

Information Theory Notes

Charles Zheng and Yuval Benjamini

December 5, 2015

These are preliminary notes.

1 Classification in high-dimension, fixed SNR regime

We observe a data point y_* which belongs to one of K classes. The distribution in the i th class is $N(\mu_i, \Omega)$. We have another dataset with r repeats per class, which we use to estimate the centroids μ_i : we obtain estimates $\hat{\mu}_i \sim N(\mu_i, r^{-1}\Omega)$. The class centroids were originally drawn i.i.d. from a multivariate normal $N(0, I)$. Furthermore Ω is unknown and have to be estimated as well: assume we have obtained estimate $\hat{\Omega}$ via some method. Without loss of generality, take the K th class to be the true class of y_* . Write $\hat{\mu}_* = \hat{\mu}_K$.

The classification rule is given by

$$\text{Estimated class} = \operatorname{argmin}_i (y_* - B\hat{\mu}_i)^T A (y_* - B\hat{\mu}_i)$$

where A and B are matrices based on $\hat{\Omega}$. The Bayes rule is given by

$$A_{\text{Bayes}} = (I + \Omega - (I + r^{-1}\Omega)^{-1})^{-1}$$

$$B_{\text{Bayes}} = (I + r^{-1}\Omega)^{-1}.$$

The “plug-in” estimates of A and B are

$$A = (I + \hat{\Omega} + (I + r^{-1}\hat{\Omega})^{-1})^{-1}$$

$$B = (I + r^{-1}\hat{\Omega})^{-1}.$$

Note that

$$(y_* - B\hat{\mu}_i)^T A(y_* - B\hat{\mu}_i) = \|A^{1/2}y_* - A^{1/2}B\hat{\mu}_i\|^2.$$

Therefore the classification rule is

$$\text{Estimated class} = \operatorname{argmin}_i Z_i,$$

where

$$Z_i = \|A^{1/2}y_* - A^{1/2}B\hat{\mu}_i\|^2.$$

We have

$$\begin{bmatrix} A^{1/2}y \\ A^{1/2}B\hat{\mu}_* \\ A^{1/2}B\hat{\mu}_i \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A^{1/2}(I + \Omega)A^{1/2} & A^{1/2}BA^{1/2} & 0 \\ A^{1/2}B(I + \frac{\Omega}{r})BA^{1/2} & & 0 \\ A^{1/2}B(I + \frac{\Omega}{r})BA^{1/2} & & \end{bmatrix} \right)$$

Therefore

$$\begin{aligned} \mathbf{E}Z_i &= \begin{cases} \operatorname{tr}[A(I + \Omega + (B(I + r^{-1}\Omega)B))] & \text{for } i \neq K \\ \operatorname{tr}[A(I + \Omega + (B(I + r^{-1}\Omega)B) - 2B)] & \text{for } i = K \end{cases}, \\ \operatorname{Cov}(Z_i, Z_j) &= \begin{cases} 2\operatorname{tr}[A(I + \Omega)]^2 & \text{for } i \neq j \neq K \\ 2\operatorname{tr}[A(I + \Omega - B)]^2 & \text{for } i = K, j \neq K \\ 2\operatorname{tr}[A(I + \Omega + B(I + r^{-1}\Omega)B)]^2 & \text{for } i = j \neq K \\ 2\operatorname{tr}[A(I + \Omega + B(I + r^{-1}\Omega)B - 2B)]^2 & \text{for } i = j = K \end{cases}. \end{aligned}$$

1.1 Equivalence of limiting MI and multivariate SNR

The mutual information between μ and y is given by

$$\text{MI} = -\log \det(I - (I + \Omega)^{-1}).$$

We will show that if $\operatorname{tr}[\Omega^{-1}] \rightarrow c$, and $\operatorname{tr}[\Omega^{-2}] \rightarrow 0$, then

$$\begin{aligned} \lim_{p \rightarrow \infty} \text{MI} &= \lim_{p \rightarrow \infty} -\log \det(I - (I + \Omega)^{-1}) \\ &= \lim_{p \rightarrow \infty} -\log \det(I - \Omega^{-1/2}(\Omega^{-1} + I)^{-1}\Omega^{-1/2}) \\ &= \lim_{p \rightarrow \infty} -\log \det(I - \Omega^{-1} + \Omega^{-2}) \\ &= \lim_{p \rightarrow \infty} -\log \det(I - \Omega^{-1}) \\ &= \lim_{p \rightarrow \infty} \log \det(I + \Omega^{-1}) \\ &= \lim_{p \rightarrow \infty} \operatorname{tr}[\Omega^{-1}] = c \end{aligned}$$

1.2 Ω known, fixed r

Suppose

$$\hat{\Omega} = \Omega.$$

Then,

$$\mathbf{E}Z_i = \begin{cases} p + 2\text{tr}[AB] & \text{for } i \neq K \\ p & \text{for } i = K \end{cases},$$

and

$$\text{Cov}(Z_i, Z_j) = \begin{cases} 2\text{tr}[I + AB]^2 & \text{for } i \neq j \neq K \\ 2p & \text{for } i = K, j \neq K \\ \text{tr}[I + 2AB]^2 & \text{for } i = j \neq K \\ 2p & \text{for } i = j = K \end{cases}.$$

Let $\gamma_i(r)$ denote the eigenvalues of AB : we have

$$\gamma_i(r) = \frac{1}{(\omega_i^2 + \omega_i)/r + \omega_i}$$

where ω_i are the eigenvalues of Ω .

The misclassification probability is

$$\text{MC} = 1 - \Pr[N(\mu(r), \sigma^2(r)) < M_{K-1}]$$

where M_{K-1} is the maximum of $K-1$ independent standard normal variates, and

$$\begin{aligned} \mu(r) &= \frac{2 \sum_{i=1}^p \gamma_i}{\sqrt{\sum_{i=1}^p 6\gamma_i^2 + 4\gamma_i}}, \\ \sigma^2(r) &= \frac{\sum_{i=1}^p \gamma_i^2 + 2\gamma_i}{\sum_{i=1}^p 3\gamma_i^2 + 2\gamma_i}. \end{aligned}$$

Under the condition that

$$\text{tr}[\Omega^{-1}] \rightarrow \text{const.}, \text{tr}[\Omega^{-2}] \rightarrow 0$$

we have

$$\begin{aligned} \mu(r) &\rightarrow \sqrt{\sum_{i=1}^p \gamma_i(r)}, \\ \sigma^2(r) &\rightarrow 1. \end{aligned}$$

1.3 Known Ω , growing r

Assume that Ω/p has a limiting spectrum, i.e. defining

$$H^{(p)} = \sum_{i=1}^p \frac{1}{p} \delta_{\omega_i/p}$$

we have

$$H^{(p)} \rightarrow H$$

for some probability measure H on \mathbb{R}^+ . Note then that

$$\text{MI} = \lim_{p \rightarrow \infty} \text{tr}[\Omega^{-1}] = \int \nu^{-1} dH(\nu).$$

Also, suppose that $\frac{r}{p} \rightarrow \kappa$. Then, define

$$\begin{aligned} \mu^*(\kappa) &= \lim_{p \rightarrow \infty} \mu(r) = \lim_{p \rightarrow \infty} \frac{2 \sum_{i=1}^p \gamma_i}{\sqrt{\sum_{i=1}^p 6\gamma_i^2 + 4\gamma_i}}, \\ &= \lim_{p \rightarrow \infty} \frac{2p \int \frac{1}{\frac{p^2 \nu^2 + p\nu}{r} + p\nu} dH(\nu)}{\sqrt{p \int 6 \left(\frac{1}{\frac{p^2 \nu^2 + p\nu}{r} + p\nu} \right)^2 + 4 \left(\frac{1}{\frac{p^2 \nu^2 + p\nu}{r} + p\nu} \right) dH(\nu)}} \\ &= \lim_{p \rightarrow \infty} \sqrt{p \int \frac{1}{\frac{p^2 \nu^2 + p\nu}{r} + p\nu} dH(\nu)} \\ &= \lim_{p \rightarrow \infty} \sqrt{\int \frac{1}{\frac{p\nu^2}{r} + (1 + 1/r)\nu} dH(\nu)} \\ &= \sqrt{\int \lim_{p \rightarrow \infty} \frac{1}{\frac{p\nu^2}{r} + (1 + 1/r)\nu} dH(\nu)} \\ &= \sqrt{\int \frac{1}{\frac{\nu^2}{\kappa} + \nu} dH(\nu)}, \end{aligned}$$

while for similar reasons

$$\lim_{p \rightarrow \infty} \sigma^2(r) = 1.$$

Thus, we have $\mu^*(\kappa)$ increasing in κ , and tending to

$$\mu^*(\infty) = \int \nu^{-1} H(d\nu)$$

as $\kappa \rightarrow \infty$. But, note that $\mu^*(\infty)$ is also the limiting mutual information.

Therefore, we get the following result.

Theorem. *Let H be a positive measure. For Ω known, if $p\Omega$ has limiting spectrum H , and $\lim r/p = \kappa$, then the asymptotic misclassification rate $MC^* = \lim_{p \rightarrow \infty} MC$ is a function only of κ , is increasing in κ , and is bounded between*

$$MC^*(0) = 1 - \Pr[N(0, 1) < M_{K-1}] < 1$$

and

$$MC^*(\infty) = 1 - \Pr[N(\mu^*(\infty), 1) < M_{K-1}] > 0,$$

where

$$\mu^*(\infty) = \int \nu^{-1} H(d\nu) = \lim_{p \rightarrow \infty} MI.$$

To summarize, the misclassification rate with a fixed number of repeats r converges to a constant, $MC^*(0) < 1$, while the Bayes misclassification rate is given by $MC^*(\infty) > 0$. To get an asymptotically constant misclassification rate between $MC^*(0)$ and $MC^*(\infty)$, one needs a number of repeats r which is of order $O(p)$.

In the case that one uses $r = \kappa p$ repeats, the asymptotic misclassification rate as a function of K matches the *Bayes* rate of a classification problem with covariance $\tilde{\Omega}$, where the eigenvalues of $\tilde{\Omega}$ are given by $\omega_i + \frac{\omega_i^2}{p\kappa}$; i.e. the eigenvalues of $\tilde{\Omega}$ are inflated compared to Ω . This means that if we were to naively interpret the obtained misclassification rates as Bayes misclassification rates, the estimated mutual information would be the true mutual information of $\tilde{\Omega}$, given by

$$\hat{MI} = \int 1/(\eta^2/\kappa + \eta) H(d\eta).$$

The asymptotic difference between the true mutual information and the naive

estimate is

$$\begin{aligned}
\text{MI} - \hat{\text{MI}} &= \int \frac{1}{\eta} - \frac{1}{\eta^2/\kappa + \eta} H(d\eta) \\
&= \int \frac{\eta^2/\kappa}{\eta^2 + \eta^3/\kappa} H(d\eta) \\
&= \int \kappa^{-1} \frac{1}{1 + \eta/\kappa} H(d\eta) \\
&\leq \int \kappa^{-1} H(d\eta) = \kappa^{-1}.
\end{aligned}$$

Hence we have a lower bound for the error of MI estimation.

Theorem. *For Ω known, if $p\Omega$ has limiting spectrum H , and $\lim r/p = \kappa$, then the asymptotic minimax risk of estimating MI based on the misclassification curve is upper bounded by*

$$\liminf_{p \rightarrow \infty} \inf_{\hat{\text{MI}}} \mathbf{E}[|\text{MI} - \hat{\text{MI}}|] \leq \frac{1}{\kappa}.$$

2 Appendix

2.1 Gaussian min probs

Define

$$f_{ng}(\mu, \sigma^2, K) = \Pr[\sigma Z_* + \mu < \max_{i=1}^K Z_i]$$

for Z_*, Z_1, \dots, Z_K i.i.d normal.

Suppose

$$\begin{bmatrix} y_* \\ y_1 \\ \vdots \\ y_{K-1} \end{bmatrix} \sim N \left(\begin{bmatrix} a \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} b & c & \dots & c \\ c & d & \dots & e \\ \dots & \dots & \ddots & \vdots \\ c & e & \dots & d \end{bmatrix} \right).$$

where $d > e > \frac{c^2}{b}$.

Then

$$\Pr[y_* < \min_{i=1}^{K-1} y_i] = 1 - f_{ng} \left(-\frac{a}{\sqrt{d-e}}, \frac{b+e-2c}{d-e}, K-1 \right).$$