Computing Science Group

STABILITY OF THE MAHALANOBIS DISTANCE: A TECHNICAL NOTE

Andrew D. Ker

CS-RR-10-20

Oxford University Computing Laboratory
Wolfson Building, Parks Road, Oxford OX1 3QD

Abstract

When we estimate the Mahalanobis distance of one point from a cluster of points, there are a number of potential sources of error. This is particularly so if the cluster is strongly correlated, which implies that its covariance matrix is ill-conditioned. Here we analyse the magnitude of the errors introduced by round-off, inversion of near-singular matrices, and empirical estimation of mean and covariance, with particular application to the problem in [1].

1 Introduction

Suppose we have a distribution of points in $\mathbb{R}^p$, which we will call the *parent distribution*, which has mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{R}^{p \times p}$. Given a point $\boldsymbol{x} \in \mathbb{R}^p$, the square of the Mahalanobis distance (from $\boldsymbol{x}$ to the parent distribution) is

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}).$$

In practice, $\boldsymbol{\mu}$ and Σ are not known, so we estimate them using a finite sample from the parent distribution. If we have n samples $\boldsymbol{Z}_1, \dots, \boldsymbol{Z}_n$ then $\boldsymbol{\mu}$ is estimated by

$$\overline{\boldsymbol{Z}} = \frac{1}{n} \sum_i \boldsymbol{Z}_i$$

and Σ by

$$\boldsymbol{\mathcal{S}} = \frac{1}{n-1} \boldsymbol{\mathcal{Z}}^T \boldsymbol{\mathcal{Z}},$$

where $\boldsymbol{\mathcal{Z}}$ is the matrix whose rows are $\boldsymbol{Z}_1 - \overline{\boldsymbol{Z}}, \dots, \boldsymbol{Z}_n - \overline{\boldsymbol{Z}}$. It is important to note that the covariance matrix Σ might be ill-conditioned (meaning that its eigenvalues span many orders of magnitude, and that it is nearly singular), when the parent distribution has very strongly correlated components. That occurs, for example, with the distributions in [1].

Computing with ill-conditioned matrices requires caution, and one must take care that small errors in the input data do not become large errors in the output. There are many potential sources of error in the above calculation including round-off, random deviation of $\overline{\boldsymbol{Z}}$ and $\boldsymbol{\mathcal{S}}$ from $\boldsymbol{\mu}$ and Σ, and numerical errors when inverting the near-singular matrix Σ. This technical note estimates the typical magnitude of error in the Mahalanobis distance calculation, and examines whether an ill-conditioned covariance matrix makes the outputs unreliable. It was prompted by the work of [1], where one needs to answer this question to have confidence in the results.

1.1 Notation

The statistics literature has developed some inconsistent notational conventions for random vectors and matrices. We will always use lower case Roman or Greek letters for

Manuscript version 4 August 2010.

constant scalars (d, λ), upper case Roman for scalar random variables (D), lower case boldface Roman or Greek for constant vectors ($\boldsymbol{x}, \boldsymbol{\mu}, \mathbf{0}$ the zero vector), upper case boldface Roman for random vectors ($\boldsymbol{E}, \boldsymbol{X}, \overline{\boldsymbol{Z}}$), upper case Greek for constant matrices (Σ, Π, I the identity matrix), calligraphic upper case for random matrices ($\boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{W}}, \boldsymbol{\mathcal{Z}}$).

$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ indicates that the random vector $\boldsymbol{X}$ has the multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ. In such cases it is implicitly assumed that Σ is symmetric positive-definite. $\mathrm{E}[-]$ is the expectation operator, $\mathrm{Var}[X]$ the variance and $\mathrm{Var}[\boldsymbol{X}]$ the covariance matrix, and $\mathrm{Tr}[\Sigma]$ denotes matrix trace. Throughout, p denotes the dimension of the space (we will not need to consider random vectors with any other dimension, so will always leave it implicit) and n the number of samples used to estimate the mean and covariance of the parent distribution.

We will use the notation $d = O(-)$ loosely, sometimes to indicate the leading terms in d and sometimes as an indication of magnitude.

2 Error analysis

We will exclusively use the squared Mahalanobis distance, to avoid square roots. We want to estimate the magnitude of the deviation

$$d = \textit{estimate of } \left[(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right] - (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}).$$

There are four principal sources of error, which will be examined separately:

$$d_1 = (\boldsymbol{x}+\boldsymbol{dx}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}+\boldsymbol{dx}-\boldsymbol{\mu}) - (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),$$

due to round-off errors in $\boldsymbol{x}$;

$$d_2 = (\boldsymbol{x}-\boldsymbol{\mu})^T\Gamma(\boldsymbol{x}-\boldsymbol{\mu}) - (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),$$

caused by incorrectly inverting Σ;

$$d_3 = (\boldsymbol{x}-\overline{\boldsymbol{Z}})^T\Sigma^{-1}(\boldsymbol{x}-\overline{\boldsymbol{Z}}) - (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),$$

due to deviation of $\overline{\boldsymbol{Z}}$ from $\boldsymbol{\mu}$;

$$d_4 = (\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\mathcal{S}}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) - (\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}),$$

due to deviation of $\boldsymbol{\mathcal{S}}$ from Σ.

We cannot consider all sources of error simultaneously because the algebra becomes too involved. There may be other sources of error, including intermediate round-off, but they should be dominated by the errors above.

We are interested in *typical* error, so we use a random $\boldsymbol{X}$ which itself it drawn from a distribution with mean $\boldsymbol{\mu}$ and covariance Σ. In the absence of any other information, we will assume that $\boldsymbol{X}$ is multivariate Gaussian

$$\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma).$$

Then the above error quantities become random variables D_i, $i = 1 \ldots 4$, and we want to measure their magnitude. We chose the mean squared error (MSE), $\mathrm{E}[D_i^2]$, which accounts for both bias and variation in the answer. Note that this is the squared error of the difference between squared distances, chosen to be convenient for manipulation rather than for having a sensible unit of measure. In Sect. 3 we will put the error magnitudes into context.

A number of facts about random variables and matrices, mostly pertaining to quadratic forms of random variables, are stated in the Appendix and used in the following calculations.

2.1 Perturbations in $\boldsymbol{X}$

We assume that Σ^{-1} and μ are known, but that the measurements $\boldsymbol{X}$ contain errors. Both D_1 and D_3 can be dealt by considering this case:

$$\begin{aligned} D &= (\boldsymbol{X} - \boldsymbol{\mu} + e\boldsymbol{E})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu} + e\boldsymbol{E}) - (\boldsymbol{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \\ &= e^2 \boldsymbol{E}^T \Sigma^{-1} \boldsymbol{E} + 2e\, \boldsymbol{E}^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \end{aligned}$$

where e separates the scale of the perturbations from their direction, which is independent of $\boldsymbol{X}$,

$$\boldsymbol{E} \sim \mathcal{N}(\mathbf{0}, \Pi).$$

For D_1, it is reasonable to suppose that $\Pi = \mathrm{I}$ (round-off errors should be uncorrelated) and e is related to the floating point epsilon. For D_3, $e\boldsymbol{E} = \overline{\boldsymbol{Z}} - \boldsymbol{\mu}$ and we can use the well-known fact

$$\overline{\boldsymbol{Z}} \sim \mathcal{N}(\boldsymbol{\mu}, \tfrac{1}{n}\Sigma)$$

to write $\Pi = \Sigma$ and $e = 1/\sqrt{n}$.

Returning to D, write $\boldsymbol{Y} = \Sigma^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$. By (2),

$$\boldsymbol{Y} \sim \mathcal{N}(\mathbf{0}, \Sigma^{-1})$$

and $\boldsymbol{Y}$ remains independent of $\boldsymbol{E}$. We can compute $\mathrm{E}[D^2]$, for arbitrary Π, as follows:

$$\begin{aligned} \mathrm{E}[D^2] &= \mathrm{E}\big[\big(e^2 \boldsymbol{E}^T \Sigma^{-1} \boldsymbol{E} + 2e\boldsymbol{E}^T \boldsymbol{Y}\big)^2\big] \\ &= e^4 \mathrm{E}\big[\big(\boldsymbol{E}^T \Sigma^{-1} \boldsymbol{E}\big)^2\big] + 4e^3 \mathrm{E}\big[\boldsymbol{E}^T \Sigma^{-1} \boldsymbol{E}\boldsymbol{E}^T \boldsymbol{Y}\big] + 4e^2 \mathrm{E}\big[\big(\boldsymbol{E}^T \boldsymbol{Y}\big)^2\big] \\ &\overset{(a)}{=} 2e^4 \mathrm{Tr}\big[(\Sigma^{-1}\Pi)^2\big] + e^4 \mathrm{Tr}[\Sigma^{-1}\Pi]^2 + 4e^2 \mathrm{Tr}[\Sigma^{-1}\Pi] \end{aligned}$$

where (a) follows by (3) and the fact that Π is symmetric (the first term), linearity of expectation and $\mathrm{E}[\boldsymbol{Y}] = \mathbf{0}$ (the second term vanishes), and (5) (the third term).

We see that $\Sigma^{-1}\Pi$ determines the mean square error. For the case of D_1, $\Sigma^{-1}\Pi = \Sigma^{-1}$ so write $\lambda = \mathrm{Tr}[\Sigma^{-1}]$. Σ^{-1} is positive-definite, so we can invoke (1) to deduce

$$\mathrm{E}[D_1^2] \le 3e^4\lambda^2 + 4e^2\lambda = O\big(4e^2\lambda\big),$$

as long as $e^2 \ll 1/\lambda$.

For the case of D_3, $\Sigma^{-1}\Pi = \mathrm{I}$ and $e = n^{-1/2}$, so

$$\mathrm{E}[D_3^2] = 2pn^{-2} + p^2 n^{-2} + 4pn^{-1} = O\big(4pn^{-1}\big),$$

as long as $p \ll n$.

2.2 Error in Σ^{-1}

We assume that Σ and μ are known, and $\boldsymbol{X}$ measured exactly, but the computation of Σ^{-1} is not precise (this is typical when Σ is ill-conditioned). Here

$$\begin{aligned} D_2 &= (\boldsymbol{X}-\boldsymbol{\mu})^T\,\Gamma(\boldsymbol{X}-\boldsymbol{\mu}) - (\boldsymbol{X}-\boldsymbol{\mu})^T\,\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \\ &= (\boldsymbol{X}-\boldsymbol{\mu})^T\,(\Gamma-\Sigma^{-1})(\boldsymbol{X}-\boldsymbol{\mu}) \end{aligned}$$

where Γ is an *approximate* inverse for Σ. Note that Γ might not be symmetric or even positive-definite, depending on the algorithm used to invert Σ.

This case is fairly simple to deal with. Write $\Gamma' = \frac{1}{2}(\Gamma + \Gamma^T)$. Then

$$\begin{aligned} \mathrm{E}[D_2^2] &\overset{(a)}{=} 2\mathrm{Tr}\big[\big((\Gamma'-\Sigma^{-1})\Sigma\big)^2\big] + \mathrm{Tr}\big[(\Gamma'-\Sigma^{-1})\Sigma\big]^2 \\ &= 2\mathrm{Tr}\big[(\Gamma'\Sigma-\mathrm{I})^2\big] + \mathrm{Tr}[\Gamma'\Sigma-\mathrm{I}]^2. \end{aligned}$$

where (a) is by (3). This expresses $\mathrm{E}[D_2^2]$ in terms of a "symmetrized" inversion error $\Gamma'\Sigma - \mathrm{I}$.

2.3 Perturbations in $\boldsymbol{\mathcal{S}}$

We assume that $\boldsymbol{\mu}$ is known, and $\boldsymbol{X}$ measured exactly, but use an estimate for Σ. With μ known the unbiased estimator for Σ is slightly different, becoming

$$\boldsymbol{\mathcal{S}} = \frac{1}{n}\boldsymbol{\mathcal{W}},$$

where $\boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{Z}}^T\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\mathcal{Z}}$ is the matrix whose rows are $\boldsymbol{Z}_1 - \boldsymbol{\mu}, \ldots, \boldsymbol{Z}_n - \boldsymbol{\mu}$[1]. Then

$$\begin{aligned} D_4 &= (\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\mathcal{S}}^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) - (\boldsymbol{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu}) \\ &= (\boldsymbol{X}-\boldsymbol{\mu})^T(n\boldsymbol{\mathcal{W}}^{-1}-\Sigma^{-1})(\boldsymbol{X}-\boldsymbol{\mu}) \end{aligned}$$

[1]It is unclear what the correct scaling for $\boldsymbol{\mathcal{S}}$ is, since the unbiased estimator for Σ does not invert to an unbiased estimator for Σ^{-1}. In fact, none of the conclusions are altered if one of the other possible scalings is chosen.

Note that $\boldsymbol{X}$ was assumed to be independent of $\boldsymbol{\mathcal{W}}$. Now $\boldsymbol{\mathcal{W}}^{-1}$ has a known distribution, the *inverse Wishart* distribution $\mathcal{IW}(\Sigma^{-1}, n)$. We can then compute $\mathrm{E}[D_4^2]$,

$$\begin{aligned}
\mathrm{E}[D_4^2] &= \mathrm{E}\big[\big((\boldsymbol{X}-\boldsymbol{\mu})^T(n\boldsymbol{\mathcal{W}}^{-1}-\Sigma^{-1})(\boldsymbol{X}-\boldsymbol{\mu})\big)^2\big] \\
&\overset{(a)}{=} \mathrm{E}\big[\big(n(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\mathcal{W}}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\big)^2\big] \\
&\quad - 2n\mathrm{E}_{\boldsymbol{X}}\Big[\mathrm{E}_{\boldsymbol{\mathcal{W}}}\big[(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\mathcal{W}}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\big]\Big] \\
&\quad + \mathrm{E}\big[\big((\boldsymbol{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\big)^2\big] \\
&\overset{(b)}{=} \mathrm{E}\big[\big(n(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{\mathcal{W}}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\big)^2\big] + \big(1-\tfrac{2n}{n-p-1}\big)\mathrm{E}\big[\big((\boldsymbol{X}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{X}-\boldsymbol{\mu})\big)^2\big] \\
&\overset{(c)}{=} \frac{p^2(n-1)^2}{(n-p-2)^2}\left(1+\frac{2(n-2)}{p(n-p-4)}\right) + \left(1-\frac{2n}{n-p-1}\right)(2p+p^2) \\
&\overset{(d)}{=} \frac{p(p+2)\big(2n^2+np^2+4np+n-(p+1)(p+3)^2\big)}{(n-p-1)(n-p-2)(n-p-4)}.
\end{aligned}$$

(a) follows because expectation can be iterated: $\mathrm{E}[Z] = \mathrm{E}_{\boldsymbol{X}}\big[\mathrm{E}_{\boldsymbol{\mathcal{W}}}[Z]\big]$. (b) uses (6) and linearity of expectation. (c) uses (7) and (4). (d) is just algebra.

This is a mess, but the leading terms are

$$\mathrm{E}[D_4^2] = O\big(2p^2n^{-1} + p^4n^{-2}\big).$$

3 Conclusion

For "typical" points, drawn from a multivariate Gaussian distribution with the same mean and covariance as the parent distribution, the true squared Mahalanobis distance

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}).$$

has expectation p, see (4). Added to the true value is some error, and we have computed the mean square of this error. The *square root* of these MSEs should therefore be compared with true values of order p.

We summarise the results:

$$\begin{aligned}
\mathrm{E}[D_1^2] &= O\big(4e^2\lambda\big), && \text{where } e \text{ is magnitude of error in } \boldsymbol{x} \\
& && \text{and } \lambda = \mathrm{Tr}[\Sigma^{-1}], \text{ assuming } e^2 \ll \lambda^{-1}, \\
\mathrm{E}[D_2^2] &= 2\mathrm{Tr}\big[(\Gamma'\Sigma - \mathrm{I})^2\big] + \mathrm{Tr}[\Gamma'\Sigma - \mathrm{I}]^2, && \text{where } \Gamma' = \tfrac{1}{2}(\Gamma + \Gamma^T) \\
& && \text{and } \Gamma \text{ is the approximate inverse to } \Sigma, \\
\mathrm{E}[D_3^2] &= O\big(4pn^{-1}\big), && \text{as long as } p \ll n, \\
\mathrm{E}[D_4^2] &= O\big(2p^2n^{-1} + p^4n^{-2}\big), && \text{as long as } p \ll n.
\end{aligned}$$

Note that the final two errors, due to the use of empirical estimates for $\boldsymbol{\mu}$ and Σ, are not affected by the true value of Σ. Contrary to intuition, we need not be concerned that an ill-conditioned Σ requires a larger sample size for estimating $\boldsymbol{\mu}$ and Σ, but we do need that n is sufficiently larger than p^2 to keep the errors small. The second error, due to inversion inaccuracy, can only be computed by looking at a particular inverse Γ and computing $\Gamma'\Sigma - \mathrm{I}$. The first error requires that the round-off errors in $\boldsymbol{x}$ be small enough that $e^2 \ll \mathrm{Tr}[\Sigma^{-1}]^{-1}$.

In the application which motivated these calculations [1], we have $p = 27$ and $n = 2000$. The true squared distances should be $O(10^1)$–$O(10^2)$. Error $d_3 = O(10^{-1})$, which is insignificant, and $d_4 = O(1)$ is reasonably small. These predicted errors were confirmed by a simple resampling bootstrap, which agreed that the relative error is usually less than 5%. We tested the approximate inverse of the covariance matrices we observed and, as long as the inversion is performed using stable Gauss-Jordan elimination which selects the largest absolute pivot at each stage [2, §9.1],

$$\mathrm{Tr}\left[(\Gamma'\Sigma - \mathrm{I})^2\right] \approx 9 \cdot 10^{-6}, \quad \mathrm{Tr}[\Gamma'\Sigma - \mathrm{I}]^2 \approx 2 \cdot 10^{-11}$$

meaning that $d_2 = O(10^{-3})$, which is certainly insignificant.

All the above errors are fixed when the matrix Σ, and the empirical data $\boldsymbol{\mathcal{Z}}$, are fixed, so we only require that the true distances not be swamped by error (which the above calculations show that they are not). Only d_1 varies with $\boldsymbol{x}$, because round-off errors will be different each time, and this makes it far more important for [1] because it varies on each iteration of the maximisation algorithm. We want to be certain that the maximisation procedure is not guided by fluctuations due to round-off errors rather than genuine improvements.

For our matrices, $\mathrm{Tr}[\Sigma^{-1}] \approx 10^{12}$. The features (vector $\boldsymbol{x}$) were computed using double-precision floating point, so we would expect e to be the result of some accumulated errors each $O(10^{16})$. We determined $e \leq 10^{-12}$, by comparing the results against some carried out using quadruple precision (128 bit floating point) arithmetic. This means that $d_1 = O(10^{-6})$. When we look at the maximization algorithm, we see that the distance changes by 10^{-5} or more in at least 98% of steps. This means that we can have confidence that the gradual improvements in distance, as discovered by the heuristic minimization algorithms in [1], are not (or at least almost never) fluctuations caused by round-off.

References

[1] V. K. Chonev and A. D. Ker, "Feature restoration and distortion metrics," to appear in *Proc. Media Watermarking, Security, and Forensics XIII*, Proc. SPIE, 2011.

[2] G. Strang, *Introduction to Linear Algebra*, 4th ed. Wellesley, 2009.

[3] X. M. Yang, X. Q. Yang, and K. L. Teo, "A matrix trace inequality," *Journal of Mathematical Analysis and Applications*, vol. 263, no. 1, pp. 327–331, 2001.

[4] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*. Academic Press, 1979.

[5] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Oct 2008, version 20081110. [Online]. Available: `http://www2.imm.dtu.dk/pubdb/p.php?3274`

[6] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, 9th ed. Dover, 1972.

A Appendix: Useful Identities

We summarise the facts, from the theory of matrices and statistics, which we used in this note.

(i) For nonnegative definite square matrices Σ and Π,

$$\mathrm{Tr}[\Sigma\Pi] \leq \mathrm{Tr}[\Sigma]\mathrm{Tr}[\Pi]. \tag{1}$$

It follows from the main result of [3].

(ii) If $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ then

$$\Gamma(\boldsymbol{X} - \boldsymbol{\mu}) \sim \mathcal{N}(\mathbf{0}, \Gamma\Sigma\Gamma^T). \tag{2}$$

A basic fact, see for example [4, 3.2.1].

(iii) If $\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ then, regardless of whether Γ is symmetric or positive-definite,

$$\mathrm{E}\big[(\boldsymbol{X}^T\Gamma\boldsymbol{X})^2\big] = 2\mathrm{Tr}\big[(\Gamma'\Sigma)^2\big] + \mathrm{Tr}[\Gamma'\Sigma]^2, \tag{3}$$

where $\Gamma' = \frac{1}{2}(\Gamma + \Gamma^T)$. Note that $\Gamma = \Gamma'$ if Γ is symmetric. Follows from [5, Eqs. (355) & (356)] and the cyclic permutation property of trace.

We also use the special case

$$\mathrm{E}\big[(\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X})^2\big] = 2p + p^2. \tag{4}$$

(iv) If $\mathrm{E}[\boldsymbol{X}] = \mathrm{E}[\boldsymbol{Y}] = \mathbf{0}$, $\mathrm{Var}[\boldsymbol{X}] = \Sigma$, $\mathrm{Var}[\boldsymbol{Y}] = \Pi$, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, then

$$\mathrm{E}[(\boldsymbol{X}^T\boldsymbol{Y})^2] = \mathrm{Tr}(\Sigma\Pi). \tag{5}$$

Source unknown. Can be proved from first principles, using linearity of expectation.

(v) If $\boldsymbol{\mathcal{W}}^{-1} \sim \mathcal{IW}(\Sigma^{-1}, n)$ then

$$\mathrm{E}[\boldsymbol{\mathcal{W}}] = \frac{1}{n-p-1}\Sigma^{-1}. \tag{6}$$

See [4, 3.8.3].

(vi) Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $\boldsymbol{\mathcal{W}}^{-1} \sim \mathcal{IW}(\Sigma^{-1}, n)$ be independent. Then

$$T^2 = n(\boldsymbol{X} - \boldsymbol{\mu})^T \boldsymbol{\mathcal{W}}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

is known as Hotelling's T-square statistic, and its distribution is given by a scaled F-distribution [4, 3.6.5]:

$$\frac{n-p}{p(n-1)} T^2 \sim \mathcal{F}_{p,\, n-p}.$$

Since $F \sim \mathcal{F}_{\nu_1,\nu_2}$ implies

$$\mathrm{E}[F^2] = \frac{\nu_2^2}{(\nu_2 - 2)^2} \left(1 + \frac{2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 4)}\right)$$

[6, 26.6.3] then we have

$$\mathrm{E}[(T^2)^2] = \frac{p^2(n-1)^2}{(n-p-2)^2} \left(1 + \frac{2(n-2)}{p(n-p-4)}\right). \tag{7}$$