An approximate joint diagonalization algorithm for off-the-grid sparse and non-sparse recovery

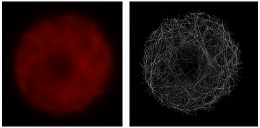
Paul Catala[1]. Joint work with J.-F. Cardoso[2], V. Duval[3] and G. Peyré[4]

GAMM 2022, Aachen, August 17 2022

[1]University of Osnabrück, [2]Institut d'Astrophysique de Paris, CNRS, [3]Inria Paris, [4]Ecole Normale Supérieure, PSL, CNRS

Motivations

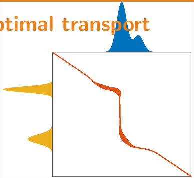
Super-resolution



source - www.cellimagelibrary.org

- recover signal μ from coarse, noisy measurements

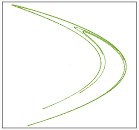
Optimal transport



Gaussian transport

- find optimal coupling between two probability distributions μ_1 and μ_2

Invariant measures

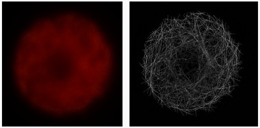


Hénon map

- find measure which is invariant under a given dynamics

A common framework

Super-resolution



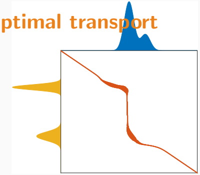
source - www.cellimagelibrary.org

- Unknown: $\mu \in \mathcal{M}_+(\mathbb{T}^d)$
- Given: $k \in \{-n, \dots, n\}^d$,

$$\hat{\mu}(k) = \int_{\mathbb{T}^d} e^{-2\imath\pi \langle k, x \rangle} \mathrm{d}\mu(x)$$

- Probl: retrieve μ from $(\hat{\mu}(k)) \in \mathbb{C}^N$

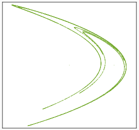
Optimal transport



Gaussian transport

- Unknown: $\gamma \in \mathcal{M}_+(\mathbb{T}^{2d})$

Invariant measures



Hénon map

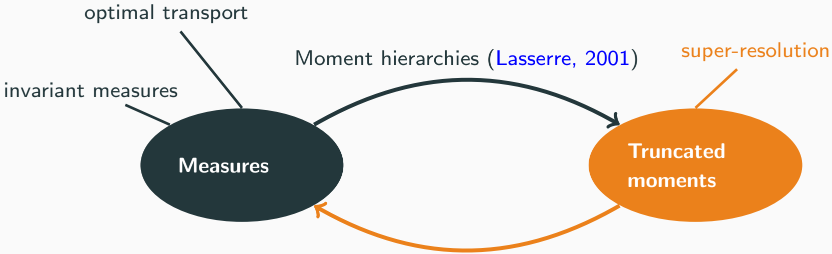
- Unknown: $\nu \in \mathcal{M}_+(\mathbb{T}^d)$

We want off-the-grid recovery algorithms (= no spatial discretization)

Roadmap

Prony (R. de Prony, 1795), ESPRIT (Roy et al., 1986), etc...
Polynomial approximations (Mhaskar, 2019),
Christoffel (Pauwels et al., 2020)

Roadmap

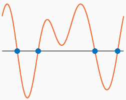


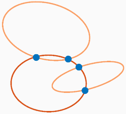
Prony (R. de Prony, 1795), ESPRIT (Roy et al., 1986), etc...
Polynomial approximations (Mhaskar, 2019),
Christoffel (Pauwels et al., 2020)

Preliminary: the sparse case

Support identification

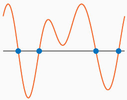
- Suppose $\mu = \sum_{k=1}^{r} \lambda_k \delta_{x_k}, \quad \lambda_k \geqslant 0, \quad x_k \in \mathbb{T}^d$.
- **Idea:** Encode $\operatorname{Supp} \mu = \mathcal{V}(\mathcal{I})$ for some ideal $\mathcal{I} \subset \mathcal{T}_n[x]$
  - $d = 1$, $\mathcal{I} = (p)$, Prony's method (R. de Prony, 1795)
  - $d > 1$, $\mathcal{I} = (p_1, \ldots, p_s)$, Stetter-Möller method (Möller and Stetter, 1995)

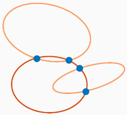




Support identification

- Suppose $\mu = \sum_{k=1}^{r} \lambda_k \delta_{x_k}, \quad \lambda_k \geqslant 0, \quad x_k \in \mathbb{T}^d$.
- **Idea:** Encode $\operatorname{Supp}\mu = \mathcal{V}(\mathcal{I})$ for some ideal $\mathcal{I} \subset \mathcal{T}_n[x]$
  - $d = 1$, $\mathcal{I} = (p)$, Prony's method (R. de Prony, 1795)
  - $d > 1$, $\mathcal{I} = (p_1, \ldots, p_s)$, Stetter-Möller method (Möller and Stetter, 1995)





- How do we find $\mathcal{I}$?

- Main ingredient: (truncated) moment matrix

$$T_n(\mu) \stackrel{\text{def.}}{=} (\hat{\mu}(k-l))_{k,l \in \{0,\ldots,n\}^d}$$

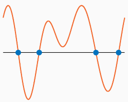
Rem. T_n is Toeplitz, and semidefinite positive since μ is nonnegative

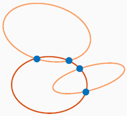
- **Theorem** (Kunis et al., 2016; Sauer, 2017).
If n is sufficiently large, then $\operatorname{Supp}\mu = \mathcal{V}((\operatorname{Ker} T_n))$

We identify a vector q to $q(x) = \sum q_k e^{-2\imath\pi\langle k, x\rangle}$

Support identification

- Suppose $\mu = \sum_{k=1}^{r} \lambda_k \delta_{x_k}, \quad \lambda_k \geqslant 0, \quad x_k \in \mathbb{T}^d$.
- **Idea:** Encode $\operatorname{Supp} \mu = \mathcal{V}(\mathcal{I})$ for some ideal $\mathcal{I} \subset \mathcal{T}_n[x]$
  - $d = 1$, $\mathcal{I} = (p)$, Prony's method (R. de Prony, 1795)
  - $d > 1$, $\mathcal{I} = (p_1, \ldots, p_s)$, Stetter-Möller method (Möller and Stetter, 1995)





- How do we find $\mathcal{I}$?

  - Main ingredient: (truncated) moment matrix

$$T_n(\mu) \overset{\text{def.}}{=} (\hat{\mu}(k-l))_{k,l \in \{0,\ldots,n\}^d}$$

Rem. T_n is Toeplitz, and semidefinite positive since μ is nonnegative

  - **Theorem** (Kunis et al., 2016; Sauer, 2017).
  If n is sufficiently large, then $\operatorname{Supp} \mu = \mathcal{V}((\operatorname{Ker} T_n))$

We identify a vector q to $q(x) = \sum q_k e^{-2\imath\pi \langle k, x \rangle}$

- A sufficient condition for "sufficiently large" is flatness (Curto and Fialkow, 1996)
  - $T_n (\succeq 0)$ is said to be flat if $\operatorname{rank} T_n = \operatorname{rank} T_{n-1}$.
  - Flatness $\Longrightarrow$ μ discrete

Multiplication matrices

- Let $\mathcal{I}_n \overset{\text{def.}}{=} (\text{Ker } T_n)$
- Computing $\mathcal{V}(\mathcal{I}_n)$ is fundamentally an eigenproblem (Stetter, 1996)

- **Definition.** The multiplication operators associated with T_n are

$$\chi_i : \begin{array}{lcl} \mathcal{T}[x]/\mathcal{I}_n & \to & \mathcal{T}[x]/\mathcal{I}_n \\ p(x) \pmod{\mathcal{I}_n} & \mapsto & e^{-2\imath\pi x_i} p(x) \pmod{\mathcal{I}_n} \end{array}$$

- **Proposition** (Laurent, 2010; Harmouch et al., 2017). Assume T_n is flat, of rank r, and let (U, Σ, U^*) be the singular value decomposition of T_{n-1}. Then in some basis

$$X_i = \Sigma^{-1} U^* T_{n-1}^{(i)} U \in \mathbb{C}^{r\times r}$$

where $T_{n-1}^{(i)}$ is the shifted matrix with entries $\hat{\mu}(k - l + e_i)$

- **Theorem** (Laurent, 2010). If T_n is flat, the matrices X_i are jointly diagonalizable: there exists $P \in \mathrm{GL}_r(\mathbb{C})$ such that

$$PX_iP^{-1} = \begin{pmatrix} e^{-2\imath\pi x_{1,i}} & & \\ & \ddots & \\ & & e^{-2\imath\pi x_{r,i}} \end{pmatrix}, \quad i = 1, \dots, r$$

Sparse recovery

Algorithm 1: Multivariate recovery for flat data

Input: T_n SDP, Toeplitz, flat matrix

Output: $x_1, \ldots, x_r \in \mathbb{T}^d$

```
for i = 1 to d do
    Compute shifted matrix T_{n-1}^{(i)}
    Compute svd T_{n-1} = U Σ U^*
    Compute multiplication matrices X_i = Σ^{-1} U^* T_{n-1}^{(i)} U
end
Compute joint diagonalization basis P
    *Diagonalize X_α = ∑ α_i X_i, for random α_i ∈ [0, 1]
Return x_{j,i} = -(1/2π) arg (P^{-1} X_i P)_{jj},   j = 1, ..., r,   i = 1, ..., d
```

* if the X_is are jointly diagonalizable, then with probability one X_α is non-derogatory (*i.e.* all eigenspaces are of dimension 1).

Approximate joint diagonalization

Non-sparse recovery

- If μ is not discrete, we essentially lose the flatness of T_n
- Guarantees of robustness in the non-flat case exist (Klep et al., 2018)
- What is the numerical perspective?

Algorithm 2: Multivariate recovery for ~~flat~~ data

Input: T_n SDP, Toeplitz, ~~flat~~ matrix

Output: $x_1, \ldots, x_r \in \mathbb{T}^d$

for $i = 1$ **to** d **do**

  Compute shifted matrix $T_{n-1}^{(i)}$

  Compute svd $T_{n-1} = U \Sigma U^*$

  Compute ~~multiplication~~ matrices $X_i = \Sigma^{-1} U^* T_{n-1}^{(i)} U$

end

Diagonalize $X_\alpha = \sum \alpha_i X_i$, for random $\alpha_i \in [0, 1]$

Return $x_{j,i} = -\frac{1}{2\pi} \arg\left(P^{-1} X_i P\right)_{jj}, \quad j = 1, \ldots, r, \quad i = 1, \ldots, d$

Non-sparse recovery

- If μ is not discrete, we essentially lose the flatness of T_n
- Guarantees of robustness in the non-flat case exist (Klep et al., 2018)
- What is the numerical perspective?

Algorithm 3: Multivariate recovery for ~~flat~~ data

Input: T_n SDP, Toeplitz, ~~flat~~ matrix

Output: $x_1, \ldots, x_r \in \mathbb{T}^d$

for $i = 1$ **to** d **do**

  Compute shifted matrix $T_{n-1}^{(i)}$

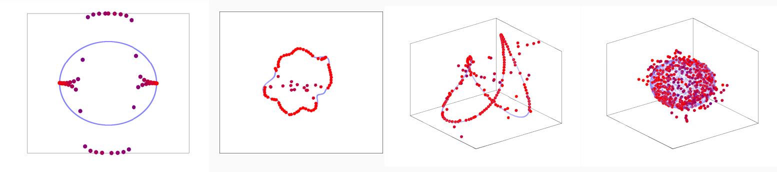
  Compute svd $T_{n-1} = U \Sigma U^*$

  Compute ~~multiplication~~ matrices $X_i = \Sigma^{-1} U^* T_{n-1}^{(i)} U$

end

Diagonalize $X_\alpha = \sum \alpha_i X_i$, for random $\alpha_i \in [0, 1]$

Return $x_{j,i} = -\frac{1}{2\pi} \arg\left(P^{-1} X_i P\right)_{jj}, \quad j = 1, \ldots, r, \quad i = 1, \ldots, d$



Non-sparse recovery

- If μ is not discrete, we essentially lose the flatness of T_n
- Guarantees of robustness in the non-flat case exist (Klep et al., 2018)
- What is the numerical perspective?

Algorithm 4: Multivariate recovery for ~~flat~~ data

Input: T_n SDP, Toeplitz, ~~flat~~ matrix

Output: $x_1, \ldots, x_r \in \mathbb{T}^d$

for $i = 1$ **to** d **do**

  Compute shifted matrix $T_{n-1}^{(i)}$

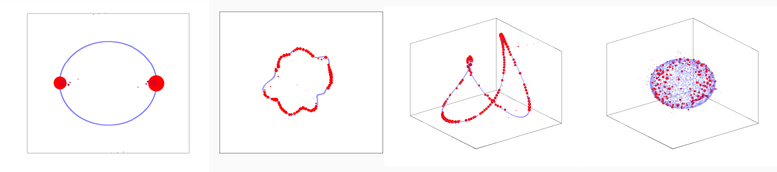
  Compute svd $T_{n-1} = U \Sigma U^*$

  Compute ~~multiplication~~ matrices $X_i = \Sigma^{-1} U^* T_{n-1}^{(i)} U$

end

Diagonalize $X_\alpha = \sum \alpha_i X_i$, for random $\alpha_i \in [0, 1]$

Return $x_{j,i} = -\frac{1}{2\pi} \arg\left(P^{-1} X_i P\right)_{jj}, \quad j = 1, \ldots, r, \quad i = 1, \ldots, d$



Diagonality Criterion

- X_i non-commuting, not jointly diagonalizable
  $\rightarrow$ find a basis in which they are "almost" diagonal
- Off-diagonal criterion to minimize

$$\mathcal{O}(P) \stackrel{\text{def.}}{=} \sum_i \sum_{\alpha \neq \beta} (P X_i P^{-1})^2_{\alpha\beta}$$

  - criterion used *e.g.* in (Cardoso and Souloumiac, 1996; Joho and Rahbar, 2002) for blind source separation, but restricted to orthogonal matrices
  - X_i are not Hermitian
  - Riemannian optimization over $\mathrm{GL}_r(\mathbb{C})$

Quasi-Newton updates

- Invertibility is maintained using updates of the form $P_{t+1} = (I_r + \mathcal{E})P_t$
- Taylor expansion: $\mathcal{O}((I + \mathcal{E})P) = \mathcal{O}(T) + \langle G(P), \mathcal{E}\rangle + \langle H(P)\mathcal{E}, \mathcal{E}\rangle + o(\|\mathcal{E}\|^2)$
  - Relative gradient: with $\underline{Y} = Y - \mathrm{Diag}(Y)$ and $Y_i = PX_iP^{-1}$

$$G(P) = \sum_i \underline{Y}_i Y_i^* - Y_i^* \underline{Y}_i$$

  - Relative Hessian: use diagonal approximation (Ablin et al., 2019). When Y_i are diagonal,

$$\tilde{H}_{pqrs}(P) = \delta_{pr}\delta_{qs} \sum_i |(Y_i)_{pp} - (Y_i)_{qq}|^2$$

$\rightarrow$ $\tilde{H}$ is sparse and positive semidefinite

- Quasi-Newton update: $P_{t+1} = (I + \alpha\mathcal{E}_t)P_t$, where α is found by linesearch and

$$\mathcal{E}_t = -(\tilde{H}(P_t) + \beta I)^{-1} \cdot G(P_t)$$

End-to-End Algorithm

Algorithm 5:

Input: $T_n \in \mathbb{C}^{N\times N}$ SDP, Toeplitz matrix, $P_0 = I_r$

for $i = 1$ **to** d **do**
    Compute svd $T_{n-1} = U\Sigma U^*$
    Compute matrices $X_i = \Sigma^{-1} U^* T_{n-1}^{(i)} U$
end
for $k = 0$ **to** $K-1$ **do**
    Compute $G(P_k; X_1, \dots, X_d)$ and $\tilde{H}(P_k; X_1, \dots, X_d)$
    Compute $\mathcal{E}_k = -(\tilde{H}(P_k) + \beta I)^{-1} \cdot G(P_k)$
    Backtracking $\min_\alpha \mathcal{O}((I + \alpha\mathcal{E}_k)P_k)$
    Update $P_{k+1} = (I + \alpha_k \mathcal{E}_k)P_k$
end
Return $x_{j,i} = -\frac{1}{2\pi} \arg (P_K^{-1} X_i P_K)_{jj},\ j = 1, \dots, r,\ i = 1, \dots, d$

Numerics

Algorithm 3:

Input: $T_n \in \mathbb{C}^{N\times N}$ SDP, Toeplitz matrix, $P_0 = I_r$

for $i = 1$ **to** d **do**

  Compute svd $T_{n-1} = U\Sigma U^*$: tolerance $\sigma_k \geqslant 10^{-3}\max(\sigma)$

  Compute matrices $X_i = \Sigma^{-1} U^* T^{(i)}_{n-1} U$

end

for $k = 0$ **to** $N-1$ **do**

  Compute $G(P_k; X_1, \dots, X_d)$ and $\tilde{H}(P_k; X_1, \dots, X_d)$

  Compute $\mathcal{E}_k = -(\tilde{H}(P_k) + \beta I)^{-1} \cdot G(P_k)$: β fixed

  Backtracking $\min_\alpha \mathcal{O}((I + \alpha\mathcal{E}_k)P_k)$: $\alpha \leftarrow \alpha/2$

  Update $P_{k+1} = (I + \alpha_k \mathcal{E}_k)P_k$

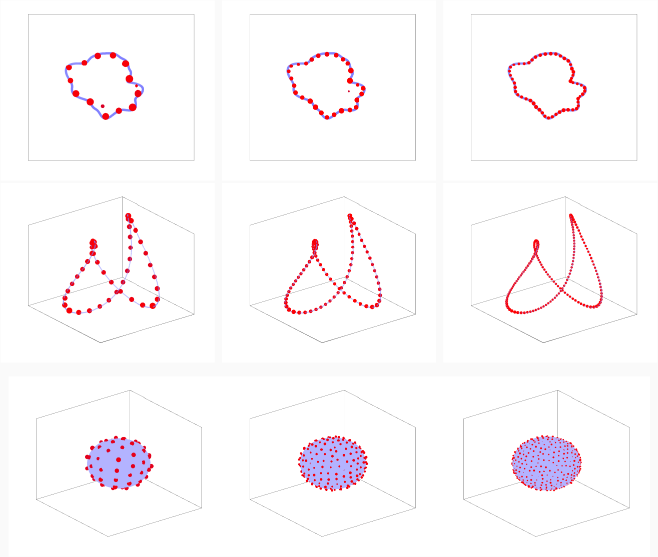
end

Return $x_{j,i} = -\frac{1}{2\pi}\arg(P_N^{-1} X_i P_N)_{jj},\ j = 1, \dots, r,\ \ i = 1, \dots, d$

Solve $V(x)\cdot a = c$, prune $a_j < 10^{-3}\max(|a|)$

Applications

Super-resolution

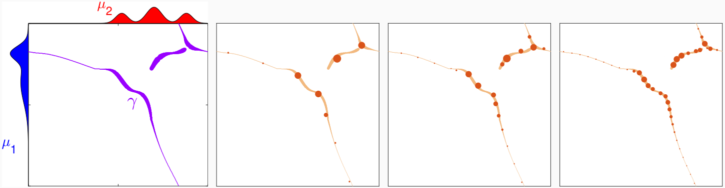


$n = 5$ $n = 10$ $n = 20$ (15 for sphere)

- We want to solve

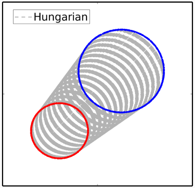
$$\min \int_{\mathbb{T}^d \times \mathbb{T}^d} c(x,y) \mathrm{d}\gamma(x,y) \quad \text{s.t.} \quad \gamma \in \Pi(\mu_1, \mu_2)$$

  - Using Lasserre's hierarchies (Lasserre, 2008), can be approximated by a SDP
  - perform the extraction on the resulting matrix

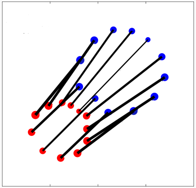


$n = 5$ $n = 10$ $n = 20$

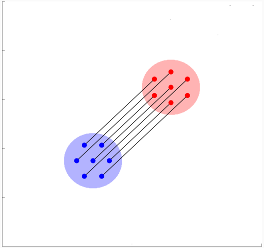
Optimal transport



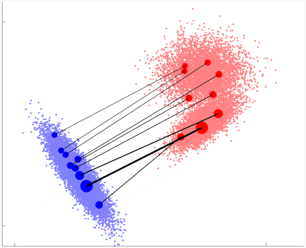
uniform distributions



$n = 10$



uniform distributions, $n = 10$

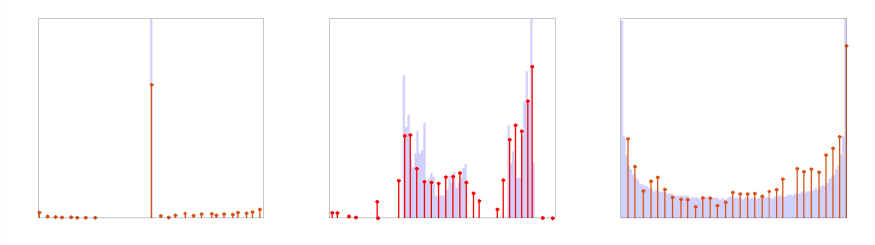


Gaussian mixtures, $n = 10$

Invariant measures

- Discrete-time system: given a system $x_{t+1} = f(x_t)$, we want to find a measure μ such that $f_\sharp \mu = \mu$
  - Logistic map, $x_{t+1} = rx_t(1 - x_t) =: f(x_t)$
  - We use Lasserre's hierarchies (Magron et al., 2019) to approximate

$$\min J(\mu) \quad \text{s.t.} \quad f_\sharp \mu = \mu \quad \text{and} \quad \mathbb{E}\mu = 1$$



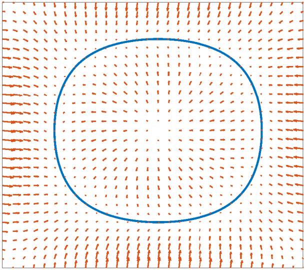
$r = 2$ $r = 3.6$ $r = 4$

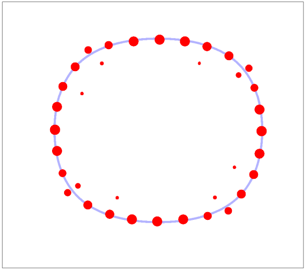
logistic map, $n = 30$

Invariant measures

- Dynamical system: given a system $\dot{x}(t) = u(x(t))$ we want to find a measure μ such that $\mathrm{div}(u\mu) = 0$
  - We consider the vector field u displayed below
  - We use Lasserre's hierarchies to approximate

$$\min J(\mu) \quad \text{s.t.} \quad \mathrm{div}(u\mu) = 0 \quad \text{and} \quad \mathbb{E}\mu = 1$$





Conclusion

- Joint diagonalization step in algebraic extraction is crucial
  $\rightarrow$ Dedicated solvers are key to make the procedure performant outside the theory
- Work hand in hand with Lasserre's hierarchies to define an integrated workflow
- application in optimal transport, invariant measures, . . .
- theoretical perspectives: convergence? geometrical interpretation?

Conclusion

- Joint diagonalization step in algebraic extraction is crucial
  $\rightarrow$ Dedicated solvers are key to make the procedure performant outside the theory
- Work hand in hand with Lasserre's hierarchies to define an integrated workflow
- application in optimal transport, invariant measures, ...
- theoretical perspectives: convergence? geometrical interpretation?

Thank you for your attention!

References

Ablin, P., Cardoso, J.-F., and Gramfort, A. (2019). Beyond pham's algorithm for joint diagonalization. In *ESANN*.

Cardoso, J. and Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM J. Mat. Anal. Appl.*, 17(1).

Curto, R. and Fialkow, L. (1996). Solution of the truncated complex moment problem for flat data. *Memoirs of the AMS*, (568).

Harmouch, J., Khalil, H., and Mourrain, B. (2017). Structured low rank decomposition of multivariate Hankel matrices. *Linear Algebra and its Applications*, 542:162 – 185.

Joho, M. and Rahbar, K. (2002). Joint diagonalization of correlation matrices by using newton methods with application to blind signal separation. In *IEEE SAM*, pages 403–407.

Klep, I., Povh, J., and Volčič, J. (2018). Minimizer extraction in polynomial optimization is robust. *SIAM J. Optim.*, 28(4):3177–3207.

Kunis, S., Peter, T., Römer, T., and Von der Ohe, U. (2016). A multivariate generalization of Prony's method. *Linear Algebra Appl.*, 490:31–47.

Lasserre, J. B. (2001). Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3):796–817.

Lasserre, J.-B. (2008). A semidefinite approach to the generalized problem of moments. *Math. Program.*, 112:65–92.

Laurent, M. (2010). Sums of squares, moment matrices and optimization over polynomials. In *Emerging Applications of Algebraic Geometry*, volume 149. Springer new York.

Magron, V., Forets, M., and Henrion, D. (2019). Semidefinite approximation of invariant measures for polynomial systems. *Disc. Cont. Dyn. Sys.*, 24(12):6745–6770.

Mhaskar, H. N. (2019). Super-resolution meets machine learning: Approximation of measures. *J. Fourier Anal. Appl.*, 25(6):3104–3122.

Möller, H. and Stetter, H. (1995). Multivariate polynomial equations with multiple zeros solved by matrix eigenproblems. *Numer. Math.*, 70:311–329.

Pauwels, E., Putinar, M., and Lasserre, J. (2020). Data analysis from empirical moments and the Christoffel function. *F. Comp. Math.*

R. de Prony, G. (1795). Essai expérimental et analytique: Sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures. *Journal de l'École Polytechnique Floréal et Plairial*, 1(cahier 22):24–76.

Roy, R., Paulraj, A., and Kailath, T. (1986). Esprit – a subspace rotation approach to estimation of parameters of cisoids in noise. *IEEE Trans. Acoustics Speech Signal Process.*, 34(5):1340–1342.

Sauer, T. (2017). Prony's method in several variables. *Numer. Math.*, 136:411–438.

Stetter, H. (1996). Matrix eigenproblems are at the heart of polynomial system solving. *ACM Sigsam Bulletin*, 30(4):22–25.