Supplementary Material: On The Stability of Feature Selection in the Presence of Feature Correlations

Konstantinos Sechidis[1], Konstantinos Papangelou[1], Sarah Nogueira[2], James Weatherall[2], and Gavin Brown[1]

[1] School of Computer Science, University of Manchester, Manchester M13 9PL, UK
{konstantinos.sechidis,konstantinos.papangelou,gavin.brown}@manchester.ac.uk
[2] Criteo, Paris, France
s.nogueira@criteo.com
[3] Advanced Analytics Centre, Global Medicines Development, AstraZeneca, Cambridge, SG8 6EE, UK.
James.Weatherall@astrazeneca.com

A Proof of Theorem 2

Theorem 2 *The effective average pairwise intersection among the M subsets can be written:*

$$\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} r_{i,j}^{\mathbb{C}} = \overline{k}_{\mathbb{C}} - \text{tr}(\mathbb{C}\mathbf{S})$$

where $\overline{k}_{\mathbb{C}} = \sum_{f=1}^{d}\sum_{f'=1}^{d} c_{f,f'}\overline{z}_{f,f'}$ the effective average number of features selected over M runs. The unbiased estimator of the covariance between Z_f and $Z_{f'}$ is $\widehat{\text{cov}}(Z_f, Z_{f'}) = \frac{M}{M-1}(\hat{p}_{f,f'} - \hat{p}_f\hat{p}_{f'})$, $\forall\ f, f' \in \{1...d\}$, while $\mathbf{S}$ is an unbiased estimator of the variance-covariance matrix of $\mathcal{Z}$.

Proof: Using Lemma 1, we can write the average pairwise effective intersection between the M feature sets as follows:

$$\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} r_{i,j}^{\mathbb{C}} = \underbrace{\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} r_{i,j}}_{\texttt{First term}}$$
$$+ \underbrace{\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'} z_{i,f} z_{j,f'}}_{\texttt{Second term}} \quad (1)$$

Now we will analyse the above two terms:
• `First term`: Using Nogueira et al. [2, Theorem 1] this term can be re-written as follows:

$$\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M} r_{i,j} = \sum_{f=1}^{d}\overline{z}_f - \sum_{f=1}^{d}\widehat{\text{var}}_{(Z_f)}, \tag{2}$$

where we used the fact that: $\bar{k} = \sum_{f=1}^{d}\sum_{i=1}^{M} z_{i,f} = \sum_{f=1}^{d}\overline{z}_f$.

• `Second term`: This term can be written as follows:

$$\begin{aligned}
&\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\ j\neq i}}^{M}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'} z_{i,f} z_{j,f'} = \\
&\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{j=1}^{M}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'} z_{i,f} z_{j,f'} \\
&\quad - \frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'} z_{i,f} z_{i,f'} = \\
&\frac{1}{M(M-1)}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'} \sum_{i=1}^{M} z_{i,f}\sum_{j=1}^{M} z_{j,f'} \\
&\quad - \frac{1}{M(M-1)}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'} \sum_{i=1}^{M} z_{i,f} z_{i,f'} = \\
&\frac{1}{M(M-1)}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'} \sum_{i=1}^{M} M\widehat{p}_f \sum_{j=1}^{M} M\widehat{p}_{f'} \\
&\quad - \frac{1}{M(M-1)}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'} \sum_{i=1}^{M} M\widehat{p}_{f,f'} = \\
&\frac{M}{(M-1)}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'}\widehat{p}_f\widehat{p}_{f'} - \frac{1}{(M-1)}\sum_{f=1}^{d}\sum_{\substack{f'=1\\ f'\neq f}}^{d} c_{f,f'}\widehat{p}_{f,f'}
\end{aligned} \tag{3}$$

Now we will introduce the notation for the unbiased estimator of the covariance between Z_f and $Z_{f'}$:

$$\widehat{\text{cov}}(Z_f, Z_{f'}) = \frac{1}{M-1}\sum_{i=1}^{M}\left(z_{i,f} - \overline{z}_f\right)\left(z_{i,f'} - \overline{z}_{f'}\right)$$

$$= \frac{M}{M-1}\left(\widehat{p}_{ff'} - \widehat{p}_f\widehat{p}_{f'}\right), \tag{4}$$

where the short-hand notation $\widehat{p}_{ff'}$ is the ML estimate of the probability $\Pr\left[Z_f = 1, Z_{f'} = 1\right]$. Re-arranging the last expression we get:

$$\widehat{p}_f\widehat{p}_{f'} = \widehat{p}_{ff'} - \frac{M-1}{M}\widehat{\mathrm{cov}}(Z_f, Z_{f'}) \tag{5}$$

By substituting eq. (5) in eq. (3) we get:

$$\begin{aligned}
&\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M}\sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}z_{i,f}z_{j,f'} = \\
&\frac{M}{(M-1)}\sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}\left(\widehat{p}_{ff'} - \frac{M-1}{M}\widehat{\mathrm{cov}}(Z_f, Z_{f'})\right) \\
&\quad - \frac{1}{(M-1)}\sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}\widehat{p}_{f,f'} = \\
&\sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}\widehat{p}_{f,f'} - \sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}\widehat{\mathrm{cov}}(Z_f, Z_{f'}) = \\
&\sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}\overline{z}_{f,f'} - \sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}\widehat{\mathrm{cov}}(Z_f, Z_{f'})
\end{aligned} \tag{6}$$

where we use the notation $\overline{z}_{f,f'}$ for the ML estimate of the expected value of $\mathbb{E}\left[Z_f Z_{f'}\right]$, which is equal to to the probability $\Pr\left[Z_f = 1, Z_{f'} = 1\right]$. Finally, by substituting the re-writing of the `First term`, eq. (2), and the re-writing of the `Second term`, eq. (6), into eq. (1), the average pairwise effective intersection can be written as follows:

$$\begin{aligned}
&\frac{1}{M(M-1)}\sum_{i=1}^{M}\sum_{\substack{j=1\\j\neq i}}^{M} r_{i,j}^{\mathbb{C}} = \sum_{f=1}^{d}\overline{z}_f - \sum_{f=1}^{d}\widehat{\mathrm{var}}_{(Z_f)} \\
&\quad + \sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}\overline{z}_{f,f'} - \sum_{f=1}^{d}\sum_{\substack{f'=1\\f'\neq f}}^{d} c_{f,f'}\widehat{\mathrm{cov}}(Z_f, Z_{f'}) \\
&= \sum_{f=1}^{d}\sum_{f'=1}^{d} c_{f,f'}\overline{z}_{f,f'} - \sum_{f=1}^{d}\sum_{f'=1}^{d} c_{f,f'}\widehat{\mathrm{cov}}(Z_f, Z_{f'}) \\
&= \overline{k}_{\mathbb{C}} - \sum_{f=1}^{d}\sum_{f'=1}^{d} c_{f,f'}\widehat{\mathrm{cov}}(Z_f, Z_{f'})
\end{aligned}$$

where we define as $\overline{k}_{\mathbb{C}} = \sum_{f=1}^{d} \sum_{f'=1}^{d} c_{f,f'} \overline{z}_{f,f'}$ the effective average number of features selected over M runs.

Using matrix notation, the average pairwise effective intersection can take the following form:

$$\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} r_{i,j}^{\mathbb{C}} = \overline{k}_{\mathbb{C}} - \mathbf{1_d^T} \left(\mathbb{C} \odot \mathbf{S} \right) \mathbf{1_d},$$

where we used the notation $\mathbf{1_d^T A 1_d}$ for the sum of all elements (i.e. grand sum) of matrix A, and $\odot$ is the Hadamard (i.e. element-wise) product between two matrices. Furthermore, $\mathbf{S}$ is an unbiased estimator of the variance-covariance matrix of $\mathcal{Z}$.

Using the identity $\mathbf{1^T A} \odot \mathbf{B1} = \text{tr}(\mathbf{AB^T})$ [1], we can re-write the Hadamard product in terms of the trace of product of two matrices, the average pairwise effective intersection can take the following form:

$$\frac{1}{M(M-1)} \sum_{i=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{M} r_{i,j}^{\mathbb{C}} = \overline{k}_{\mathbb{C}} - \text{tr}(\mathbb{C}\mathbf{S}).$$

B Proof of Theorem 3

Theorem 3 *Under the Null Model, the variance/covariance matrix of $\mathcal{Z}$ is given by:*

$$\mathbf{\Sigma^0} = \begin{bmatrix} \text{var}\left(Z_1 \middle| H_0\right) & \dots & \text{cov}\left(Z_1, Z_d \middle| H_0\right) \\ \vdots & \ddots & \vdots \\ \text{cov}\left(Z_d, Z_1 \middle| H_0\right) & \dots & \text{var}\left(Z_d \middle| H_0\right) \end{bmatrix},$$

where the main diagonal elements are given by: $\text{var}\left(Z_f \middle| H_0\right) = \frac{\overline{k}}{d}\left(1 - \frac{\overline{k}}{d}\right)$ *and the off-diagonal elements, $f \neq f'$ are:* $\text{cov}\left(Z_f, Z_{f'} \middle| H_0\right) = \frac{\overline{k^2} - \overline{k}}{d^2 - d} - \frac{\overline{k}^2}{d^2}$

Proof: We will spit the proof in two parts, firstly we will derive the main-diagonal elements (variances) and then the cross-diagonal (covariances).

• `Main diagonal elements - variances`: By definition, the variance of a Bernoulli random variable Z_f can be written as follows:

$$\text{var}\left(Z_f \middle| H_0\right) = p_f^0 (1 - p_f^0), \tag{7}$$

where p_f^0 captures the probability of whether or not the feature was selected, under the null model of feature selection. The main challenge is to calculate p_f^0, which is equal:

$$p_f^0 = \mathbb{E}_{Z_f}\left[Z_f \middle| H_0\right],$$

Using the law of total expectation we can rewrite the last expression as

$$p_f^0 = \mathbb{E}_K \left[\mathbb{E}_{Z_f} \left[Z_f \middle| H_0, K \right] \right] \tag{8}$$

where K is a random variable that captures the sum of d, possibly dependent, bernoulli distributions: $K = \sum_{f=1}^{d} Z_f$. At this point we will calculate the inner expectation:

$$\mathbb{E}_{Z_f} \left[Z_f \middle| H_0, K \right] = \Pr \left[Z_f = 1 \middle| H_0, K \right]$$

Since we have $K = k$ bits set to 1, the probability is equal:

$$\Pr \left[Z_f = 1 \middle| H_0, K \right] = \frac{\#\{\text{bit-strings with } K \text{ 1s and } Z_f = 1\}}{\#\{\text{bit-strings with } K \text{ 1s}\}}.$$

The denominator is equal to $\binom{d}{K}$. For the numerator we know $Z_f = 1$, which means we have $K-1$ bits left to set to 1 from the remaining $d-1$ bits. Therefore the numerator is equal to $\binom{d-1}{K-1}$. Replacing these two terms in the previous equation we get that :

$$\mathbb{E}_{Z_f} \left[Z_f \middle| H_0, K \right] = \Pr \left[Z_f = 1 \middle| H_0, K \right] = \frac{\binom{d-1}{K-1}}{\binom{d}{K}} = \frac{K}{d}$$

Replacing the last expression in eq. (8) we get

$$p_f^0 = \mathbb{E}_K \left[\mathbb{E}_{Z_f} \left[Z_f \middle| H_0, K \right] \right] = \frac{\mathbb{E}_K \left[K \right]}{d} \tag{9}$$

Thus, under the null model of feature selection, all the features have an equal probability of being selected.

Substituting these probabilities in eq. (7), the variance under the null model of feature selection can be written as follows:

$$\text{var} \left(Z_f \middle| H_0 \right) = \frac{\mathbb{E}_K \left[K \right]}{d} \left(1 - \frac{\mathbb{E}_K \left[K \right]}{d} \right).$$

Under the definition of the null model of feature selection, the expectation $\mathbb{E}_K \left[K \right]$ can be estimated by the matrix $\mathcal{Z}$, i.e. we can use the sample mean $\overline{k}$, which is an unbiased estimator. Thus, the variances under the null model of feature selection are given by:

$$\text{var} \left(Z_f \middle| H_0 \right) = \frac{\overline{k}}{d} \left(1 - \frac{\overline{k}}{d} \right)$$

which is the same as the expression derived by Nogueira et al. [2].

• `Off diagonal elements - covariances`: By definition, the covariance of two bernoulli random variables Z_f and $Z_{f'}$ can be written as follows:

$$\text{cov}\left(Z_f, Z_{f'} \middle| H_0\right) = p^0_{ff'} - p^0_f p^0_{f'}, \tag{10}$$

where p^0_f captures the probability of whether or not the feature was selected, under the null model of feature selection, while $p^0_{ff'}$ the joint probability that both features were selected under the null model. Earlier we calculate the marginal probability p^0_f, while now we will calculate the joint $p^0_{ff'}$, which is equal:

$$p^0_{ff'} = \mathbb{E}_{Z_f Z_{f'}}\left[Z_f Z_{f'} \middle| H_0\right]$$

Using the law of total expectation we can rewrite the last expression as

$$p^0_{ff'} = \mathbb{E}_K\left[\mathbb{E}_{Z_f Z_{f'}}\left[Z_f Z_{f'} \middle| H_0, K\right]\right] \tag{11}$$

At this point we will calculate the inner expectation:

$$\mathbb{E}_{Z_f Z_{f'}}\left[Z_f Z_{f'} \middle| H_0, K\right] = \Pr\left[Z_f = 1, Z_{f'} = 1 \middle| H_0, K\right]$$

Since we have $K = k$ bits set to 1, the probability is equal:

$$\Pr\left[Z_f = 1, Z_{f'} = 1 \middle| H_0, K\right] = \frac{\#\{\text{bit-strings with } K \text{ 1s and } Z_f = 1, Z_{f'} = 1\}}{\#\{\text{bit-strings with } K \text{ 1s}\}}.$$

The denominator is equal to $\binom{d}{K}$. For the numerator we know $Z_f = 1$ and $Z_{f'} = 1$, which means we have $K - 2$ bits left to set to 1 from the remaining $d-2$ bits. Therefore the numerator is equal to $\binom{d-2}{K-2}$. Replacing these two terms in the previous equation we get that:

$$\begin{aligned}\mathbb{E}_{Z_f Z_{f'}}\left[Z_f Z_{f'} \middle| H_0, K\right] &= \Pr\left[Z_f = 1, Z_{f'} = 1 \middle| H_0, K\right] \\ &= \frac{\binom{d-2}{K-2}}{\binom{d}{K}} = \frac{K(K-1)}{d(d-1)}\end{aligned}$$

Replacing the last expression in eq. (11) we get

$$\begin{aligned}p^0_{ff'} &= \mathbb{E}_K\left[\mathbb{E}_{Z_f Z_{f'}}\left[Z_f Z_{f'} \middle| H_0, K\right]\right] \\ &= \frac{\mathbb{E}_K\left[K^2\right] - \mathbb{E}_K\left[K\right]}{d^2 - d}\end{aligned} \tag{12}$$

Thus, under the null hypothesis, all the pairs of features have an equal probability of being selected.

Substituting the probabilities of eqs. (9) and (12) in eq. (10), the covariance under the null model of feature selection can be written as follows:

$$\text{cov}\left(Z_f, Z_{f'} \middle| H_0\right) = \frac{\mathbb{E}_K\left[K^2\right] - \mathbb{E}_K\left[K\right]}{d^2 - d} - \left(\frac{\mathbb{E}_K\left[K\right]}{d}\right)^2.$$

Under the definition of the null mode of feature selection, the expectations over the variable K can be estimate by the matrix $\mathcal{Z}$. Thus, the covariances under the null model of feature selection are given by:

$$\operatorname{cov}\left(Z_f, Z_{f'} \middle| H_0\right) = \frac{\overline{k^2} - \overline{k}}{d^2 - d} - \frac{\overline{k}^2}{d^2}.$$

References

1. Horn, R.A., Johnson, C.R.: Matrix analysis. Cambridge university press (1990)
2. Nogueira, S., Sechidis, K., Brown, G.: On the stability of feature selection algorithms. Journal of Machine Learning Research 18(174), 1–54 (2018), http://jmlr.org/papers/v18/17-514.html