Appendix

We give an outline of the proof of the early stopping result. In the following, that result is called Corollary 2.

We use the same notation as in the body of the article: k is the number of bins, n is the "current" number of epochs, N is a number of epochs larger than n, and Ψ_{ν_1,ν_2} is the cumulative distribution function of an F random variable with degrees of freedom ν_1 and ν_2.

The situation we consider is that we have observed n epochs $x_1, \dots, x_n$, with each x_i being a k-dimensional vector of numbers representing the ith epoch after binning. We then ask, if an additional $N - n$ epochs are added to the original sample to get an extended sample $x_1, \dots x_n, x_{n+1}, \dots, x_N$ of size N, what values can the extended sample's Hotelling p-value take?

Let T_r^2 and p_r be the Hotelling T^2 statistic and the associated p-value, respectively, of the sample $x_1, \dots, x_r$, for $r = n$ and $r = N$.

We consider the original sample $x_1, \dots, x_n$ to be fixed, so the values of T_n^2 and p_n are fixed. We assume $T_n^2 > 0$, which is equivalent to assuming $p_n < 1$, and we make the usual assumption that covariance matrices of the data are positive definite.

Under these assumptions, we have the following results.

Theorem 1. The maximum possible value of T_N^2 is

$$\left(\frac{N-1}{n}\right)\left(\left(\frac{N}{n-1}\right)T_n^2 + N - n\right).$$

Corollary 1. The minimum possible value of p_N is

$$1 - \Psi_{k,N-k}\left[\left(\frac{N-k}{n}\right)\left(\left(\frac{N}{n-k}\right)\Psi_{k,n-k}^{-1}(1-p_n) + \frac{N-n}{k}\right)\right].$$

Corollary 2. If $0 < q < 1$ and

$$p_n > \ 1 - \Psi_{k,n-k}\left[\left(\frac{n-k}{N}\right)\left(\left(\frac{n}{N-k}\right)\Psi_{k,N-k}^{-1}(1-q) - \frac{N-n}{k}\right)\right],$$

then $p_N > q$.

As an aside, it is easy to show that the minimum possible value of T_N^2 is 0 and the maximum possible value of p_N is 1.

To prove Theorem 1, we start by considering the simplest case of $N = n + 1$ and $k = 1$, that is, one additional epoch and one-dimensional data. Note that with one-dimensional data, the Hotelling T^2 statistic is the square of the one-sample t-test statistic.

For each r, let $\bar{x}_r$, s_r^2 and t_r^2 be the sample mean, sample variance and squared t statistic, respectively, of the first r epochs. Note that the assumption that t_n^2 is non-zero means that $\bar{x}_n$ is non-zero.

We want an expression for t_{n+1}^2 in which the only variable quantity is x_{n+1}, keeping in mind that the first n epochs are assumed to be fixed. By definition we have

$$t_{n+1}^2 = \frac{(n+1)\bar{x}_{n+1}^2}{s_{n+1}^2}, \qquad (A1)$$

so we want to express $\bar{x}_{n+1}$ and s_{n+1}^2 in terms of fixed quantities and x_{n+1}. The required expressions are

$$\bar{x}_{n+1} = \frac{n\bar{x}_n + x_{n+1}}{n+1}$$

and

$$s_{n+1}^2 = \frac{(n+1)(n-1)s_n^2 + n(\bar{x}_n - x_{n+1})^2}{n(n+1)}.$$

Substituting these into (A1) gives

$$t_{n+1}^2 = \frac{n(n\bar{x}_n + x_{n+1})^2}{(n+1)(n-1)s_n^2 + n(\bar{x}_n - x_{n+1})^2}, \qquad (A2)$$

and we define $g(x_{n+1})$ to be the right-hand side of (A2), so $t_{n+1}^2 = g(x_{n+1})$.

We want to maximise the function g, so we look for points at which its derivative is zero. The derivative of g can be written as

$$g'(x_{n+1}) = \frac{2n(n+1)(n\bar{x}_n + x_{n+1})((n-1)s_n^2 + n\bar{x}_n^2 - n\bar{x}_n x_{n+1})}{[(n+1)(n-1)s_n^2 + n(\bar{x}_n - x_{n+1})^2]^2},$$

so the solutions of $g'(x_{n+1}) = 0$ are $x_{n+1} = -n\bar{x}_n$ and

$$x_{n+1} = \bar{x}_n + \frac{(n-1)s_n^2}{n\bar{x}_n}. \qquad (A3)$$

At the first of these solutions g is 0, and since g can't be negative, this must be a minimum. It can be verified that at the second solution, the second derivative of g is negative, so that is where g is maximum.

The maximum possible value of t_{n+1}^2 therefore occurs when x_{n+1} has the value in (A3). Substituting this value into (A2), and denoting the maximum possible value of t_{n+1}^2 by $\max t_{n+1}^2$, gives

$$\max t_{n+1}^2 = \left(\frac{n+1}{n-1}\right) t_n^2 + 1, \qquad (A4)$$

which is Theorem 1 for $N = n + 1$ and $k = 1$.

We now consider the case with $N = n + 1$ and $k > 1$. Using the notation $t^2(v_1, \dots, v_r)$ to mean the squared one-sample t statistic of the univariate sample $v_1, \dots, v_r$, it is well-known (e.g., Johnson & Wichern, 2007) that

$$T_r^2 = \max_{a \neq 0} t^2(a'x_1, \dots, a'x_r),$$

with a and the x_i being viewed as $k \times 1$ matrices and a' denoting the transpose of a.

Using this fact together with (A4) gives

$$\max T_{n+1}^2 = \left(\frac{n+1}{n-1}\right) T_n^2 + 1,$$

which is Theorem 1 for $N = n + 1$. From this, the result for general $N > n$ can be proved by induction.

Corollary 1 can be obtained by using the fundamental relation

$$p_r = 1 - \Psi_{k,r-k}\left(\frac{r-k}{k(r-1)}T_r^2\right)$$

(Johnson & Wichern, 2007) for $r = n$ and $r = N$ together with Theorem 1, and by noting that the p-value is minimised when the corresponding T^2 is maximised.

Corollary 2 can be obtained by rearranging Corollary 1.

Reference:

Johnson, R. A., & Wichern, D. W. (2007). Applied multivariate statistical analysis (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.