INFORMATION-ESTIMATION RELATIONSHIP in MISMATCHED GAUSSIAN CHANNELS

Saloua Chlaily, Chengfang Ren, Pierre-Olivier Amblard, Olivier Michel, Pierre Comon, and Christian Jutten

***Abstract*—In this paper, we investigated the connection between information and estimation measures for mismatched Gaussian models. In addition to the input prior mismatch we take into account the noise mismatch and establish a new relation between relative entropy and excess mean square error. The derived formula shows that the input prior mismatch may be cancelled by the noise mismatch. Finally, an example illustrates the impact of model mismatches on estimation accuracy.**

***Index Terms*—Relative entropy, excess MSE, mutual information, estimation, mismatched Gaussian channel, mismatched input.**

I. Introduction

In information theory, entropy and mutual information are two fundamental information measures. Several works showed that these measures have some connections with estimation measures. Indeed, a well known relationship between entropy and Fisher information is given by de Bruijn identity [1, p.672]. Another link between mutual information and Minimal Mean Square Error (MMSE) has been established by Palomar *et al.* [2] and Guo *et al.* [3] in linear Gaussian communication channels. From an estimation point of view, the latter relation is interesting since MMSE characterizes the optimal achievable performance in Mean Square Error (MSE) sense, which is usually well defined in Bayesian estimation by the MSE of MMSE estimator. However, few works have investigated the mismatched scenario, despite its growing interest in practical applications. In fact in an actual estimation scenario, one cannot ensure the estimation model is correctly specified due to measurement or estimation imperfections [4], [5], [6]. This is particularly important when using several measurements coming from different sensors or devices, a situation referred to as multimodality [7], [8]. Assumptions usually made about the links between modalities may be false, such as independence of noises. Therefore, the influence of model mismatches on estimation performance deserves to be explored.

Mismatched models could also occur in communication channels. A recent work of Verdú [9] investigated the case of mismatched input and established a relationship between the excess MSE and relative entropy. This result was generalized to vector Gaussian channels by Chen and Lafferty [10]. A similar relationship has been proposed by Guo for non-Gaussian additive noise [11]. However, all these works focus on the cases where the channel noise is correctly specified. Thus, the relations they proposed do not hold for mismatched channel noise. The purpose of this paper is to derive a new relationship between the relative entropy and excess MSE in the case of mismatched Gaussian channel noise and mismatched inputs. A mismatched channel noise occurs when the true pdf of the observation noise differs from the pdf assumed to model data. This phenomenon is generally due to channel calibration default or to a simplification of observed data model. A recalibration procedure could help to correct mismatches but it is often computationally expensive and could complexify the data model. This is the reason why it is sometimes helpful to study estimation and information measures under mismatched contexts rather than trying to correct mismatches.

The mismatched MSE was also analysed from a statistical physics perspective by Merhav and Huleihel for mismatched channel noise [12] and mismatched channel matrix [13]. The authors specifically consider the problem of estimating a codeword transmitted over a white Gaussian channel. In this paper we consider a quite different scenario: the problem of estimating a Gaussian signal using two correlated complementary modalities. In contrast to [12] and [13], we explore a double mismatch, *i.e.* input prior mismatch and channel noise mismatch, which could be used to decrease the mismatched MSE.

This paper is organized as follows. We review in section II some of the existing connections between estimation theory and information theory under exact model and mismatched input priors. Our original contributions are presented in Section III and IV. In Section III we state the new main theorems that relate relative entropy and excess MSE, and mutual information and MSE under mismatched Gaussian channels. In Section IV we study the impact of wrong links between two modalities on excess MSE on a simple example. Section V concludes the paper.

II. Background

For notational convenience, scalar random variables are denoted by lowercase letters, *e.g.,* z. Vector random variables are denoted by bold lowercase letters, *e.g.,* $\mathbf{z}$. Matrices are denoted by bold uppercase letters, *e.g.,* $\mathbf{A}$. The exact pdf of a random variable $\mathbf{z}$ is denoted by $p(\mathbf{z})$ and the assumed one by $q(\mathbf{z})$. Even though these notations are not rigorous, they

This work has been partly supported by the CHESS project, ERC AdG-2012-320684, and partly by the DECODA project, ERC AdG-2013-320594.

S. Chlaily, P-O Amblard, O. Michel, P. Comon, and C. Jutten are with GIPSA-Lab, UMR CNRS 5216, Grenoble Campus, BP46, F-38402 Saint-Martin-d'Héres, France (E-mail: first-name.last-name@gipsa-lab.grenoble-inp.fr).

C. Ren is with SONDRA - CentraleSupélec, 91192 Gif-sur-Yvette cedex, France (E-mail: chengfang.ren@centralesupelec.fr).

are largely utilized in information theory literature in order to simplify mathematical expressions (see [1]).

Let us consider the general linear Gaussian channel defined in [2]:

$$\mathbf{x} \quad = \quad \mathbf{Hs} + \mathbf{n} \tag{1}$$

where $\mathbf{s} \in \mathbb{R}^m$ is the input random vector, $\mathbf{x} \in \mathbb{R}^l$ is the observed random vector, $\mathbf{n} \in \mathbb{R}^l$ is an additive noise vector and $\mathbf{H} \in \mathbb{R}^{l\times m}$ is the channel matrix. $\mathbf{s}$ and $\mathbf{n}$ are assumed to be independent. If the noise vector is correctly specified to be normally distributed, namely $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, and the assumed prior pdf of $\mathbf{s}$ corresponds to the true prior pdf, $p(\mathbf{s})$, then Palomar and Verdú [2] provide the following relationship between the mutual information and the MMSE

$$\nabla_{\mathbf{H}} I(\mathbf{s};\mathbf{x}) = \mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{M}_{p,p} \tag{2}$$

where $\nabla_{\mathbf{H}}$ denotes the gradient[1] operator with respect to (w.r.t.) $\mathbf{H}$ and $I(\mathbf{s};\mathbf{x})$ denotes the mutual information between $\mathbf{s}$ and $\mathbf{x}$ defined by

$$\begin{aligned} I(\mathbf{s};\mathbf{x}) \quad &= \quad \int \log\left(\frac{p(\mathbf{x},\mathbf{s})}{p(\mathbf{x})p(\mathbf{s})}\right) p(\mathbf{x},\mathbf{s}) d\mathbf{x} d\mathbf{s} \\ &\triangleq D_{KL}(p(\mathbf{x},\mathbf{s})||p(\mathbf{x})p(\mathbf{s})). \end{aligned} \tag{3}$$

The latter is a measure of mutual dependence between $\mathbf{s}$ and $\mathbf{x}$, and can be viewed as the Kullback Leibler divergence between the joint pdf $p(\mathbf{x},\mathbf{s})$ and the product $p(\mathbf{x})p(\mathbf{s})$ of the marginal pdf's. The MMSE is defined by the $m \times m$ matrix:

$$\mathbf{M}_{p,p} = \mathbb{E}_p\left[(\mathbf{s} - \mathbb{E}_p[\mathbf{s}|\mathbf{x}])(\mathbf{s} - \mathbb{E}_p[\mathbf{s}|\mathbf{x}])^T\right] \tag{4}$$

which is the exact covariance of the exact MMSE estimator error, namely of $\mathbf{s} - \widehat{\mathbf{s}}_p(\mathbf{x})$ where $\widehat{\mathbf{s}}_p(\mathbf{x}) = \mathbb{E}_p[\mathbf{s}|\mathbf{x}]$. The index p in $\mathbb{E}_p[\mathbf{s}]$ (resp. $\mathbb{E}_p[\mathbf{s}|\mathbf{x}]$) means the expectation is taken w.r.t. the true pdf $p(\mathbf{s})$ (resp. the true conditional pdf $p(\mathbf{s}|\mathbf{x})$).

Observation model (1) is investigated further by Verdù, Guo, Chen and Lafferty [9], [10], [11] in mismatched contexts. They consider an input prior mismatch, *i.e.* the exact input pdf $p(\mathbf{s})$ and assumed input pdf $q(\mathbf{s})$ are different. In this case, they establish the following relationship

$$\nabla_{\mathbf{H}} D_{KL}(p(\mathbf{x})||q(\mathbf{x})) = \mathbf{\Sigma}^{-1}\mathbf{H}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p}) \tag{5}$$

where

$$\mathbf{M}_{p,q} = \mathbb{E}_p\left[(\mathbf{s} - \mathbb{E}_q[\mathbf{s}|\mathbf{x}])(\mathbf{s} - \mathbb{E}_q[\mathbf{s}|\mathbf{x}])^T\right] \tag{6}$$

is the exact covariance of the mismatched MMSE estimator error. The estimator $\widehat{\mathbf{s}}_q(\mathbf{x}) = \mathbb{E}_q[\mathbf{s}|\mathbf{x}]$ is the so-called mismatched (or the quasi-) MMSE estimator. Relation (5) characterizes the excess estimation error due to mismatched model assumption. This excess quantity is directly related to the gradient of the Kullback-Leibler divergence between the exact and the assumed pdf's of observed data.

Our purpose is to extend relation (5) in the case where, in addition to the input prior mismatch, the channel noise is also mismatched, *i.e.*, the exact noise pdf $p(\mathbf{n})$ and the assumed noise pdf $q(\mathbf{n})$ are different. This scenario seems more realistic since both mismatches, input prior and channel noise, could occur in practice for instance in multimodal estimation. Unfortunately, relation (5) does not hold in this case. Therefore, we propose in this paper a general expression which depicts and quantifies the excess estimation error for mismatched Gaussian models.

[1]The gradient of a function $a(.)$ w.r.t. $\mathbf{H}$ is defined entry-wise by $\{\nabla_{\mathbf{H}} a(.)\}_{i,j} = \frac{\partial a(.)}{\partial h_{i,j}}$ where $h_{i,j}$ denotes element (i,j) of $\mathbf{H}$.

III. Mismatched Gaussian channels

A. Statement and main result

Let us consider the linear vector channel defined in (1). We suppose now this model is misspecified in the sense that the prior pdf of signal $\mathbf{s}$ and the pdf of noise $\mathbf{n}$ are both mismatched. Let us denote respectively the exact pdf's of $\mathbf{s}$ and $\mathbf{n}$ by $p(\mathbf{s})$ and $p(\mathbf{n})$ which are different of their assumed pdf's $q(\mathbf{s})$ and $q(\mathbf{n})$.

The exact and assumed signals are both zero mean Gaussian distributed but with different covariance matrices, *i.e.* $p(\mathbf{s}) = \mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$ and $q(\mathbf{s}) = \mathcal{N}(\mathbf{0}, \widehat{\mathbf{\Gamma}})$ with $\mathbf{\Gamma} \neq \widehat{\mathbf{\Gamma}}$. Similar mismatch is considered for the channel noise, *i.e.* $p(\mathbf{n}) = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and $q(\mathbf{n}) = \mathcal{N}(\mathbf{0}, \widehat{\mathbf{\Sigma}})$ with $\mathbf{\Sigma} \neq \widehat{\mathbf{\Sigma}}$. The last hypothesis implies that the conditional pdf's of observations are given by $p(\mathbf{x}|\mathbf{s}) = \mathcal{N}(\mathbf{Hs}, \mathbf{\Sigma})$ and $q(\mathbf{x}|\mathbf{s}) = \mathcal{N}(\mathbf{Hs}, \widehat{\mathbf{\Sigma}})$.

The main difference between this scenario and the one proposed in [10], which was recalled in Section II, is the mismatch on the channel noise. In our case, we obtain the following theorem:

Theorem III.1. *Consider the aforementioned mismatched communication model. Then*

$$\nabla_{\mathbf{H}} D_{KL}(p(\mathbf{x})||q(\mathbf{x})) = \widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p}) + \mathbf{R}, \tag{7}$$

where $\mathbf{R}$ *is a residual term given by*

$$\mathbf{R} = (\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma} - \mathbf{I})\left(\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{M}_{p,p} - \widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}\mathbf{M}_{q,q}\right) \tag{8}$$

and $\mathbf{M}_{q,q} = \mathbb{E}_q[(\mathbf{s} - \mathbb{E}_q[\mathbf{s}|\mathbf{x}])(\mathbf{s} - \mathbb{E}_q[\mathbf{s}|\mathbf{x}])^T]$ *is the covariance of the mismatched MMSE estimator error, computed with the mismatched pdf's* $q(\mathbf{s})$ *and* $q(\mathbf{s}|\mathbf{x})$.

Proof. see Appendix A □

The residual term is a weighted difference between the MMSE under distribution p and the MMSE under distribution q. This term vanishes if the channel noise is correctly specified, *i.e.* $\widehat{\mathbf{\Sigma}} = \mathbf{\Sigma}$, and (7) reduces to (5). The above theorem can be interpreted as an extension of Chen and Lafferty result [10] for mismatched Gaussian channel noise. Even though $\mathbf{\Gamma}$ and $\widehat{\mathbf{\Gamma}}$ do not appear explicitly in (7) they occur in the relative entropy and the MMSE's (cf Appendix A).

Example: we consider a scalar channel with $\mathbf{H} = \sqrt{\lambda}$, $p(s) = \mathcal{N}(0, \gamma^2)$, $q(s) = \mathcal{N}(0, \hat{\gamma}^2)$, $p(n) = \mathcal{N}(0, \sigma^2)$ and $q(n) = \mathcal{N}(0, \hat{\sigma}^2)$.

By applying theorem III.1 and using $2\sqrt{\lambda}\,\frac{dD_{KL}}{d\lambda} = \frac{dD_{KL}}{d\sqrt{\lambda}}$, we have

$$2\frac{d}{d\lambda} D_{KL}(p(x)||q(x)) = \frac{1}{\hat{\sigma}^2}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p}) + \frac{1}{\sqrt{\lambda}}\mathbf{R}, \tag{9}$$

Then, by integrating both sides of (9) w.r.t λ, it follows

$$2\left(D_{KL}(p(s)||q(s)) - D_{KL}(p(n)||q(n))\right) = \int_0^{+\infty} \left(\frac{1}{\hat{\sigma}^2}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p}) + \frac{1}{\sqrt{\lambda}}\mathbf{R} \right) d\lambda. \quad (10)$$

Equation (10) states that the area defined by $(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})/\hat{\sigma}^2 + \mathbf{R}/\sqrt{\lambda}$ is equal to the difference between input prior mismatch and noise pdf mismatch. The noise mismatch introduces an opposite effect to the signal mismatch. This suggests that both mismatches, input prior and channel noise, may compensate each other. Hence, this compensation may enhance the channel quality and the estimation accuracy. If $p(n) = q(n)$ equation (10) reduces to the relation proposed by Verdú in [9].

B. *Mismatched entropy and mutual information*

As we have seen beforehand, when considering a mismatched model, it is no longer appropriate to use standard measures borrowed from estimation theory. For instance, $\mathbf{M}_{p,p}$ must be replaced by $\mathbf{M}_{p,q}$ to quantify the MSE for mismatched models. The same applies for information theory measures, mainly entropy and mutual information. Indeed, the classical entropy $h(\mathbf{x}) \triangleq -\mathbb{E}_p[\log p(\mathbf{x})]$, defined by the average[2] of Shannon information, namely $-\log p(\mathbf{x})$, quantifies the uncertainty of $\mathbf{x}$ when data pdf $p(\mathbf{x})$ is correctly specified. However, if the observation model is mismatched, *i.e.* the assumed pdf of $\mathbf{x}$ is $q(\mathbf{x})$ instead of $p(\mathbf{x})$, then the Shannon information should be modified accordingly. In case of mismatched model given by section III-A, the Shannon information is now equal to $-\log q(\mathbf{x})$ and the mismatched entropy is defined as follows:

$$h_{p,q}(\mathbf{x}) = -\mathbb{E}_p[\log q(\mathbf{x})] \quad (11)$$

In coding theory, $h_{p,q}$ is usually used to quantify the additional required bits to code an event considering a wrong pdf [1, p.115].

Similarly, the expression of the mutual information (3) does not hold under mismatched models. A natural extension could be derived by using the notion of mismatched entropy, since mutual information is also defined as the difference between two entropies $I(\mathbf{s};\mathbf{x}) \triangleq h(\mathbf{x}) - h(\mathbf{x}|\mathbf{s})$. Thus, the mismatched mutual information is given by

$$I_{p,q}(\mathbf{s};\mathbf{x}) \triangleq h_{p,q}(\mathbf{x}) - h_{p,q}(\mathbf{x}|\mathbf{s}) = \mathbb{E}_p\left[\log \frac{q(\mathbf{x}|\mathbf{s})}{q(\mathbf{x})}\right]$$
$$= I(\mathbf{s};\mathbf{x}) - D_{KL}(p(\mathbf{x}|\mathbf{s})||q(\mathbf{x}|\mathbf{s})) + D_{KL}(p(\mathbf{x})||q(\mathbf{x})), \quad (12)$$

where $D_{KL}(p(\mathbf{x}|\mathbf{s})||q(\mathbf{x}|\mathbf{s})) \triangleq \int\int \log\left(\frac{p(\mathbf{x}|\mathbf{s})}{q(\mathbf{x}|\mathbf{s})}\right) p(\mathbf{x},\mathbf{s}) d\mathbf{x} d\mathbf{s}$. The nonnegativity of $I_{p,q}(\mathbf{s};\mathbf{x})$ is not guaranteed in general, contrary to the classical mutual information. Using Theorem III.1, a relation between mutual information and MSE is provided by:

Theorem III.2. *Consider the model presented in (1), following the Gaussian assumptions, $I_{p,q}(\mathbf{s};\mathbf{x})$ is related to $\mathbf{M}_{p,q}$, by the following formula*

$$\nabla_{\mathbf{H}} I_{p,q}(\mathbf{s};\mathbf{x}) = \widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}\,\mathbf{M}_{p,q} + \left(\mathbf{I} - \widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}\right)\widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}\,\mathbf{M}_{q,q} \quad (13)$$

[2]The average is always taken under the true data distribution

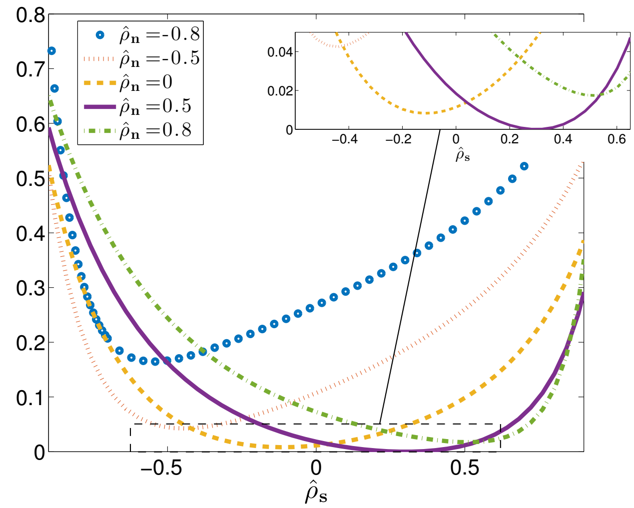

Fig. 1. Variation of Tr$(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})$ as a function of $\hat{\rho}_\mathbf{s}$, for different values of $\hat{\rho}_\mathbf{n}$. Plotted for $\rho_\mathbf{s} = 0.3$, $\rho_\mathbf{n} = 0.5$, $\alpha_1 = 1$ and $\alpha_2 = 1.5$. Tr$(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})$ vanishes when no mismatch occurs *i.e.* $\hat{\rho}_\mathbf{n} = \rho_\mathbf{n}$ (solid line) and $\hat{\rho}_\mathbf{s} = \rho_\mathbf{s}$. However, when the noise correlation is mismatched $\hat{\rho}_\mathbf{n} \neq \rho_\mathbf{n}$, Tr$(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})$ is cancelled by a signal correlation mismatch rather than by the exact signal correlation $\rho_\mathbf{s}$. For instance, the dashed line ($\hat{\rho}_\mathbf{n} = 0$) is almost zero at $\hat{\rho}_\mathbf{s} = -0.1$ while $\rho_\mathbf{s} = 0.3$.

Note that $\mathbf{\Gamma}$ and $\widehat{\mathbf{\Gamma}}$ take action in $I_{p,q}$, $\mathbf{M}_{p,q}$ and $\mathbf{M}_{q,q}$. The second term on the right-hand side of (13) vanishes if only the input prior is mismatched, *i.e* $\mathbf{\Gamma} \neq \widehat{\mathbf{\Gamma}}$ and $\mathbf{\Sigma} = \widehat{\mathbf{\Sigma}}$. Straightforwardly, (13) reduces to (2) if $\mathbf{\Sigma} = \widehat{\mathbf{\Sigma}}$ and $\mathbf{\Gamma} = \widehat{\mathbf{\Gamma}}$. The above theorem is an extension to mismatched models of the relation (2) proposed by Palomar and Verdú [2].

Proof. see Appendix B □

IV. Example: Model mismatch and Excess MSE

As a simple example, easy to interpret, let $m = 2$, $l = 2$ and define $\mathbf{H} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}$, where $\alpha_2 \leq \alpha_1$. Model (1) becomes:

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix}, \quad (14)$$

where s_1, s_2, n_1 and n_2 are all zero mean, unit-variance Gaussian random variables. We suppose that signals and noises are independent but s_1 and s_2 (resp. n_1 and n_2) are correlated and denote the correlation coefficient $\rho_\mathbf{s}$ (resp. $\rho_\mathbf{n}$). It can be shown that the estimation accuracy improves if x_1 and x_2 are used jointly, *i.e.* the correlations between signals and between noises are exploited, since x_j also bears information about s_i $(i, j \in \{1, 2\}, i \neq j)$. But, what happens if we mismatch $\rho_\mathbf{s}$ and $\rho_\mathbf{n}$? Let $\hat{\rho}_\mathbf{s}$ and $\hat{\rho}_\mathbf{n}$ be the assumed signal and noise correlations, different from the exact signal and noise correlations $\rho_\mathbf{s}$ and $\rho_\mathbf{n}$. Namely, $\mathbf{\Gamma} = \begin{bmatrix} 1 & \rho_\mathbf{s} \\ \rho_\mathbf{s} & 1 \end{bmatrix}$, $\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho_\mathbf{n} \\ \rho_\mathbf{n} & 1 \end{bmatrix}$, $\widehat{\mathbf{\Gamma}} = \begin{bmatrix} 1 & \hat{\rho}_\mathbf{s} \\ \hat{\rho}_\mathbf{s} & 1 \end{bmatrix}$ and $\widehat{\mathbf{\Sigma}} = \begin{bmatrix} 1 & \hat{\rho}_\mathbf{n} \\ \hat{\rho}_\mathbf{n} & 1 \end{bmatrix}$. Accordingly, the mismatched estimator of s_i using both modalities may be written as:

$$\widehat{\mathbf{s}}_q(i) = \frac{\alpha_i(1+\alpha_j) - \alpha_j(\hat{\rho}_{\mathbf{n}} + \alpha_1\alpha_2\hat{\rho}_{\mathbf{s}})}{(1+\alpha_1^2)(1+\alpha_2^2) - (\hat{\rho}_{\mathbf{n}} + \alpha_1\alpha_2\hat{\rho}_{\mathbf{s}})^2}\, x_i + \frac{(\alpha_i\hat{\rho}_{\mathbf{n}} - \alpha_j\hat{\rho}_{\mathbf{s}})}{((1+\alpha_1^2)(1+\alpha_2^2) - (\hat{\rho}_{\mathbf{n}} + \alpha_1\alpha_2\hat{\rho}_{\mathbf{s}})^2)^2}\, x_j \quad (15)$$

The coefficient of x_j vanishes for $\hat{\rho}_{\mathbf{n}} = \frac{\alpha_j}{\alpha_i}\hat{\rho}_{\mathbf{s}}$, so that the estimator of s_i is only based on x_i. This means that if $\hat{\rho}_{\mathbf{n}} = \frac{\alpha_j}{\alpha_i}\hat{\rho}_{\mathbf{s}}$, x_j is redundant to x_i for estimating s_i and any additional information held by x_j about s_i is ignored. If in addition $\alpha_1 = \alpha_2$, $\hat{\rho}_{\mathbf{n}} = \frac{\alpha_j}{\alpha_i}\hat{\rho}_{\mathbf{s}}$ becomes $\hat{\rho}_{\mathbf{n}} = \hat{\rho}_{\mathbf{s}}$, then x_i is also presumed to be redundant w.r.t x_j about s_j and the observations are used separately *i.e.* without any interaction. The diagonal elements of matrices $\mathbf{M}_{p,p}$ and $\mathbf{M}_{p,q}$ are given by:

$$\mathbf{M}_{p,p}(i,i) = \frac{(1-\rho_{\mathbf{n}}^2) + (1-\rho_{\mathbf{s}}^2)\alpha_j^2}{(1+\alpha_1^2)(1+\alpha_2^2) - (\rho_{\mathbf{n}} + \alpha_1\alpha_2\rho_{\mathbf{s}})^2} \quad (16)$$

$$\begin{aligned}\mathbf{M}_{p,q}(i,i) &= \frac{(1-\hat{\rho}_{\mathbf{n}}^2) + (1-\hat{\rho}_{\mathbf{s}}^2)\alpha_j^2}{(1+\alpha_1^2)(1+\alpha_2^2) - (\hat{\rho}_{\mathbf{n}} + \alpha_1\alpha_2\hat{\rho}_{\mathbf{s}})^2} \quad (17)\\ &+(\alpha_i\hat{\rho}_{\mathbf{n}} - \alpha_j\hat{\rho}_{\mathbf{s}})\frac{(1+\alpha_j^2)\left(\alpha_j(\rho_{\mathbf{s}} - \hat{\rho}_{\mathbf{s}}) - \alpha_i(\rho_{\mathbf{n}} - \hat{\rho}_{\mathbf{n}})\right)}{((1+\alpha_1^2)(1+\alpha_2^2) - (\hat{\rho}_{\mathbf{n}} + \alpha_1\alpha_2\hat{\rho}_{\mathbf{s}})^2)^2}\\ &+(\alpha_i\hat{\rho}_{\mathbf{n}} - \alpha_j\hat{\rho}_{\mathbf{s}})\frac{\alpha_j(\hat{\rho}_{\mathbf{s}}\rho_{\mathbf{n}} - \rho_{\mathbf{s}}\hat{\rho}_{\mathbf{n}})(\hat{\rho}_{\mathbf{n}} + \alpha_1\alpha_2\hat{\rho}_{\mathbf{s}})}{((1+\alpha_1^2)(1+\alpha_2^2) - (\hat{\rho}_{\mathbf{n}} + \alpha_1\alpha_2\hat{\rho}_{\mathbf{s}})^2)^2}\end{aligned}$$

We show in Fig. 1 the variation of $\mathrm{Tr}\left(\mathbf{M}_{p,q} - \mathbf{M}_{p,p}\right)$ as a function of $\hat{\rho}_{\mathbf{s}}$ for different values of $\hat{\rho}_{\mathbf{n}}$. The absence of noise mismatch *i.e.* $\hat{\rho}_{\mathbf{n}} = \rho_{\mathbf{n}}$ (solid line) does not guarantee the minimal excess MSE for all $\hat{\rho}_{\mathbf{s}}$. On the contrary, a noise mismatch may partially or almost completely compensate for the signal mismatch. Indeed, we remark that for a given signal mismatch, there exists a relevant noise mismatch that induces the minimum excess MSE. These results are consistent with equation (10). Fig. 1 shows that excess MSE increases for strong $\hat{\rho}_{\mathbf{s}}$, as we rely wrongly on the information given by the other modality.

V. Conclusion

We extended the relationship proposed by Chen and Lafferty between relative entropy and excess MSE under mismatched input and mismatched Gaussian channels. The generalization of mutual information and MMSE relationship follows easily. Although our results were restricted to Gaussian inputs, they are still prominent. In fact, our results pointed out the possible benefit of a double mismatch, *i.e* input prior mismatch and channel noise mismatch, in enhancing the estimation accuracy. Both mismatches may partially or completely cancel each other leading to good results as does the true model. Future works are focused on the extension of these results to arbitrary distributed inputs and noises.

Appendix A
Proof of main theorem

The exact and mismatched pdf's of observed data are Gaussian and are given respectively by
$p(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{H}\mathbf{\Gamma}\mathbf{H}^T + \mathbf{\Sigma})$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{0}, \mathbf{H}\widehat{\mathbf{\Gamma}}\mathbf{H}^T + \widehat{\mathbf{\Sigma}})$. For notational convenience, let us denote $\mathbf{\Omega} = \mathbf{H}\mathbf{\Gamma}\mathbf{H}^T + \mathbf{\Sigma}$ and $\widehat{\mathbf{\Omega}} = \mathbf{H}\widehat{\mathbf{\Gamma}}\mathbf{H}^T + \widehat{\mathbf{\Sigma}}$. Straightforwardly, the Kullback-Leibler divergence between two centered Gaussian distributions is given by

$$D_{KL}\left(p(\mathbf{x})||q(\mathbf{x})\right) = \frac{1}{2}\Big(\mathrm{Tr}\big(\widehat{\mathbf{\Omega}}^{-1}\mathbf{\Omega}\big) - \log\big|\widehat{\mathbf{\Omega}}^{-1}\mathbf{\Omega}\big| - l\Big) \quad (19)$$

where $\mathrm{Tr}(\mathbf{A})$ and $|\mathbf{A}|$ denote respectively the trace and the determinant of matrix $\mathbf{A}$ and l is the size of vector $\mathbf{x}$. Then, we need to calculate gradient of the aforementioned expression. Using results from [14], it can be shown that

$$\frac{1}{2}\Big(\nabla_{\mathbf{H}}\log\big|\widehat{\mathbf{\Omega}}\big| - \nabla_{\mathbf{H}}\log\big|\mathbf{\Omega}\big|\Big) = \widehat{\mathbf{\Omega}}^{-1}\mathbf{H}\widehat{\mathbf{\Gamma}} - \mathbf{\Omega}^{-1}\mathbf{H}\mathbf{\Gamma},$$

$$\text{and} \qquad \frac{1}{2}\nabla_{\mathbf{H}}\mathrm{Tr}\big(\widehat{\mathbf{\Omega}}^{-1}\mathbf{\Omega}\big) = -\widehat{\mathbf{\Omega}}^{-1}\mathbf{\Omega}\widehat{\mathbf{\Omega}}^{-1}\mathbf{H}\widehat{\mathbf{\Gamma}} + \widehat{\mathbf{\Omega}}^{-1}\mathbf{H}\mathbf{\Gamma}.$$

On the other hand, a closed-form expression can be derived for both estimators, the MMSE estimator and the mismatched MMSE estimator, in linear Gaussian channel with Gaussian prior [15, theorem 10.3]:

$$\widehat{\mathbf{s}}_p(\mathbf{x}) \triangleq \mathbb{E}_p\left[\mathbf{s}|\mathbf{x}\right] = \mathbf{\Gamma}\mathbf{H}^T\mathbf{\Omega}^{-1}\mathbf{x} \quad (20)$$

$$\widehat{\mathbf{s}}_q(\mathbf{x}) \triangleq \mathbb{E}_q\left[\mathbf{s}|\mathbf{x}\right] = \widehat{\mathbf{\Gamma}}\mathbf{H}^T\widehat{\mathbf{\Omega}}^{-1}\mathbf{x} \quad (21)$$

Therefore, the MSE of those estimators and the mismatched MSE are respectively :

$$\begin{aligned}\mathbf{M}_{p,p} &= \mathbf{\Gamma} - \mathbf{\Gamma}\mathbf{H}^T\mathbf{\Omega}^{-1}\mathbf{H}\mathbf{\Gamma}, \quad (22)\\ \mathbf{M}_{q,q} &= \widehat{\mathbf{\Gamma}} - \widehat{\mathbf{\Gamma}}\mathbf{H}^T\widehat{\mathbf{\Omega}}^{-1}\mathbf{H}\widehat{\mathbf{\Gamma}}, \quad (23)\\ \mathbf{M}_{p,q} &= \mathbf{M}_{q,q} + \mathbf{M}_{q,q}\widehat{\mathbf{\Gamma}}^{-1}(\mathbf{\Gamma} - \widehat{\mathbf{\Gamma}})\widehat{\mathbf{\Gamma}}^{-1}\mathbf{M}_{q,q}\\ &\quad + \mathbf{M}_{q,q}\mathbf{H}^T\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{\Sigma} - \widehat{\mathbf{\Sigma}})\widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}\mathbf{M}_{q,q} \quad (24)\end{aligned}$$

By applying consecutively Woodbury matrix identity and positive definite identity [16, eq. 185] on $\mathbf{M}_{p,p}$, $\mathbf{M}_{q,q}$ and $\mathbf{M}_{p,q}$ we obtain

$$\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{M}_{p,p} = \mathbf{\Omega}^{-1}\mathbf{H}\mathbf{\Gamma} = \frac{1}{2}\nabla_{\mathbf{H}}\log\big|\mathbf{\Omega}\big|, \quad (25)$$

$$\widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}\mathbf{M}_{q,q} = \widehat{\mathbf{\Omega}}^{-1}\mathbf{H}\widehat{\mathbf{\Gamma}} = \frac{1}{2}\nabla_{\mathbf{H}}\log\big|\widehat{\mathbf{\Omega}}\big|, \quad (26)$$

$$\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{H}\mathbf{M}_{p,q} - \mathbf{\Sigma}\widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}\mathbf{M}_{q,q}) = \frac{1}{2}\nabla_{\mathbf{H}}\mathrm{Tr}(\widehat{\mathbf{\Omega}}^{-1}\mathbf{\Omega}). \quad (27)$$

By combining previous results, $\nabla_{\mathbf{H}}D_{KL}\left(p(\mathbf{x})||q(\mathbf{x})\right)$ is well formulated by equation (7).

Appendix B
Proof of theorem 3.2.

As we have mentioned in subsection III-B, the mismatched mutual information is related to the classical mutual information by equation (12). By applying the gradient operator w.r.t. $\mathbf{H}$, we find that:

$$\nabla_{\mathbf{H}}I_{p,q}(\mathbf{s};\mathbf{x}) = \nabla_{\mathbf{H}}I(\mathbf{s};\mathbf{x}) + \nabla_{\mathbf{H}}D_{KL}(p(\mathbf{x})||q(\mathbf{x})) - \nabla_{\mathbf{H}}D_{KL}(p(\mathbf{x}|\mathbf{s})||q(\mathbf{x}|\mathbf{s}))$$

Note that $p(\mathbf{x}|\mathbf{s}) = \mathcal{N}(\mathbf{H}\mathbf{s}, \mathbf{\Sigma})$ and $q(\mathbf{x}|\mathbf{s}) = \mathcal{N}(\mathbf{H}\mathbf{s}, \widehat{\mathbf{\Sigma}})$, thus $\nabla_{\mathbf{H}}D_{KL}(p(\mathbf{x}|\mathbf{s})||q(\mathbf{x}|\mathbf{s})) = \mathbf{0}$. Since the gradient of the classical mutual information satisfies relation (2) and $\nabla_{\mathbf{H}}D_{KL}(p(\mathbf{x})||q(\mathbf{x}))$ is given by theorem (III.1). Consequently,

$$\nabla_{\mathbf{H}}I_{p,q}(\mathbf{s};\mathbf{x}) = \widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}\mathbf{M}_{p,q} + \Big(\mathbf{I} - \widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}\Big)\widehat{\mathbf{\Sigma}}^{-1}\mathbf{H}\,\mathbf{M}_{q,q}.$$

REFERENCES

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, NY, USA, 1991.

[2] D.P. Palomar and S. Verdú, "Gradient of mutual information in linear vector gaussian channels," *Information Theory, IEEE Transactions on*, vol. 52, no. 1, pp. 141–154, Jan 2006.

[3] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in gaussian channels," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.

[4] M. Akcakaya and A. Nehorai, "MIMO radar detection and adaptive design under a phase synchronization mismatch," *IEEE Transactions on Signal Processing*, vol. 58, no. 10, pp. 4994–5005, Oct 2010.

[5] M. Greco, S. Fortunati, and F. Gini, "Maximum likelihood covariance matrix estimation for complex elliptically symmetric distributions under mismatched conditions," *Signal Processing*, vol. 104, pp. 381 – 386, 2014.

[6] C. Ren, M. N. E. Korso, J. Galy, E. Chaumette, P. Larzabal, and A. Renaux, "Performance bounds under misspecification model for mimo radar application," in *Signal Processing Conference (EUSIPCO), 2015 23rd European*, Aug 2015, pp. 514–518.

[7] D. Lahat, T. Adalı, and C. Jutten, "Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, Aug. 2015.

[8] S. Chlaily, P-O. Amblard, O. Michel, and C. Jutten, "Impact of Noise Correlation on Multimodality," in *24th European Signal Processing Conference (EUSIPCO 2016)*, Budapest, Hungary, Aug. 2016.

[9] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3712–3720, Aug 2010.

[10] M. Chen and J. Lafferty, "Mismatched estimation and relative entropy in vector gaussian channels," in *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, July 2013, pp. 2845–2849.

[11] D. Guo, "Relative entropy and score function: New information-estimation relationships through arbitrary additive perturbation," in *2009 IEEE International Symposium on Information Theory*, June 2009, pp. 814–818.

[12] N. Merhav, "Optimum estimation via gradients of partition functions and information measures: A statistical-mechanical perspective," *IEEE Transactions on Information Theory*, vol. 57, no. 6, pp. 3887–3898, June 2011.

[13] W. Huleihel and N. Merhav, "Analysis of mismatched estimation errors using gradients of partition functions," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2190–2216, April 2014.

[14] T. P. Minka, "Old and new matrix algebra useful for statistics," Tech. Rep., 2001.

[15] S.M. Kay, *Fundamentals of Statistical Signal Processing: Detection theory*, Prentice-Hall PTR, 1998.

[16] K. B. Petersen and M. S. Pedersen, "The matrix cookbook," Nov 2012.