Calculating Optimistic Likelihoods Using (Geodesically) Convex Optimization

Viet Anh Nguyen†, Soroosh Shafieezadeh-Abadeh†, Man-Chung Yue‡, Daniel Kuhn† and Wolfram Wiesemann§

†EPFL, ‡The Hong Kong Polytechnic University, §Imperial College London

Contributions

MLE problem for i.i.d. data points

$$\hat{x}_1^M \triangleq \hat{x}_1, \ldots, \hat{x}_M \in \mathbb{R}^n$$

and candidate distributions $\mathbb{P}_c = \mathcal{N}(\mu_c, \Sigma_c)$:

$$c^\star \in \arg\max_{c \in \mathcal{C}} \left\{ \ell(\hat{x}_1^M, \mathbb{P}_c) \triangleq -\frac{1}{M}\sum_{m=1}^{M} (\hat{x}_m - \mu_c)^\top \Sigma_c^{-1} (\hat{x}_m - \mu_c) - \log\det \Sigma_c \right\}$$

Motivation:

- MLE problem is fundamental in hypothesis testing and discriminant analysis
- The parameters (μ_c, Σ_c) of $\mathbb{P}_c$ are uncertain
- Ignoring this uncertainty leads to poor out-of-sample performance

Contributions:

- We propose an **optimistic likelihood** (OL) problem over **Fisher-Rao** (FR) and **Kullback-Leibler** (KL) ambiguity sets containing normal distributions
- For FR ambiguity sets, the OL problem reduces to a **geodesically convex** problem
- We devise a **Riemannian gradient descent** algorithm
- For KL ambiguity sets, the OL problem reduces to a one dimensional convex problem

Fisher-Rao Distance

Parametric distributions with density function $p_\theta(x)$

- For any $\theta \in \Theta$, the Fisher information matrix
$$I_\theta = \mathbb{E}_x[\nabla_\theta \log(p_\theta(x)) \nabla_\theta \log(p_\theta(x))^\top]$$
defines an inner product $\langle \cdot, \cdot \rangle_\theta$ on the tangent space $T_\theta \Theta$ as
$$\langle \zeta_1, \zeta_2 \rangle_\theta = \zeta_1^T I_\theta \zeta_2, \quad \forall \zeta_1, \zeta_2 \in T_\theta \Theta$$
- The set of $\{\langle \cdot, \cdot \rangle_\theta\}_{\theta \in \Theta}$ defines a Riemannian metric called the FR metric
- The FR metric is invariant under transformations on the data space
- The FR distance on Θ is a geodesic distance defined as
$$d(\theta_0, \theta_1) = \inf_\gamma \int_0^1 \sqrt{\langle \gamma'(t), \gamma'(t) \rangle_{\gamma(t)}}\, dt$$
- The infimum is over smooth curves $\gamma : [0,1] \to \Theta$ with $\gamma(0) = \theta_0$ and $\gamma(1) = \theta_1$

Proposition 1. [Atkinson and Mitchell (1981)] For the family of Gaussian distributions with identical mean and and covariance matrices Σ_0, Σ_1, we have

$$d(\Sigma_0, \Sigma_1) = \frac{1}{\sqrt{2}} \left\| \log(\Sigma_1^{-\frac{1}{2}} \Sigma_0 \Sigma_1^{-\frac{1}{2}}) \right\|_F$$

Kullback-Leibler Divergence

For distributions $\mathbb{P}_0$ and $\mathbb{P}_1$ with density functions $p_0(x)$ and $p_1(x)$, we have

$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_1) = \int_{-\infty}^{\infty} p_0(x) \log\left(\frac{p_0(x)}{q_1(x)}\right) \mathrm{d}x$$

When $\mathbb{P}_0 = \mathcal{N}(\mu_0, \Sigma_0)$ and $\mathbb{P}_1 = \mathcal{N}(\mu_1, \Sigma_1)$, the KL divergence coincides with

$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_1) = \frac{1}{2}\left(\mathrm{Tr}\left[\Sigma_1^{-1}\Sigma_0\right] + \log\det(\Sigma_1 \Sigma_0^{-1}) - n + (\mu_0 - \mu_1)^\top \Sigma_1^{-1} (\mu_0 - \mu_1) \right).$$

Optimistic Likelihood Problems

We consider the optimistic likelihood problem

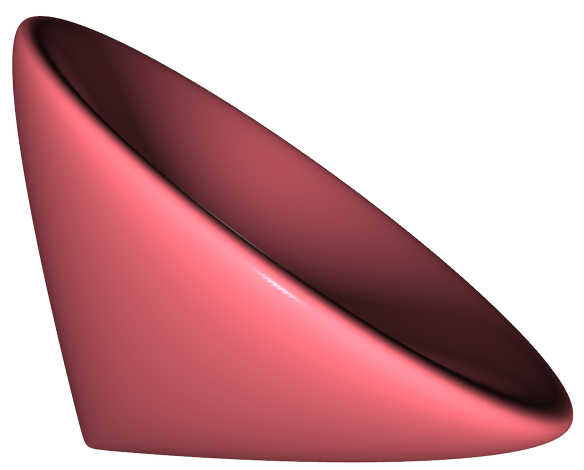
$$\textbf{OL:} \quad \max_{\mathbb{P} \in \mathcal{P}} \ell(\hat{x}_1^M, \mathbb{P}) \quad \text{with} \quad \mathcal{P} = \left\{ \mathbb{P} \in \mathcal{M} : \varphi(\hat{\mathbb{P}}, \mathbb{P}) \le \rho \right\}.$$

- Candidate Gaussian distribution: $\hat{\mathbb{P}} = \mathcal{N}(\hat{\mu}, \hat{\Sigma})$
- $\mathcal{M}$ is the family of Gaussian distributions with fixed mean $\hat{\mu}$
- $\varphi(\cdot,\cdot)$ is the dissimilarity measure $\Rightarrow$ FR distance or KL divergence
- ρ is the size of the ambiguity set

OL Problem under the FR Distance

The **OL** problem reduces to $\min_{\Sigma \in \mathcal{B}^{\mathrm{FR}}} L(\Sigma)$, where

$$L(\Sigma) \triangleq \langle S, \Sigma^{-1} \rangle + \log\det \Sigma$$
$$\mathcal{B}^{\mathrm{FR}} \triangleq \{ \Sigma \in \mathbb{S}_{++}^n : d(\Sigma, \hat{\Sigma}) \le \rho \}$$
$$S = M^{-1} \textstyle\sum_{m=1}^{M} (\hat{x}_m - \hat{\mu})(\hat{x}_m - \hat{\mu})^\top$$



Theorem 1. ▸ $\mathcal{B}^{\mathrm{FR}}$ is a geodesically convex set

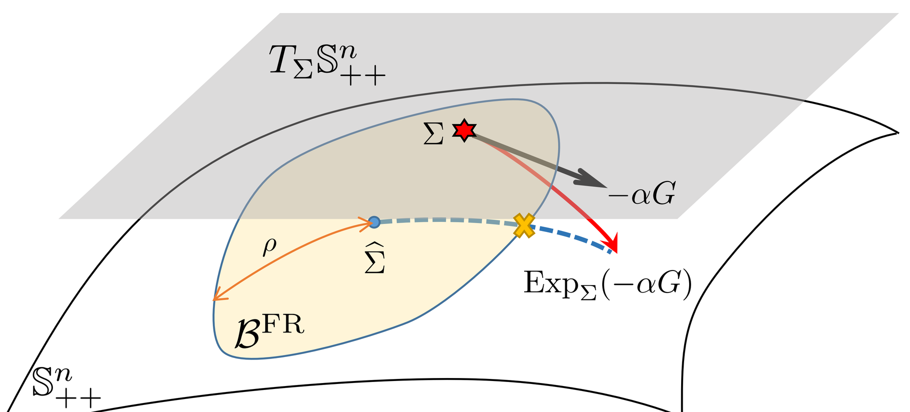
▸ $L(\cdot)$ is a geodesically convex function over $\mathbb{S}_{++}^n$

▸ $L(\cdot)$ is geodesically β-smooth and σ-strongly on $\mathcal{B}^{\mathrm{FR}}$ with

$$\beta = \frac{2\lambda_{\max}(S)}{\lambda_{\min}(\hat{\Sigma}) \exp(-\sqrt{2}\rho)}, \quad \text{and} \quad \sigma = \frac{2\lambda_{\min}(S)}{\lambda_{\max}(\hat{\Sigma}) \exp(\sqrt{2}\rho)}.$$

Projected Geodesic Gradient Descent

1. Start from a feasible point Σ
2. Follow the Riemannian gradient G
3. Project back to S_{++}^n manifold
   $\Rightarrow$ use the exponential map $\mathrm{Exp}_\Sigma(-\alpha G)$
4. Project back to $\mathcal{B}^{\mathrm{FR}}$

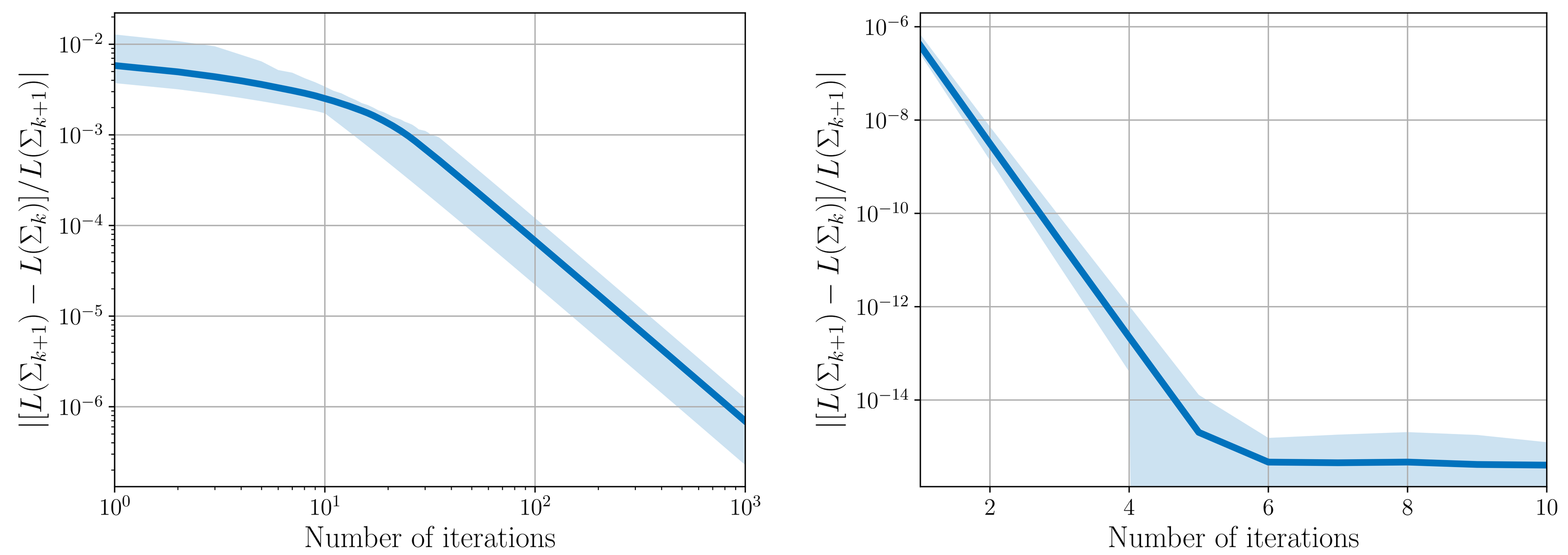


Based on [Zhang and Sra (2016)]

Proposition 2. If $d(\Sigma', \hat{\Sigma}) = \rho' > \rho$, then

$$\mathrm{Proj}_{\mathcal{B}^{\mathrm{FR}}}(\Sigma') = \hat{\Sigma}^{\frac{1}{2}} \left(\hat{\Sigma}^{-\frac{1}{2}} \Sigma' \hat{\Sigma}^{-\frac{1}{2}} \right)^{\frac{\rho}{\rho'}} \hat{\Sigma}^{\frac{1}{2}}$$

Theorem 2. With a constant stepsize $\alpha_k = \mathcal{O}(1/\sqrt{K})$, the projected geodesic gradient descent converges with the sublinear rate $\mathcal{O}(1/\sqrt{K})$.

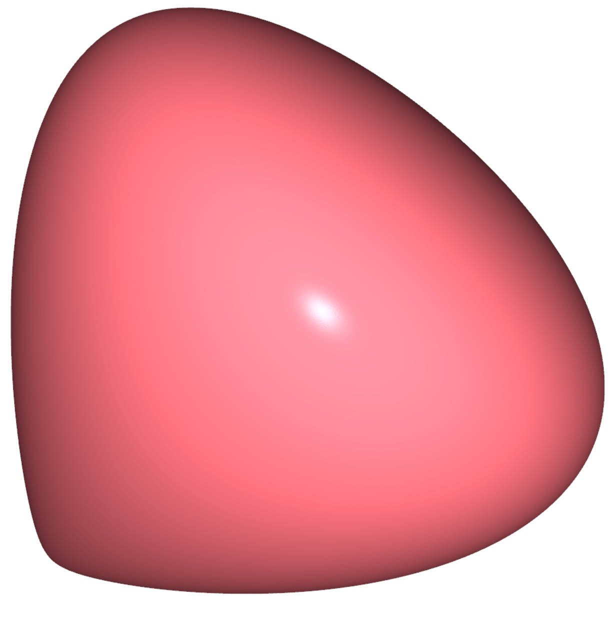


(a) Convergence for $S \succeq 0$ (b) Convergence for $S \succ 0$

OL Problem under the KL Divergence

The **OL** problem reduces to

$$\min_{\Sigma \succ 0} \ \mathrm{Tr}\left[S\Sigma^{-1}\right] + \log\det \Sigma$$
$$\text{s.t.} \ \ \mathrm{Tr}\left[\Sigma^{-1}\hat{\Sigma}\right] + \log\det(\Sigma \hat{\Sigma}^{-1}) - n \le 2\rho.$$



- Both objective and constraint are non-convex
- Convexification by substitution $X \leftarrow \Sigma^{-1}$
- Further reduction to one dimensional problem

Theorem 3. The **OL** problem is solved by

$$\Sigma^\star = S + \gamma^\star \hat{\Sigma},$$

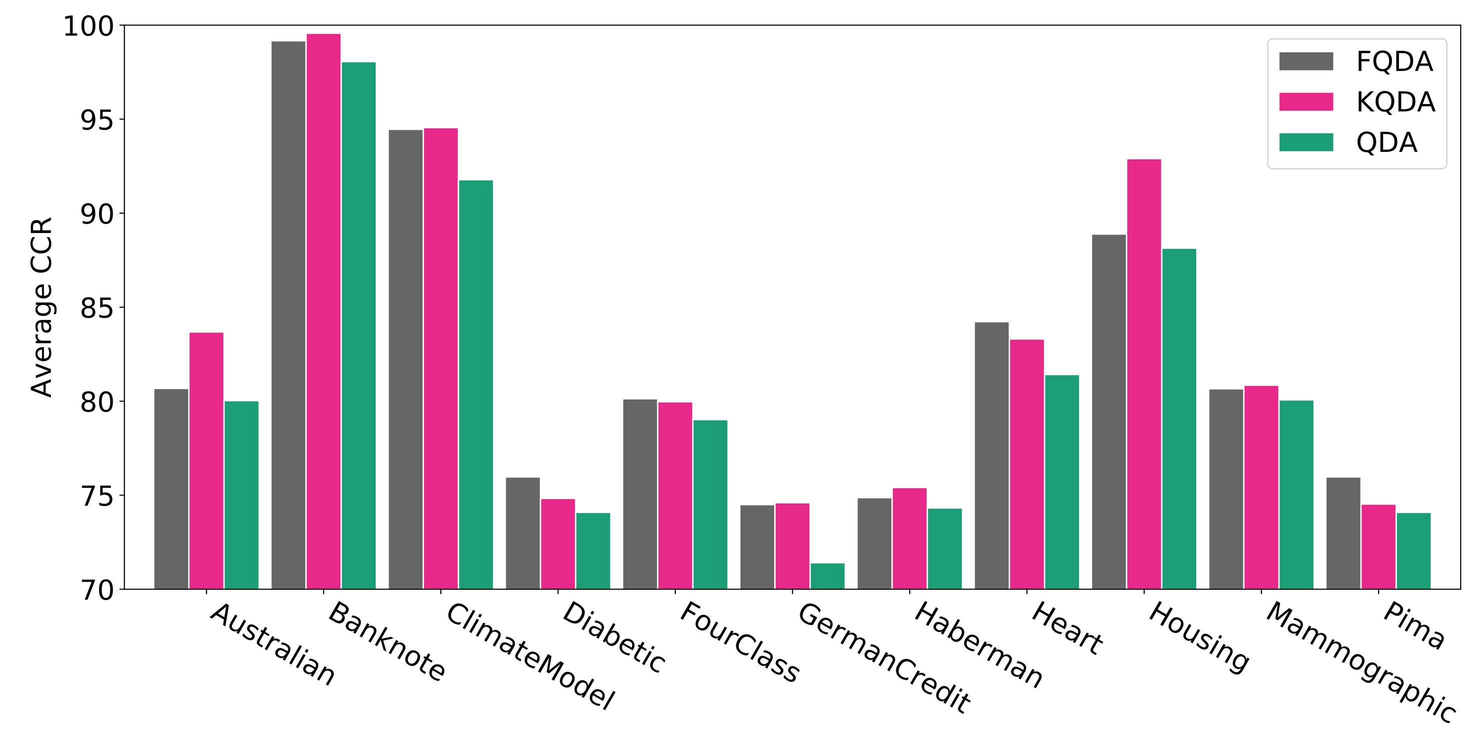
where $\gamma^\star$ is the solution of

$$\min_{\gamma^\star > 0} \ \gamma^\star(2\rho + \log\det \hat{\Sigma}) + n(1+\gamma^\star)\log(1+\gamma^\star) - (1+\gamma^\star)\log\det(S + \gamma^\star \hat{\Sigma}).$$

Flexible Discriminant Rules

- Classification problem with $Y \in \mathcal{C}, \mathcal{C} = \{1, \ldots, C\}$
- Bayes' Theorem implies that $\mathbb{P}(Y = c | X = x) \propto \pi_c \cdot f_c(x)$
- Assumption: $\hat{\mathbb{P}}_c = \mathcal{N}(\hat{\mu}_c, \hat{\Sigma}_c)$ and $\hat{\pi}_c = N_c/N$
- $\hat{\mu}_c$ and $\hat{\Sigma}_c$ are estimated from training data
- QDA rule: $\mathcal{C}_{\mathrm{QDA}}(x) \in \arg\max_{c \in \mathcal{C}} \left\{ \frac{1}{2}\ell(x, \hat{\mathbb{P}}_c) + \log(\hat{\pi}_c) \right\}$
- Our suggestion: $\mathcal{C}_{\mathrm{flex}}(x) \in \arg\max_{c \in \mathcal{C}} \ \max_{\mathbb{P} \in \mathcal{P}_c} \left\{ \frac{1}{2}\ell(x, \mathbb{P}) + \log(\hat{\pi}_c) \right\}$

Empirical Experiments (UCI dataset): average correct classification rates



References

- C. Atkinson and A. F. Mitchell. Rao's distance measure. Sankhya: The Indian Journal of Statistics, Series A, 43(3):345-365, 1981.
- H. Zhang and S. Sra. First-order methods for geodesically convex optimization. In Conference on Learning Theory, pages 1617-1638, 2016.