Mismatched Estimation and Relative Entropy in Vector Gaussian Channels

Minhua Chen and John Lafferty
Department of Statistics &
Department of Computer Science
University of Chicago
Chicago, IL 60637 USA
Email: {minhua,lafferty}@galton.uchicago.edu

***Abstract*— We derive a novel relation between mismatched estimation and relative entropy (KL divergence) in vector Gaussian channels under the mean squared estimation criterion. This relation includes as special cases several previous results connecting estimation theory and information theory. A direct proof is provided, together with a verification using Gaussian inputs. An interesting relationship between the KL divergence and Fisher divergence is derived as a direct consequence of our work. The relations established here are potentially useful for inference in graphical models and the design of information systems.**

I. Introduction

The relationship between estimation theory and information theory is a classical topic, which has been recently revisited in a new light [1], [2], [3]. This recent work shows how the gradient of information-theoretic quantities with respect to a Gaussian channel parameter can be expressed in terms of optimal estimation-theoretic quantities in closed-form.

Given a scalar Gaussian channel

$$y = \sqrt{\gamma}x + n \tag{1}$$

where x is the input signal with source distribution $p(x)$, γ is the signal-to-noise ratio of the channel, $n \sim \mathcal{N}(n; 0, 1)$ is additive Gaussian noise, and y is the channel output, the input-output conditional distribution is $p(y|x) = \mathcal{N}(y; \sqrt{\gamma}x, 1)$. It was proved in [1] that under a finite variance constraint on $p(x)$,

$$\frac{d}{d\gamma} I(X;Y) = \frac{1}{2}\text{mmse}(\gamma) \tag{2}$$

where $I(X;Y)$ is the mutual information between the input and output, and mmse(γ) is the Minimum Mean Squared Error (MMSE) of estimating x given y, expressed as

$$\text{mmse}(\gamma) = \int\!\!\int p(x)p(y|x)\|x - x_p(y)\|^2 dxdy$$

where $x_p(y) = \int x \cdot p(x|y)dx$ is the MMSE estimator. Equation (2) reveals an intimate relation between MMSE and mutual information, two core quantities in estimation theory and information theory. The fact that equation (2) is true for any input distribution $p(x)$ makes the result very general in nature. A direct implication is a new definition of mutual information via MMSE as [4]

$$I(X;Y) = \frac{1}{2}\int_0^\gamma \text{mmse}(\gamma)d\gamma.$$

Another important result in this direction was developed in [2], where the scalar Gaussian channel in (1) is generalized as a vector channel:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}. \tag{3}$$

Here $\mathbf{x} \in \mathbf{R}^s$ is a vector input signal with source distribution $p(\mathbf{x})$, $\mathbf{H} \in \mathbf{R}^{m\times s}$ is the channel matrix, $\mathbf{n} \sim \mathcal{N}(\mathbf{n}; \mathbf{0}, \mathbf{\Sigma}_n)$ is additive Gaussian noise with $m \times m$ covariance matrix $\mathbf{\Sigma}_n$, and $\mathbf{y}$ is the channel output. Now the channel input-output conditional distribution becomes $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mathbf{H}\mathbf{x}, \mathbf{\Sigma}_n)$. It was shown in [2] that the relation in equation (2) can be generalized to this vector Gaussian channel as

$$\nabla_{\mathbf{H}} I(\mathbf{X};\mathbf{Y}) = \mathbf{\Sigma}_n^{-1}\mathbf{H}\mathbf{M}_{p,p} \tag{4}$$

where

$$\mathbf{M}_{p,p} = \int\!\!\int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})(\mathbf{x} - \mathbf{x}_p(\mathbf{y}))(\mathbf{x} - \mathbf{x}_p(\mathbf{y}))^\top d\mathbf{x}d\mathbf{y} \tag{5}$$

is the MMSE matrix with $\mathbf{x}_p(\mathbf{y}) = \int \mathbf{x} \cdot p(\mathbf{x}|\mathbf{y})d\mathbf{x}$. Since many information processing systems can be modeled as a vector Gaussian channel, equation (4) can be applied to many situations where the goal is to design a channel matrix $\mathbf{H}$ such that the mutual information $I(\mathbf{X};\mathbf{Y})$ is optimized. Such applications include precoder design in communication systems [5], projection matrix optimization in compressive sensing [6] and linear discriminant analysis for classification problems [7].

A more recent result [3] extends [1] to a mismatched estimation scenario, although still under the scalar Gaussian channel assumption (1). Suppose another input distribution $q(x)$, which is different from the true input distribution $p(x)$, is assumed to estimate x from y. The resulting mismatched estimator $x_q(y) = \int x \cdot q(x|y)dx$ has mean squared error (MSE)

$$\text{mse}_q(\gamma) = \int\!\!\int p(x)p(y|x)\|x - x_q(y)\|^2 dxdy.$$

It was proved in [3] that the following identity holds:

$$\frac{d}{d\gamma}D(p(y)\|q(y)) = \frac{1}{2}(\mathrm{mse}_q(\gamma) - \mathrm{mmse}(\gamma)). \tag{6}$$

The KL divergence $D(p(y)\|q(y))$, or relative entropy, on the lefthand side is an information-theoretic measure of the mismatch; the excess mean squared error $(\mathrm{mse}_q(\gamma) - \mathrm{mmse}(\gamma))$ on the righthand side is an estimation-theoretic measure of the mismatch. Hence this equation reveals a fundamental connection between the two areas. It is also demonstrated in [3] that equation (2) is a special case of equation (6). Equation (6) implies a new representation of the relative entropy:

$$D(p(y)\|q(y)) = \frac{1}{2}\int_0^\gamma (\mathrm{mse}_q(\gamma) - \mathrm{mmse}(\gamma))d\gamma.$$

These results have spurred signficant further work. A simple proof of the entropy-power inequality was obtained in [8] by directly applying results in [1]. Also building on the work of [1], the functional properties of MMSE and mutual information were studied in detail in [9]. The MMSE crossing property developed in [4] was generalized to parallel vector Gaussian channels in [10]. Pointwise relations between estimation-theoretic and information-theoretic random quantities were derived in [11] to shed new light on their expectation counterparts. Based on the work of [2], the Hessian of mutual information in the vector Gaussian channel was derived in [12] to assess the concavity properties of mutual information and differential entropy. The results in [2] were further generalized in [13], where the gradient of mutual information is represented by conditional marginal input distributions given the outputs. In [14], a simple relationship between optimum estimation and certain information measures was derived via the use of partition functions. As an extension to [3], the relationship between causal and non-causal mismatched estimation was analyzed in a continuous-time Additive White Gaussian Noise (AWGN) channel in [15]. Furthermore, the relations between mutual information, relative entropy and mismatched estimation were studied under the Poisson channel assumption in [16].

In this paper, the relation between mismatched estimation and relative entropy in [3] is generalized to vector Gaussian channels, similar to the way [1] was generalized to [2]. Specifically, we show that in vector Gaussian channels, the gradient with respect to the channel matrix of the relative entropy between the channel output distributions is directly linked to the excess mean squared error due to the mismatched input distribution. Moreover, the proof provided in this paper is simpler than that in [3], which uses a local perturbation analysis of the KL divergence. We then relate our result to previous work (see Table I), and show that it reduces to previous results under special specifications. We also derive a relationship between the KL divergence and the Fisher divergence. Potential applications of our result are discussed, which include inference in graphical models and the design of information systems.

TABLE I
RESULTS IN THE RELEVANT PAPERS.

Paper	Mismatch Estimation	Vector Gaussian Channel
[1]	×	×
[2]	×	√
[3]	√	×
this paper	√	√

II. MAIN RESULT

We begin by giving a detailed description of the mismatched estimation problem. Suppose in an information processing system, an input signal $\mathbf{x} \in \mathbf{R}^s$ with source distribution $p(\mathbf{x})$ goes through the vector Gaussian channel in (3) and produces an output signal $\mathbf{y} \in \mathbf{R}^m$ with marginal distribution $p(\mathbf{y}) = \int p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}$.

In practice, the true input distribution $p(\mathbf{x})$ may not be known exactly, or may too computationally expensive to be used; instead, another input distribution $q(\mathbf{x})$ is assumed (see Figure 1 for an illustration). This results in a mismatched estimator

$$\mathbf{x}_q(\mathbf{y}) = \int \mathbf{x}\, q(\mathbf{x}|\mathbf{y})d\mathbf{x}$$

where $q(\mathbf{x}|\mathbf{y}) = q(\mathbf{x})p(\mathbf{y}|\mathbf{x})/q(\mathbf{y})$, with $q(\mathbf{y})$ given by $q(\mathbf{y}) = \int q(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}$. The corresponding mismatched MSE matrix is

$$\mathbf{M}_{p,q} = \iint p(\mathbf{x})p(\mathbf{y}|\mathbf{x})(\mathbf{x}-\mathbf{x}_q(\mathbf{y}))(\mathbf{x}-\mathbf{x}_q(\mathbf{y}))^\top d\mathbf{x}d\mathbf{y} \tag{7}$$

where the first subscript in $\mathbf{M}_{p,q}$ denotes the true input distribution and the second subscript denotes the assumed one. Notice that the true distribution $p(\mathbf{x})$ is still used to integrate out $\mathbf{x}$ in the above expression. It is straightforward to see that $\mathbf{M}_{p,p}$ expressed in (5) corresponds to the MMSE matrix in the matched case, $q(\mathbf{x}) = p(\mathbf{x})$. The excess mean squared error matrix is computed as

$$\begin{aligned}&\mathbf{M}_{p,q} - \mathbf{M}_{p,p} \qquad (8)\\ &\quad = \iint p(\mathbf{x})p(\mathbf{y}|\mathbf{x})(\mathbf{x}_p(\mathbf{y})-\mathbf{x}_q(\mathbf{y}))(\mathbf{x}_p(\mathbf{y})-\mathbf{x}_q(\mathbf{y}))^\top d\mathbf{x}d\mathbf{y}\end{aligned}$$

which is nonnegative definite. The question to be answered in this paper is how $(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})$ is related to the KL divergence $D(p(\cdot)\|q(\cdot))$.

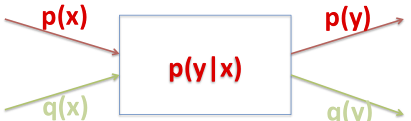

Fig. 1. Illustration of mismatched estimation in information processing systems; $p(\mathbf{x})$ denotes the true input distribution and $q(\mathbf{x})$ the assumed distribution.

Theorem 1: For an arbitrary input distribution $p(\mathbf{x})$, an assumed input distribution $q(\mathbf{x})$ and a vector Gaussian channel $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}; \mathbf{Hx}, \mathbf{\Sigma}_n)$ in (3), we have

$$\nabla_{\mathbf{H}} D(p(\mathbf{y})\|q(\mathbf{y})) = \mathbf{\Sigma}_n^{-1}\mathbf{H}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})$$

where $D(p(\mathbf{y})\|q(\mathbf{y})) = \int p(\mathbf{y})\log(p(\mathbf{y})/q(\mathbf{y}))d\mathbf{y}$ is the KL divergence, $\mathbf{M}_{p,p}$ is the MMSE matrix defined in (5), and $\mathbf{M}_{p,q}$ is the mismatched MSE matrix defined in (7).

This is the main result of the paper. In order to prove this theorem, the following two lemmas from [12], [2] are introduced. For completeness of presentation, we include them here.

Lemma 1: $\nabla_{\mathbf{H}} p(\mathbf{y}|\mathbf{x}) = -\nabla_{\mathbf{y}} p(\mathbf{y}|\mathbf{x})\mathbf{x}^\top$

Proof: Since $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y};\mathbf{H}\mathbf{x},\mathbf{\Sigma}_n)$, we have $\nabla_{\mathbf{y}}\log p(\mathbf{y}|\mathbf{x}) = -\mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})$, and $\nabla_{\mathbf{H}}\log p(\mathbf{y}|\mathbf{x}) = \mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})\mathbf{x}^\top = (-\nabla_{\mathbf{y}}\log p(\mathbf{y}|\mathbf{x}))\mathbf{x}^\top$. Multiplying both sides with $p(\mathbf{y}|\mathbf{x})$, we obtain the result in the lemma.

Lemma 2: $\nabla_{\mathbf{y}}\log p(\mathbf{y}) = -\mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}_p(\mathbf{y}))$

Proof:

$$\begin{aligned}
&\nabla_{\mathbf{y}}\log p(\mathbf{y}) = (\nabla_{\mathbf{y}}p(\mathbf{y}))/p(\mathbf{y})\\
&= \int(\nabla_{\mathbf{y}}p(\mathbf{y}|\mathbf{x}))p(\mathbf{x})d\mathbf{x}/p(\mathbf{y})\\
&= \int p(\mathbf{y}|\mathbf{x})(\nabla_{\mathbf{y}}\log p(\mathbf{y}|\mathbf{x}))p(\mathbf{x})d\mathbf{x}/p(\mathbf{y})\\
&= -\int p(\mathbf{y}|\mathbf{x})\mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})p(\mathbf{x})d\mathbf{x}/p(\mathbf{y})\\
&= -\int p(\mathbf{x}|\mathbf{y})\mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x})d\mathbf{x} = -\mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}_p(\mathbf{y})).
\end{aligned}$$

Note that this lemma is also true if $p(\cdot)$ is replaced with $q(\cdot)$.

Using the above two lemmas and the method of integration by parts (IBP), we now prove the following.

Lemma 3:

$$\begin{aligned}
\nabla_{\mathbf{H}}\Big(-\int p(\mathbf{y})\log q(\mathbf{y})d\mathbf{y}\Big) = \mathbf{\Sigma}_n^{-1}\mathbf{H}\iint\Big(&\mathbf{x}\mathbf{x}^\top - \mathbf{x}_q(\mathbf{y})\mathbf{x}_p^\top(\mathbf{y})\\
&-\mathbf{x}_p(\mathbf{y})\mathbf{x}_q^\top(\mathbf{y}) + \mathbf{x}_q(\mathbf{y})\mathbf{x}_q^\top(\mathbf{y}))p(\mathbf{x})p(\mathbf{y}|\mathbf{x})\Big)d\mathbf{x}d\mathbf{y}.
\end{aligned}$$

Proof: The lefthand side can be expressed as

$$\begin{aligned}
&\nabla_{\mathbf{H}}\Big(-\int p(\mathbf{y})\log q(\mathbf{y})d\mathbf{y}\Big)\\
&= -\int(\nabla_{\mathbf{H}}p(\mathbf{y}))\log q(\mathbf{y})d\mathbf{y} - \int p(\mathbf{y})\nabla_{\mathbf{H}}\log q(\mathbf{y})d\mathbf{y} \qquad (9)
\end{aligned}$$

The first term in (9) can be expressed as

$$\begin{aligned}
&-\int(\nabla_{\mathbf{H}}p(\mathbf{y}))\log q(\mathbf{y})d\mathbf{y}\\
&= -\iint(\nabla_{\mathbf{H}}p(\mathbf{y}|\mathbf{x}))p(\mathbf{x})d\mathbf{x}\cdot\log q(\mathbf{y})d\mathbf{y}\\
&\overset{\text{(Lemma 1)}}{=} \int\Big(\int(\nabla_{\mathbf{y}}p(\mathbf{y}|\mathbf{x}))\log q(\mathbf{y})d\mathbf{y}\Big)\cdot\mathbf{x}^\top p(\mathbf{x})d\mathbf{x}\\
&\overset{\text{(IBP)}}{=} \int\Big(-\int(\nabla_{\mathbf{y}}\log q(\mathbf{y}))p(\mathbf{y}|\mathbf{x})d\mathbf{y}\Big)\cdot\mathbf{x}^\top p(\mathbf{x})d\mathbf{x}\\
&\overset{\text{(Lemma 2)}}{=} \iint\mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}_q(\mathbf{y}))\mathbf{x}^\top p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y}\\
&= \mathbf{\Sigma}_n^{-1}\mathbf{H}\iint(\mathbf{x}\mathbf{x}^\top - \mathbf{x}_q(\mathbf{y})\mathbf{x}_p^\top(\mathbf{y}))p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y}.
\end{aligned}$$

The method of integration by parts (IBP) is used in the above derivation, *i.e.*, $\int\mathbf{u}d\mathbf{v} = \mathbf{u}\mathbf{v} - \int\mathbf{v}d\mathbf{u}$. In the above case, the first term $\mathbf{u}\mathbf{v} = \log q(\mathbf{y})\cdot p(\mathbf{y}|\mathbf{x})$ vanishes at infinity, and only the second term remains. Similar technique was used in [12], [2].

The second term in (9) can be expressed as

$$\begin{aligned}
&-\int p(\mathbf{y})\nabla_{\mathbf{H}}\log q(\mathbf{y})d\mathbf{y}\\
&= -\int(\nabla_{\mathbf{H}}q(\mathbf{y}))\cdot(p(\mathbf{y})/q(\mathbf{y}))d\mathbf{y}\\
&= -\iint(\nabla_{\mathbf{H}}p(\mathbf{y}|\mathbf{x}))q(\mathbf{x})d\mathbf{x}\cdot(p(\mathbf{y})/q(\mathbf{y}))d\mathbf{y}\\
&\overset{\text{(Lemma 1)}}{=} \iint(\nabla_{\mathbf{y}}p(\mathbf{y}|\mathbf{x}))\mathbf{x}^\top q(\mathbf{x})d\mathbf{x}\cdot(p(\mathbf{y})/q(\mathbf{y}))d\mathbf{y}\\
&= \int\nabla_{\mathbf{y}}\Big(\int p(\mathbf{y}|\mathbf{x})\mathbf{x}^\top q(\mathbf{x})d\mathbf{x}\Big)\cdot(p(\mathbf{y})/q(\mathbf{y}))d\mathbf{y}\\
&= \int(\nabla_{\mathbf{y}}(q(\mathbf{y})\mathbf{x}_q^\top(\mathbf{y})))\cdot(p(\mathbf{y})/q(\mathbf{y}))d\mathbf{y}\\
&\overset{\text{(IBP)}}{=} -\int(\nabla_{\mathbf{y}}(p(\mathbf{y})/q(\mathbf{y})))\cdot(q(\mathbf{y})\mathbf{x}_q^\top(\mathbf{y}))d\mathbf{y}\\
&= -\int(\nabla_{\mathbf{y}}\log(p(\mathbf{y})/q(\mathbf{y})))\cdot(p(\mathbf{y})/q(\mathbf{y}))\cdot(q(\mathbf{y})\mathbf{x}_q^\top(\mathbf{y}))d\mathbf{y}\\
&\overset{\text{(Lemma 2)}}{=} -\int(-\mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}_p(\mathbf{y}))+\mathbf{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}_q(\mathbf{y})))\mathbf{x}_q^\top(\mathbf{y})p(\mathbf{y})d\mathbf{y}\\
&= \mathbf{\Sigma}_n^{-1}\mathbf{H}\int(\mathbf{x}_q(\mathbf{y})-\mathbf{x}_p(\mathbf{y}))\mathbf{x}_q^\top(\mathbf{y})p(\mathbf{y})d\mathbf{y}\\
&= \mathbf{\Sigma}_n^{-1}\mathbf{H}\iint(\mathbf{x}_q(\mathbf{y})\mathbf{x}_q^\top(\mathbf{y})-\mathbf{x}_p(\mathbf{y})\mathbf{x}_q^\top(\mathbf{y}))p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y}.
\end{aligned}$$

The method of integration by parts (IBP) is used again here, and the term $(p(\mathbf{y})/q(\mathbf{y}))\cdot(q(\mathbf{y})\mathbf{x}_q^\top(\mathbf{y}))$ also vanishes at infinity. Combining these two terms, (9) reduces to the result of the lemma. With Lemma 3, it is straightforward to prove the main theorem.

Proof of Theorem 1:
According to the definition of KL divergence,

$$D(p(\mathbf{y})\|q(\mathbf{y})) = \Big(-\int p(\mathbf{y})\log q(\mathbf{y})d\mathbf{y}\Big) - \Big(-\int p(\mathbf{y})\log p(\mathbf{y})d\mathbf{y}\Big).$$

The gradient of the first term is given in Lemma 3, and the gradient of the second term can be derived from Lemma 3 by replacing $q(\cdot)$ with $p(\cdot)$:

$$\begin{aligned}
&\nabla_{\mathbf{H}}\Big(-\int p(\mathbf{y})\log p(\mathbf{y})d\mathbf{y}\Big)\\
&\qquad = \mathbf{\Sigma}_n^{-1}\mathbf{H}\iint(\mathbf{x}\mathbf{x}^\top - \mathbf{x}_p(\mathbf{y})\mathbf{x}_p^\top(\mathbf{y}))p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y}.
\end{aligned}$$

This is the main result derived in [2] (expressed in (4)), since $\nabla_{\mathbf{H}}I(\mathbf{X};\mathbf{Y}) = \nabla_{\mathbf{H}}(h(\mathbf{Y}) - h(\mathbf{Y}|\mathbf{X})) = \nabla_{\mathbf{H}}h(\mathbf{Y})$ for the vector Gaussian channel. And finally

$$\begin{aligned}
&\nabla_{\mathbf{H}}D(p(\mathbf{y})\|q(\mathbf{y}))\\
&= \nabla_{\mathbf{H}}\Big(-\int p(\mathbf{y})\log q(\mathbf{y})d\mathbf{y}\Big) - \nabla_{\mathbf{H}}\Big(-\int p(\mathbf{y})\log p(\mathbf{y})d\mathbf{y}\Big)\\
&= \mathbf{\Sigma}_n^{-1}\mathbf{H}\iint(\mathbf{x}_p(\mathbf{y})-\mathbf{x}_q(\mathbf{y}))(\mathbf{x}_p(\mathbf{y})-\mathbf{x}_q(\mathbf{y}))^\top p(\mathbf{x})p(\mathbf{y}|\mathbf{x})d\mathbf{x}d\mathbf{y}\\
&= \mathbf{\Sigma}_n^{-1}\mathbf{H}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})
\end{aligned}$$

where the identity (8) is used in the last equation.

III. Remarks

Similar to (6), Theorem 1 reveals a fundamental relationship between estimation theory and information theory.

A. Verification of Theorem 1 via Two Gaussian Inputs

Theorem 1 holds for arbitrary input distribution $p(\mathbf{x})$ and arbitrary assumed distribution $q(\mathbf{x})$. When $p(\mathbf{x})$ and $q(\mathbf{x})$ are Gaussian, all quantities in Theorem 1 can be computed analytically. Hence we can verify Theorem 1 in this special case. Details are included in the appendix.

B. Reduction to [3]

When $s = m = 1$, the vector Gaussian channel reduces to a scalar Gaussian channel. Defining $H = \sqrt{\gamma}$, $\Sigma_n = 1$, and applying Theorem 1, we obtain $\nabla_\gamma D(p(\mathbf{y})\|q(\mathbf{y})) = \sqrt{\gamma}(M_{p,q} - M_{p,p})/(2\sqrt{\gamma}) = (M_{p,q} - M_{p,p})/2 = (\mathrm{mse}_q(\gamma) - \mathrm{mmse}(\gamma))/2$, which is equivalent to the result in [3] (see equation (6)). The case of vector input ($s = m > 1$) was also discussed in Section IV of [3], but the channel was always limited to multiplication by a scalar rather than a full matrix.

C. Reduction to [2]

From the proof, we already see that the result in [2] is a special case of Theorem 1. This fact can be illustrated in another way. $I(\mathbf{X};\mathbf{Y})$ can be expressed as $I(\mathbf{X};\mathbf{Y}) = \iint p(\mathbf{x})p(\mathbf{y}|\mathbf{x})\log(p(\mathbf{y}|\mathbf{x})/p(\mathbf{y}))d\mathbf{y}d\mathbf{x} = \int p(\mathbf{x})D(p(\mathbf{y}|\mathbf{x})\|p(\mathbf{y}))d\mathbf{x}$. Since $p(\mathbf{y}|\mathbf{x}) = \int\delta(\tilde{\mathbf{x}}-\mathbf{x})p(\mathbf{y}|\tilde{\mathbf{x}})d\tilde{\mathbf{x}}$, we can view $\delta(\tilde{\mathbf{x}} - \mathbf{x})$ as the true input distribution and $p(\tilde{\mathbf{x}})$ as the assumed input. Applying Theorem 1, we obtain

$$\begin{aligned}
&\nabla_{\mathbf{H}} D(p(\mathbf{y}|\mathbf{x})\|p(\mathbf{y})) = \boldsymbol{\Sigma}_n^{-1}\mathbf{H}(\mathbf{M}_{\delta,p} - \mathbf{M}_{\delta,\delta}) \\
&= \boldsymbol{\Sigma}_n^{-1}\mathbf{H}\iint\delta(\tilde{\mathbf{x}}-\mathbf{x})p(\mathbf{y}|\tilde{\mathbf{x}})(\tilde{\mathbf{x}}-\mathbf{x}_p(\mathbf{y}))(\tilde{\mathbf{x}}-\mathbf{x}_p(\mathbf{y}))^\top d\tilde{\mathbf{x}}d\mathbf{y} \\
&= \boldsymbol{\Sigma}_n^{-1}\mathbf{H}\int p(\mathbf{y}|\mathbf{x})(\mathbf{x}-\mathbf{x}_p(\mathbf{y}))(\mathbf{x}-\mathbf{x}_p(\mathbf{y}))^\top d\mathbf{y}
\end{aligned}$$

since the MMSE matrix $\mathbf{M}_{\delta,\delta}$ is zero in this case. Hence

$$\begin{aligned}
\nabla_{\mathbf{H}} I(\mathbf{X};\mathbf{Y}) &= \int p(\mathbf{x})\nabla_{\mathbf{H}} D(p(\mathbf{y}|\mathbf{x})\|p(\mathbf{y}))d\mathbf{x} \\
&= \boldsymbol{\Sigma}_n^{-1}\mathbf{H}\iint p(\mathbf{x})p(\mathbf{y}|\mathbf{x})(\mathbf{x}-\mathbf{x}_p(\mathbf{y}))(\mathbf{x}-\mathbf{x}_p(\mathbf{y}))^\top d\mathbf{x}d\mathbf{y} \\
&= \boldsymbol{\Sigma}_n^{-1}\mathbf{H}\mathbf{M}_{p,p}
\end{aligned}$$

which is the main result in [2] (see equation (4)). Since (4) reduces to (2) in the scalar case, Theorem 1 also reduces to the result in [1].

D. Non-Gaussianity

Define μ_1 and $\boldsymbol{\Sigma}_1$ to be the mean and covariance matrix of $\mathbf{x}$ computed by the true input distribution $p(\mathbf{x})$ (not necessarily Gaussian), and let $q(\mathbf{x}) = \mathcal{N}(\mathbf{x};\mu_1,\boldsymbol{\Sigma}_1)$. Then $D(p(\mathbf{y})\|q(\mathbf{y}))$ measures the non-Gaussianity [17] of $\mathbf{Y}$. Since $\mathbf{x}_q(\mathbf{y})$ can be analytically expressed [18] as $\mathbf{x}_q(\mathbf{y}) = (\mathbf{H}^\top\boldsymbol{\Sigma}_n^{-1}\mathbf{H} + \boldsymbol{\Sigma}_1^{-1})^{-1}(\mathbf{H}^\top\boldsymbol{\Sigma}_n^{-1}\mathbf{y} + \boldsymbol{\Sigma}_1^{-1}\mu_1)$, we have $\mathbf{x} - \mathbf{x}_q(\mathbf{y}) = (\mathbf{H}^\top\boldsymbol{\Sigma}_n^{-1}\mathbf{H} + \boldsymbol{\Sigma}_1^{-1})^{-1}(-\mathbf{H}^\top\boldsymbol{\Sigma}_n^{-1}(\mathbf{y} - \mathbf{H}\mathbf{x}) + \boldsymbol{\Sigma}_1^{-1}(\mathbf{x} - \mu_1))$, and according to (7), we can derive $\mathbf{M}_{p,q} = (\mathbf{H}^\top\boldsymbol{\Sigma}_n^{-1}\mathbf{H} + \boldsymbol{\Sigma}_1^{-1})^{-1}$. Hence in this special case, our Theorem 1 coincides with Theorem 8 in [2].

E. Relation to Fisher Divergence

The Fisher divergence [19], or relative Fisher information [3], is defined as

$$\mathbf{J}(p(\mathbf{y})\|q(\mathbf{y})) = \int p(\mathbf{y})\nabla_{\mathbf{y}}\log\frac{p(\mathbf{y})}{q(\mathbf{y})}\cdot\nabla_{\mathbf{y}}^\top\log\frac{p(\mathbf{y})}{q(\mathbf{y})}d\mathbf{y} \quad (10)$$

which is a matrix induced from the Fisher information. Applying Lemma 2 to the above equation, we obtain

$$\begin{aligned}
\nabla_{\mathbf{y}}\log\frac{p(\mathbf{y})}{q(\mathbf{y})} &= -\boldsymbol{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}_p(\mathbf{y})) + \boldsymbol{\Sigma}_n^{-1}(\mathbf{y}-\mathbf{H}\mathbf{x}_q(\mathbf{y})) \\
&= \boldsymbol{\Sigma}_n^{-1}\mathbf{H}(\mathbf{x}_p(\mathbf{y}) - \mathbf{x}_q(\mathbf{y})).
\end{aligned}$$

Plugging into (10) and using Theorem 1, we obtain

$$\begin{aligned}
\mathbf{J}(p(\mathbf{y})\|q(\mathbf{y})) =& \boldsymbol{\Sigma}_n^{-1}\mathbf{H}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})\mathbf{H}^\top\boldsymbol{\Sigma}_n^{-1} \\
=& (\nabla_{\mathbf{H}} D(p(\mathbf{y})\|q(\mathbf{y})))\cdot\mathbf{H}^\top\boldsymbol{\Sigma}_n^{-1}.
\end{aligned}$$

Hence a direct relationship is made between Fisher divergence and KL divergence under the vector Gaussian channel assumption. A similar result for the scalar Gaussian channel was obtained in [19], [3]. Since Fisher divergence is used in score matching [19], [20], the novel relation we recover here may be useful for statistical inference of graphical models.

F. Optimization

KL divergence is a widely used distance metric for distributions. It is often used as the objective function to optimize model parameters, especially in statistical inference [21], [22] and information system design [23]. Thanks to Theorem 1, the gradient of KL divergence can be evaluated using estimation-theoretic quantities, which are generally easy to compute for many input distributions (e.g., Gaussian mixture models). Hence gradient descent can be applied to find the optimal channel matrix $\mathbf{H} \leftarrow \mathbf{H} - \xi\nabla_{\mathbf{H}} D(p(\mathbf{y})\|q(\mathbf{y}))$ where ξ is the step size. It is expected that applications will arise in this direction, just as (4) is applied to many real-world problems [5], [6], [7].

IV. Conclusion

In this paper, a novel relation between mismatched estimation and relative entropy in vector Gaussian channels is established. Future research directions include extending the current result to continuous-time Additive White Gaussian Noise (AWGN) channels [15], and finding applications where it provides a useful tool. For ease of presentation, we have restricted our analysis to the real domain; an extension to the complex domain should be straightforward.

Appendix

Suppose the input distributions are Gaussian, *i.e.*, $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_1, \mathbf{\Sigma}_1)$ and $q(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \mu_2, \mathbf{\Sigma}_2)$, and the vector Gaussian channel is specified in (3). Then the output distributions are $p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{H}\mu_1, \mathbf{\Omega}_1^{-1})$ and $q(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mathbf{H}\mu_2, \mathbf{\Omega}_2^{-1})$, where $\mathbf{\Omega}_1 = (\mathbf{H}\mathbf{\Sigma}_1\mathbf{H}^\top + \mathbf{\Sigma}_n)^{-1}$ and $\mathbf{\Omega}_2 = (\mathbf{H}\mathbf{\Sigma}_2\mathbf{H}^\top + \mathbf{\Sigma}_n)^{-1}$. Consequently,

$$\begin{aligned} &D(p(\mathbf{y})\|q(\mathbf{y})) \\ &= \frac{1}{2}\Big(\mathrm{tr}(\mathbf{\Omega}_2\mathbf{\Omega}_1^{-1}) + (\mu_2-\mu_1)^\top\mathbf{H}^\top\mathbf{\Omega}_2\mathbf{H}(\mu_2-\mu_1) \\ &\quad - \log\det(\mathbf{\Omega}_1^{-1}) + \log\det(\mathbf{\Omega}_2^{-1}) - m\Big). \end{aligned}$$

Directly taking the gradient of the above equation, we obtain

$$\begin{aligned} &\nabla_{\mathbf{H}} D(p(\mathbf{y})\|q(\mathbf{y})) \\ &= \mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_1 - \mathbf{\Omega}_2\mathbf{\Omega}_1^{-1}\mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2 && (11) \\ &\quad + \mathbf{\Omega}_2\mathbf{H}(\mu_2-\mu_1)(\mu_2-\mu_1)^\top \\ &\quad - \mathbf{\Omega}_2\mathbf{H}(\mu_2-\mu_1)(\mu_2-\mu_1)^\top\mathbf{H}^\top\mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2 \\ &\quad - \mathbf{\Omega}_1\mathbf{H}\mathbf{\Sigma}_1 + \mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2 \\ &= (\mathbf{\Omega}_2 - \mathbf{\Omega}_1)\mathbf{\Omega}_1^{-1}(\mathbf{\Omega}_1\mathbf{H}\mathbf{\Sigma}_1 - \mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2) \\ &\quad + \mathbf{\Omega}_2\mathbf{H}(\mu_2-\mu_1)(\mu_2-\mu_1)^\top(\mathbf{I} - \mathbf{H}^\top\mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2). && (12) \end{aligned}$$

On the other hand, the posterior distributions $p(\mathbf{x}|\mathbf{y})$ and $q(\mathbf{x}|\mathbf{y})$ can be derived analytically as Gaussians [18], with posterior mean $\mathbf{x}_p(\mathbf{y}) = \mu_1 + \mathbf{\Sigma}_1\mathbf{H}^\top\mathbf{\Omega}_1(\mathbf{y} - \mathbf{H}\mu_1)$ and $\mathbf{x}_q(\mathbf{y}) = \mu_2 + \mathbf{\Sigma}_2\mathbf{H}^\top\mathbf{\Omega}_2(\mathbf{y} - \mathbf{H}\mu_2)$. Hence

$$\begin{aligned} \mathbf{x}_p(\mathbf{y}) - \mathbf{x}_q(\mathbf{y}) &= (\mathbf{I} - \mathbf{\Sigma}_2\mathbf{H}^\top\mathbf{\Omega}_2\mathbf{H})(\mu_1-\mu_2) \\ &\quad + (\mathbf{\Sigma}_1\mathbf{H}^\top\mathbf{\Omega}_1 - \mathbf{\Sigma}_2\mathbf{H}^\top\mathbf{\Omega}_2)(\mathbf{y} - \mathbf{H}\mu_1). \end{aligned}$$

Using the above expression and the identity

$$\mathbf{H}\mathbf{\Sigma}_2\mathbf{H}^\top\mathbf{\Omega}_2 = \mathbf{H}\mathbf{\Sigma}_2\mathbf{H}^\top(\mathbf{H}\mathbf{\Sigma}_2\mathbf{H}^\top + \mathbf{\Sigma}_n)^{-1} = \mathbf{I} - \mathbf{\Sigma}_n\mathbf{\Omega}_2,$$

the righthand side of Theorem 1 becomes

$$\begin{aligned} &\mathbf{\Sigma}_n^{-1}\mathbf{H}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p}) \\ &= \mathbf{\Sigma}_n^{-1}\mathbf{H}\int(\mathbf{x}_p(\mathbf{y}) - \mathbf{x}_q(\mathbf{y}))(\mathbf{x}_p(\mathbf{y}) - \mathbf{x}_q(\mathbf{y}))^\top p(\mathbf{y})d\mathbf{y} \\ &= \mathbf{\Sigma}_n^{-1}\mathbf{H}\Big((\mathbf{I} - \mathbf{\Sigma}_2\mathbf{H}^\top\mathbf{\Omega}_2\mathbf{H})(\mu_1-\mu_2)(\mu_1-\mu_2)^\top(\mathbf{I} - \mathbf{\Sigma}_2\mathbf{H}^\top\mathbf{\Omega}_2\mathbf{H})^\top \\ &\quad + (\mathbf{\Sigma}_1\mathbf{H}^\top\mathbf{\Omega}_1 - \mathbf{\Sigma}_2\mathbf{H}^\top\mathbf{\Omega}_2)\mathbf{\Omega}_1^{-1}(\mathbf{\Sigma}_1\mathbf{H}^\top\mathbf{\Omega}_1 - \mathbf{\Sigma}_2\mathbf{H}^\top\mathbf{\Omega}_2)^\top\Big) \\ &= \mathbf{\Sigma}_n^{-1}\Big((\mathbf{H} - (\mathbf{I} - \mathbf{\Sigma}_n\mathbf{\Omega}_2)\mathbf{H})(\mu_1-\mu_2)(\mu_1-\mu_2)^\top(\mathbf{I} - \mathbf{H}^\top\mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2) \\ &\quad + ((\mathbf{I} - \mathbf{\Sigma}_n\mathbf{\Omega}_1) - (\mathbf{I} - \mathbf{\Sigma}_n\mathbf{\Omega}_2))\mathbf{\Omega}_1^{-1}(\mathbf{\Omega}_1\mathbf{H}\mathbf{\Sigma}_1 - \mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2)\Big) \\ &= \mathbf{\Omega}_2\mathbf{H}(\mu_2-\mu_1)(\mu_2-\mu_1)^\top(\mathbf{I} - \mathbf{H}^\top\mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2) \\ &\quad + (\mathbf{\Omega}_2 - \mathbf{\Omega}_1)\mathbf{\Omega}_1^{-1}(\mathbf{\Omega}_1\mathbf{H}\mathbf{\Sigma}_1 - \mathbf{\Omega}_2\mathbf{H}\mathbf{\Sigma}_2). \end{aligned}$$

Notice that the first term is the mismatch due to the mean, and the second term is the mismatch due to the covariance matrix. Comparing with equation (12), we obtain $\nabla_{\mathbf{H}} D(p(\mathbf{y})\|q(\mathbf{y})) = \mathbf{\Sigma}_n^{-1}\mathbf{H}(\mathbf{M}_{p,q} - \mathbf{M}_{p,p})$. Hence the special case of Gaussian inputs is verified.

References

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

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

[4] D. Guo, Y. Wu, S. Shamai, and S. Verdú, "Estimation in Gaussian noise: Properties of the minimum mean-square error," *Information Theory, IEEE Transactions on*, vol. 57, no. 4, pp. 2371–2385, april 2011.

[5] C. Xiao, Y. Zheng, and Z. Ding, "Globally optimal linear precoders for finite alphabet signals over complex vector Gaussian channels," *Signal Processing, IEEE Transactions on*, vol. 59, no. 7, pp. 3301–3314, july 2011.

[6] W. Carson, M. Chen, M. Rodrigues, R. Calderbank, and L. Carin, "Communications-inspired projection design with application to compressive sensing," *SIAM Journal on Imaging Sciences*, vol. 5, no. 4, pp. 1185–1212, 2012.

[7] M. Chen, W. Carson, M. Rodrigues, R. Calderbank, and L. Carin, "Communications inspired linear discriminant analysis," in *International Conference on Machine Learning*, 2012.

[8] S. Verdú and D. Guo, "A simple proof of the entropy-power inequality," *Information Theory, IEEE Transactions on*, vol. 52, no. 5, pp. 2165–2166, may 2006.

[9] Y. Wu and S. Verdú, "Functional properties of minimum mean-square error and mutual information," *Information Theory, IEEE Transactions on*, vol. 58, no. 3, pp. 1289–1301, 2012.

[10] R. Bustin, M. Payaro, D. P. Palomar, and S. Shamai (Shitz), "On mmse crossing properties and implications in parallel vector Gaussian channels," *Information Theory, IEEE Transactions on*, vol. 59, no. 2, pp. 818–844, feb. 2013.

[11] K. Venkat and T. Weissman, "Pointwise relations between information and estimation in Gaussian noise," *Information Theory, IEEE Transactions on*, vol. 58, no. 10, pp. 6264–6281, oct. 2012.

[12] M. Payaro and D. Palomar, "Hessian and concavity of mutual information, differential entropy, and entropy power in linear vector Gaussian channels," *Information Theory, IEEE Transactions on*, vol. 55, no. 8, pp. 3613–3628, aug. 2009.

[13] D. P. Palomar and S. Verdu, "Representation of mutual information via input estimates," *Information Theory, IEEE Transactions on*, vol. 53, no. 2, pp. 453–470, feb. 2007.

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

[15] T. Weissman, "The relationship between causal and noncausal mismatched estimation in continuous-time awgn channels," *Information Theory, IEEE Transactions on*, vol. 56, no. 9, pp. 4256–4273, sept. 2010.

[16] R. Atar and T. Weissman, "Mutual information, relative entropy, and estimation in the poisson channel," *Information Theory, IEEE Transactions on*, vol. 58, no. 3, pp. 1302 –1318, march 2012.

[17] M. Pinsker, V. Prelov, and S. Verdú, "Sensitivity of channel capacity," *Information Theory, IEEE Transactions on*, vol. 41, no. 6, pp. 1877–1888, nov 1995.

[18] C. Bishop, *Pattern recognition and machine learning*. Springer, New York, 2006.

[19] S. Lyu, "Interpretation and generalization of score matching," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 359–366.

[20] A. Hyvarinen, "Estimation of non-normalized statistical models by score matching," *Journal of Machine Learning Research*, vol. 6, no. 1, pp. 695–708, 2006.

[21] G. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[22] S. Lyu, "Unifying non-maximum likelihood learning objectives with minimum kl contraction," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[23] A. Seghouane, "A kullback-leibler divergence approach to blind image restoration," *Image Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2078–2083, july 2011.