PCA Subspaces are not always optimal for Bayesian Learning

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

Bayesian Neural Networks are often sought after for their strong and trustworthy predictive power. However, inference in these models is often computationally expensive and can be reduced using dimensionality reduction where the key goal is to find an appropriate subspace in which to perform the inference, while retaining significant predictive power. In this work, we propose a theoretical comparative study of the Principal Component Analysis versus the random projection for Bayesian Linear Regression. We find that the PCA is not always the optimal dimensionality reduction method and that the random projection can actually be superior, especially in cases where the data distribution is shifted and the labels have a small norm. We then confirm these results experimentally. Therefore, this work suggests to consider dimension reduction by random projection for Bayesian inference when noisy data are expected.

1 Introduction

Bayesian methods and especially Bayesian Neural Networks (BNN) [16, 18, 7] are often sought after for their strong and trustworthy predictive power. However, inference in these models is often computationally expensive, be it via Laplace inference [16, 14, 13], variational inference [11, 1, 6], Markov chain Monte Carlo [18, 19, 9, 8, 10], or ensemble-based inference [3, 5, 4]. To reduce the cost of this inference, different methods of dimensionality reduction have been studied where the key goal is to find an appropriate subspace in which to perform the inference, while retaining significant predictive power. This is similar to approaches known from Gaussian Processes [e.g., 2]. For BNNs, the methods based on the Principal Component Analysis (PCA) of the Stochastic Gradient Descent (SGD) trajectory or random projections seem to provide promising computational results as discussed in Maddox et al. [17] and Izmailov et al. [15].

In this work, we propose a theoretical comparative analysis of these different dimensionality reduction methods. Namely, we will focus on the comparison between the PCA of the SGD trajectory and the random projection. Since deep learning models are theoretically hard to study, we will focus on Bayesian Linear Regression, which offers the advantage of having a tractable posterior distribution. Moreover, we will use predictive inference distribution as the criterion for comparison. We find, possibly surprisingly, that the PCA is not always the optimal dimensionality reduction method and that the random projection can actually be superior, especially in cases where the data is noisy and the labels have a small norm.

2 Methods

In this section we will introduce the problem and notation. Let us consider training inputs $X \in \mathbb{R}^{n \times d}$ with associated labels $Y \in \mathbb{R}^n$ and test inputs $X^* \in \mathbb{R}^{n \times d}$ with labels $Y^* \in \mathbb{R}^n$. We will denote

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

the singular values of X as $(r_i)_{i\in 1,...,d}$ and the ones of X^* as $(r_i^*)_{i\in 1,...,d}$. The parameters of our model are $\theta \in \mathbb{R}^d$, the noise variance is σ^2, and the prior variance is λ^2. The main problem we are interested in here is to project the data into a subspace of dimension $k < d$.

2.1 Assumptions

Assumption A: We assume a classical Bayesian Linear Regression model with $d \leq n$ and homoscedastic Gaussian noise, such that $Y = X\theta + \nu$, $\nu \sim \mathcal{N}(0, \sigma^2 I)$ and Gaussian prior with parameter λ independent of X such that $\theta \sim \mathcal{N}(0, \lambda^2 I)$.

Assumption B1: $\sigma = \lambda = 1$.

2.2 Predictive Distribution in the Global Space

The predictive distribution in the global space $\Pr(Y^*|X^*, X, Y)$ is Gaussian: $\mathcal{N}_{global} := Y^*|X^*, X, Y \sim \mathcal{N}(\mu_S, S)$. Using Bayesian model averaging (the proof is detailed in Appendix 5), we obtain

$$S = (I - X^*(X^\top X + X^{*\top} X^* + I)^{-1} X^{*\top})^{-1} \tag{1}$$

$$\mu_S = SX^*(X^\top X + X^{*\top} X^* + I)^{-1} X^\top Y \tag{2}$$

2.3 Projected distribution

Let $P_E \in \mathbb{R}^{d\times k}$ be the matrix such that its columns are generating the subspace E. We have $P_E^\top P_E = I_k$ and $P_E P_E^\top = H$ a projection matrix ($H^2 = H$). Then, we can take both formulas for the global space 1,2 and multiply X and X^* by P_E to the right to compute the distribution of $\mathcal{N}_E := Y^*|X^* P_E, XP_E, Y \sim \mathcal{N}(\mu_E, S_E)$ with $\mu_E \in \mathbb{R}^n$ and $S_E \in \mathbb{R}^{n\times n}$.

PCA projection Izmailov et al. [15] proposed to use a PCA on the SGD trajectory to select the subspace. As described in Gur-Ari et al. [12], this method is similar to keeping the eigenvectors associated with the largest eigenvalues of the Hessian, which in our Bayesian Linear Regression setup is similar to performing the PCA on X. If we use the Singular Value Decomposition (SVD) of X: $X = URV$ with U, V being two orthogonal matrices of dimension $n \times n$ and $d \times d$ and R a diagonal matrix of dimension $n \times d$ containing the singular values $(r_i)_{i\in 1,...,d}$ of X. Rearranging U, R, V, we will assume that R contains the singular values in increasing order. Thus, the projection matrix P_{PCA} for the PCA method is a submatrix of V containing the k eigenvectors associated to the k largest eigenvalues r_i.

Random Projection Instead of using PCA, we can alternatively project into a random subspace. To do so, we can construct a matrix P_{rand} of dimension $d \times k$ containing d independent Gaussian vectors of dimension k: $\epsilon_1, .., \epsilon_d \sim \mathcal{N}(0, I_k)$ and $P_{rand} = (\epsilon_1, ..., \epsilon_d)^\top$.

2.4 KL-Divergence as Comparison Tool

A good projection is a projection whose predictive distribution is as close as possible to the distribution in the global space, that is, whose expectation and covariance matrices are as close as possible to μ_S and S respectively. Different tools can be used to compare these distributions and we have chosen the *Kullback-Leibler (KL) divergence* which offers the nice advantage of being tractable and simple for Gaussian distributions. It gives for a subspace E:

$$D_E := D_{\mathrm{KL}}(\mathcal{N}_E \| \mathcal{N}_{global}) = \frac{1}{2}(\mathrm{Tr}(S^{-1} S_E) + (\mu_S - \mu_E)^\top S^{-1} (\mu_S - \mu_E) - n + \ln(\frac{|S|}{|S_E|})) \tag{3}$$

Our problem therefore consists of comparing D_{PCA} with D_{rand}.

3 Results

3.1 PCA is not perfect: a counter-example

By the Eckart-Young theorem, PCA is the best low-rank approximation in terms of the Frobenius norm. However, here we instead care about the KL divergence between the global space and the subspace, therefore the Eckart-Young theorem does not apply and PCA is not necessarily the best low-rank approximation. Here, we will derive a simple counter-example which proves that PCA is not optimal. To do so, we assume that $n = 3$, $d = 2$, and $X = X^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$, $P_{PCA} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and we compare it with the projection in the span of $\gamma = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. As detailed in Appendix 6 if we choose the labels to be $y_2 = 0$ and $y_1 = 0$:

$$D_\gamma \approx 0.08 \geq D_{PCA} \approx 0.07$$

However, if instead we take $y_2 = 1$ and $y_1 = 0$:

$$D_\gamma \approx 0.17 \leq D_{PCA} \approx 0.24$$

This shows that the PCA is not a generally optimal solution.

3.2 Study in Expectation

3.2.1 The KL Divergence for PCA

Theorem 1 *Using the SVD decomposition of $X = URV$ and noting that $Q := U^\top Y$ we can derive the following equation under assumptions A, B1. The proof is given in Appendix 7.*

$$D_{PCA} = \frac{1}{2}(\sum_{i=k+1}^{d} -\frac{r_i^{*2}}{1+r_i^2+r_i^{*2}} - \log(1-\frac{r_i^{*2}}{r_i^2+r_i^{*2}+1}) + Q_i^2\frac{r_i^2 r_i^{*2}}{(1+r_i^2+r_i^{*2})(1+r_i^2)}) \quad (4)$$

3.2.2 The KL Divergence for Random Projection

Assumption B2: $\sigma = \lambda = 1$ and d is large enough.
Assumption C: If $P_{rand} = (\epsilon_1, \ldots, \epsilon_d)^\top$, we assume that $\sum_{i=1}^{d} r_i^2 \epsilon_i \epsilon_i^\top$ and $\sum_{i=1}^{d} r_i^{*2} \epsilon_i \epsilon_i^\top$ are respectively equal to their expectations $\sum_{i=1}^{d} r_i^2 I_k$ and $\sum_{i=1}^{d} r_i^{*2} I_k$.
Under assumption A, B2, C, we can derive the KL divergence for the random projection in equation 5:

$$D_{rand} = \frac{1}{2}(\sum_{i=1}^{d} -\frac{r_i^{*2}}{1+r_i^2+r_i^{*2}} - \log(1-\frac{r_i^{*2}}{1+r_i^2+r_i^{*2}}) - \hat{\Sigma}_{rand}\frac{r_i^{*4}\epsilon_i^\top\epsilon_i}{1+r_i^2+r_i^{*2}} + \sum_{i,j=1}^{d} Q_i Q_j G_{ij}) \quad (5)$$

with $\hat{\Sigma}_{rand} := (1 + \sum_{i=1}^{d} r_i^2 \epsilon_i^2 + 2\sum_{i=1}^{d} r_i^{*2}\epsilon_i^2)^{-1}$ and G a matrix the coefficients of which are detailed in the proof in Appendix 8.

3.2.3 Comparison of PCA and Random Projection

To compare the PCA with the random projection we chose to evaluate their expected behaviour. Hence, we need to compare the values of D_{PCA} with the expectation of D_{rand} according to ϵ: $\mathbb{E}_\epsilon(D_{rand})$. We first observe with equations 4 and 5 that both KL divergences behave quite similarly. Yet, PCA seems to be more dependent on the data X as seen in the equation 4 with the sum beginning at $k + 1$. Thus, if the testing data X^* is perturbed along the other axes of the projection, PCA should underperform compared to the random projection which does not depend on the choice of the projection eigenvector. In the following, we will assume that the testing data X^* is perturbed with a small value δ and compare D_{PCA} with the $E_\epsilon(D_{rand})$ as functions of δ. We will thus make the following assumptions:
Assumption D: X is the identity, i.e., $X = \text{diag}(1, ..., 1)$ and $P_{PCA} = I^{d\times k}$. We then perturb the test data X^* by adding a small perturbation δ to the $(k+1)$-th singular value of X: $X^* = \text{diag}(1, .., 1, 1+\delta, 1, .., 1)$.

Small-norm outputs: The KL divergences of the two methods seem to have different behaviours depending on the norm of the Y outputs. We will therefore separate different cases starting by focusing firstly on small-norm outputs.
Assumption E1: $\|Y\|_\infty \leq 0.1$ with high probability.

We can prove (Appendix 9) using assumptions A, B2, C, D, E1 that if we denote $f(\delta) := D_{PCA} - \mathbb{E}_\epsilon(D_{rand})$:

$$f(0) \lessapprox 0 \text{ and } f \text{ is an increasing function} \tag{6}$$

Result 6 shows that without perturbation, PCA is better than random projection, however the latter is more responsive to perturbed data. We will now study experimentally the behaviour of both KL divergences as a function of δ and for small-norm outputs Y. To do so, we will generate vectors $\beta \sim \mathcal{N}(0, 0.01I_d)$ and obtain outputs such that $Y = X\beta$. Averaging over 10,000 runs, we obtain Figure 1, where we notice that the difference between both KL divergences increases with δ as proven in equation 6. Moreover, we can indeed see that for $\delta = 0$, i.e., without perturbation, $f(\delta) \leq 0$ and therefore PCA is better. As δ increases, $f(\delta)$ becomes quickly positive as shown on equation 6 and random projection is then better. Hence, PCA is better without perturbation but random projection reacts better to perturbations in δ for small outputs Y.

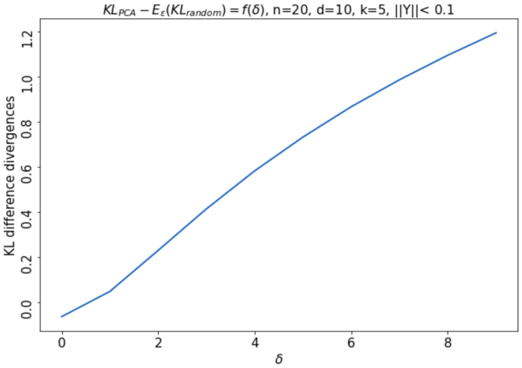

Figure 1: D_{rand} and D_{PCA} as function of δ for $n = 20$ with $\|Y\|_\infty \leq 0.1$.

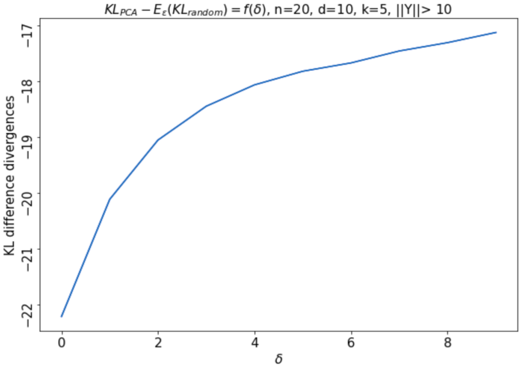

Figure 2: D_{rand} and D_{PCA} as function of δ for $n = 20$ with $\|Y\|_\infty \geq 10$.

Large-norm outputs:
Assumption E2: $\|Y\|_\infty \geq 10$ with high probability.
Using Assumption E2, we can reconsider the KL divergence for a subspace E to be only:

$$D_E = \frac{1}{2}(\mu_S - \mu_E)^\top S^{-1}(\mu_S - \mu_E) \tag{7}$$

The behaviour of this term is much more difficult to study than in the previous paragraph because it involves the outputs Y whose distribution is unknown. However, we are still able to prove that for large perturbations, the PCA performs better than the random projection as seen in the equation 8 proven in Appendix 10 under Assumptions A, B2, C2, D, E2.

$$\lim_{\delta \to \infty} D_{PCA} - \mathbb{E}_\epsilon(D_{rand}) = -\sum_{i=1}^{k} \frac{Q_i^2}{12} \leq 0 \tag{8}$$

We repeated the same experiment as above with large-norm outputs, that is, we generate vectors $\beta \sim \mathcal{N}(10, 0.01I_d)$ and obtain outputs such that $Y = X\beta$. Again averaging over 10,000 runs, we obtain Figure 2, where we can notice that $f(\delta)$ increases and is always negative, thus the PCA is always better than the random projection for large-norm outputs.

4 Conclusion

In this study, we compared two dimensionality reduction methods for Bayesian linear regression: the PCA of the data (or similarly the SGD trajectory) and the random projection. We showed experimentally and theoretically that the PCA is better for noiseless data and also for large-norm outputs. However, for small-norm outputs and noisy data, the random projection can be superior.

References

[1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks, 2015.

[2] Joaquin Quiñonero Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *J. Mach. Learn. Res.*, 6:1939–1959, 2005. URL `http://jmlr.org/papers/v6/quinonero-candela05a.html`.

[3] Kamil Ciosek, Vincent Fortuin, Ryota Tomioka, Katja Hofmann, and Richard Turner. Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations*, 2019.

[4] Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are bayesian. *arXiv preprint arXiv:2106.11642*, 2021.

[5] Francesco D'Angelo, Vincent Fortuin, and Florian Wenzel. On stein variational neural network ensembles. *arXiv preprint arXiv:2106.10760*, 2021.

[6] Michael W Dusenberry, Ghassen Jerfel, Yeming Wen, Yi-an Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable Bayesian neural nets with rank-1 factors. *arXiv preprint arXiv:2005.07186*, 2020.

[7] Vincent Fortuin. Priors in bayesian deep learning: A review. *arXiv preprint arXiv:2105.06868*, 2021.

[8] Vincent Fortuin, Adrià Garriga-Alonso, Mark van der Wilk, and Laurence Aitchison. Bnnpriors: A library for bayesian neural network inference with different prior distributions. *Software Impacts*, 9:100079, 2021.

[9] Vincent Fortuin, Adrià Garriga-Alonso, Florian Wenzel, Gunnar Rätsch, Richard Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. *arXiv preprint arXiv:2102.06571*, 2021.

[10] Adrià Garriga-Alonso and Vincent Fortuin. Exact langevin dynamics with stochastic gradients. *arXiv preprint arXiv:2102.01691*, 2021.

[11] Alex Graves. Practical variational inference for neural networks. *Advances in neural information processing systems*, 24, 2011.

[12] Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace, 2018.

[13] Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Mohammad Emtiyaz Khan. Scalable marginal likelihood estimation for model selection in deep learning. *arXiv preprint arXiv:2104.04975*, 2021.

[14] Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of bayesian neural nets via local linearization. In *International Conference on Artificial Intelligence and Statistics*, pages 703–711. PMLR, 2021.

[15] Pavel Izmailov, Wesley J. Maddox, Polina Kirichenko, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Subspace inference for bayesian deep learning, 2019.

[16] David J.C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[17] Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, and Andrew Gordon Wilson. A simple baseline for bayesian uncertainty in deep learning, 2019.

[18] Radford M. Neal. *Bayesian learning for neural networks*, volume 118. Springer, 1996.

[19] Ruqi Zhang, Chunyuan Li, Jianyi Zhang, Changyou Chen, and Andrew Gordon Wilson. Cyclical stochastic gradient mcmc for bayesian deep learning. *arXiv preprint arXiv:1902.03932*, 2019.

Supplementary Material

5 Appendix: Proof of equations 1 and 2

We use the notation $cste(.)$ for constant depending on the parameters $(.)$.

5.1 Assumptions

Assumption 1: We take the classical Bayesian Linear Regression model:

$$Y = X\theta + \nu$$

with homeostatic Gaussian noise:

$$\nu \sim \mathcal{N}(0, \sigma^2 I) \tag{9}$$

Finally we have:

$$Y \sim \mathcal{N}(X\theta, \sigma^2 I)$$

And therefore our log-likelihood is:

$$-2 \log \Pr(Y|X, \theta) = \frac{(Y - X\theta)^\top (Y - X\theta)}{\sigma^2} + cste \tag{10}$$

Assumption 2: We only considers Gaussian prior of parameter λ independent of X such that:

$$\theta \sim \mathcal{N}(0, \lambda^2 I)$$

5.2 Computation of the distributions in the global space

5.2.1 Posterior distribution

Now that the model is fixed, we want to find the best weights θ given a dataset of inputs and outputs (X, Y). With a Bayesian perspective we want to compute the posterior $\Pr(\theta|X, Y)$. Then we will be able to compute the predictive distribution, *i.e* the distribution of unseen data.
Using Bayes' rule and assumption 5.1 we obtain:

$$\Pr(\theta|X, Y) \propto \Pr(Y|X, \theta) \Pr(\theta) \tag{11}$$

Both right terms are Gaussian, thus our posterior is also Gaussian. Hence, $\exists \Sigma \in \mathbb{R}^{d \times d}$ and $\exists \mu \in \mathbb{R}^d$ such that:

$$-2 \log \Pr(\theta|X, Y) = (\theta - \mu)^\top \Sigma^{-1} (\theta - \mu) + cste = \theta^\top \Sigma^{-1} \theta - 2\theta^\top \Sigma^{-1} \mu + cste \tag{12}$$

Using assumption 5.1 and equations 10, 12 and 11:

$$\begin{aligned} -2 \log \Pr(\theta, |X, Y) &= \frac{(Y - X\theta)^\top (Y - X\theta)}{\sigma^2} + \frac{\theta^\top \theta}{\lambda^2} + cste \\ &= \theta^\top (\frac{I_p}{\lambda^2} + \frac{X^\top X}{\sigma^2})\theta - 2\theta^\top \frac{X^\top Y}{\sigma^2} + cste \\ &= \theta^\top \Sigma^{-1} \theta - 2\theta^\top \Sigma^{-1} \mu + cste \end{aligned}$$

By equalizing the terms in $\theta^\top$ and $\theta^\top \theta$ we obtain:

$$\Sigma = (\frac{X^\top X}{\sigma^2} + \frac{I}{\lambda^2})^{-1} \tag{13}$$

$$\mu = \Sigma \frac{X^\top Y}{\sigma^2} \tag{14}$$

5.2.2 Predictive distribution

Now we have the keys to compute the predictive distribution, *i.e* for new data Y^*, X^* compute $\Pr(Y^*|X^*, X, Y)$ which is Gaussian. To take some notation, $Y^*|X^*, X, Y \sim \mathcal{N}(\mu_S, S)$. To do so we will use the Bayesian model averaging technique:

$$\Pr(Y^*|X^*, X, Y) = \int \Pr(Y^*|X^*, \theta, X, Y) \Pr(\theta|X, Y) d\theta$$

We first focus on the term in the integral

$$C(\theta) := \Pr(Y^*|X^*, \theta, X, Y) \Pr(\theta|X, Y) \sim \mathcal{N}(\hat{\mu}, \hat{\Sigma}) \tag{15}$$

and we need now to find $\hat{\mu}, \hat{\Sigma}$.

i. Find $\hat{\mu}, \hat{\Sigma}$: To do so, we will separate the constant part between $cste$, which depends neither on Y^* nor on θ, and $cste_{Y^*}$ which is a sum of terms depending from Y^*. Intuitively $cste_Y^*$ will actually represent the predictive distribution.

$$\begin{aligned}
-2\log C(\theta) &= -2\log[\Pr(Y^*|X^*, \theta, X, Y) \Pr(\theta|X, Y)] \\
&= (\theta - \mu)^\top \Sigma^{-1} (\theta - \mu) + \frac{(Y^* - X^*\theta)^\top (Y^* - X^*\theta)}{\sigma^2} + cste \\
&= \theta^\top (\Sigma^{-1} + \frac{X^{*\top} X^*}{\sigma^2})\theta - 2\theta^\top (\Sigma^{-1}\mu + \frac{X^{*\top} Y^*}{\sigma^2}) + \frac{Y^{*\top} Y^*}{\sigma^2} + cste \\
&= (\theta - \hat{\mu})^\top \hat{\Sigma}^{-1} (\theta - \hat{\mu}) + cste_{Y^*} + cste \\
&= \theta^\top \hat{\Sigma}^{-1} \theta - 2\theta^\top \hat{\Sigma}^{-1} \hat{\mu} + \hat{\mu}^\top \hat{\Sigma}^{-1} \hat{\mu} + cste_{Y^*} + cste
\end{aligned}$$

where we successively used equations [10,12], separate the terms in θ and $\theta^\top \theta$, used the definition of the distribution of $C(\theta)$.

By taking the third and last lines and by equalizing the terms in $\theta^\top \theta$ and $\theta^\top$, we obtain:

$$\hat{\Sigma} = (\Sigma^{-1} + \frac{X^{*\top} X^*}{\sigma^2})^{-1} \tag{16}$$

$$\hat{\mu} = \hat{\Sigma}(\Sigma^{-1}\mu + \frac{X^{*\top} Y^*}{\sigma^2}) \tag{17}$$

Thus,

$$C(\theta) \propto e^{\frac{-1}{2}(\theta - \hat{\mu})^\top \hat{\Sigma}^{-1} (\theta - \hat{\mu}) + cste_{Y^*}} \tag{18}$$

ii. Integration

$$\begin{aligned}
\Pr(Y^*|X^*, X, Y) &= \int \Pr(Y^*|X^*, \theta, X, Y) \Pr(\theta|X, Y) d\theta \\
&= \int C(\theta) d\theta \\
&\propto \int e^{\frac{-1}{2}[(\theta - \hat{\mu})^\top \hat{\Sigma}^{-1} (\theta - \hat{\mu}) + cste_{Y^*}]} d\theta \\
&\propto e^{\frac{-1}{2} cste_{Y^*}} \text{ using the normalization property}
\end{aligned}$$

This proves that $cste_Y^*$ contains all the information about the distribution of $Y^*|X^*, X, Y$. Hence,

$$cste_Y^* = (Y^* - \mu_S)^\top S^{-1} (Y^* - \mu_S) \tag{19}$$

iii. Find $cste_Y^*$: By taking the previous computation of $-2\log C(\theta)$ and equalizing again the third and last lines we can obtain:

$$\frac{Y^{*\top} Y^*}{\sigma^2} = \hat{\mu}^\top \hat{\Sigma}^{-1} \hat{\mu} + cste_{Y^*} + cste$$

Thus,

$$\begin{aligned} cste_{Y^*} &= \frac{Y^{*\top}Y^*}{\sigma^2} - \hat{\mu}^\top\hat{\Sigma}^{-1}\hat{\mu} + cste \\ &= \frac{Y^{*\top}Y^*}{\sigma^2} - (\Sigma^{-1}\mu + \frac{X^{*\top}Y^*}{\sigma^2})^\top\hat{\Sigma}\hat{\Sigma}^{-1}\hat{\Sigma}(\Sigma^{-1}\mu + \frac{X^{*\top}Y^*}{\sigma^2}) + cste \\ &= \frac{Y^{*\top}Y^*}{\sigma^2} - (\frac{X^tY}{\sigma^2} + \frac{X^{*\top}Y^*}{\sigma^2})^\top\hat{\Sigma}(\frac{X^tY}{\sigma^2} + \frac{X^{*\top}Y^*}{\sigma^2}) + cste \\ &= \frac{Y^{*\top}Y^*}{\sigma^2} - \frac{Y^{*\top}X^*}{\sigma^2}\hat{\Sigma}\frac{X^{*\top}Y^*}{\sigma^2} - 2\frac{Y^{*\top}X^*}{\sigma^2}\hat{\Sigma}\frac{X^tY}{\sigma^2} + cste \\ &= Y^{*\top}(\frac{I}{\sigma^2} - \frac{X^*\hat{\Sigma}X^{*\top}}{\sigma^4})Y^* - 2\frac{Y^{*\top}X^*\hat{\Sigma}X^tY}{\sigma^4} + cste \\ &= (Y^* - \mu_S)^\top S^{-1}(Y^* - \mu_S) + cste \end{aligned}$$

Therefore by equalizing terms in Y^* and $Y^{*\top}Y^*$ we obtain:

$$S = (\frac{I}{\sigma^2} - \frac{X^*\hat{\Sigma}X^{*\top}}{\sigma^4})^{-1} \tag{20}$$

$$\mu_S = S(\frac{X^*\hat{\Sigma}X^\top Y}{\sigma^4}) \tag{21}$$

6 Appendix: Proof of the Counter Example 3.1

In this paragraph we will find a projection which can be more precise than the projection in the eigen vectors. To do so, we assume that $\sigma = 1$, $\lambda = 1$ and

$$X = X^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \tag{22}$$

Moreover we are projecting only in 1 dimension. Thus we first project in the span of the first eigen vector P:

$$P = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{23}$$

and we compare it with the projection in the span of ϵ:

$$\epsilon = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{24}$$

First we compute all matrices from paragraph 5.2:
using equation 13 and 16.

$$\hat{\Sigma} = (2\frac{X^\top X}{\sigma^2} + \frac{I}{\lambda^2})^{-1} = (2\frac{I_2}{1} + \frac{I_2}{1})^{-1} = \frac{1}{3}I_2 \tag{25}$$

using equation 1:

$$S = (\frac{I}{\sigma^2} - \frac{X^*\hat{\Sigma}X^{*\top}}{\sigma^4})^{-1} = (\frac{I}{1} - \frac{X^*X^{*\top}}{3})^{-1} = (\begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix})^{-1} = \begin{pmatrix} \frac{3}{2} & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{26}$$

using equation 2

$$\mu_S = SX^\top\hat{\Sigma}XY = \frac{1}{3}\begin{pmatrix} \frac{3}{2} & 0 & 0 \\ 0 & \frac{3}{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}Y = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix}Y \tag{27}$$

Now for the span of P with the exact same computation as above but with only one vector:

$$\hat{\Sigma}_P = \frac{1}{3} \tag{28}$$

$$S_P = \begin{pmatrix} \frac{3}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{29}$$

$$\mu_P = S_P X_P^\top \hat{\Sigma}_P X_P Y = \frac{1}{3} \begin{pmatrix} \frac{3}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} Y = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} Y \tag{30}$$

Now for ϵ:

$$\hat{\Sigma}_\epsilon = (2\frac{\epsilon^\top X^\top X \epsilon}{\sigma^2} + \frac{I}{\lambda^2})^{-1} = (2\frac{\epsilon^\top \epsilon}{1} + \frac{I}{1})^{-1} = (2\frac{2}{1} + \frac{I}{1})^{-1} = \frac{1}{5} \tag{31}$$

$$S_\epsilon = (\frac{I}{\sigma^2} - \frac{X^* \epsilon \hat{\Sigma}_\epsilon \epsilon^\top X^{*\top}}{\sigma^4})^{-1} = (\frac{I}{1} - \frac{X^* \epsilon \epsilon^\top X^{*\top}}{5})^{-1} = \begin{pmatrix} \frac{4}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{4}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{32}$$

$$\mu_\epsilon = S_\epsilon (X^* \epsilon) \hat{\Sigma}_\epsilon (X \epsilon)^\top Y = \frac{1}{5} \begin{pmatrix} \frac{4}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{4}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} Y = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 \\ 0 & 0 & 0 \end{pmatrix} Y \tag{33}$$

Now we can compute the KL divergence terms:

$$(\mu_S - \mu_P)^\top S^{-1} (\mu_S - \mu_P) = Y^\top \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 \end{pmatrix} Y = \frac{y_2^2}{6}$$

$$\mathrm{Tr}(S^{-1} S_P) = \mathrm{Tr} \begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{3}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \frac{8}{3}$$

$$\log(\frac{|S|}{|S_P|}) = \log(\frac{3}{2})$$

Now for ϵ:

$$(\mu_S - \mu_\epsilon)^\top S^{-1} (\mu_S - \mu_\epsilon) = Y^\top \begin{pmatrix} \frac{1}{6} & \frac{-1}{3} & 0 \\ \frac{-1}{3} & \frac{1}{6} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{6} & \frac{-1}{3} & 0 \\ \frac{-1}{3} & \frac{1}{6} & 0 \\ 0 & 0 & 0 \end{pmatrix} Y$$

$$= \frac{1}{54}(5y_1^2 + 5y_2^2 - 8y_1 y_2)$$

$$\mathrm{Tr}(S^{-1} S_\epsilon) = \mathrm{Tr} \begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{4}{3} & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{4}{3} & 0 \\ 0 & 0 & 1 \end{pmatrix} = \frac{25}{9}$$

$$\log(\frac{|S|}{|S_\epsilon|}) = \log \frac{27}{20}$$

Finally:

$$D_P \approx \frac{y_2^2}{6} + 0.07$$

$$D_\epsilon \approx \frac{1}{54}(5y_1^2 + 5y_2^2 - 8y_1 y_2) + 0.08$$

If we take $y_2 = 0$ and $y_1 = 0$:

$$D_\epsilon \approx 0.08 \geq D_P \approx 0.07$$

If we take $y_2 = 1$ and $y_1 = 0$:

$$D_\epsilon \approx 0.17 \leq D_P \approx 0.24$$

7 Appendix: Computation of equation 4

First we are taking the Singular Value Decomposition of X and X^*:

$$X = URV \text{ and } X^* = UR^*V$$

with $R = \operatorname{diag}(r_i)_{i=1,..,n} \in \mathbb{R}^{n \times d}$, $R^* = \operatorname{diag}(r_i^*)_{i=1,..,n} \in \mathbb{R}^{n \times d}$ and U, V 2 orthogonal matrices of respectively size n and d. Using equation 1 we obtain:

$$S = UD_SU^\top \tag{34}$$

with

$$D_S = (I - R^*(R^\top R + R^{*\top}R^* + I)^{-1}R^{*\top})^{-1}$$

Using equation 2 we obtain with the notation $Q = U^\top Y$:

$$\mu_S = UD_SR^*(R^\top R + R^{*\top}R^* + I)^{-1}R^\top Q \tag{35}$$

For the PCA projection we can notice with the notation $P = P_{PCA}$ that P is containing the first k eigenvectors of $X^\top X$ which are the first k lines of V, even if it means rearranging the order of its lines. Thus,

$$PP^\top = V^\top(\begin{matrix} I_k & 0 \\ 0 & 0 \end{matrix})V$$

$PP^\top$ and $X^\top X$ are symmetric, have the same eigen vectors and thus they commute. Therefore, we have:

$$(P^\top X^\top XP + P^\top X^{*\top}X^*P + I)^{-1} = P^\top((X^\top X + X^{*\top}X^* + I))^{-1}P$$

Hence,

$$S_{PCA} = UD_{PCA}U^\top \tag{36}$$

with

$$\begin{aligned} D_{PCA} &= (I - R^*VPP^\top V^\top(R^\top R + R^{*\top}R^* + I)^{-1}VPP^\top V^\top R^{*\top})^{-1} \\ &= (I - R^*(R^\top R + R^{*\top}R^* + I)_k^{-1}R^{*\top})^{-1} \end{aligned}$$

with the notation $_k$ which is equivalent to the multiplication by $(\begin{matrix} I_k & 0 \\ 0 & 0 \end{matrix})$ or the selection to the first k diagonal terms. Moreover,

$$\mu_{PCA} = UD_{PCA}R^*(R^\top R + R^{*\top}R^* + I)_k^{-1}R^\top Q \tag{37}$$

Thus,

$$\begin{aligned} \mu_S - \mu_{PCA} &= U(D_SR^*(R^\top R + R^{*\top}R^* + I)^{-1} - D_{PCA}R^*(R^\top R + R^{*\top}R^* + I)_k^{-1})R^\top Q \\ &:= UD_kR^\top Q \end{aligned}$$

Therefore,

$$\begin{aligned} (\mu_S - \mu_{PCA})^\top S^{-1}(\mu_S - \mu_{PCA}) &= Q^\top RD_kD_S^{-1}D_kR^\top Q \\ &:= Q^\top \Delta Q \end{aligned}$$

If we look at a diagonal coefficients of Δ: if $i \in$ the chosen K eigen vectors

$$\Delta_i = 0$$

Otherwise:

$$\begin{aligned} \Delta_i &= r_i^2(\frac{d_i^S r_i^*}{1 + r_i^2 + r_i^{*2}})^2 \frac{1}{d_i^S} \\ &= \frac{r_i^2 r_i^{*2}}{(1 + r_i^2 + r_i^{*2})^2(1 - \frac{r_i^{*2}}{1 + r_i^2 + r_i^{*2}})} \\ &= \frac{r_i^2 r_i^{*2}}{(1 + r_i^2 + r_i^{*2})(1 + r_i^2)} \end{aligned}$$

Then,

$$(\mu_S - \mu_K)^\top S^{-1}(\mu_S - \mu_K) = \sum_{i \in \bar{K}} q_i^2 \Delta_i$$

Moreover,

$$\begin{aligned}
\mathrm{Tr}(S^{-1}S_{PCA}) &= \mathrm{Tr}(D_S^{-1}D_{PCA}) \\
&= k + \sum_{i \in \bar{K}} 1 - \frac{r_i^{*2}}{1 + r_i^2 + r_i^{*2}} + n - d \\
&= n - \sum_{i \in \bar{K}} \frac{r_i^{*2}}{1 + r_i^2 + r_i^{*2}}
\end{aligned}$$

$$\log(\frac{|S|}{|S_K|}) = -\sum_{i \in \bar{K}} \log(1 - \frac{r_i^{*2}}{r_i^2 + r_i^{*2} + 1})$$

Finally,

$$D_{PCA} = \frac{1}{2}(\sum_{i \in \bar{K}} -\frac{r_i^{*2}}{1 + r_i^2 + r_i^{*2}} - \log(1 - \frac{r_i^{*2}}{r_i^2 + r_i^{*2} + 1}) + Q_i^2 \frac{r_i^2 r_i^{*2}}{(1 + r_i^2 + r_i^{*2})(1 + r_i^2)})$$

8 Appendix: Computation of equation 5

In this part we assume A, B2, C and still $\sigma = \lambda = 1$. We still have for the global space the equations 34 and 35.

Now we need to compute the different terms of D_{rand}. The projection is $\epsilon := P_{rand} = (\epsilon_1, .., \epsilon_d)^\top$ with $\forall i \in 1, ., d : \epsilon_i \sim \mathcal{N}(0, I_k)$. First we compute:

$$\begin{aligned}
(\epsilon^\top X^\top X \epsilon + \epsilon^\top X^{*\top} X^* \epsilon + I_k)^{-1} &= (\epsilon^\top V^\top R^\top R V \epsilon + \epsilon^\top V^\top R^{*\top} R^* V \epsilon + I_k)^{-1} \\
&= (\epsilon^\top V^\top R^\top R V \epsilon + \epsilon^\top V^\top R^{*\top} R^* V \epsilon + I_k)^{-1} \\
&= (\tilde{\epsilon}^\top R^\top R \tilde{\epsilon} + \tilde{\epsilon}^\top R^{*\top} R^* \tilde{\epsilon} + I_k)^{-1} \\
&= (\sum_{i=1}^{d} r_i^2 \tilde{\epsilon}_i \tilde{\epsilon}_i^\top + r_i^{*2} \tilde{\epsilon}_i \tilde{\epsilon}_i^\top + I_k)^{-1} \\
&= \frac{I_k}{1 + \sum_{i=1}^d (r_i^2 + r_i^{*2})} \\
&:= \Sigma I_k
\end{aligned}$$

where we used the notation: $\forall i \in 1, .., d, \tilde{\epsilon}_i := V \epsilon_i \sim \mathcal{N}(0, I_k)$ and assumption C.

Then:

$$\begin{aligned}
S_{rand} &= (I_n - X^* \epsilon(\epsilon^\top X^\top X \epsilon + \epsilon^\top X^{*\top} X^* \epsilon + I_k)^{-1} X^{*\top})^{-1} \\
&= U(I - \Sigma R^* \tilde{\epsilon} \tilde{\epsilon}^\top R^{*\top})^{-1} U^\top \\
&= U(I + \Sigma R^* \tilde{\epsilon}(I_k + \tilde{\epsilon}^\top R^{*\top} \Sigma R^* \tilde{\epsilon})^{-1} \tilde{\epsilon}^\top R^{*\top}) U^\top \\
&= U(I + \Sigma R^* \tilde{\epsilon}(I_k + \Sigma \sum_{i=1}^{d} r_i^{*2} \tilde{\epsilon}_i \tilde{\epsilon}_i^\top)^{-1} \tilde{\epsilon}^\top R^{*\top}) U^\top \\
&= U(I + (\frac{1}{1 + \sum_{i=1}^d (r_i^2 + r_i^{*2)}}) \frac{1}{(1 + \frac{1}{1 + \sum_{i=1}^d (r_i^2 + r_i^{*2})} \sum_{i=1}^d r_i^{*2})} R^* \tilde{\epsilon} \tilde{\epsilon}^\top R^{*\top}) U^\top \\
&= U(I + \frac{1}{(1 + \sum_{i=1}^d r_i^2 + 2\sum_{i=1}^d r_i^{*2})} R^* \tilde{\epsilon} \tilde{\epsilon}^\top R^{*\top}) U^\top
\end{aligned}$$

where we successively used equation 1, $\tilde{\epsilon} = V\epsilon$, Woodbury's identity and the previous result. Thus,

$$S_{rand} = U(I + \frac{1}{(1 + \sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} R^* \tilde{\epsilon}\tilde{\epsilon}^\top R^{*\top})U^\top \tag{38}$$

Then using equation 2 and the projection with ϵ we have:

$$\begin{aligned}
\mu_{rand} &= S_{rand}(X^* \epsilon \Sigma \epsilon^\top X^\top Y) \\
&= S_{rand} U(R^* \tilde{\epsilon} \Sigma \tilde{\epsilon}^\top R^\top U^\top Y) \\
&= U(I + \frac{1}{(1 + \sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} R^* \tilde{\epsilon}\tilde{\epsilon}^\top R^{*\top})(R^* \tilde{\epsilon} \Sigma \tilde{\epsilon}^\top R^\top U^\top Y) \\
&= \Sigma U(I + \frac{1}{(1 + \sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} R^* \tilde{\epsilon}\tilde{\epsilon}^\top R^{*\top}) R^* \tilde{\epsilon}\tilde{\epsilon}^\top R^\top Q
\end{aligned}$$

where we used successively $\tilde{\epsilon} = V\epsilon$, orthogonality of U and $Q = U^\top Y$.

As ϵ and $\tilde{\epsilon}$ are identically distributed, we will rename in the following $\tilde{\epsilon}$ as ϵ. Therefore using 35 and previous result we obtain:

$$\mu_S - \mu_{rand} = UFQ$$

with the notation: $F := ((I - R^*(R^\top R + R^{*\top} R^* + I)^{-1} R^{*\top})^{-1} R^*(R^\top R + R^{*\top} R^* + I)^{-1} - \Sigma(I + \frac{1}{(1+\sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} R^* \epsilon\epsilon^\top R^{*\top}) R^* \epsilon\epsilon^\top) R^\top$.

Then we have $\forall i, j \in 1, .., p$:

$$F_{ij} = \begin{cases} \frac{r_i r_i^*}{1+r_i^2} - \Sigma(1 + \frac{1}{(1+\sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} r_i^{*2} \epsilon_i^\top \epsilon_i) r_i^* r_i \epsilon_i^\top \epsilon_i \text{ if } i = j \\ -\Sigma(1 + \frac{1}{(1+\sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} r_i^* r_j^* \epsilon_i^\top \epsilon_j) r_i^* r_j \epsilon_i^\top \epsilon_j \text{ else} \end{cases}$$

$$\begin{aligned}
(\mu_S - \mu_{rand})^\top S^{-1} (\mu_S - \mu_{rand}) &= Q^\top F^\top U^\top U D_S^{-1} U^\top U F Q \\
&= Q^\top F(I - R^*(R^\top R + R^{*\top} R^* + I)^{-1} R^{*\top}) F Q \\
&:= Q^\top G Q
\end{aligned}$$

Then we have $\forall i, j \in 1, .., d$:

$$G_{ij} = \begin{cases} (1 - \frac{r_i^{*2}}{1+r_i^2+r_i^{*2}}) F_{ii}^2 + \sum_{k \neq i} (1 - \frac{r_k^{*2}}{1+r_k^2+r_k^{*2}}) F_{ik}^2 \text{ if } i = j \\ F_{ij}(F_{jj}(1 - \frac{r_j^{*2}}{1+r_j^2+r_j^{*2}}) + F_{ii}(1 - \frac{r_i^{*2}}{1+r_i^2+r_i^{*2}})) + \sum_{k \neq i,j} (1 - \frac{r_k^{*2}}{1+r_k^2+r_k^{*2}}) F_{ik} F_{kj} \text{ else} \end{cases}$$

Finally we have:

$$(\mu_S - \mu_{rand})^\top S^{-1} (\mu_S - \mu_{rand}) = \sum_{i,j=1}^{d} Q_i Q_j G_{ij}$$

Then,

$$\begin{aligned}
\mathrm{Tr}(S^{-1} S_{rand}) &= \mathrm{Tr}(D_S^{-1}(I + \frac{1}{(1 + \sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} R^* \epsilon\epsilon^\top R^{*\top})) \\
&= \sum_{i=1}^{d} (1 - \frac{r_i^{*2}}{1 + r_i^2 + r_i^{*2}})(1 + \frac{1}{(1 + \sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} r_i^{*2} \epsilon_i^\top \epsilon_i) + n - d
\end{aligned}$$

using equations 34, 38.

Moreover,

$$\begin{aligned}
\log \frac{|S|}{|S_{rand}|} &= \log |D_S| - \log |I + \frac{1}{(1 + \sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} R^* \epsilon\epsilon^\top R^{*\top}| \\
&= \sum_{i=1}^{d} - \log(1 - \frac{r_i^{*2}}{1 + r_i^2 + r_i^{*2}}) - \mathrm{Tr}(\frac{1}{(1 + \sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} R^* \epsilon\epsilon^\top R^{*\top}) \\
&= \sum_{i=1}^{d} - \log(1 - \frac{r_i^{*2}}{1 + r_i^2 + r_i^{*2}}) - \frac{1}{(1 + \sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})} r_i^{*2} \epsilon_i^\top \epsilon_i
\end{aligned}$$

using that $\log|I+X| \approx \mathrm{Tr}(X)$ which is verified under assumption B2.
Finally with $\hat{\Sigma}_{rand} := \frac{1}{(1+\sum_{i=1}^{d} r_i^2 + 2\sum_{i=1}^{d} r_i^{*2})}$,

$$D_{rand} = \frac{1}{2}(\sum_{i,j=1}^{d} Q_i Q_j G_{ij} + \sum_{i=1}^{d} -\log(1-\frac{r_i^{*2}}{1+r_i^2+r_i^{*2}}) - \frac{r_i^{*2}}{1+r_i^2+r_i^{*2}} - \hat{\Sigma}_{rand}\frac{r_i^{*4}\epsilon_i^\top \epsilon_i}{1+r_i^2+r_i^{*2}}) \tag{39}$$

9 Appendix: Proof of equation 6

For small outputs equations 4 and 5 become under assumption C2:

$$D_{PCA} = \frac{1}{2}(\sum_{i=k+1}^{d} -\frac{r_i^{*2}}{1+r_i^2+r_i^{*2}} - \log(1-\frac{r_i^{*2}}{r_i^2+r_i^{*2}+1})) \tag{40}$$

and

$$D_{rand} = \frac{1}{2}(\sum_{i=1}^{d} -\frac{r_i^{*2}}{1+r_i^2+r_i^{*2}} - \log(1-\frac{r_i^{*2}}{1+r_i^2+r_i^{*2}}) - \hat{\Sigma}_{rand}\frac{r_i^{*4}\epsilon_i^\top \epsilon_i}{1+r_i^2+r_i^{*2}}) \tag{41}$$

Thus,

$$\begin{aligned}
&D_{PCA} - E_\epsilon(D_{rand}) \\
&= \frac{1}{2}(\sum_{i=1}^{k} \log(1-\frac{r_1^{*2}}{1+r_1^2+r_1^{*2}}) + \frac{r_1^{*2}}{1+r_1^2+r_1^{*2}} + E_\epsilon(\hat{\Sigma}_{rand}\sum_{i=1}^{d}\frac{r_i^{*4}\epsilon_i^\top\epsilon_i}{1+r_i^2+r_i^{*2}})) \\
&= \frac{1}{2}(\sum_{i=1}^{k} \log(\frac{2}{3}) + \frac{1}{3} + E_\epsilon(\hat{\Sigma}_{rand}\sum_{i \neq k+1}^{d}\frac{\epsilon_i^\top\epsilon_i}{3} + \hat{\Sigma}_{rand}\frac{(1+\delta)^4\epsilon_1^\top\epsilon_1}{2+(1+\delta)^2})) \\
&= \frac{k}{2}(\log(\frac{2}{3}) + \frac{1}{3} + \hat{\Sigma}_{rand}\sum_{i \neq k+1}^{d}\frac{k}{3} + \hat{\Sigma}_{rand}\frac{k(1+\delta)^4}{2+(1+\delta)^2}) \\
&= \frac{k}{2}(\log(\frac{2}{3}) + \frac{1}{3} + \frac{(d-1)}{12(1+3d+4\delta+\delta^2)} + \frac{1}{4(1+3d+4\delta+\delta^2)}\frac{(1+\delta)^4}{2+(1+\delta)^2}) \\
&:= f(\delta)
\end{aligned}$$

where we successively used assumption C2, D2, definition of ϵ. We have $f(0) = \frac{k}{2}(\log(\frac{2}{3})) + \frac{1}{3} + \frac{d}{12(1+3d)})$. As $d \geq 1$:

$$-0.024 * k \leq f(0) \leq \frac{1}{2}(\log(\frac{2}{3}) + \frac{1}{3} + \frac{1}{36}) \leq 0 \tag{42}$$

Moreover:

$$\begin{aligned}
\frac{df}{d\delta} &\propto 18\delta^6 + (81+27d)\delta^5 + (244+134d)\delta^4 + (374d+418)\delta^3 + (584d+334)\delta^2 \qquad (43) \\
&+ (447d+93)\delta + 126d \geq 0 \qquad (44)
\end{aligned}$$

Hence, f is an increasing function.

10 Appendix: Proof of equation 8

For large outputs equations 4 and 5 become:

$$D_{PCA} = \frac{1}{2}(\sum_{i=k+1}^{d} \frac{r_i^2 r_i^{*2}}{(1+r_i^2+r_i^{*2})(1+r_i^2)})$$

$$D_{rand} = \frac{1}{2}(\sum_{i,j=1}^{d} Q_i Q_j G_{ij})$$

Moreover:

$$\lim_{\delta \to \infty} \hat{\Sigma}_{rand} = 0$$

and

$$\lim_{\delta \to \infty} \Sigma = 0$$

Thus $\forall i, j \in 1, .., d$:

$$\lim_{\delta \to \infty} F_{ij} = \lim_{\delta \to \infty} \begin{cases} \frac{r_i r_i^*}{1+r_i^2} \text{ if } i = j \\ 0 \text{ else} \end{cases}$$

Hence,

$$\lim_{\delta \to \infty} G_{ij} = \lim_{\delta \to \infty} \begin{cases} (\frac{r_i^2 r_i^{*2}}{(1+r_i^2)(1+r_i^2+r_i^{*2})}) \text{ if } i = j \\ 0 \text{ else} \end{cases}$$

Therefore:

$$\lim_{\delta \to \infty} \mathop{\mathbb{E}}_{\epsilon}(D_{rand}) = \lim_{\delta \to \infty} \frac{1}{2}(\sum_{i=1}^{d} Q_i^2 (\frac{r_i^2 r_i^{*2}}{(1+r_i^2)(1+r_i^2+r_i^{*2})}))$$

Thus:

$$\lim_{\delta \to \infty} D_{PCA} - \mathop{\mathbb{E}}_{\epsilon}(D_{rand}) = -\sum_{i=1}^{k} \frac{Q_i^2 r_i^4}{2(1+2r_i^2)(1+r_i^2))}$$

As X is the identity we obtain finally,

$$\lim_{\delta \to \infty} D_{PCA} - \mathop{\mathbb{E}}_{\epsilon}(D_{rand}) = -\sum_{i=1}^{k} \frac{Q_i^2}{12}$$