Supplementary Material for Distributed non-disclosive validation of predictive models by a modified ROC-GLM

Daniel Schalk[1,3,4], Verena S. Hoffmann[2,3], Bernd Bischl[1,4] and Ulrich Mansmann[1,2,3]

[1] Department of Statistics, LMU Munich, Munich
[2] Institute for Medical Information Processing, Biometry and Epidemiology, LMU Munich, Munich
[3] DIFUTURE (DataIntegration for Future Medicine, www.difuture.de), LMU Munich, Munich
[4] Munich Center for Machine Learning (MCML)

Appendix

A.1. Decomposition of score vector and Fisher information

$$\begin{aligned}\mathcal{V}(\hat{\theta}_m) &= \left[\frac{\partial \ell_\theta(\mathcal{D})}{\partial \theta}\right]_{\theta=\hat{\theta}_m} = \left[\frac{\partial \sum_{i=1}^n \log(f_Y(y_i, x_i))}{\partial \theta}\right]_{\theta=\hat{\theta}_m} \\ &= \left[\sum_{i=1}^n \frac{\partial \log(f_Y(y_i, x_i))}{\partial \theta}\right]_{\theta=\hat{\theta}_m} = \sum_{k=1}^K \left[\sum_{(y,x)\in\mathcal{D}^{(k)}} \frac{\partial \log(f_Y(y, x))}{\partial \theta}\right]_{\theta=\hat{\theta}_m} \\ &= \sum_{k=1}^K \left[\frac{\partial \log(\ell_\theta(\mathcal{D}^{(k)}))}{\partial \theta}\right]_{\theta=\hat{\theta}_m} = \sum_{k=1}^K \mathcal{V}_k(\hat{\theta}_m)\end{aligned}$$

$$\begin{aligned}\mathcal{I}(\hat{\theta}_m) &= \left[\frac{\partial \mathcal{V}(\theta)}{\partial \theta}\right]_{\theta=\hat{\theta}_m} = \left[\frac{\partial \sum_{k=1}^K \mathcal{V}_k(\hat{\theta}_m)}{\partial \theta}\right]_{\theta=\hat{\theta}_m} = \\ &= \sum_{k=1}^K \left[\frac{\partial \mathcal{V}_k(\hat{\theta}_m)}{\partial \theta}\right]_{\theta=\hat{\theta}_m} = \sum_{k=1}^K \mathcal{I}_k(\hat{\theta}_m)\end{aligned}$$

A.2. Illustration of the Gaussian mechanism

The noise r added in the Gaussian mechanism to obtain a noise representation of scores $\hat{f}(x)$ highly depends on the three values of $\Delta_2(f)$, ε, and δ. In our approach, we first examine the value of $\Delta_2(f)$ and then set ε and δ accordingly to not introduce too much noise and hence worsen the accuracy. Figure 1 and Figure 2 visualize the Gaussian mechanism for $\Delta_2(f) \in \{0.01, 0.05\} = 0.01$, $\varepsilon \in \{0.1, 0.3, 0.5\}$ and $\delta \in \{0.1, 0.3, 0.5\}$ to roughly judge how the noise distorts the score values $f(x)$. The figures contain the score values (upper labels), the value of the variance τ^2 (which is based on $\Delta_2(f)$, ε, and δ), the respective density of the normal distribution, and the noisy score values (lower labels). Both figures also visualizes how the Gaussian mechanism changes the order of the scores depending on the noise.

A.3. Number of observations per bin for the calibration curves

Site k	**Bin**									
	$(0, .1]$	$(.1, .2]$	$(.2, .3]$	$(.3, .4]$	$(.4, .5]$	$(.5, .6]$	$(.6, .7]$	$(.7, .8]$	$(.8, .9]$	$(0.9, 1]$
1	**12**	**11**	**13**	3	2	**7**	**5**	0	0	0
2	**11**	**14**	**9**	1	4	**5**	2	0	0	0
3	**13**	**12**	**12**	**5**	3	4	**7**	1	0	0
4	**8**	**6**	**9**	**5**	**9**	**6**	**5**	0	0	0
5	**13**	**13**	**10**	1	**6**	**5**	**9**	1	0	0
$\sum$	**57**	**56**	**53**	**15**	**24**	**27**	**28**	2	0	0

Table 1: Number of observations per bin. Values in these bins are shared only if the numbers per bin are larger than 5. The values for which this applies are highlighted as bold numbers.

References

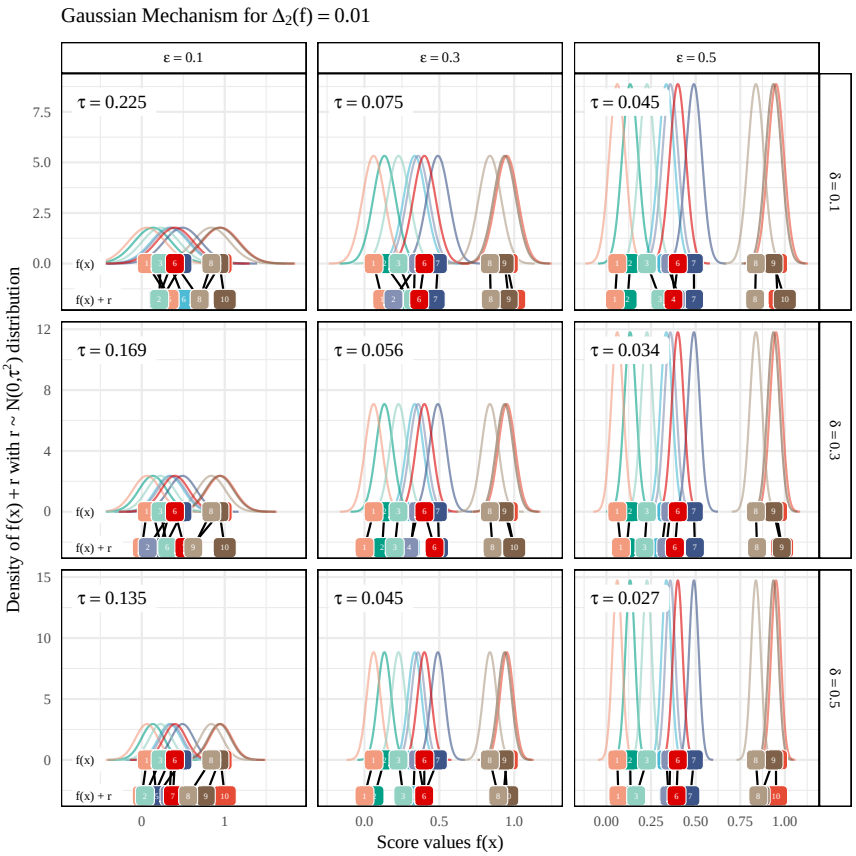

Figure 1: Visualization of how the Gaussian mechanism adds noise to the original score values $f(x)$ (upper labels) to obtain a noisy representation $f(x) + r$ (lower labels) with $r \sim \mathcal{N}(0, \tau^2)$. At each point in $f(x)$, the corresponding density of $N(f(x), \tau^2$ is added to visualize how the Gaussian mechanism shuffles the order (lines between the two labels of $f(x)$ and $f(x) + r$) of the score values, depending on the variance τ^2. The whole mechanism depends on the ℓ_2-sensitivity, which is set to 0.01 here.

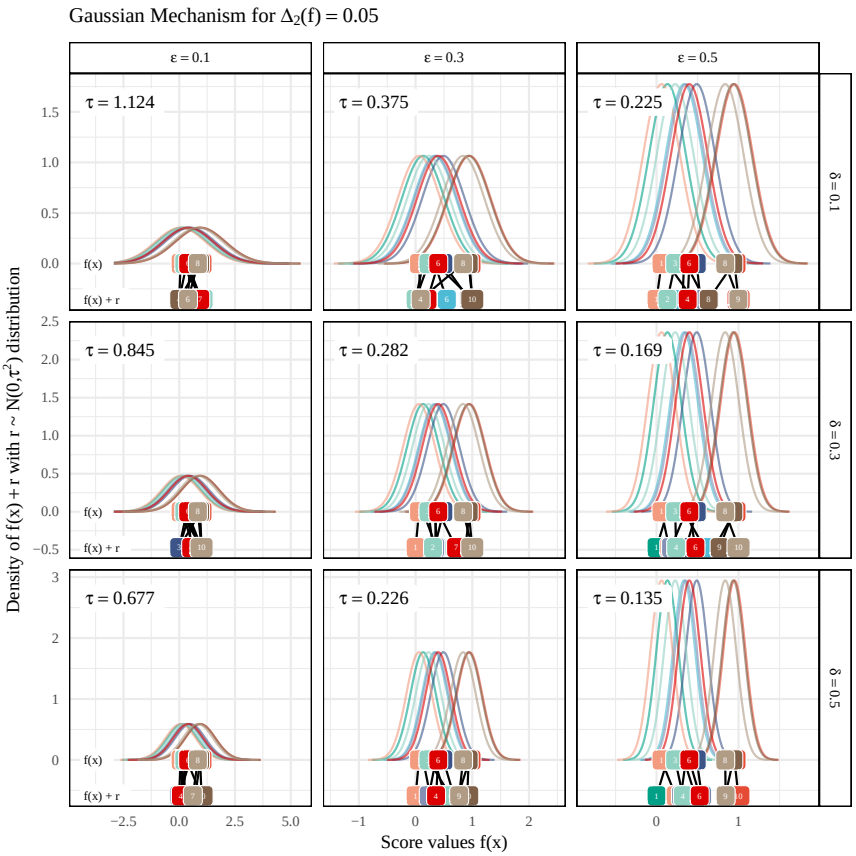

Figure 2: Visualization of how the Gaussian mechanism adds noise to the original score values $f(x)$ (upper labels) to obtain a noisy representation $f(x) + r$ (lower labels) with $r \sim \mathcal{N}(0, \tau^2)$. At each point in $f(x)$, the corresponding density of $N(f(x), \tau^2$ is added to visualize how the Gaussian mechanism shuffles the order (lines between the two labels of $f(x)$ and $f(x) + r$) of the score values, depending on the variance τ^2. The whole mechanism depends on the ℓ_2-sensitivity, which is set to 0.05 here.