LDP two-sample chi-squared test

1 Setting

- $\boldsymbol{Y}_i \overset{iid}{\sim} multi(n_1, \boldsymbol{p}_{\boldsymbol{Y}}), \boldsymbol{Z}_i \overset{iid}{\sim} multi(n_2, \boldsymbol{p}_{\boldsymbol{Z}})$ with k categories
- One-hot vector form i.e. random vectors with dependent Bernoulli random variable entries
- Allow for $n_1 \neq n_2$

2 Generalized Randomized Response and two sample Pearson chi-square statistic

2.1 Privacy mechanism: Generalized Randomized Response

Definition 2.1 (Generalized Randomized Response (Theorem 5.4. of Gaboardi and Rogers [1]))**.** *For a multinomial random vector $\boldsymbol{Y}_i \overset{iid}{\sim} multi(n_1, \boldsymbol{p}_{\boldsymbol{Y}})$, we define*

$$\mathbb{P}\big(\mathcal{M}_{\mathtt{GenRR}}(\boldsymbol{Y}_i) = \boldsymbol{y}' | \boldsymbol{Y}_i = \boldsymbol{y}\big) := \begin{cases} \dfrac{\exp(\alpha)}{\exp(\alpha)+k-1} & \text{if } \boldsymbol{y}' = \boldsymbol{y} \\ \dfrac{1}{\exp(\alpha)+k-1} & \text{if } \boldsymbol{y}' \neq \boldsymbol{y}. \end{cases}$$

Then $\tilde{\boldsymbol{Y}}_i := \mathcal{M}_{\mathtt{GenRR}}(\boldsymbol{Y}_i)$ is a multinomial random vector with probability vector

$$\tilde{\boldsymbol{p}}_{\boldsymbol{Y}} := \boldsymbol{p}_{\boldsymbol{Y}} \frac{\exp(\alpha)}{\exp(\alpha)+k-1} + (1-\boldsymbol{p}_{\boldsymbol{Y}}) \frac{1}{\exp(\alpha)+k-1}.$$

Since $e^\alpha > 1$ for $\alpha > 0$, the probability of sending the original category is a little bit higher than sending the other category. Gaboardi and Rogers [1] constructs a private goodness-of-fit test based on a chi-square statistic evaluated on $\tilde{\boldsymbol{Y}}_i$'s. They demonstrate that the limiting distribution is chi-square distribution both under the null and alternative.

2.2 Two sample chi-square statistic

We extend the goodness-of-fit test by Gaboardi and Rogers [1] into two-sample testing by privatizing the raw samples $\boldsymbol{Z}_i \overset{iid}{\sim} multi(n_2, \boldsymbol{p}_{\boldsymbol{Z}})$ into $\tilde{\boldsymbol{Z}}_j := \mathcal{M}_{\mathtt{GenRR}}(\boldsymbol{Z}_j)$. Under the null, $\mathcal{M}_{\mathtt{GenRR}}(\boldsymbol{Y}_i)$ and $\mathcal{M}_{\mathtt{GenRR}}(\boldsymbol{Z}_j)$ follow multinomial distributions with the same probability vector. Therefore, the usual two-sample chi-square test statistic

$$T_\chi := \sum_{\ell=1}^{k} \frac{\big(n_2 \sum_{i=1}^{n_1} \tilde{\boldsymbol{Y}}_i(\ell) - n_1 \sum_{j=1}^{n_1} \tilde{\boldsymbol{Z}}_j(\ell)\big)^2}{n_1 n_2 (n_1+n_2) \sum_{j=1}^{n_1} \big(\tilde{\boldsymbol{Y}}_j(\ell) + \tilde{\boldsymbol{Z}}_j(\ell)\big)}$$

converges to a chi-square distribution with degree of freedom $k-1$ and yields a valid test with size γ. This test statistic is from Van der Vaart's book Asymptotic Statistics, pp. 253.

3 Bit flip privatization and related test statisitc

3.1 Bit flip privatization

We next consider another LDP algorithm $\mathcal{M}_{bit} : \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\} \to \{0,1\}^k$, which is the Algorithm 4 of Gaboardi and Rogers [1]. It flips each bit with some biased probability. The Algorithm 1 is

Algorithm 1 Bit Flip Local Randomizer: $\mathcal{M}_{bit}$

Input: $\boldsymbol{y} \in \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_k\}$, α.
  for $\ell \in [k]$ **do**
    Set $\tilde{\boldsymbol{y}}_\ell = \boldsymbol{y}_\ell$ with probability $e^{\alpha/2}/(e^{\alpha/2}+1)$, otherwise $\tilde{\boldsymbol{y}}_\ell = 1 - \boldsymbol{y}_\ell$
  end for
  Output: $\tilde{\boldsymbol{y}}$

α-LDP (Theorem 5.5 of Gaboardi and Rogers [1]).)

3.2 test statistic

We first review the one-sample test statistic of Gaboardi and Rogers [1] and expand it into two-sample statistic.

3.2.1 Review of one-sample statistic

We first review how Gaboardi and Rogers [1] builds goodness-of-fit chi-square statistic for histogram of bit-flipped observations. We start with applying CLT to the bit-flipped obsevations.

Lemma 3.1 (Applying CLT to bit-flipped observations, Lemma 5.7 of [1])**.** *When $Y_i \overset{iid}{\sim} multinomial(\boldsymbol{p}, 1)$, the mean vector and covariance matrix of the flipped observation are computed as follows:*

$$\tilde{\boldsymbol{p}} := \mathbb{E}(\mathcal{M}_{bit}(Y_1)) = \frac{\big(\exp(\alpha/2) - 1\big)\boldsymbol{p} + 1}{\exp(\alpha/2) + 1}, \text{ and} \tag{1}$$

$$\Sigma_{\tilde{\boldsymbol{p}}} := Var(\mathcal{M}_{bit}(Y_1)) = \left(\frac{\exp(\alpha/2) - 1}{\exp(\alpha/2) + 1}\right)^2 \big(diag(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\big) + \frac{\exp(\alpha/2)}{(\exp(\alpha/2)+1)^2} I_d, \tag{2}$$

For any $\alpha > 0$ and $\boldsymbol{p} > 0$, $\Sigma_{\tilde{\boldsymbol{p}}}$ is positive definite and one of its eigenvector is one-vector. Denote the histogram of flipped observations as

$$\tilde{\boldsymbol{H}} := \sum_{i=1}^{n} \mathcal{M}_{bit}(Y_i). \tag{3}$$

By the CLT for i.i.d random vectors, we get the following asymptotic distribution:

$$\sqrt{n}(\tilde{\boldsymbol{H}}/n - \tilde{\boldsymbol{p}}) \xrightarrow{d} N\big(0, \Sigma_{\tilde{\boldsymbol{p}}}\big) \tag{4}$$

In non-private chi-square test, we apply CLT and multiply by $diag(\boldsymbol{p})^{-1/2}$ to turn the covariance matrix on the RHS into a projector matrix. Here in the private setting, we also need a scaling matrix to turn the covariance matrix into a projector matrix. Gaboardi and Rogers [1] proposes $\tilde{\boldsymbol{p}}^{-1/2}\Pi$, where $\Pi := I_k - \frac{1}{k}\mathbf{1}\mathbf{1}^\top$. The properties of Π are as follows:

1. It is symmetric idempotent (a projecter matrix).
2. Its null space is $span\{\mathbf{1}\}$, so when multiplied to a symmetric matrix, it deletes an eigenvector $\mathbf{1}$.

$$\begin{aligned}\Pi x = 0 &\iff x = (1/k)\mathbf{1}\mathbf{1}^\top x \\ &\iff x = (1/k)\mathbf{1}(\mathbf{1}^\top x) = ((\mathbf{1}^\top x)/k)\mathbf{1} = c\mathbf{1}\end{aligned}$$

By multiplying $\Sigma_{\tilde{\boldsymbol{p}}}^{-1/2}\Pi$ to the LHS vector of (4), we get

$$\sqrt{n}\Sigma_{\tilde{\boldsymbol{p}}}^{-1/2}\Pi(\tilde{\boldsymbol{H}}/n - \tilde{\boldsymbol{p}}) \xrightarrow{d} N\big(0, \Sigma_{\tilde{\boldsymbol{p}}}^{-1/2}\Pi\Sigma_{\boldsymbol{p}}\Pi\Sigma_{\tilde{\boldsymbol{p}}}^{-1/2}\big). \tag{5}$$

The next lemma specifies the property of the covariance matrix $\Sigma_{\tilde{\boldsymbol{p}}}^{-1/2}\Pi\Sigma_{\boldsymbol{p}}\Pi\Sigma_{\tilde{\boldsymbol{p}}}^{-1/2}$.

Lemma 3.2. *Let $\Sigma \in \mathbb{R}^{k\times k}$ be a symmetric positive definite matrix one of whose eigenvector is $\mathbf{1}$. Then we can diagonalize as $\Sigma = BDB^T$. Let $\Pi := I_k - \frac{1}{k}\mathbf{1}\mathbf{1}^\top$. Then the following matrix is the identity matrix except one of the entries on the diagonal is zero.*

$$\Sigma^{-1/2}\Pi\Sigma\Pi\Sigma^{-1/2} = \Sigma^{-1/2}\Pi BDB^\top\Pi\Sigma^{-1/2} \tag{6}$$

Now we invoke the following classical theorem to derive an asymptotic chi-square distribution with degree of freedom $k-1$.

Theorem 3.1 (Ferguson (1996))**.** *If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \Sigma)$ and Σ is a projection matrix of rank ν an $\Sigma\boldsymbol{\mu} = \boldsymbol{\mu}$ then $\boldsymbol{X}^\top\boldsymbol{X} \sim \chi^2_\nu(\boldsymbol{\mu}^\top\boldsymbol{\mu})$.*

3.3 Extension to two sample test statistic

We extend the previous result to two-sample setting. For simplicity, we follow the equal sample size setting of Gaboardi and Rogers [1]. Now we have two collections of raw data $\{\boldsymbol{Y}_i\}_{i\in[n]}$ and $\{\boldsymbol{Z}_j\}_{j\in[n]}$ generated from multinomial distributions with probability vectors $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$, respectiveley. According to Lemma 3.1, flipped observations of these samples have following moments:

$$\begin{aligned}
\tilde{\boldsymbol{p}}_Y &:= \mathbb{E}(\mathcal{M}_{bit}(Y_1)) = \frac{\big(\exp(\alpha/2)-1\big)\boldsymbol{p}_{\boldsymbol{Y}} + 1}{\exp(\alpha/2)+1}, \\
\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}} &:= Var(\mathcal{M}_{bit}(Y_1)) = \left(\frac{\exp(\alpha/2)-1}{\exp(\alpha/2)+1}\right)^2 \big(diag(\boldsymbol{p}_{\boldsymbol{Y}}) - \boldsymbol{p}_{\boldsymbol{Y}}\boldsymbol{p}_{\boldsymbol{Y}}^\top\big) + \frac{\exp(\alpha/2)}{(\exp(\alpha/2)+1)^2}I_d \\
\tilde{\boldsymbol{p}}_Z &:= \mathbb{E}(\mathcal{M}_{bit}(Z_1)) = \frac{\big(\exp(\alpha/2)-1\big)\boldsymbol{p}_{\boldsymbol{Z}} + 1}{\exp(\alpha/2)+1}, \\
\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}} &:= Var(\mathcal{M}_{bit}(Z_1)) = \left(\frac{\exp(\alpha/2)-1}{\exp(\alpha/2)+1}\right)^2 \big(diag(\boldsymbol{p}_{\boldsymbol{Z}}) - \boldsymbol{p}_{\boldsymbol{Z}}\boldsymbol{p}_{\boldsymbol{Z}}^\top\big) + \frac{\exp(\alpha/2)}{(\exp(\alpha/2)+1)^2}I_d
\end{aligned}$$

We also denote the histograms of flipped observations as

$$\tilde{\boldsymbol{H}_{\boldsymbol{Y}}} := \sum_{i=1}^{n} \mathcal{M}_{bit}(Y_i) \text{ and } \tilde{\boldsymbol{H}_{\boldsymbol{Z}}} := \sum_{i=1}^{n} \mathcal{M}_{bit}(Z_i). \tag{7}$$

Then we have the following asymptotic distribution:

$$\sqrt{n}\left(\left(\frac{\tilde{\boldsymbol{H}_{\boldsymbol{Y}}}}{n} - \frac{\tilde{\boldsymbol{H}_{\boldsymbol{Z}}}}{n}\right) - (\tilde{\boldsymbol{p}}_Y - \tilde{\boldsymbol{p}}_Z)\right) \xrightarrow{d} N(0, \Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}} + \Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}}). \tag{8}$$

According to Lemma 3.1, $\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}}$ and $\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}}$ are positive-definite. Since the set of symmetric positive-definite matrices is closed under nonnegative linear combination, $\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}} + \Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}}$ is also symmetric positive definite. Lemma 3.1 also implies that both of $\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}}$ and $\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}}$ have eigenvector $\mathbf{1}$. Therefore, $\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}} + \Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}}$ also has eigenvector $\mathbf{1}$. Thus we can invoke Lemma 3.2 to modify the asymptotic distribution (8). Let us denote $\tilde{\Sigma} := \Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}} + \Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}}$. Then we have the following asymptotic distribution:

$$\sqrt{n}\tilde{\Sigma}^{-1/2}\Pi\left(\left(\frac{\tilde{\boldsymbol{H}_{\boldsymbol{Y}}}}{n} - \frac{\tilde{\boldsymbol{H}_{\boldsymbol{Z}}}}{n}\right) - (\tilde{\boldsymbol{p}}_Y - \tilde{\boldsymbol{p}}_Z)\right) \xrightarrow{d} N(0, \tilde{\Sigma}^{-1/2}\Pi\tilde{\Sigma}\Pi\tilde{\Sigma}^{-1/2}). \tag{9}$$

Since $\|\mathcal{M}_{bit}(Y_i)\|_2 \leq \sqrt{k}$ and $\|\mathcal{M}_{bit}(Z_j)\|_2 \leq \sqrt{k}$, the sample covariance matrices

$$\hat{\Sigma}_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}} := \frac{1}{n}\sum_{i=1}^{n}(\mathcal{M}_{bit}(Y_i) - \tilde{\boldsymbol{H}_{\boldsymbol{Y}}}/n)(\mathcal{M}_{bit}(Y_i) - \tilde{\boldsymbol{H}_{\boldsymbol{Y}}}/n)^{\top}$$
$$\hat{\Sigma}_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}} := \frac{1}{n}\sum_{j=1}^{n}(\mathcal{M}_{bit}(Z_j) - \tilde{\boldsymbol{H}_{\boldsymbol{Z}}}/n)(\mathcal{M}_{bit}(Z_j) - \tilde{\boldsymbol{H}_{\boldsymbol{Z}}}/n)^{\top}$$

converge in probability to $\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}}$ and $\Sigma_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}}$, respectively (Corolloary 6.20 of Wainwright [2]). Let us denote $\hat{\Sigma} := \hat{\Sigma}_{\tilde{\boldsymbol{p}}_{\boldsymbol{Y}}} + \hat{\Sigma}_{\tilde{\boldsymbol{p}}_{\boldsymbol{Z}}}$. Since matrix inversion and matrix square root is continuous mapping on the space of positive symmetric definite matrices, we have

$$\hat{\Sigma}^{-1/2}\tilde{\Sigma}^{-1/2} \xrightarrow{p} I_k. \tag{10}$$

Therefore, by the Slutsky's theorem, we have

$$\sqrt{n}\hat{\Sigma}^{-1/2}\Pi\left(\left(\frac{\tilde{\boldsymbol{H}_{\boldsymbol{Y}}}}{n} - \frac{\tilde{\boldsymbol{H}_{\boldsymbol{Z}}}}{n}\right) - (\tilde{\boldsymbol{p}}_Y - \tilde{\boldsymbol{p}}_Z)\right) \xrightarrow{d} N(0, \tilde{\Sigma}^{-1/2}\Pi\tilde{\Sigma}\Pi\tilde{\Sigma}^{-1/2}). \tag{11}$$

Under the Under the null hypothesis of $\boldsymbol{p}_{\boldsymbol{Y}} = \boldsymbol{p}_Z$, we have $\tilde{\boldsymbol{p}}_Y - \tilde{\boldsymbol{p}}_Z = 0$. Therefore, we have

$$\sqrt{n}\hat{\Sigma}^{-1/2}\Pi\left(\frac{\tilde{\boldsymbol{H}_{\boldsymbol{Y}}}}{n} - \frac{\tilde{\boldsymbol{H}_{\boldsymbol{Z}}}}{n}\right) \xrightarrow{d} N(0, \tilde{\Sigma}^{-1/2}\Pi\tilde{\Sigma}\Pi\tilde{\Sigma}^{-1/2}). \tag{12}$$

Finally, we invoke Theorem 3.1 to obtain the following asymptotic null distribution

$$n\left(\frac{\tilde{\boldsymbol{H}_{\boldsymbol{Y}}}}{n} - \frac{\tilde{\boldsymbol{H}_{\boldsymbol{Z}}}}{n}\right)^{\top}\Pi\hat{\Sigma}^{-1}\Pi\left(\frac{\tilde{\boldsymbol{H}_{\boldsymbol{Y}}}}{n} - \frac{\tilde{\boldsymbol{H}_{\boldsymbol{Z}}}}{n}\right) \xrightarrow{d} \chi^2_{(k-1)}, \tag{13}$$

and we define the lefthand side of (13) as our test statistic.

References

[1] Gaboardi, M. and Rogers, R. (2018). Local private hypothesis testing: Chi-square tests. *Proceedings of the 35th International Conference on Machine Learning*, 80:1626–1635.

[2] Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.