Mismatched Estimation of Symmetric Rank-One Matrices Under Gaussian Noise

Conference Paper

Author(s):
Pourkamali, Farzad; Macris, Nicolas

Publication date:
2022-03-02

Permanent link:
https://doi.org/10.3929/ethz-b-000535287

Rights / license:
In Copyright - Non-Commercial Use Permitted
Mismatched Estimation of Symmetric Rank-One Matrices Under Gaussian Noise

Farzad Pourkamali and Nicolas Macris
SMILS, EPFL, Lausanne, Switzerland
Emails: {farzad.pourkamali, nicolas.macris}@epfl.ch

Abstract—We consider the estimation of an $n$-dimensional vector $s$ from noisy element-wise measurements of $ss^T$, a problem that frequently arises in statistics and machine learning. We investigate a mismatched Bayesian inference setting in which the statistician is unaware of some of the parameters. For the particular case of Gaussian priors for the vector $s$ and additive noise, we derive the complete exact analytic expression for the asymptotic mean squared error (MSE) in the large system size limit. Our formulas demonstrate that estimation is still possible in the mismatched case. Also, the minimum MSE (MMSE) can be achieved by selecting a non-trivial set of parameters beyond the matched case. Our technique is based on the asymptotic behavior of spherical integrals and can be used as long as the statistician chooses a rotationally invariant prior.

I. INTRODUCTION

Many problems in machine learning and statistics can be expressed as estimating a low-rank matrix from its noisy observation. Examples are sparse PCA [1], the spiked Wigner model, community detection [2]. For the rank-one symmetric case, the problem is formulated as follows: a vector $s \in \mathbb{R}^n$ is generated with i.i.d. elements distributed according to $s_i \sim \mathcal{N}^*$, the matrix $ss^T$ is observed through an element-wise additive white gaussian noise channel. The goal is to estimate the vector $s$ upon observing the noisy version of $ss^T$.

The statistical and computational limits of this problem have been extensively studied. Most works have so far considered the "Bayes-optimal" setting, in which the prior $\mathcal{P}^*$ and possibly other hyper-parameters (e.g., SNR) are known to the statistician. In the Bayes-optimal setting, computing the mutual information enables us to compute the minimum mean squared error (MMSE) and derive the information-theoretical limits of the estimation. The analytical but highly non-rigorous replica and cavity methods rooted in statistical physics have been used to derive expressions for the mutual information between the true signal and the observation matrix [3]. These expressions were already rigorously derived in early work [4] for binary signals using Guerra-Toninelli interpolation [5]. Later the problem has been studied in much detail for general signals, in [6] using approximate message passing (AMP) and spatial coupling, in [7] by Guerra-Toninelli interpolation and Aizenman-Sims-Starr methods. Further, [8], [9] used the adaptive interpolation method to rigorously prove the limiting expressions of mutual information and MMSE. All these methods crucially rely on the assumption that the prior and the parameters of the estimation problem are known to the statistician. The Bayes law then induces remarkable identities that enable the analysis to proceed. In the present case, we lack such identities.

Despite the vast amount of work on this problem in the Bayes-optimal setting, to the best of our knowledge, there is no rigorous result for the mismatched case corresponding to the realistic situation where the statistician does not know the true prior or/and hyper-parameters, and can only make assumptions about them. Mismatched inference for the scalar and vector estimation problems has been considered in [10], [11]. In particular, [10] proved a result relating the MSE in the mismatched inference to the relative entropy of the true prior and the statistician’s prior. We follow this work and define the MSE similarly (up to natural modification for the matrix case).

The main contribution of this paper is to compute the asymptotic mismatched MSE for the rank-one matrix estimation problem in the large $n$ limit. Our approach uses results on the spherical integrals from the mathematical physics literature [12]. A primary assumption in our method that would be difficult to dispense of, is the rotational invariance of the statistician's prior. Despite this restriction, we can study non-rotation invariant true priors, non-symmetric matrix estimation, higher-ranks (finite w.r.t $n \to +\infty$). In this short note, we limit ourselves to the theoretical limits of mismatched estimation for the case of Gaussian priors (both for the true and the statistician's) and postpone the detailed study of the more general cases to a forthcoming detailed work. As will become clear in section III, already under this limited setting, the phase transitions phenomenology is quite rich.

The rest of the paper is organized as follows. In Section II, we introduce the setting and formulate the problem. Section III describes the main result and discusses it in several special cases, followed by the proof sketch of the main theorem in Section IV. Lastly, we conclude the paper with some remarks and possible future directions for this line of work.

II. PROBLEM SETTING

Suppose the ground-truth vector $s \in \mathbb{R}^n$ is generated with i.i.d. elements from $\mathcal{P}^* = \mathcal{N}(0, \sigma^2)$, the observed matrix is

$$Y = \sqrt{\frac{\lambda}{n}} ss^T + Z$$

where we call (with an abuse of language) $\lambda \in \mathbb{R}_+$ the signal-to-noise-ratio (SNR), and the noise matrix $Z$ is a symmetric matrix with i.i.d. $\mathcal{N}(0,1)$ off-diagonal and $\mathcal{N}(0,2)$ diagonal entries. This model is called the Spiked-Wigner model. The
purpose of the scaling factor $\frac{1}{n}$ is to make the inference problem neither trivially easy nor completely impossible in the large system limit.

The statistician is aware that the channel is additive Gaussian and that the true prior is a centered Gaussian, but he does not know the values $\lambda$ and $\sigma$. He assumes values $\lambda'$ and $\sigma'$ as the SNR and the prior variance. Following the Bayesian estimation principle, he chooses the posterior mean as the estimate of the SNR and the prior variance. Therefore almost everywhere uniform convergence is difficult to establish from general principles. However, we conjecture that it holds and that eq. (2) holds almost everywhere (i.e., except possibly at phase transition lines).

Remark 2. One can see that the normalized MSE, i.e. $\sigma^{-4}\text{MSE}_n$, can be expressed as a function of the three dimensionless variables $\lambda\sigma^4, \lambda'\sigma'^4, \frac{\sigma^2}{\sigma'^2}$. This allows us to study the problem for the case $\sigma'^2 = 1$ and easily generalize the analysis to other cases by rescaling the parameters.

The MSE is illustrated for the case of $\sigma = 1, \lambda = 2$ in Fig. 1. The observed behavior is generic for $\lambda\sigma^4 > 1$. We observe one phase transition line and an intermediate region where estimation better than chance is possible, in the sense that the MSE is smaller than $\sigma^4$. We refer to the caption of Fig. 1 for details. In the case $\sigma = 1$ and $\lambda < 1$, or more generally $\lambda\sigma^4 < 1$, it is easy to see from Eq. (2) that the intermediate region disappears and the MSE is always greater or equal to $\sigma^4$ (the phase transition line is still present technically speaking).

![Fig. 1: Plot of MSE according to Eq. (2) for $\sigma = 1, \lambda = 2$. The solid leftmost (red) curve is a phase transition line (the MSE is continuous but the derivative is discontinuous). On the left of this curve $MSE = \sigma^4 = 1$. In the intermediate region between the solid leftmost (red) curve and the dashed (red) curve the MSE takes values less than $\sigma^4 = 1$. In this intermediate region estimation better than chance is possible. On the dotted (green) curve the MSE attains the value $\sigma^4 = 1$ and the dashed (red) curve is taken as the boundary of the region. The MSE equals $\sigma^4 = 1$ on the dashed (red) line and takes higher values in the region on the right hand side of this line. Note that this is not a phase transition line, and the MSE is a perfectly analytic function there. The analytical expressions of the phase transition line, as well as dotted and dashed lines can easily be written down from eqs. (2) and (3). For $\sigma = 1, \lambda = 2$ the dotted (resp. dashed) curves have horizontal asymptotes $\lambda' = 8$ (resp. $\lambda' = 2$). See also figures 2 and 3 in [14]).](image-url)
A. Inference with Matched SNR

Suppose that the statistician fully knows the channel and can choose \( \lambda' = \lambda \). The asymptotic mismatched MSE in the limit \( n \to \infty \) is:

\[
\text{if } \sigma' \leq \sigma, \text{ MSE} = \left\{ \begin{array}{ll}
\sigma^2 + \frac{1}{n(\lambda^2 - \frac{1}{\sigma})} & \text{if } \lambda \leq \frac{1}{\sigma^2} \\
\sigma^2 + \frac{1}{n(\lambda^2 - \frac{1}{\sigma^2})} & \text{if } \lambda \geq \frac{1}{\sigma^2} 
\end{array} \right.
\]

\[
\text{if } \sigma' \geq \sigma, \text{ MSE} = \left\{ \begin{array}{ll}
\sigma^4 + \frac{1}{n(\lambda^2 - \frac{1}{\sigma^2})} & \text{if } \lambda \leq \frac{1}{\sigma^2} \\
\sigma^4 + \frac{1}{n(\lambda^2 - \frac{1}{\sigma^2})} & \text{if } \lambda \geq \frac{1}{\sigma^2} 
\end{array} \right.
\]

For \( \sigma = 1 \) the MSE is plotted as a function of SNR for various values of \( \sigma' \) in Fig. 2. When \( \sigma' > \sigma \), we observe that the MSE increases as the SNR increases (a similar behavior occurs in Fig. 1 in [10] for the scalar case). Although this happens when we are still in the regime of small SNR and estimation is impossible, we find this behavior rather counterintuitive.

![Figure 2: Behavior of the MSE for matched SNR \( \lambda' = \lambda \).](image)

Remark 3. For \( \sigma' = \sigma, \lambda' = \lambda \), we are in the Bayes optimal setting and we recover the minimum MSE (MMSE) in the limit \( n \to \infty \):

\[
\text{MMSE} = \left\{ \begin{array}{ll}
\sigma^2 & \text{if } \lambda \leq \frac{1}{\sigma^2} \\
\frac{1}{n(\lambda^2 - \frac{1}{\sigma^2})} & \text{if } \lambda \geq \frac{1}{\sigma^2} 
\end{array} \right.
\]

This expression is well known and derived previously by a host of different approaches (see [11], [2], [6]–[8]).

As a sanity check of our result for the matched SNR case, with a bit of work we can check explicitly that

\[
\int_0^\infty \text{MSE}(\sigma, \sigma', \lambda, \lambda) - \text{MMSE}(\sigma, \lambda) \, d\lambda = 4D_{KL}(\mathcal{N}(0, \sigma^2), \mathcal{N}(0, \sigma^2)\quad(3)
\]

where \( D_{KL} \) denotes the Kullback-Leibler divergence. This sum-rule for vector channels is derived in [10] (with a factor of 2 instead of 4 in the vector case).

IV. Analysis

A. Mismatched free Energy and MSE

From the statistician’s point of view, the posterior distribution reads up to a normalizing factor

\[
P(x|Y) \propto e^{-\frac{1}{2}\|x\|^2 - \frac{1}{n}Yx^T \sigma^2 P(x)}
\]

\[
\propto e^{-\frac{1}{2}||x||^2 + \frac{1}{n}Tr(Yx^T \sigma^2 P(x)}
\]

where \( P \) is the normal distribution with iid entries and variance \( \sigma^2 \). In deriving the second line, we use the fact that \( \|Y\|_F^2 \) is a constant (because it is being conditioned on). Note that, \( Y \) is symmetric and the upper (or lower) part is distributed as \( (Y_{ij})_{i<j} \sim \mathcal{N}(\frac{1}{n}s_is_j, 1) \), and the diagonal \( (Y_{ii}) \sim \mathcal{N}(\frac{1}{n}s_i^2, 2) \).

The partition function is defined as the normalization factor of the last expression

\[
Z(Y) = \int dx \, e^{-\frac{1}{2n}||x||^2 + \frac{1}{2n}Yx^T P(x)} \quad(5)
\]

and the mismatched free energy is defined as

\[
f_n(\sigma, \sigma', \lambda, \lambda') = -\frac{1}{n}E_{p_x, p_{z_x}}[\ln Z(Y)] \quad(6)
\]

Now we state a lemma relating the mismatched free energy to MSE. This lemma does not require any assumption on priors, and holds as long as the noise is additive Gaussian. Keep in mind that both mismatched free energy and MSE are functions of \( \sigma, \sigma', \lambda, \lambda' \), but for simplicity of notation, we drop the arguments.

Lemma 1.

\[
\frac{df_n}{d\lambda} = -\frac{1}{n}E_{p_x, p_{z_x}}[\ln Z(Y)] \quad(7)
\]

Remark 4. Eq. (7) generalizes the classical I-MMSE relation. Here the mismatched free energy cannot be related to a mutual information. However, note that, in the special case where \( \lambda' = \lambda \) Eq. (7) simplifies slightly and combining with the I-MMSE relation, we obtain that the difference of MSE and MMSE is directly related to a derivative of a relative entropy, equivalent to relations discussed in detail in [10] for vector channels.

Proof of lemma. We have

\[
\frac{df_n}{d\lambda} = \frac{1}{n}E_{p_x, p_{z_x}}[\ln Z(Y)]
\]

and by using a standard Gaussian integration by parts trick,

\[
\frac{df_n}{d\lambda} = \frac{1}{n}E_{p_x, p_{z_x}}[\ln Z(Y)]
\]

Putting these two equations together, the left-hand side of eq. (7) is equal to

\[
\frac{1}{n}E_{p_x, p_{z_x}}[\ln Z(Y)] = \frac{1}{n}E_{p_x, p_{z_x}}[\ln Z(Y)] = \frac{1}{n}E_{p_x, p_{z_x}}[\ln Z(Y)]
\]

Thus, the problem is reduced to computing the (mismatched) free energy. The main idea is to exploit the rotational invariance of the normal distribution that the statistician chooses.
Changing variables $x \to Ux$, for an orthogonal matrix $U \in \mathbb{R}^{n \times n}$, the integral in eq. (5) becomes ($|\det U| = 1$):

$$Z(Y) = \int dx \ e^{-\frac{1}{2} \|Ux\|^2 + \frac{1}{2} \sqrt{n} \ \langle YUxx^T \rangle U^T P(Ux)}$$

$$= \int dx P(x) e^{-\frac{1}{2} \|x\|^2 + \frac{1}{2} \sqrt{n} \ \langle YUxx^T \rangle U^T}$$

Since this holds for any orthogonal matrix $U$, we can take the expectation over the Haar measure on the group of $n \times n$ orthogonal matrices.

$$Z(Y) = \int dx P(x) e^{-\frac{1}{2} \|x\|^2} \int DU e^{\frac{1}{2} \sqrt{n} \ \langle YUxx^T \rangle U^T}$$

where $DU$ denotes the Haar measure.

In the next subsection, we will discuss computing the inner integral in eq. (8).

### B. Spherical Integrals

The spherical integral is defined as:

$$I_n(A, B) = \int DU e^{\frac{1}{2} \langle AUB^T \rangle}$$

where $A, B \in \mathbb{R}^{n \times n}$, and $DU$ denotes the Haar measure over the orthogonal matrices. Note that, this definition can also be extended to the unitary matrices. In the mathematical physics literature, such integrals are often called Harish-Chandra-Itzykson-Zuber (HCIZ) integrals. The interest for these objects dates to the work of the mathematician Harish Chandra [15], and they have been extensively studied and developed in physics and mathematics. In particular, [12] derived the asymptotics of spherical integrals when the rank of matrix $B$ is $O(1)$ w.r.t $n$. We will apply this result for the rank-one $B$ to our problem. For simplicity of notation, we denote the integral by $I_n(\theta, A)$, where $\theta$ is the only non-zero eigenvalue of $B$.

From the definition (9), one may notice that the integral only depends on the eigenvalues of $A, B$. So, it is natural to expect that the asymptotic of the integral depends on the limiting spectral measure of the matrix $A$. The result of [12] is based on the hypothesis that the spectral measure $\mu_A$ converges weakly towards a compactly supported measure $\mu$, and the minimum and maximum eigenvalues of $A$ converge to the finite values $\gamma_{\min}, \gamma_{\max}$, respectively.

For a probability measure $\mu$, the Hilbert (or Stieljes) transform is the map $H_\mu : \mathbb{R} \setminus \text{supp}(\mu) \to \mathbb{R}$, $H_\mu(z) = \int \frac{1}{z - \mu(t)} \ d\mu(t)$. This map is invertible, and denoting its inverse by $H^{-1}_\mu(\cdot)$, for $z$ in range of $H_\mu$ we define the $R$-transform of a probability measure $\mu$ as $R_\mu(z) = H^{-1}_\mu(z) - \frac{1}{2}$.

**Theorem 2** (Guionnet and Maida [12]), Suppose $\mu_A$ converges weakly towards $\mu$. Let $H_{\min} = \lim_{z \to \gamma_{\min}} H_\mu(z)$, $H_{\max} = \lim_{z \to \gamma_{\max}} H_\mu(z)$. Then:

$$\lim_{n \to \infty} -\frac{1}{n} \ln I_n(\theta, A) = \theta \nu(\theta) - \frac{1}{2} \int \ln(1 + 2\theta \nu(\theta) - 2\theta) \ d\mu(t)$$

where

$$\nu(\theta) = \begin{cases} R_\mu(2\theta) & \text{if } H_{\min} \leq 2\theta \leq H_{\max} \\ \gamma_{\min} - \frac{1}{2} & \text{if } 2\theta > H_{\max} \\ \gamma_{\max} - \frac{1}{2} & \text{if } 2\theta < H_{\min} \end{cases}$$

### C. Computing Free Energy

To apply the result from [12], we can rewrite the spherical integral in eq. (8) as

$$I_n(\frac{\sqrt{n}}{2\pi} \|x\|^2, \frac{Y}{\sqrt{n}}) = \int DU e^{\frac{1}{2} Tr (\frac{Y}{\sqrt{n}} Uxx^T U^T)}$$

where

$$\frac{Y}{\sqrt{n}} = \frac{\sqrt{n}}{2\pi} \sigma^2 + \frac{1}{\sqrt{n}} Z$$

and $\frac{1}{\sqrt{n}} Z$ is the suitably normalized Wishart matrix, whose limiting spectral measure is the semi-circle law with density $d\mu_{SC} = \frac{1}{4\sqrt{\delta - x^2}}$. At the same time, the spectral measure of $\frac{1}{\sqrt{n}} Z$ converges almost surely (a.s.) as $n \to \infty$. We have $H_{\mu_{SC}}(z) = \frac{1}{2}(z - \sqrt{z^2 - 4})$ and $R_{\mu_{SC}}(z) = z$.

Let $\gamma_{\min}$ and $\gamma_{\max}$ be the smallest and the largest eigenvalues of $\frac{Y}{\sqrt{n}}$, from the results in [17], as $n \to \infty$ we have (a.s.)

$$\gamma_{\min} = -2, \gamma_{\max} = \begin{cases} 2 & \text{if } \sqrt{\lambda^2} \leq 1 \\ \frac{1}{\sqrt{\lambda^2}} & \text{if } \sqrt{\lambda^2} > 1 \end{cases}$$

So,

$$H_{\min} = -1, H_{\max} = \begin{cases} 1 & \text{if } \sqrt{\lambda^2} \leq 1 \\ \frac{1}{\sqrt{\lambda^2}} & \text{if } \sqrt{\lambda^2} > 1 \end{cases}$$

**Theorem 3**. For all $\sigma, \sigma', \lambda, \lambda'$ positive, the asymptotic free energy of the mismatched inference model is given in eq. (11).

**Proof sketch.** Eq. (8) can be written as

$$Z(Y) = \int dx P(x) e^{\frac{1}{2\pi n} \|x\|^2 + \ln I_n(\frac{\sqrt{n}}{2\pi} \|x\|^2, \frac{1}{\sqrt{n}} Y)}$$

Since $P(x)$ is Gaussian, the integrand in (12) is a function of $\|x\|$, so we can use spherical coordinates to reduce the integral in (12) to a one-dimensional integral.

$$Z(Y) = \frac{2^{-\frac{5}{2}} + 1}{\Gamma(\frac{3}{2}) \sigma^{\frac{5}{2}}} \times \int_0^{+\infty} dp \rho^{-\frac{1}{2}} e^{-\frac{\rho}{2\sigma^2} - \frac{\lambda^2}{\rho} s + \ln I_n(\frac{\sqrt{n}}{2\pi} \|s\|^2, \frac{1}{\sqrt{n}} Y)}$$

where $\rho := \|x\|$, and $\Gamma(.)$ is the Gamma function. Changing variable $\frac{\rho n}{2} \to \rho$, we obtain

$$Z(Y) = \frac{2^{-\frac{5}{2}} + 1}{\Gamma(\frac{3}{2}) \sigma^{\frac{5}{2}}} \times \int_0^{+\infty} dp \ e^{-\frac{p}{\sigma^2} - \frac{1}{\rho^2} \ln p + \frac{\lambda^2}{\rho^2} s - J_n(\frac{\sqrt{n}}{2\pi} \|s\|^2, \frac{1}{\sqrt{n}} Y)}$$

(13)

where $J_n(\theta, \frac{1}{\sqrt{n}} Y) \equiv \frac{1}{n} I_n(\theta, \frac{1}{\sqrt{n}} Y)$. By Theorem 2, $J_n(\theta, \frac{1}{\sqrt{n}} Y)$ converges to a deterministic function $J(\theta, \mu_{SC}(\gamma_{\max}))$. We are interested in $\lim_{n \to \infty} f_n = \lim_{n \to \infty} \mathbb{E}[\frac{1}{n} \ln Z(Y)]$. The prefactors in (13) are independent of $Y$. and
\[
\lim_{n \to \infty} f_n(\sigma, \sigma', \lambda, \lambda') = \begin{cases} 
\frac{-4}{\lambda '} + \frac{1}{\sqrt{\lambda \sigma^2}} - \frac{3}{4} + \ln \lambda \sigma' \\
\frac{1}{4} \ln \sqrt{\lambda \sigma^2} \sigma'^2 - \frac{1}{4} \ln \lambda - \frac{\lambda^4}{4} + \sqrt{\frac{\lambda}{\lambda '}} \frac{\sigma^2}{2 \sigma'^2} + \frac{1}{2} \ln \lambda 
\end{cases}
\]

if \( \lambda^4 \leq 1 \), and \( \lambda \sigma' \geq 1 \)

\[
\frac{1}{2} \lambda^4 > 1, \text{ and } \sqrt{\lambda \lambda '} \geq \frac{1}{\sigma '}
\]

if o.w.

---

**ACKNOWLEDGMENT**

The work of F. P. is supported by the SNSF grant no 200021E 175541. N.M is thankful to Jean Barbier and Emanuele Mingione for numerous discussions.

**REFERENCES**


