Sparse Projections onto the Simplex

Anastasios Kyrillidis[†] Stephen Becker[‡] Volkan Cevher[†] Christoph Koch[§]

[†]Laboratory for Information and Inference Systems (EPFL)
[‡]Laboratory Jacques Louis-Lions (Université Pierre et Marie Curie)
[§]Data Analysis, Theory and Application Laboratory (EPFL)
{anastasios.kyrillidis, volkan.cevher}@epfl.ch, stephen.becker@upmc.fr

Motivation examples

- **Quantum tomography (QT):**
  - **In plain words:** learn a density matrix $\mathbf{X}^\star \in \mathbb{C}^{d\times d}$ s.t. $\mathbf{X}^\star \succeq 0$, $\text{rank}(\mathbf{X}^\star) = r$ and $\text{tr}(\mathbf{X}^\star) = 1$ from a set of measurements.
  - **In ML:** Given a few samples $\mathbf{y} = \boldsymbol{\mathcal{A}}(\mathbf{X}^\star) + \boldsymbol{\eta}$, solve: $\min_{\{\mathbf{X}:\, \mathbf{X}\succeq 0,\ \text{rank}(\mathbf{X})=r,\ \text{tr}(\mathbf{X})=1\}} \|\boldsymbol{\mathcal{A}}(\mathbf{X}) - \mathbf{y}\|_2^2$.
  - **Key ingredient:** $\widehat{\mathbf{B}} \in \arg\min_{\{\mathbf{B}:\, \mathbf{B}\succeq 0,\ \text{rank}(\mathbf{B})=r,\ \text{tr}(\mathbf{B})=1\}} \|\mathbf{B} - \mathbf{W}\|_F^2$.
- **Markowitz portfolio optimization (MPO):**
  - **In plain words:** find a pdf over n assets to maximize the returns and minimize the risk.
  - **In ML:** learn a *normalized* vector $\boldsymbol{\beta}^* \in \mathbb{R}^n$ that minimizes a return-adjusted risk.
  - **Desiderata:** $\boldsymbol{\beta}^* \rightarrow$ sparse due to: (i) **robustness** and, (ii) **transaction fees are expensive**.
- **Sparse (Gaussian) kernel density estimation (sKDE):**
  - **In plain words:** find a finite number of Gaussian kernel functions (and their centers) such that their combination adequately explains a given pdf $f(\cdot)$.
  - **In ML:** learn a *normalized* vector $\boldsymbol{\beta}^* \in \mathbb{R}_+^n$ s.t. $\widehat{f}(\mathbf{x}) = \sum_i \beta_i^* \kappa_\sigma(\cdot)$ minimizes $\mathbb{E}\|\widehat{f}(\cdot) - f(\cdot)\|_2^2$.
  - **Desiderata:** $\boldsymbol{\beta}^* \rightarrow$ sparse due to: (i) **robustness** and, (ii) **interpretability of results**.

Optimization criterion

- We can solve QT and sKDE by:

$$\boldsymbol{\beta}^* \in \mathop{\arg\min}_{\boldsymbol{\beta}\in\Delta_\lambda^+\cap\Sigma_s} f(\boldsymbol{\beta}),$$

where $\Delta_\lambda^+ = \{\boldsymbol{\beta}\in\mathbb{R}^n : \beta_i \geq 0, \sum_i \beta_i = \lambda\}$ and $\Sigma_s = \{\boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 \leq s\}$.

- **Conventional wisdom:** let's sparsify $\boldsymbol{\beta}^*$ using ℓ_1-norm constraint/regularizer!... Unluckily:

ℓ_1-norm conflicts with Δ_1^+.

- Dropping the non-negative constraints, we can solve (1) for MPO using:

$$\Delta_\lambda = \Big\{\boldsymbol{\beta}\in\mathbb{R}^n : \sum_i \beta_i = \lambda\Big\}.$$

Sparse projection onto Δ_λ^+

Given $\mathbf{w}\in\mathbb{R}^n$:

$$(\mathcal{P}^{\mathcal{S}}): \quad \boldsymbol{\beta}^* \in \mathop{\arg\min}_{\boldsymbol{\beta}\in\Delta_\lambda^+\cap\Sigma_s} \|\boldsymbol{\beta} - \boldsymbol{w}\|_2^2$$

- Problem $\mathcal{P}^{\mathcal{S}}$ can be equivalently nested as:

$$\boldsymbol{\beta}^* \in \mathop{\arg\min}_{\mathcal{S}:\mathcal{S}\in\Sigma_s} \min_{\boldsymbol{\beta}_{\mathcal{S}}\in\Delta_\lambda^+,\, \boldsymbol{\beta}_{\setminus\mathcal{S}}=0} g(\boldsymbol{\beta}, \mathbf{w}),$$

where $g(\boldsymbol{\beta}, \mathbf{w}) = \|(\boldsymbol{\beta} - \boldsymbol{w})_{\mathcal{S}}\|_2^2 + \|(\boldsymbol{w})_{\setminus\mathcal{S}}\|_2^2$.

- Given $\mathcal{S}^*$, $(\boldsymbol{\beta}^*)_{\mathcal{S}^*} = \Big[w_i + \frac{1}{|\mathcal{S}^*|}\big(\lambda - \sum_{i\in\mathcal{S}^*} w_i\big)\Big]_+$.
- Thus, we need to solve the set maximization:

$$\mathcal{S}^* \in \mathop{\arg\max}_{\mathcal{S}:|\mathcal{S}|\leq s} F(\mathcal{S}). \quad (1)$$

Greedy selector and simplex projector (GSSP)

1. $\mathcal{S}^* = \text{supp}\left(\mathcal{P}_{L_s}(\mathbf{w})\right)$,
2. $(\boldsymbol{\beta}^*)_{\mathcal{S}^*}$ as above and, $(\boldsymbol{\beta}^*)_{\setminus\mathcal{S}^*} = 0$.

$\mathcal{P}_{L_s}$ keeps the s-largest entries (**not in magnitude**).

Theorem: *GSSP Algorithm provably solves the sparse projection onto Δ_λ^+ problem.*

- **Complexity:** $\mathcal{O}(n\log_2(n))$.

Sparse projection onto Δ_λ

Given $\mathbf{w}\in\mathbb{R}^n$:

$$(\mathcal{P}^{\mathcal{H}}): \quad \boldsymbol{\beta}^* \in \mathop{\arg\min}_{\boldsymbol{\beta}\in\Delta_\lambda\cap\Sigma_s} \|\boldsymbol{\beta} - \boldsymbol{w}\|_2^2$$

- Similarly, given $\mathcal{S}^*$, $(\boldsymbol{\beta}^*)_{\mathcal{S}^*} = w_i + \frac{1}{|\mathcal{S}^*|}\big(\lambda - \sum_{i\in\mathcal{S}^*} w_i\big) =: \mathcal{P}_\lambda(\mathbf{w}_{|\mathcal{S}^*})$.
- How to solve this projection?

Greedy selector and hyperplane projector (GSHP)

1. $\ell = 1$, $\mathcal{S} = j$, $\quad j \in \arg\max_i (\lambda w_i)$.
2. Repeat: $\ell \leftarrow \ell + 1$, $\mathcal{S} \leftarrow \mathcal{S}\cup j$, where

$$j \in \mathop{\arg\max}_{i\in\mathcal{N}\setminus\mathcal{S}} \left| w_i - \frac{\sum_{j\in\mathcal{S}} w_j - \lambda}{\ell - 1}\right|,$$

until $\ell = k$.

3. Set $\mathcal{S}^* \leftarrow \mathcal{S}$.
4. $\boldsymbol{\beta}_{|\mathcal{S}^*} = \mathcal{P}_\lambda(\mathbf{w}_{|\mathcal{S}^*})$, $\boldsymbol{\beta}_{|(\mathcal{S}^*)^c} = 0$.

- The algorithm for Δ_λ^+ (GSSP) is not applicable in this problem...
- GSHP selects the index of the largest element with the same sign as λ (Step 1). It then grows the index set one at a time by finding the farthest element from the current mean, as adjusted by λ (Step 2).

- What about guarantees?

Theorem: *GSHP Algorithm provably solves the sparse projection onto Δ_λ problem.*

- **Complexity:** $\mathcal{O}(n\log_2(n))$.

Experimental results

- **Sparse Gaussian KDE:** $f(x) = \frac{1}{5}\sum_{i=1}^5 \kappa_{\sigma_i}(\mu_i, x)$ where $\sigma_i = (7/9)^i$ and $\mu_i = 14(\sigma_i - 1)$.
- Other methods lead to wrong number of kernels.
- Nonconvex approach works even if s is slightly over-estimated.

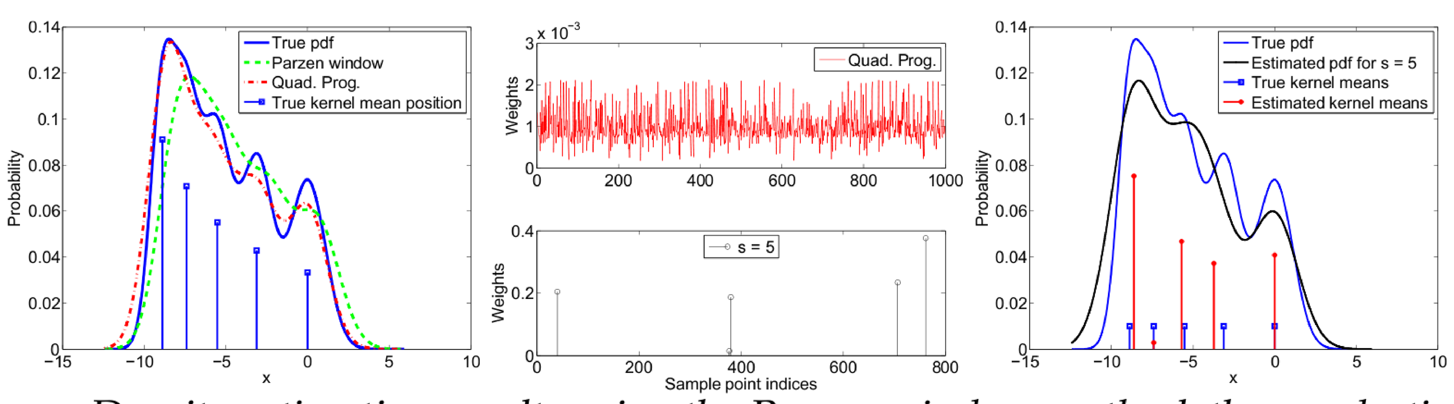
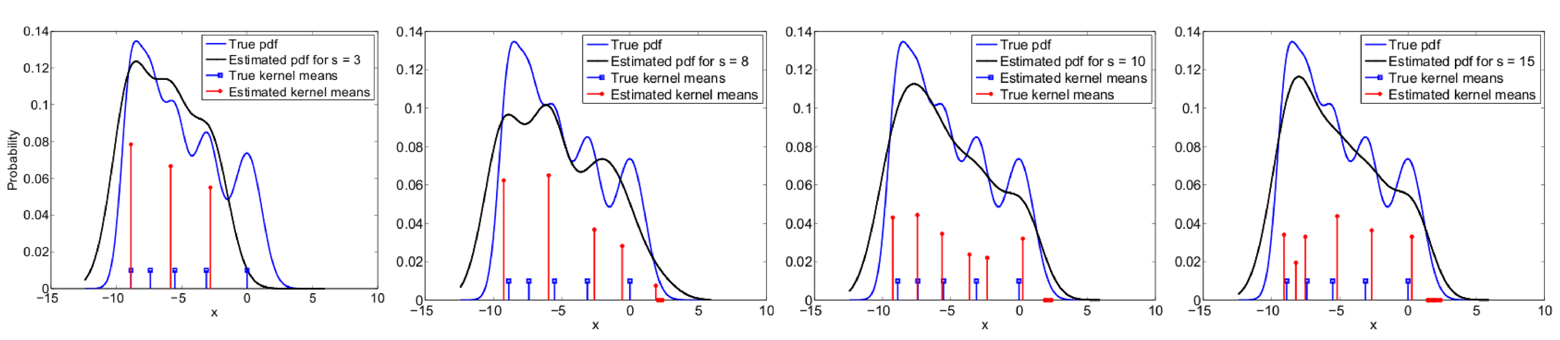


Figure: *Density estimation results using the Parzen window method, the quadratic programming and our approach for* $s = 5$.



(a) $s = 3$ (b) $s = 8$ (c) $s = 10$ (d) $s = 15$

Figure: *Density estimation results for various sparsity levels* s*. Red spikes depict the estimated kernel means and their relative contibution to the Gaussian mixture.*

- **Quantum tomography:** We use the simple gradient descent algorithm:

$$\boldsymbol{\beta}^{i+1} = \mathcal{P}(\boldsymbol{\beta}^i - \mu^i\nabla f(\boldsymbol{\beta}^i)), \text{ where } \mu^i = 3/\|\boldsymbol{\mathcal{A}}\|^2 \text{ and } f(\boldsymbol{\beta}) = \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X})\|_2^2.$$

- $\mathbf{X}^\star$: randomly generated with $\text{rank}(\mathbf{X}^\star) = 2$. We assume $r = 2$ is known.
- Convex counterpart: $\min_{\{\mathbf{X}:\, \mathbf{X}\succeq 0,\ \|\mathbf{X}\|_*\leq 1\}} \|\mathbf{y} - \boldsymbol{\mathcal{A}}(\mathbf{X})\|_2^2$ using `TFOCS` package.
- **Left figure**: QT wth 8 qubits and 30 dB SNR. Each point is the median relative error $\|\mathbf{X} - \mathbf{X}^\star\|_F^2 / \|\mathbf{X}^\star\|_F^2$ vs. # measurements over 10 Monte Carlo iterations.
- **Right figure**: QT wth 7 qubits, no noise. Same configuration as above.