Appendix – Robust Bayesian Classification Using an Optimistic Score Ratio

Viet Anh Nguyen [1] **Nian Si** [1] **Jose Blanchet** [1]

A. Proof of Section 1

Proof of Proposition 1.1. To ease the exposition, we use the following notational shorthands $\mathbb{B}_c = \mathbb{B}_{\rho_c}(\widehat{\mathbb{P}}_c) \cap \mathcal{P}$ for $c \in \{0,1\}$ and

$$f_c^{\max}(x) = \sup_{f_c \in \mathbb{B}_c} f_c(x), \quad f_c^{\min}(x) = \inf_{f_c \in \mathbb{B}_c} f_c(x) \qquad \forall c \in \{0,1\}.$$

If action $a = 1$ is chosen, then the worst-case probability of mis-classification is

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{B}} \mathbb{P}(Y = 0 | X = x) &= \left\{ \begin{array}{cl} \sup & \dfrac{f_0(x)\pi_0}{f_0(x)\pi_0 + f_1(x)\pi_1} \\ \text{s.t.} & f_0 \in \mathbb{B}_0,\ f_1 \in \mathbb{B}_1 \end{array} \right. \\
&= \sup_{f_0 \in \mathbb{B}_0} \sup_{f_1 \in \mathbb{B}_1} \frac{f_0(x)\pi_0}{f_0(x)\pi_0 + f_1(x)\pi_1} \\
&= \sup_{f_0 \in \mathbb{B}_0} \frac{f_0(x)\pi_0}{f_0(x)\pi_0 + f_1^{\min}(x)\pi_1} \qquad \text{(A.1a)} \\
&= \frac{f_0^{\max}(x)\pi_0}{f_0^{\max}(x)\pi_0 + f_1^{\min}(x)\pi_1}, \qquad \text{(A.1b)}
\end{aligned}$$

where equality (A.1a) holds because $\pi_1 > 0$, thus for any $f_0 \in \mathbb{B}_0$, the optimal choice of f_1 for the inner supremum problem will minimize $f_1(x)$ over all $f_1 \in \mathbb{B}_1$. Equality (A.1b) holds because $f_1^{\min}(x)\pi_1 > 0$, thus it is optimal to choose f_0 that maximizes $f_0(x)$ over all $f_0 \in \mathbb{B}_0$. Using similar lines of arguments, if action $a = 0$ is chosen, then the worst-case probability of mis-classification is

$$\begin{aligned}
\sup_{\mathbb{P} \in \mathcal{B}} \mathbb{P}(Y = 1 | X = x) &= \left\{ \begin{array}{cl} \sup & \dfrac{f_1(x)\pi_1}{f_0(x)\pi_0 + f_1(x)\pi_1} \\ \text{s.t.} & f_0 \in \mathbb{B}_0,\ f_1 \in \mathbb{B}_1 \end{array} \right. \\
&= \frac{f_1^{\max}(x)\pi_1}{f_0^{\min}(x)\pi_0 + f_1^{\max}(x)\pi_1}.
\end{aligned}$$

Thus, by comparing the two values of the worst-case probability, action $a = 1$ is optimal whenever

$$\frac{f_0^{\max}(x)\pi_0}{f_0^{\max}(x)\pi_0 + f_1^{\min}(x)\pi_1} \le \frac{f_1^{\max}(x)\pi_1}{f_0^{\min}(x)\pi_0 + f_1^{\max}(x)\pi_1},$$

which in turn is equivalent to the condition

$$\frac{f_1^{\max}(x)}{f_0^{\max}(x)} \ge \frac{f_0^{\min}(x)\pi_0^2}{f_1^{\min}(x)\pi_1^2}. \qquad \text{(A.2)}$$

By setting the right-hand side of (A.2) to a threshold $\tau(x) \in \mathbb{R}_+$, we arrive at the postulated result. □

[1] Stanford University. Correspondence to: Viet Anh Nguyen <viet-anh.nguyen@stanford.edu>.

Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020. Copyright 2020 by the author(s).

As the proof reveals in (A.2), the optimal threshold $\tau(x)$ in the statement of Proposition 1.1 admits an explicit expression

$$\tau(x) = \frac{f_0^{\min}(x)\pi_0^2}{f_1^{\min}(x)\pi_1^2}.$$

This value $\tau(x)$ can be found by evaluating the minimum likelihood values $f_0^{\min}(x)$ and $f_0^{\min}(x)$. Unfortunately, it remains intractable to find the exact values of $f_0^{\min}(x)$ and $f_1^{\min}(x)$. To demonstrate this fact, we consider the Gaussian parametric setting as in Section 4, and evaluating the minimum likelihood in this case is equivalent to solving

$$\begin{array}{ll} \min & -(\mu - x)^\top \Sigma^{-1}(\mu - x) - \log\det\Sigma \\ \text{s. t.} & \mu \in \mathbb{R}^d,\ \Sigma \in \mathbb{S}_{++}^d \\ & (\mu - \widehat{\mu})^\top \Sigma^{-1}(\mu - \widehat{\mu}) + \operatorname{Tr}\big[\widehat{\Sigma}\Sigma^{-1}\big] - \log\det\widehat{\Sigma}\Sigma^{-1} - d \le \rho \end{array} \tag{A.3}$$

for some $\widehat{\mu} \in \mathbb{R}^d$, $\widehat{\Sigma} \in \mathbb{S}_{++}^d$ and $\rho \ge 0$. Problem (A.3) is the minimization counterpart of the maximization problem (11a), it is also non-convex, however, we are not aware of any tractable approach to solve (A.3).

B. Proofs of Section 2

Proof of Theorem 2.2. Throughout this proof, we use $\xrightarrow{dist.}$ and $\xrightarrow{p.}$ to denote the convergence in distribution and in probability, respectively. For n sufficiently big, $\widehat{\Sigma}_n$ defined as in (7) is invertible with probability 1. In this case, we find

$$\begin{aligned} \mathbb{D}\big((\widehat{\mu}_n, \widehat{\Sigma}_n) \parallel (m, S)\big) &= \operatorname{Tr}\big[\widehat{\Sigma}_n S^{-1}\big] + (m - \widehat{\mu}_n)^\top S^{-1}(m - \widehat{\mu}_n) - d - \log\det(\widehat{\Sigma}_n S^{-1}) \\ &= \frac{1}{n}\sum_{t=1}^n \widehat{\eta}_t^\top \widehat{\eta}_t - d - \log\det(S^{-\frac{1}{2}}\widehat{\Sigma}_n S^{-\frac{1}{2}}), \end{aligned} \tag{A.4}$$

where $\widehat{\eta}_t = S^{-\frac{1}{2}}(\widehat{\xi}_t - m)$ is the isotropic transformation of $\widehat{\xi}_t$ for each $t = 1, \ldots, n$. Furthermore, denote by $\bar{\mu}_n$ the sample average of $\widehat{\eta}_1, \ldots, \widehat{\eta}_n$ defined as

$$\bar{\mu}_n = \frac{1}{n}\sum_{t=1}^n \widehat{\eta}_t = S^{-\frac{1}{2}}(\widehat{\mu}_n - m).$$

By adding and subtracting $\log\det\big(n^{-1}\sum_{t=1}^n \widehat{\eta}_t\widehat{\eta}_t^\top\big)$ into (A.4), we have

$$\begin{aligned} &\mathbb{D}\big((\widehat{\mu}_n, \widehat{\Sigma}_n) \parallel (m, S)\big) = \\ &\underbrace{\left(\log\det\Big(\frac{1}{n}\sum_{t=1}^n \widehat{\eta}_t\widehat{\eta}_t^\top\Big) - \log\det(S^{-\frac{1}{2}}\widehat{\Sigma}_n S^{-\frac{1}{2}})\right)}_{\text{(A)}} + \underbrace{\left(\frac{1}{n}\sum_{t=1}^n \widehat{\eta}_t^\top\widehat{\eta}_t - d - \log\det\Big(\frac{1}{n}\sum_{t=1}^n \widehat{\eta}_t\widehat{\eta}_t^\top\Big)\right)}_{\text{(B)}}. \end{aligned}$$

We analyze the 2 terms (A) and (B) separately. First, rewrite

$$\begin{aligned} S^{-\frac{1}{2}}\widehat{\Sigma}_n S^{-\frac{1}{2}} &= S^{-\frac{1}{2}}\Big(\frac{1}{n}\sum_{t=1}^n (\widehat{\xi}_t - \widehat{\mu}_n)(\widehat{\xi}_t - \widehat{\mu}_n)^\top\Big)S^{-\frac{1}{2}} \\ &= S^{-\frac{1}{2}}\Big(\frac{1}{n}\sum_{t=1}^n (\widehat{\xi}_t - m + m - \widehat{\mu}_n)(\widehat{\xi}_t - m + m - \widehat{\mu}_n)^\top\Big)S^{-\frac{1}{2}} \\ &= \frac{1}{n}\sum_{t=1}^n \Big(\widehat{\eta}_t\widehat{\eta}_t^\top + S^{-\frac{1}{2}}(m - \widehat{\mu}_n)\widehat{\eta}_t^\top + \widehat{\eta}_t(m - \widehat{\mu}_n)^\top S^{-\frac{1}{2}} + S^{-\frac{1}{2}}(m - \widehat{\mu}_n)(m - \widehat{\mu}_n)^\top S^{-\frac{1}{2}}\Big) \\ &= \frac{1}{n}\sum_{t=1}^n \Big(\widehat{\eta}_t\widehat{\eta}_t^\top - \bar{\mu}_n\widehat{\eta}_t^\top - \widehat{\eta}_t\bar{\mu}_n^\top + \bar{\mu}_n\bar{\mu}_n^\top\Big) \\ &= \left(\frac{1}{n}\sum_{t=1}^n \widehat{\eta}_t\widehat{\eta}_t^\top\right) - \bar{\mu}_n\bar{\mu}_n^\top. \end{aligned}$$

Then, (A) becomes

$$\log\det\big(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top\big) - \log\det(S^{-\frac{1}{2}}\widehat{\Sigma}_n S^{-\frac{1}{2}}) = -\log\det\left(I_d - \left(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top\right)^{-1}\left(\bar{\mu}_n\bar{\mu}_n^\top\right)\right).$$

By the weak law of large numbers, as $n\uparrow\infty$, we find

$$\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top \xrightarrow{p.} I_d \quad \text{and} \quad \frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t \xrightarrow{p.} 0.$$

By the central limit theorem, we find as $n\uparrow\infty$

$$\sqrt{n}\bar{\mu}_n \xrightarrow{dist.} H, \quad \sqrt{n}\left(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top - I_d\right) \xrightarrow{dist.} Z,$$

where the random vector H and the random matrix Z are defined as in the statement of the theorem. By Slutsky's theorem (van der Vaart, 1998, Theorem 2.8), we find

$$n\left(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top\right)^{-1}\left(\bar{\mu}_n\bar{\mu}_n^\top\right) \xrightarrow{dist.} HH^\top.$$

By the delta method (van der Vaart, 1998, Theorem 3.1), we have

$$n\times\underbrace{\Big(\log\det\big(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top\big) - \log\det(S^{-\frac{1}{2}}\widehat{\Sigma}_n S^{-\frac{1}{2}})\Big)}_{\text{(A)}} \xrightarrow{dist.} \operatorname{Tr}\left[HH^\top\right] = H^\top H. \tag{A.5}$$

Now, we are ready to analyze (B). Using a Taylor expansion for the log-determinant function around I_d, we find

$$\begin{aligned}&\log\det\big(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top\big)\\ &= \log\det(I_d) + \operatorname{Tr}\left[\left(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top - I_d\right)\right] - \frac{1}{2}\operatorname{Tr}\left[\left(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top - I_d\right)^2\right] + o\left(\operatorname{Tr}\left[\left(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top - I_d\right)^2\right]\right).\end{aligned}$$

Therefore, by the second-order delta method, we have

$$n\times\underbrace{\Big(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t^\top\widehat{\eta}_t - d - \log\det\big(\frac{1}{n}\sum_{t=1}^{n}\widehat{\eta}_t\widehat{\eta}_t^\top\big)\Big)}_{\text{(B)}} \xrightarrow{dist.} \frac{1}{2}\operatorname{Tr}\left[Z^2\right]. \tag{A.6}$$

Finally, by combining the limits from (A.5) and (A.6), we obtain the postulated result. □

Proof of Corollary 2.3. From Theorem 2.2, we have as $n\uparrow\infty$

$$n\times\frac{1}{2}\operatorname{KL}\big(\mathcal{N}(\widehat{\mu}_n,\widehat{\Sigma}_n)\parallel\mathcal{N}(m,S)\big) = n\times\mathbb{D}\big((\widehat{\mu}_n,\widehat{\Sigma}_n)\parallel(m,S)\big) \to H^\top H + \frac{1}{2}\operatorname{Tr}\left[Z^2\right] \quad \text{in distribution,}$$

where the random vector H and the random matrix Z are defined as in the statement of Theorem 2.2. In the Gaussian setting, the elements of the isotropic random vector η are i.i.d. standard univariate normal random variables. Therefore, we have

$$\operatorname{cov}(Z_{jk}, Z_{j'k'}) = \begin{cases}\mathbb{E}_{\mathbb{P}}\left[(\eta_j)^4\right] - 1 & \text{if } j = k = j' = k',\\ 1 & \text{if } j < k, (j = j', k = k' \text{ or } j = k', j' = k),\\ 0 & \text{otherwise.}\end{cases}$$

Recall that $\mathbb{E}_{\mathbb{P}}\left[(\eta_j)^4\right] = 3$, which gives $\operatorname{cov}(Z_{jj}, Z_{jj}) = 2$. Hence, $\frac{1}{2}\operatorname{Tr}\left[Z^2\right]$ follows $\chi^2\left(d(d+1)/2\right)$ and $H^\top H$ follows $\chi^2\left(d\right)$. Finally, since H and Z are independent in the Gaussian case, we have $H^\top H + \frac{1}{2}\operatorname{Tr}\left[Z^2\right]$ follows $\chi^2\left(d\right) + \chi^2\left(d(d+1)/2\right) = \chi^2\left(d(d+3)/2\right)$. □

C. Proofs of Section 3

We first prove the compactness property of the uncertainty set $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$.

Lemma C.1 (Compactness of $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$)**.** *For any $\widehat{\mu} \in \mathbb{R}^d$, $\widehat{\Sigma} \in \mathbb{S}^d_{++}$ and $\rho \in \mathbb{R}_+$, the set $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ written as*

$$\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma}) = \{(\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{S}^d_{++} : \operatorname{Tr}\big[\widehat{\Sigma}\Sigma^{-1}\big] - \log\det(\widehat{\Sigma}\Sigma^{-1}) - d + (\mu - \widehat{\mu})^\top \Sigma^{-1}(\mu - \widehat{\mu}) \le \rho\}$$

is compact.

Proof of Lemma C.1. If $\rho = 0$ then $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is a singleton $\{(\widehat{\mu}, \widehat{\Sigma})\}$ and the claim holds trivially. For the rest of the proof, we consider when $\rho > 0$. Pick an arbitrary $(\mu, \Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$, it is obvious that Σ should satisfy

$$\operatorname{Tr}\big[\widehat{\Sigma}^{\frac{1}{2}}\Sigma^{-1}\widehat{\Sigma}^{\frac{1}{2}}\big] - \log\det(\widehat{\Sigma}^{\frac{1}{2}}\Sigma^{-1}\widehat{\Sigma}^{\frac{1}{2}}) - d \le \rho,$$

which implies that Σ is bounded. To see this, suppose that $\{\Sigma_k\}_{k\in\mathbb{N}}$ is a sequence of positive definite matrices and $\{\sigma_k\}_{k\in\mathbb{N}}$ is the corresponding sequence of the minimum eigenvalues of $\{\widehat{\Sigma}^{-\frac{1}{2}}\Sigma_k^{-1}\widehat{\Sigma}^{-\frac{1}{2}}\}_{k\in\mathbb{N}}$. Because the function $\sigma \mapsto \sigma - \log\sigma - 1$ is non-negative for every $\sigma > 0$, we find

$$\operatorname{Tr}\big[\widehat{\Sigma}^{\frac{1}{2}}\Sigma_k^{-1}\widehat{\Sigma}^{\frac{1}{2}}\big] - \log\det(\widehat{\Sigma}^{\frac{1}{2}}\Sigma_k^{-1}\widehat{\Sigma}^{\frac{1}{2}}) - d \ge \sigma_k - \log\sigma_k - 1.$$

If Σ_k tends to infinity, then σ_k tends to 0, and in this case $\sigma_k - \log\sigma_k - 1 \to +\infty$. This implies that Σ should be bounded in the sense that $\Sigma \preceq \bar{\sigma} I_d$ for some finite positive constant $\bar{\sigma}$. Using an analogous argument, we can show that Σ is lower bounded in the sense that $\Sigma \succeq \underline{\sigma} I_d$ for some finite positive constant $\underline{\sigma}$. As a consequence, μ is also bounded because μ should satisfy $\underline{\sigma}\|\mu - \widehat{\mu}\|_2^2 \le \rho$. We now can rewrite $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ as

$$\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma}) = \{(\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{S}^d_{++} : \underline{\sigma}\|\mu - \widehat{\mu}\|_2^2 \le \rho,\ \underline{\sigma} I_d \preceq \Sigma \preceq \bar{\sigma} I_d,\ \mathbb{D}\big((\widehat{\mu}, \widehat{\Sigma}) \parallel (\mu, \Sigma)\big) \le \rho\},$$

which implies that $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is a closed set because $\mathbb{D}\big((\widehat{\mu}, \widehat{\Sigma}) \parallel (\,\cdot\,, \cdot\,)\big)$ is a continuous function over (μ, Σ) when Σ ranges over $\underline{\sigma} I_d \preceq \Sigma \preceq \bar{\sigma} I_d$. This observation coupled with the boundedness of (μ, Σ) established previously completes the proof. □

For a fixed $\widehat{\mu} \in \mathbb{R}^d$, $x \in \mathbb{R}^d$ and $\varepsilon \in \mathbb{R}_+$, define the following function $g : \mathbb{S}^d_{++} \to \mathbb{R}_+$ as

$$g(\Omega) \triangleq \left\{ \begin{array}{ll} \min & (\mu - x)^\top \Omega (\mu - x) \\ \text{s.t.} & \mu \in \mathbb{R}^d,\ (\mu - \widehat{\mu})^\top \Omega (\mu - \widehat{\mu}) \le \varepsilon, \end{array} \right. \tag{A.7}$$

where the dependence of g on $\widehat{\mu}$, x and ε has been made implicit to avoid clutter. The objective function of problem (A.7) is continuous in μ and the feasible set of problem (A.7) is compact because $\Omega \in \mathbb{S}^d_{++}$, which justify the minimization operator of problem (A.7). The next lemma asserts that the value $g(\Omega)$ coincides with the optimal value of a univariate convex optimization problem.

Lemma C.2 (Reformulation of g)**.** *For any $\widehat{\mu} \in \mathbb{R}^d$, $x \in \mathbb{R}^d$, $\varepsilon \in \mathbb{R}_+$ and $\Omega \in \mathbb{S}^d_{++}$, the value $g(\Omega)$ coincides with the optimal value of the univariate convex optimization problem*

$$\max_{\lambda \ge 0} \left\{ \frac{\lambda}{1 + \lambda}(x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}) - \lambda\varepsilon \right\}. \tag{A.8}$$

Moreover, denote by $\lambda^\star$ the unique optimal solution of the maximization problem (A.8), then the unique minimizer $\mu^\star$ of problem (A.7) is $\mu^\star = (x + \lambda^\star\widehat{\mu})/(1 + \lambda^\star)$. Furthermore, we have

$$\left\{ \begin{array}{ll} \lambda^\star = 0,\ g(\Omega) = 0 & \textit{if } \varepsilon \ge (x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}), \\ \lambda^\star = \big(\sqrt{\varepsilon (x - \widehat{\mu})^\top \Omega (x - \widehat{\mu})} - \varepsilon\big)/\varepsilon,\ g(\Omega) = \big(\sqrt{\varepsilon} - \sqrt{(x - \widehat{\mu})^\top \Omega (x - \widehat{\mu})}\big)^2 & \textit{otherwise.} \end{array} \right.$$

Proof of Lemma C.2. Using a change of variables $y \leftarrow \mu - \widehat{\mu}$ and a change of parameters $w \leftarrow x - \widehat{\mu}$, problem (A.7) can be recast in the following equivalent form

$$g(\Omega) = \left\{ \begin{array}{ll} \min & (y - w)^\top \Omega (y - w) \\ \text{s.t.} & y \in \mathbb{R}^d,\ y^\top \Omega y \le \varepsilon, \end{array} \right. \tag{A.9}$$

which is a convex optimization problem. Assume momentarily that $\varepsilon > 0$. By invoking a duality argument, we find

$$\begin{aligned} g(\Omega) &= \min_{y} \max_{\lambda \geq 0} \ (y - w)^\top \Omega (y - w) + \lambda \big(y^\top \Omega y - \varepsilon \big) \\ &= \max_{\lambda \geq 0} \ w^\top \Omega w - \lambda \varepsilon + \min_{y} \ \{ (1+\lambda) y^\top \Omega y - 2 y^\top \Omega w \} \qquad \text{(A.10a)} \\ &= \max_{\lambda \geq 0} \ \frac{\lambda}{1+\lambda} w^\top \Omega w - \lambda \varepsilon, \qquad \text{(A.10b)} \end{aligned}$$

where the interchanging of the inf-sup operators are justified because the feasible set of the primal problem (A.9) is non-empty and compact (Bertsekas, 2009, Proposition 5.5.4). For any $\lambda \geq 0$, the minimizer of the inner minimization problem in (A.10a) is

$$y^\star(\lambda) = \frac{w}{1+\lambda}.$$

Furthermore, this minimizer $y^\star(\lambda)$ is unique for any $\lambda \geq 0$ because the objective function of the inner minimization over y in (A.10a) is strictly convex in y. Substituting this optimal solution into the objective of (A.10a) leads to (A.10b), and substituting the value of w by $x - \widehat{\mu}$ leads to the reformulation (A.9).

We now study the maximizer $\lambda^\star$ of problem (A.10b). The Karush-Kuhn-Tucker condition asserts that there exists $\gamma^\star \in \mathbb{R}_+$ such that $(\lambda^\star, \gamma^\star)$ satisfy the system of algebraic equations

$$\begin{cases} (1+\lambda^\star)^{-2} w^\top \Omega w - \gamma^\star &= \ \varepsilon \\ \gamma^\star \lambda^\star &= \ 0 \\ \gamma^\star \ \geq \ 0, \ \lambda^\star &\geq \ 0. \end{cases}$$

If $w^\top \Omega w \leq \varepsilon$, then $\lambda^\star = 0$. If $w^\top \Omega w > \varepsilon$, then

$$\lambda^\star = \sqrt{\frac{w^\top \Omega w}{\varepsilon}} - 1.$$

In both cases, $\lambda^\star$ is unique. Substituting the value of $\lambda^\star$ into the objective function of (A.10b) gives the analytical expression for $g(\Omega)$.

We note that when $\varepsilon = 0$, we have $g(\Omega) = (x - \widehat{\mu})^\top \Omega (x - \widehat{\mu})$. The expressions for $\lambda^\star$ remain still valid in this case by taking the limit as $\varepsilon \downarrow 0$. Finally, the uniqueness of $\mu^\star$ follows from the uniqueness of $\lambda^\star$ and $y^\star(\lambda)$ obtained previously. The proof is thus completed. □

We are now ready to prove Theorem 3.1 in the main text.

Proof of Theorem 3.1. The optimistic nonparametric score evaluation problem can be decomposed using a two-layer formulation (6) as

$$\sup_{\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}})} \mathbb{Q}(\{x\}) = \sup_{(\mu, \Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})} \ \sup_{\mathbb{Q} \in \mathcal{M}(\mu, \Sigma)} \mathbb{Q}(\{x\}).$$

Using the result from Marshall & Olkin (1960) or Bertsimas & Popescu (2005, Theorem 6.1) to reformulate the inner supremum problem, we have

$$\sup_{\mathbb{Q} \in \mathcal{M}(\mu, \Sigma)} \mathbb{Q}(\{x\}) = \frac{1}{1 + (\mu - x)^\top \Sigma^{-1} (\mu - x)},$$

where the supremum is attained thanks to Bertsimas & Popescu (2005, Theorem 6.2) because the set $\{x\}$ is a singleton, and hence it is closed. This establishes equality (8a), where the maximization operator in the right hand side of (8a) is justified because $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ is compact by Lemma C.1 and the objective function is continuous over $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$.

It remains to find the optimal solution $(\mu^\star, \Sigma^\star)$ that solves the maximization problem (8a). If $x = \widehat{\mu}$ then the optimal value of problem (8a) is trivially 0. It suffices to consider the case when $x \neq \widehat{\mu}$. Define $\overline{\rho} \triangleq \rho + d + \log\det \widehat{\Sigma}$. Using a reparametrization $\Omega \leftarrow \Sigma^{-1}$, the maximizer $(\mu^\star, \Sigma^\star)$ also solves

$$\begin{array}{ll} \min & (\mu - x)^\top \Omega (\mu - x) \\ \text{s.t.} & \mu \in \mathbb{R}^d, \ \Omega \in \mathbb{S}^d_{++} \\ & (\mu - \widehat{\mu})^\top \Omega (\mu - \widehat{\mu}) + \operatorname{Tr}\big[\widehat{\Sigma}\Omega\big] - \log\det\Omega \leq \overline{\rho}. \end{array} \qquad \text{(A.11)}$$

This optimization problem with decision variables (μ, Ω) is still a non-convex optimization problem because of the multiplication terms between μ and Ω. However, it can be re-expressed as

$$\begin{array}{ll} \min & \begin{array}{ll} \min & (\mu - x)^\top \Omega (\mu - x) \\ \text{s.t.} & \mu \in \mathbb{R}^d,\ (\mu - \widehat{\mu})^\top \Omega (\mu - \widehat{\mu}) \le \overline{\rho} - \operatorname{Tr}\left[\widehat{\Sigma}\Omega\right] + \log\det\Omega \end{array} \\ \text{s.t.} & \Omega \in \mathbb{S}^d_{++},\ \operatorname{Tr}\left[\widehat{\Sigma}\Omega\right] - \log\det\Omega \le \overline{\rho}, \end{array}$$

where we note that the constraint $\operatorname{Tr}\left[\widehat{\Sigma}\Omega\right] - \log\det\Omega \le \overline{\rho}$ is redundant, but it is added to ensure that the inner problem over μ is feasible for any feasible value of Ω in the outer problem. Applying Lemma C.2 to solve the inner problem over μ for any given $\Omega \in \mathbb{S}^d_{++}$, problem (A.11) is equivalent to

$$\begin{array}{ll} \min & \max\limits_{\lambda \ge 0} \ -\lambda(\overline{\rho} - \operatorname{Tr}\left[\widehat{\Sigma}\Omega\right] + \log\det\Omega) + \frac{\lambda}{1+\lambda}(x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}) \\ \text{s.t.} & \Omega \in \mathbb{S}^d_{++},\ \operatorname{Tr}\left[\widehat{\Sigma}\Omega\right] - \log\det\Omega \le \overline{\rho}. \end{array}$$

For any $\widehat{\Sigma} \in \mathbb{S}^d_{++}$ and $\overline{\rho} = \rho + d + \log\det\widehat{\Sigma} \in \mathbb{R}$, the feasible set $\{\Omega \in \mathbb{S}^d_{++} : \operatorname{Tr}\left[\widehat{\Sigma}\Omega\right] - \log\det\Omega \le \overline{\rho}\}$ is compact[1] and convex. Moreover, the objective function is convex in Ω and concave in λ. Applying Sion's minimax theorem (Sion, 1958), we can interchange the operators and obtain an equivalent problem

$$\max_{\lambda \ge 0} \ \begin{array}{ll} \min & -\lambda(\overline{\rho} - \operatorname{Tr}\left[\widehat{\Sigma}\Omega\right] + \log\det\Omega) + \frac{\lambda}{1+\lambda}(x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}) \\ \text{s.t.} & \Omega \in \mathbb{S}^d_{++}\ ,\ \operatorname{Tr}\left[\widehat{\Sigma}\Omega\right] - \log\det\Omega \le \overline{\rho}. \end{array}$$

For any $\lambda \ge 0$, we can use a duality argument to reformulate the inner minimization, and we obtain the equivalent problem

$$\begin{aligned} &\max_{\lambda \ge 0} \inf_{\Omega \in \mathbb{S}^d_{++}} \max_{\nu \ge 0} \ -(\lambda + \nu)\overline{\rho} - (\lambda + \nu)\log\det\Omega + (\lambda + \nu)\operatorname{Tr}\left[\Omega\widehat{\Sigma}\right] + \frac{\lambda}{1+\lambda}(x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}) \\ = &\max_{\substack{\lambda \ge 0 \\ \nu \ge 0}} \inf_{\Omega \in \mathbb{S}^d_{++}} -(\lambda + \nu)\overline{\rho} - (\lambda + \nu)\log\det\Omega + (\lambda + \nu)\operatorname{Tr}\left[\Omega\widehat{\Sigma}\right] + \frac{\lambda}{1+\lambda}(x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}), \end{aligned}$$

where the interchange of the infimum operator with the innermost maximum operator is justified thanks to Bertsekas (2009, Proposition 5.5.4). Using a change of variables $\gamma \leftarrow \lambda + \nu$, problem (A.11) is equivalent to

$$\max_{\gamma \ge \lambda \ge 0} \left\{ \varphi(\gamma, \lambda) \triangleq \inf_{\Omega \in \mathbb{S}^d_{++}} -\gamma\overline{\rho} - \gamma\log\det\Omega + \gamma\operatorname{Tr}\left[\Omega\widehat{\Sigma}\right] + \frac{\lambda}{1+\lambda}(x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}) \right\}.$$

If $\gamma = \lambda = 0$, we have $\varphi(0, 0) = 0$. For any $\lambda \ge 0$ and $\gamma \ge \lambda$ such that $\gamma > 0$, the inner minimization admits the optimal solution

$$\Omega^\star(\lambda, \gamma) = \left(\widehat{\Sigma} + \frac{\lambda}{\gamma(1+\lambda)}(x - \widehat{\mu})(x - \widehat{\mu})^\top\right)^{-1}. \tag{A.12}$$

Furthermore, because $\gamma > 0$, the inner minimization problem has a strictly convex objective function over $\mathbb{S}^d_{++}$, in this case, the minimizer $\Omega^\star(\lambda, \gamma)$ is unique. By substituting the value of the minimizer $\Omega^\star(\lambda, \gamma)$, we obtain

$$\begin{aligned} \varphi(\gamma, \lambda) = &\ -\gamma\rho + \gamma\log\det\left(I_d + \frac{\lambda}{\gamma(1+\lambda)}\widehat{\Sigma}^{-\frac{1}{2}}(x - \widehat{\mu})(x - \widehat{\mu})^\top\widehat{\Sigma}^{-\frac{1}{2}}\right) \\ = &-\gamma\rho + \gamma\log\left(1 + \frac{\lambda}{\gamma(1+\lambda)}(x - \widehat{\mu})^\top\widehat{\Sigma}^{-1}(x - \widehat{\mu})\right) \end{aligned}$$

where $\widehat{\Sigma}^{-\frac{1}{2}}$ denotes the inverse of the unique principal square root of $\widehat{\Sigma}$. In the second equality, we have used Bernstein (2009, Fact 2.16.3) which implies that

$$\det(I_d + ab^\top) = 1 + b^\top a \qquad \forall (a, b) \in \mathbb{R}^d \times \mathbb{R}^d.$$

[1] Compactness follows from a reasoning similar to the proof of Lemma C.1, thus the details are omitted.

In the next step, we show that for any $\gamma \geq \lambda$, the optimal solution for the variable λ is $\lambda^\star(\gamma) = \gamma$. To this end, rewrite the above optimization problem as a two-layer optimization problem

$$\max_{\gamma \geq 0} \; \max_{\substack{\lambda \geq 0 \\ \lambda \leq \gamma}} \; \varphi(\gamma, \lambda).$$

This claim is trivial if $\gamma = 0$ because in this case, the only feasible solution for λ is $\lambda^\star(0) = 0$. If $\gamma > 0$, the gradient of φ in the variable λ satisfies

$$\frac{\partial \varphi}{\partial \lambda} = \frac{\gamma (x - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (x - \widehat{\mu})}{(1+\lambda)(\gamma(1+\lambda) + \lambda (x - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (x - \widehat{\mu}))} \geq 0 \quad \forall \lambda \in [0, \gamma],$$

which implies that at optimality, we have $\lambda^\star(\gamma) = \gamma$. Thus, we can eliminate the variable λ and obtain the equivalent univariate optimization problem

$$\max_{\gamma \geq 0} \; - \gamma \rho + \gamma \log \Big(1 + \frac{1}{1+\gamma} (x - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (x - \widehat{\mu}) \Big).$$

Converting this problem into a minimization problem gives the formulation (8d). By studying the objective function of (8d) and its gradient and Hessian[2], one can verify that this objective function is strictly convex and it tends to infinity as γ goes to infinity. This implies that the minimizer $\gamma^\star$ of (8d) exists and is unique. Let $\gamma^\star$ be the minimizer of (8d), one can reconstruct $\Sigma^\star$ from (A.12) and $\mu^\star$ from Lemma C.2, which gives the expression (8c). This observation completes the proof. □

D. Proof of Section 4

Proof of Theorem 4.1. Evaluating the optimistic score under the Gaussian assumption is equivalent to solving a non-convex minimization problem

$$\min \left\{ (\mu - x)^\top \Sigma^{-1} (\mu - x) + \log\det \Sigma : (\mu, \Sigma) \in \mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma}) \right\}, \tag{A.13}$$

where the minimization operator is justified by the compactness of the uncertainty set $\mathcal{U}_\rho(\widehat{\mu}, \widehat{\Sigma})$ in Lemma C.1. Define $\overline{\rho} \triangleq \rho + d + \log\det \widehat{\Sigma}$. Using a reparametrization $\Omega \leftarrow \Sigma^{-1}$, problem (A.13) admits an equivalent formulation

$$\begin{array}{ll} \min & \begin{array}{ll} \min & (\mu - x)^\top \Omega (\mu - x) - \log\det \Omega \\ \text{s.t.} & \mu \in \mathbb{R}^d, \; (\mu - \widehat{\mu})^\top \Omega (\mu - \widehat{\mu}) \leq \overline{\rho} - \operatorname{Tr}\big[\widehat{\Sigma}\Omega\big] + \log\det \Omega \end{array} \\ \text{s.t.} & \Omega \in \mathbb{S}^d_{++}, \; \operatorname{Tr}\big[\widehat{\Sigma}\Omega\big] - \log\det \Omega \leq \overline{\rho}, \end{array}$$

where we emphasize that the constraint $\operatorname{Tr}\big[\widehat{\Sigma}\Omega\big] - \log\det \Omega \leq \overline{\rho}$ is redundant to ensure the feasibility of the inner problem over μ for each admissible Ω. Applying Lemma C.2 to solve the inner problem over μ for any given $\Omega \in \mathbb{S}^d_{++}$, problem (A.13) is equivalent to

$$\left\{ \begin{array}{ll} \min & \max\limits_{\lambda \geq 0} \; -\lambda(\overline{\rho} - \operatorname{Tr}\big[\widehat{\Sigma}\Omega\big]) - (\lambda + 1) \log\det \Omega + \frac{\lambda}{1+\lambda} (x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}) \\ \text{s.t.} & \Omega \in \mathbb{S}^d_{++}, \; \operatorname{Tr}\big[\widehat{\Sigma}\Omega\big] - \log\det \Omega \leq \overline{\rho}. \end{array} \right.$$

Follow a similar steps as in the proof of Theorem 3.1, we find that problem (A.13) is equivalent to

$$\max_{\gamma \geq \lambda \geq 0} \left\{ \varphi(\gamma, \lambda) \triangleq \inf_{\Omega \in \mathbb{S}^d_{++}} -\gamma \overline{\rho} - (\gamma + 1) \log\det \Omega + \gamma \operatorname{Tr}\big[\Omega \widehat{\Sigma}\big] + \frac{\lambda}{1+\lambda} (x - \widehat{\mu})^\top \Omega (x - \widehat{\mu}) \right\}.$$

For any $\lambda \geq 0$ and $\gamma \geq \lambda$ such that $\gamma > 0$, the inner minimization admits the optimal solution

$$\Omega^\star(\lambda, \gamma) = \Big(\frac{\gamma}{1+\gamma} \widehat{\Sigma} + \frac{\lambda}{(1+\gamma)(1+\lambda)} (x - \widehat{\mu})(x - \widehat{\mu})^\top \Big)^{-1}, \tag{A.14}$$

[2] The closed form expressions can be found in Section E.

By substituting the value of the minimizer $\Omega^\star(\lambda,\gamma)$, we obtain

$$\varphi(\gamma,\lambda) = (d+\log\det\widehat{\Sigma}) - \gamma\rho - d(\gamma+1)\log\Big(1+\frac{1}{\gamma}\Big) + (1+\gamma)\log\Big(1+\frac{\lambda(x-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(x-\widehat{\mu})}{\gamma(1+\lambda)}\Big),$$

If $\gamma=\lambda=0$, we have $\varphi(0,0)=-\infty$ because the objective value in this case tends to $-\infty$ as Ω tends to $+\infty$. Thus without loss of optimality, we can omit the variable $\gamma=\lambda=0$ from the outer maximization problem because this set of solution is never optimal. Problem (A.13) is hence equivalent to the following two-layer optimization problem

$$\max_{\gamma>0}\ \max_{0\le\lambda\le\gamma}\ \varphi(\gamma,\lambda), \tag{A.15}$$

where we emphasize that the feasible set for γ is over the open set $(0,+\infty)$. For any $\gamma>0$, the gradient of φ in λ satisfies

$$\frac{\partial\varphi}{\partial\lambda} = \frac{\gamma(1+\gamma)(x-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(x-\widehat{\mu})}{\gamma(1+\lambda)+\lambda(x-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(x-\widehat{\mu})} \ge 0 \qquad \forall\lambda\in[0,\gamma],$$

which implies that the inner maximization problem in (A.15) admits the optimal solution $\lambda^\star(\gamma)=\gamma$. We thus have

$$\max_{\gamma>0}\ (d+\log\det\widehat{\Sigma}) - \gamma\rho - d(\gamma+1)\log\Big(1+\frac{1}{\gamma}\Big) + (1+\gamma)\log\Big(1+\frac{(x-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(x-\widehat{\mu})}{(1+\gamma)}\Big)$$

Dropping the constant term in the objective function and converting the problem into the minimization form results in problem (11d) . By studying the objective function of (11d) and its gradient and Hessian[3], one can verify that this objective function is strictly convex and it tends to infinity as γ goes to infinity. This implies that the minimizer $\gamma^\star$ of (11d) exists and is unique. Let $\gamma^\star$ be the minimizer of (11d), one can reconstruct $\Sigma^\star$ from (A.14) and $\mu^\star$ from Lemma C.2, which give expression (11c). This finishes the proof. □

E. Calculations of the Gradients and Hessians

Throughout this section, we use the shorthand $\alpha=(x-\widehat{\mu})^\top\widehat{\Sigma}^{-1}(x-\widehat{\mu})\ge 0$. Denote momentarily by $\varphi_1:\mathbb{R}_+\to\mathbb{R}$ the objective function of problem (8d), that is,

$$\varphi_1(\gamma) = \gamma\rho - \gamma\log\Big(1+\frac{\alpha}{1+\gamma}\Big).$$

The gradient and Hessian of φ_1 are

$$\begin{aligned}\frac{\partial\varphi_1}{\partial\gamma} &= \rho - \log\Big(1+\frac{\alpha}{1+\gamma}\Big) + \frac{\gamma\alpha}{(1+\gamma)[1+\gamma+\alpha]},\\ \frac{\partial^2\varphi_1}{\partial\gamma^2} &= \frac{\alpha(2+2\gamma+2\alpha+\alpha\gamma)}{(1+\gamma)^2(1+\gamma+\alpha)^2} \ge 0.\end{aligned}$$

Now, denote momentarily by $\varphi_2:\mathbb{R}_{++}\to\mathbb{R}$ the objective function of problem (11d), that is,

$$\varphi_2(\gamma) = \gamma\rho + d(\gamma+1)\log\Big(1+\frac{1}{\gamma}\Big) - (1+\gamma)\log\Big(1+\frac{\alpha}{(1+\gamma)}\Big).$$

The gradient and Hessian of φ_2 are

$$\begin{aligned}\frac{\partial\varphi_2}{\partial\gamma} &= \rho + d\left[\log\Big(1+\frac{1}{\gamma}\Big) - \frac{1}{\gamma}\right] - \left[\log\Big(1+\frac{\alpha}{1+\gamma}\Big) - \frac{\alpha}{1+\gamma+\alpha}\right],\\ \frac{\partial^2\varphi_2}{\partial\gamma^2} &= \frac{d}{\gamma^2(1+\gamma)} + \frac{\alpha^2}{(1+\gamma+\alpha)^2(1+\gamma)} \ge 0.\end{aligned}$$

[3] The closed form expressions can be found in Section E.

References

Bernstein, D. S. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

Bertsekas, D. *Convex Optimization Theory*. Athena Scientific, 2009.

Bertsimas, D. and Popescu, I. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005.

Marshall, A. W. and Olkin, I. Multivariate Chebyshev inequalities. *The Annals of Mathematical Statistics*, 31(4):1001–1014, 1960.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press, 1998.