu_1)} \int_{\mathbb{R} \times \mathbb{R}} c(x, y) m(x, y) dx dy, \quad (1)$$

where $c(x, y)$ is the cost of transporting one unit of mass from location x to y . In particular, when $c(x, y) = |x - y|^2$, the optimal cost gives rise to the 2-Wasserstein metric $W_2(\mu_0, \mu_1) = \sqrt{\mathcal{T}_2(\mu_0, \mu_1)}$ where

$$\mathcal{T}_2(\mu_0, \mu_1) := \inf_{m \in \mathcal{M}(\mu_0, \mu_1)} \int_{\mathbb{R} \times \mathbb{R}} |x - y|^2 m(x, y) dx dy. \quad (2)$$

Consider now the family of matrix-valued functions

$$\mathcal{F} := \left\{ \boldsymbol{\mu}(\cdot) \mid \text{for } x \in \mathbb{R}, \boldsymbol{\mu}(x)^* = \boldsymbol{\mu}(x) \in \mathbb{C}^{n \times n}, \right. \\ \left. \boldsymbol{\mu}(x) \geq 0, \text{tr} \left(\int_{\mathbb{R}} \boldsymbol{\mu}(x) dx \right) = 1 \right\}.$$

These are Hermitian, positive semi-definite matrix-valued functions on \mathbb{R} normalized so that their trace integrates to 1. They will be referred to as matrix-valued densities and can be thought of as a generalization of probability density functions. The scalar-valued $\text{tr}(\boldsymbol{\mu})$ represents mass at location x . Thus, all elements in \mathcal{F} have the same total mass over the support. In order to define a suitable generalization of Kantorovich formulation of OMT to matrix-valued densities we need a “joint” measure to live in a considerably bigger space - a tensor product which is discussed next.

Consider a $n^2 \times n^2$ (positive definite) matrix $\boldsymbol{\mu}$ as an element in $\mathcal{L}(\mathcal{H}_0 \otimes \mathcal{H}_1)$, where \mathcal{L} represents the space of linear operators, and both \mathcal{H}_0 and \mathcal{H}_1 are, for our purposes simply both identified with \mathbb{C}^n . The partial traces $\text{tr}_{\mathcal{H}_0}$ and $\text{tr}_{\mathcal{H}_1}$, or tr_0 and tr_1 for brevity, are linear maps

$$\begin{aligned} \text{tr}_1 : \mathcal{L}(\mathcal{H}_0 \otimes \mathcal{H}_1) &\rightarrow \mathcal{L}(\mathcal{H}_0) : \boldsymbol{\mu} \mapsto \text{tr}_1(\boldsymbol{\mu}) \\ \text{tr}_0 : \mathcal{L}(\mathcal{H}_0 \otimes \mathcal{H}_1) &\rightarrow \mathcal{L}(\mathcal{H}_1) : \boldsymbol{\mu} \mapsto \text{tr}_0(\boldsymbol{\mu}) \end{aligned}$$

defined uniquely by the property that on simple products they act as follows:

$$\text{tr}_1(\boldsymbol{\mu}_0 \otimes \boldsymbol{\mu}_1) = \text{tr}(\boldsymbol{\mu}_1)\boldsymbol{\mu}_0 \text{ and } \text{tr}_0(\boldsymbol{\mu}_0 \otimes \boldsymbol{\mu}_1) = \text{tr}(\boldsymbol{\mu}_0)\boldsymbol{\mu}_1$$

for any $\boldsymbol{\mu}_0 \in \mathcal{L}(\mathcal{H}_0)$ and $\boldsymbol{\mu}_1 \in \mathcal{L}(\mathcal{H}_1)$. Alternatively, $\boldsymbol{\mu} \in \mathcal{L}(\mathcal{H}_0 \otimes \mathcal{H}_1)$ can be represented by a matrix $[\boldsymbol{\mu}_{ik,\ell m}]$ of size $n^2 \times n^2$ as it maps a basis element $u_i \otimes v_k \in \mathcal{H}_0 \otimes \mathcal{H}_1$ to $\sum_{\ell,m} \boldsymbol{\mu}_{ik,\ell m} u_\ell \otimes v_m$. Then, the partial trace e.g., $\text{tr}_1(\boldsymbol{\mu})$ is the represented by the $n \times n$ matrix with (i, ℓ) -th entry $\sum_k \boldsymbol{\mu}_{ik,\ell k}$, for $1 \leq i, \ell \leq n$. Likewise the (k, m) -th entry of $\text{tr}_0(\boldsymbol{\mu})$ is $\sum_i \boldsymbol{\mu}_{ik,im}$, for $1 \leq k, m \leq n$. See [9] for the significance of partial trace in the context of quantum mechanics.

Consider now two matrix-valued density functions $\boldsymbol{\mu}_0, \boldsymbol{\mu}_1 \in \mathcal{F}$. We seek

$$\boldsymbol{m}(x, y) \text{ a } n^2 \times n^2 \text{ positive semi-definite matrix,} \quad (3a)$$

for (x, y) in the specified support (e.g., $\mathbb{R} \times \mathbb{R}$, or $[0, 2\pi]^2$), such that

$$\boldsymbol{m}_0(x, y) := \text{tr}_1(\boldsymbol{m}(x, y)), \quad \boldsymbol{m}_1(x, y) := \text{tr}_0(\boldsymbol{m}(x, y)), \quad (3b)$$

$$\int_{\mathbb{R}} \boldsymbol{m}_0(x, y) dy = \boldsymbol{\mu}_0(x), \quad \int_{\mathbb{R}} \boldsymbol{m}_1(x, y) dx = \boldsymbol{\mu}_1(y), \quad (3c)$$

and we denote

$$\mathcal{M}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1) := \left\{ \boldsymbol{m} \mid (3a) - (3c) \text{ are satisfied} \right\}.$$

A suitable transportation cost can be defined as a functional on a joint density in $\mathcal{M}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)$, just as in the scalar case. However, in contrast to the scalar case, besides penalizing transport of mass between two points x and y , we need to

impose a penalty on a corresponding rotation as well (see [1] for more discussion).

We first consider a scalar “mass transference” cost

$$\min_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)} \int_{\mathbb{R} \times \mathbb{R}} c(x, y) \text{tr}(\boldsymbol{m}(x, y)) dx dy. \quad (4)$$

This coincides with the optimal transportation cost between scalar-valued densities $\text{tr}(\boldsymbol{\mu}_0)$ and $\text{tr}(\boldsymbol{\mu}_1)$. Thus, if $\text{tr}(\boldsymbol{\mu}_0(x)) = \text{tr}(\boldsymbol{\mu}_1(x))$, the optimal value of (4) is zero since it reduces to optimal transport between identical scalar marginals. Thus, (4) fails to quantify mismatch of directionality between the given matrix-valued marginals. A term that penalizes directionality mismatch is introduced next.

Assume that the marginals are positive definite pointwise. Define the *normalized partial traces*

$$\begin{aligned} \underline{\text{tr}}_0(\boldsymbol{m}(x, y)) &:= \text{tr}_0(\boldsymbol{m}(x, y)) / \text{tr}(\boldsymbol{m}(x, y)) \\ \underline{\text{tr}}_1(\boldsymbol{m}(x, y)) &:= \text{tr}_1(\boldsymbol{m}(x, y)) / \text{tr}(\boldsymbol{m}(x, y)). \end{aligned}$$

Their difference captures the directional mismatch between the two partial traces, and then

$$\text{tr}(\|(\underline{\text{tr}}_0 - \underline{\text{tr}}_1)\boldsymbol{m}(x, y)\|_{\mathbb{F}}^2 \boldsymbol{m}(x, y))$$

quantifies the rotational mismatch, and thus we arrive at the cost functional

$$\text{tr} \left((c(x, y) + \lambda \|(\underline{\text{tr}}_0 - \underline{\text{tr}}_1)\boldsymbol{m}(x, y)\|_{\mathbb{F}}^2) \boldsymbol{m}(x, y) \right),$$

with $\lambda > 0$, to weigh in the relative significance of the linear and rotational penalties, and define as our Wasserstein matricial distance,

$$\min_{\boldsymbol{m} \in \mathcal{M}(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1)} \int_{\mathbb{R} \times \mathbb{R}} \text{tr} \left((c + \lambda \|(\underline{\text{tr}}_0 - \underline{\text{tr}}_1)\boldsymbol{m}\|_{\mathbb{F}}^2) \boldsymbol{m} \right) dx dy$$

with $c(x, y) = (x - y)^2$, which in fact turns out to be a convex optimization problem. Details are provided in [1].

REFERENCES

- [1] L. Ning, T. T. Georgiou, and A. Tannenbaum, “On matrix-valued monge–kantorovich optimal mass transport,” *Automatic Control, IEEE Transactions on*, vol. 60, no. 2, pp. 373–382, 2015.
- [2] C. Villani, *Topics in optimal transportation*. American Mathematical Society, 2003, vol. 58.
- [3] L. Ambrosio, “Lecture notes on optimal transport problems,” *Mathematical aspects of evolving interfaces*, pp. 1–52, 2003.
- [4] S. Rachev and L. Rüschendorf, *Mass Transportation Problems: Theory*. Springer Verlag, 1998, vol. 1.
- [5] J. Benamou and Y. Brenier, “A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem,” *Numerische Mathematik*, vol. 84, no. 3, pp. 375–393, 2000.
- [6] R. J. McCann, “A convexity principle for interacting gases,” *Advances in mathematics*, vol. 128, no. 1, pp. 153–179, 1997.
- [7] W. Gangbo and R. J. McCann, “The geometry of optimal transportation,” *Acta Mathematica*, vol. 177, no. 2, pp. 113–161, 1996.
- [8] X. Jiang, Z. Luo, and T. Georgiou, “Geometric methods for spectral analysis,” *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1064–1074, 2012.
- [9] D. Petz, *Quantum Information Theory and Quantum Statistics (Theoretical and Mathematical Physics)*. Berlin: Springer, 2008.