Partially Linear Bayesian Estimation with Application to Sparse Approximations

Tomer Michaeli, Daniel Sigalov and Yonina C. Eldar, *Senior Member, IEEE*

***Abstract*—We address the problem of estimating a random vector X from two sets of measurements Y and Z, such that the estimator is linear in Y. We show that the partially linear minimum mean squared error (PLMMSE) estimator does not require knowing the joint distribution of X and Y in full, but rather only its second-order moments. This renders it of potential interest in various applications. We further show that the PLMMSE method is minimax-optimal among all estimators that solely depend on the second-order statistics of X and Y. Finally, we demonstrate our approach in the context of recovering a vector, which is sparse in a unitary dictionary, from two sets of noisy measurements. We show that in this setting PLMMSE estimation has a clear computational advantage, while its performance is comparable to state-of-the-art algorithms.**

***Index Terms*—Bayesian estimation, minimum mean squared error, linear estimation.**

I. Introduction

Bayesian estimation is concerned with the prediction of a random quantity X based on a set of observations Y, which are statistically related to X. It is well known that the estimator minimizing the mean squared error (MSE) is given by the conditional expectation $\hat{X} = \mathbb{E}[X|Y]$. There are various scenarios, however, in which the minimal MSE (MMSE) estimator cannot be used. This can either be due to implementation constraints, because of the fact that no closed form expression for $\mathbb{E}[X|Y]$ exists, or due to lack of complete knowledge of the joint distribution of X and Y. In these cases, one often resorts to linear estimation. The appeal of the linear MMSE (LMMSE) estimator comes from the fact that it possesses an easily implementable closed form expression, which merely requires knowledge of the joint first- and second-order moments of X and Y.

For example, the amount of computation required for calculating the MMSE estimate of a jump-Markov Gaussian random process from its noisy version grows exponentially in time [1]. By contrast, the LMMSE estimator in this setting possesses a simple recursive implementation, similar to the Kalman filter [2]. A similar problem arises in the area of sparse representations, in which the use of sparsity-inducing Gaussian mixture priors and of Laplacian priors is very common. The complexity of calculating the MMSE estimator under the former prior is exponential in the vector's dimension, calling for approximate solutions [3]. The MMSE estimator under the latter prior does not possess a closed form expression [4], which has motivated the use of alternative estimation strategies such as the maximum a-posteriori (MAP) method.

In practical situations, the reasons for not using the MMSE estimator may only apply to a subset of the measurements. In these cases, it may be desirable to construct an estimator that is linear in part of the measurements and nonlinear in the rest. For example, in multi-view regression problems, the goal is to construct an estimator of X based on two sets of features Y and Z [5]. In these applications, one may be given a large training set of examples $\{x_i, z_i\}$ drawn independently from the joint distribution $F_{XZ}(x,z)$ of X and Z and only a small number of examples $\{x_i, y_i\}$ drawn from $F_{XY}(x,y)$. Thus, $F_{XZ}(x,z)$ can be approximated from the first training set with great accuracy, for example using nonparametric techniques. However, due to its small cardinality, the second set can only be used to estimate the cross-covariance matrix $\mathbf{\Gamma}_{XY}$ of X and Y, but not the entire distribution $F_{XY}(x,y)$. This implies that the MMSE estimator $\mathbb{E}[X|Y,Z]$ cannot be computed and we must settle for estimators that do not require knowing F_{XYZ}. As we will see, one such approach is minimization of the MSE over the class of estimators that are linear in Y.

Partially linear estimation was studied in the statistical literature in the context of regression [6]. In this line of research, it is assumed that the conditional expectation $g(y,z) = \mathbb{E}[X|Y=y, Z=z]$ is linear in y. The goal, then, is to approximate $g(x,y)$ from a set of examples $\{x_i, y_i, z_i\}$ drawn independently from the joint distribution of X, Y and Z. In this correspondence, our goal is to derive the partially linear MMSE (PLMMSE) estimator. Namely, we do not make any assumptions on the structure of the MMSE estimate $\mathbb{E}[X|Y,Z]$, but rather look for the estimator that minimizes the MSE among all functions $g(x,y)$ that are linear in y.

The correspondence is organized as follows. In Section II we present the PLMMSE estimator and discuss some of its properties. In Section III, we show that the PLMMSE method is optimal among all estimators that solely rely on the second-order statistics of X and Y. Finally, we conclude in Section IV with a numerical simulation demonstrating the usefulness of our approach in the context of recovering a sparse signal from noisy measurements.

II. Partially Linear Estimation

We denote random variables (RVs) by capital letters. The pseudo-inverse of a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^\dagger$. The mean $\mathbb{E}[X]$ of an RV X is denoted μ_X and the auto-covariance matrix $\mathbb{C}\text{ov}(X) = \mathbb{E}[(X-\mu_X)(X-\mu_X)^T]$ of X is denoted $\mathbf{\Gamma}_{XX}$. Similarly, $\mathbf{\Gamma}_{XY}$ stands for the cross-covariance matrix $\mathbb{C}\text{ov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)^T]$ of two RVs X and Y. The joint cumulative distribution function of X and Y is written $F_{XY}(x,y) = \mathbb{P}(X \le x, Y \le y)$, where the inequalities are element-wise. By definition, the marginal distribution of X is $F_X(x) = F_{XY}(x,\infty)$. In our setting, X is the quantity to be estimated and Y and Z are two sets of measurements thereof. The RVs X, Y and Z take values in $\mathbb{R}^M$, $\mathbb{R}^N$ and $\mathbb{R}^Q$, respectively. The MSE of an estimator $\hat{X}$ of X is defined as $\mathbb{E}[\|X-\hat{X}\|^2]$.

We begin by considering the most general form of a partially linear estimator of X based on Y and Z, which is given by

$$\hat{X} = \boldsymbol{A}(Z)Y + b(Z). \tag{1}$$

Here $\boldsymbol{A}(z)$ is a matrix-valued function and $b(z)$ is a vector-valued function, so that the realization z of Z is used to choose one of a family of linear estimators of x based on y.

This work has been submitted to IEEE Transactions on Signal Processing for possible publication.

This work was supported in part by the Israel Science Foundation under Grant no. 1081/07 and by the European Commission in the framework of the FP7 Network of Excellence in Wireless COMmunications NEWCOM++ (contract no. 216715).

T. Michaeli and Y. C. Eldar are with the department of electrical engineering and D. Sigalov is with the interdepartmental program for applied mathematics, Technion–Israel Institute of Technology, Haifa, Israel 32000 (phone: +972-4-8294798, +972-4-8294682, fax: +972-4-8295757, e-mail: tomermic@tx.technion.ac.il, dansigal@tx.technion.ac.il, yonina@ee.technion.ac.il). Y. C. Eldar is also a Visiting Professor at Stanford, CA USA.

Theorem 1 *Consider estimators of X having the form* (1)*, for some (Borel measurable) functions $\boldsymbol{A}:\mathbb{R}^Q\to\mathbb{R}^{M\times N}$ and $b:\mathbb{R}^Q\to\mathbb{R}^M$. Then the estimator minimizing the MSE within this class is given by*

$$\hat{X}=\boldsymbol{\Gamma}_{XY|Z}\boldsymbol{\Gamma}_{YY|Z}^{\dagger}(Y-\mathbb{E}[Y|Z])+\mathbb{E}[X|Z], \tag{2}$$

where $\boldsymbol{\Gamma}_{XY|Z}=\mathbb{E}[(X-\mathbb{E}[X|Z])(Y-\mathbb{E}[Y|Z])^T|Z]$ denotes the cross-covariance of X and Y given Z and $\boldsymbol{\Gamma}_{YY|Z}=\mathbb{E}[(Y-\mathbb{E}[Y|Z])(Y-\mathbb{E}[Y|Z])^T|Z]$ is the auto-covariance of Y given Z.

Proof: See Appendix A. ■

Note that (2) is indeed of the form of (1) with $\boldsymbol{A}(Z)=\boldsymbol{\Gamma}_{XY|Z}\boldsymbol{\Gamma}_{YY|Z}^{\dagger}$ and $b(Z)=\mathbb{E}[X|Z]-\boldsymbol{\Gamma}_{XY|Z}\boldsymbol{\Gamma}_{YY|Z}^{\dagger}\mathbb{E}[Y|Z]$. As can be seen, although the MMSE solution among the class of estimators (1) has a simple form, it requires knowing the conditional covariance $\boldsymbol{\Gamma}_{XY|Z}$, which limits its applicability. In particular, this solution cannot be applied in cases where we merely know the unconditional covariance $\boldsymbol{\Gamma}_{XY}$, such as in the multi-view regression scenario described in Section I.

To relax this restriction, we next consider *separable* partially linear estimation. Namely, we seek to minimize the MSE among all functions of the form

$$\hat{X}=\boldsymbol{A}Y+b(Z), \tag{3}$$

where $\boldsymbol{A}$ is a deterministic matrix and $b(z)$ is a vector-valued function.

Theorem 2 *Consider estimators of X having the form* (3)*, for some matrix $\boldsymbol{A}\in\mathbb{R}^{M\times N}$ and (Borel measurable) function $b:\mathbb{R}^Q\to\mathbb{R}^M$. Then the estimator minimizing the MSE within this class is given by*

$$\hat{X}=\boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}W+\mathbb{E}[X|Z], \tag{4}$$

where

$$W=Y-\mathbb{E}[Y|Z]. \tag{5}$$

Proof: See Appendix B. ■

Note again that (4) is of the form of (3) with $\boldsymbol{A}=\boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}$ and $b(Z)=\mathbb{E}[X|Z]-\boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}\mathbb{E}[Y|Z]$. The major advantage of this solution with respect to the non-separable estimator (1), is that the only required knowledge regarding the statistical relation between X and Y is of second-order type. Specifically, as we show in Appendix C, (4) can be equivalently written as

$$\hat{X}=\left(\boldsymbol{\Gamma}_{XY}-\boldsymbol{\Gamma}_{\hat{X}_Z\hat{Y}_Z}\right)\left(\boldsymbol{\Gamma}_{YY}-\boldsymbol{\Gamma}_{\hat{Y}_Z\hat{Y}_Z}\right)^{\dagger}\left(Y-\hat{Y}_Z\right)+\hat{X}_Z, \tag{6}$$

where we denoted $\hat{X}_Z=\mathbb{E}[X|Z]$ and $\hat{Y}_Z=\mathbb{E}[Y|Z]$. Therefore, all we need to know in order to be able to compute the separable PLMMSE estimator (4) is the covariance matrix $\boldsymbol{\Gamma}_{XY}$, the conditional expectation $\mathbb{E}[X|Z]$ and the marginal joint cumulative distribution function F_{YZ} of Y and Z. This is illustrated in Fig. 1. In fact, as we show in Section III, in addition to being optimal among all partially linear methods, the PLMMSE solution (4) is also optimal in a minimax sense among all estimation strategies that rely solely on the quantities appearing in Fig. 1.

The intuition behind (4) is similar to that arising in dynamic estimation schemes, such as the Kalman filter. Specifically, we begin by constructing the estimate $\mathbb{E}[X|Z]$ of X based on the measurements Z, which minimizes the MSE among all functions of Z. Next, we would like to account for Y. However, since Z has already been accounted for, we first need to subtract from Y all variations caused by Z. This is done by constructing the RV W of (5), which can be thought of as the *innovation* associated with the measurements Y with respect to the initial estimate $\mathbb{E}[X|Z]$. Finally, since we want an estimate that is partially linear in Y, we update our initial estimate with the LMMSE estimate of X based on W.

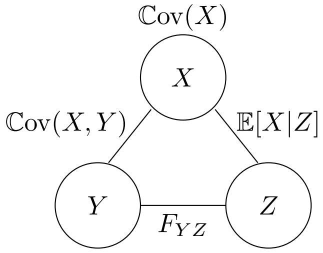

Fig. 1: The statistical knowledge required for computing the PLMMSE estimator (4).

Before proving the minimax-optimality of the PLMMSE estimator, it is insightful to examine several special cases, as we do next.

a) Independent Measurements: Consider first the case in which Y and Z are statistically independent. In this setting, $W=Y-\mu_Y$ and therefore the PLMMSE estimator (4) becomes

$$\hat{X}=\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\mu_Y)+\mathbb{E}[X|Z]=\hat{X}_Y^{\mathrm{L}}+\hat{X}_Z-\mu_X, \tag{7}$$

where $\hat{X}_Y^{\mathrm{L}}$ denotes the LMMSE estimate of X from Y. Thus, in this setting, the PLMMSE estimate reduces to a linear combination of the LMMSE estimate $\hat{X}_Y^{\mathrm{L}}$ and the MMSE estimate $\hat{X}_Z$. The need for subtracting the expectation of X arises from the fact that both $\hat{X}_Y^{\mathrm{L}}$ and $\hat{X}_Z$ account for it.

b) Z is Independent of X and Y: Suppose next that both X and Y are statistically independent of Z. Thus, in addition to the fact that $W=Y-\mu_Y$, we also have $\mathbb{E}[X|Z]=\mu_X$. Consequently, the PLMMSE solution (4) reduces to the LMMSE estimate of X given Y:

$$\hat{X}=\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{YY}^{\dagger}(Y-\mu_Y)+\mu_X=\hat{X}_Y^{\mathrm{L}}. \tag{8}$$

c) Y is Uncorrelated with X and Independent of Z: Consider the situation in which X and Z are statistically independent and X and Y are uncorrelated. Then $W=Y-\mu_Y$, and also $\boldsymbol{\Gamma}_{XW}=\boldsymbol{\Gamma}_{XY}=0$ so that (4) becomes the MMSE estimate of X from Z:

$$\hat{X}=\mathbb{E}[X|Z]=\hat{X}_Z. \tag{9}$$

d) X is Independent of Z: In situations where X and Z are statistically independent, one may be tempted to conclude that the PLMMSE estimator should not be a function of Z. However, this is not necessarily the case. Specifically, although the second term in (4) becomes the constant $\mathbb{E}[X|Z]=\mu_X$ in this setting, it is easily verified that $\boldsymbol{\Gamma}_{XW}=\boldsymbol{\Gamma}_{XY}$, so that the first term in (4) does not vanish unless X is uncorrelated with Y. As a consequence, the PLMMSE estimator can be written as

$$\hat{X}=\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{WW}^{\dagger}Y+\mu_X-\boldsymbol{\Gamma}_{XY}\boldsymbol{\Gamma}_{WW}^{\dagger}\mathbb{E}[Y|Z], \tag{10}$$

in which the last term is a function of Z. This should come as no surprise, though, because if, for instance, $Y=X+Z$, then the optimal estimate is $\hat{X}=Y-Z$, even if X and Z are independent. This solution is clearly a function of Z.

e) X is Uncorrelated with Y: A similar phenomenon occurs when X and Y are uncorrelated. Indeed in this case, $\boldsymbol{\Gamma}_{XW}=-\boldsymbol{\Gamma}_{\hat{X}_Z\hat{Y}_Z}$, so that the first term in (4) does not vanish unless $\hat{X}_Z$ is uncorrelated with $\hat{Y}_Z$. Consequently, the estimator (4) can be expressed as

$$\hat{X}=-\boldsymbol{\Gamma}_{\hat{X}_Z\hat{Y}_Z}\boldsymbol{\Gamma}_{WW}^{\dagger}Y+\boldsymbol{\Gamma}_{\hat{X}_Z\hat{Y}_Z}\boldsymbol{\Gamma}_{WW}^{\dagger}\mathbb{E}[Y|Z]+\mathbb{E}[X|Z], \tag{11}$$

in which the first term is clearly a linear function of Y.

f) Additive Noise: Perhaps the most widely studied measurement model corresponds to linear distortion and additive noise. Specifically, suppose that

$$Y = \boldsymbol{H}X + U, \quad Z = \boldsymbol{G}X + V, \tag{12}$$

where $\boldsymbol{H} \in \mathbb{R}^{N \times M}$ and $\boldsymbol{G} \in \mathbb{R}^{Q \times M}$ are given matrices and U and V are zero-mean RVs such that X, U and V are mutually independent. As we show in Section IV, there are situations in which the distribution of X is such that the complexity of computing the MMSE estimator $\mathbb{E}[X|Y, Z]$ is huge, whereas the complexity of computing $\mathbb{E}[X|Z]$ is modest. In these cases one may prefer to resort to PLMMSE estimation. This method does not correspond to a convex combination of the LMMSE estimate of X from Y and the MMSE estimate of X from Z, as might be suspected. Indeed, substituting $Y = \boldsymbol{H}X + U$, we have that $\boldsymbol{\Gamma}_{XY} = \boldsymbol{\Gamma}_{XX}\boldsymbol{H}^T$ and $\boldsymbol{\Gamma}_{YY} = \boldsymbol{H}\boldsymbol{\Gamma}_{XX}\boldsymbol{H}^T + \boldsymbol{\Gamma}_{UU}$. Furthermore, $\mathbb{E}[Y|Z] = \boldsymbol{H}\mathbb{E}[X|Z]$, so that $\boldsymbol{\Gamma}_{\hat{X}_Z \hat{Y}_Z} = \boldsymbol{\Gamma}_{\hat{X}_Z \hat{X}_Z}\boldsymbol{H}^T$ and $\boldsymbol{\Gamma}_{\hat{Y}_Z \hat{Y}_Z} = \boldsymbol{H}\boldsymbol{\Gamma}_{\hat{X}_Z \hat{X}_Z}\boldsymbol{H}^T$. Consequently, the PLMMSE estimator (6) becomes

$$\hat{X} = \boldsymbol{A}Y + (\boldsymbol{I} - \boldsymbol{A}\boldsymbol{H})\mathbb{E}[X|Z], \tag{13}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = (\boldsymbol{\Gamma}_{XX} - \boldsymbol{\Gamma}_{\hat{X}_Z \hat{X}_Z})\boldsymbol{H}^T \left(\boldsymbol{H}(\boldsymbol{\Gamma}_{XX} - \boldsymbol{\Gamma}_{\hat{X}_Z \hat{X}_Z})\boldsymbol{H}^T + \boldsymbol{\Gamma}_{UU}\right)^{\dagger}. \tag{14}$$

We see that, as opposed to a convex combination of $\hat{X}_Z$ and $\hat{X}_Y^{\mathrm{L}}$, the PLMMSE method reduces to a combination of $\hat{X}_Z$ and Y. Furthermore, the weights of this combination are matrices rather than scalars.

III. Partial Knowledge of Statistical Relations

As discussed in Section II, one of the appealing properties of the PLMMSE solution is that it does not require knowing the entire joint distribution of X and Y, but rather only its second-order moments. However, the fact that the PLMMSE estimator is merely determined by $\mathbb{E}[X|Z]$, $\mathbb{C}\mathrm{ov}(X, Y)$ and $F_{YZ}(y, z)$, does not yet imply that it is optimal among all methods that rely solely on these quantities. The question of optimality of an estimator with respect to partial knowledge regarding the joint distribution of the signal and measurements was recently addressed in [7]. One of the notions of optimality considered there, which we adopt here as well, follows from a worst-case perspective. Specifically, for any estimator $\hat{X} = g(Y, Z)$, there may be distributions $F_{XYZ}(x, y, z)$ consistent with our knowledge such that the MSE is high and there may be distributions leading to low MSE. We consider an estimator as optimal if its worst-case MSE over the set of all feasible distributions is minimal. For example, it was shown in [7] that the LMMSE estimator $\hat{X}_Y^{\mathrm{L}}$ attains the minimal possible worst-case MSE over the set of distributions $F_{XY}(x, y)$ with given first- and second-order moments.

In the next theorem we show that the PLMMSE method is optimal in the sense that its worst-case MSE over the set of all distributions $F_{XYZ}(x, y, z)$ complying with the knowledge appearing in Fig. 1 is minimal.

Theorem 3 *Let $\mathcal{A}$ be the set of probability distributions of (X, Y, Z) satisfying*

$$\begin{aligned} &\mathbb{C}\mathrm{ov}(X) = \boldsymbol{\Gamma}_{XX}, \;\; \mathbb{C}\mathrm{ov}(X, Y) = \boldsymbol{\Gamma}_{XY}, \;\; \mathbb{E}[X|Z] = g(Z), \\ &F_{XYZ}(\infty, y, z) = F_{YZ}(y, z), \end{aligned} \tag{15}$$

where $\boldsymbol{\Gamma}_{XX}$ and $\boldsymbol{\Gamma}_{XY}$ are given matrices, $g(z)$ is a given function and $F_{YZ}(y, z)$ is a given cumulative distribution function. Then, among all estimators of X based on Y and Z, the PLMMSE method (4) *has the minimal worst-case MSE*

$$\sup_{F_{XYZ} \in \mathcal{A}} \mathbb{E}_{F_{XYZ}} \left[\left\| X - \hat{X} \right\|^2 \right], \tag{16}$$

over the set $\mathcal{A}$.

Proof: See Appendix D. ■

IV. Application to Sparse Approximations

We now demonstrate the usefulness of the PLMMSE estimator in the context of sparse approximations. Specifically, consider the situation in which X is known to be sparsely representable in a unitary dictionary $\boldsymbol{\Psi} \in \mathbb{R}^{M \times M}$ in the sense that

$$X = \boldsymbol{\Psi}A \tag{17}$$

for some RV A that is sparse with high probability. More concretely, we assume, as in [3], that the elements of A are given by

$$A_i = S_i B_i, \quad i = 1, \ldots M, \tag{18}$$

where the RVs $\{S_i\}$ and $\{B_i\}$ are statistically independent and distributed as $\mathbb{P}(S_i = 1) = 1 - \mathbb{P}(S_i = 0) = p_i$ and $B_i \sim \mathcal{N}(0, \sigma_{B_i}^2)$.

Assume the signal X is observed through two linear systems, as in (12), where $\boldsymbol{H}$ is an arbitrary matrix, $\boldsymbol{G} = \alpha\boldsymbol{I}$ for some constant $\alpha \neq 0$, and U and V are Gaussian RVs with $\boldsymbol{\Gamma}_{UU} = \boldsymbol{\Gamma}_{VV} = \sigma^2\boldsymbol{I}$. This setting can be cast in the standard sparse approximation form as

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} \boldsymbol{H} \\ \alpha\boldsymbol{I} \end{pmatrix} X + \begin{pmatrix} U \\ V \end{pmatrix}. \tag{19}$$

It is well known that the expression for the MMSE estimate $\mathbb{E}[X|Y, Z]$ in this case generally comprises 2^M summands, which correspond to the different possibilities of sparsity patterns in A [3]. This renders the computation of the MMSE estimate prohibitively expensive even for modest values of M and consequently various approaches have been devised to approximate this solution by a small number of terms (see *e.g.,* [3] and references therein). For example, the fast Bayesian matching pursuit (FBMP) algorithm developed in [3] employs a search in the tree representing all sparsity patterns in order to choose the terms participating in the approximation.

There are some special cases, however, in which the MMSE estimate possesses a simple structure, which can be implemented efficiently. As we show next, one such case is when both the channel's response and the dictionary over which X is sparse correspond to orthogonal matrices. Since in our setting $\boldsymbol{\Psi}$ is unitary and $\boldsymbol{G} = \alpha\boldsymbol{I}$, we can efficiently compute the MMSE estimate $\mathbb{E}[X|Z]$ of X from Z. This implies that, instead of resorting to schemes for approximating $\mathbb{E}[X|Y, Z]$, we can employ the PLMMSE estimator of X based on Y and Z, which possesses a closed form expression (see (13)) in this situation. This technique is particularly effective when the SNR of the observation Y is much worse than that of Z, since the MMSE estimate $\mathbb{E}[X|Y, Z]$ in this case is close to being partially linear in Y. Such a setting is demonstrated in Section IV-C.

A. MMSE Estimation of a Sparse Signal in a Unitary Dictionary

In our setting

$$Z = \alpha X + V = \alpha\boldsymbol{\Psi}A + V, \tag{20}$$

with A of (18). Since $\boldsymbol{\Psi}$ is unitary, it is invertible, and thus the RV

$$\tilde{Z} = \boldsymbol{\Psi}^T Z \tag{21}$$

carries the same information on X as Z does, so that

$$\mathbb{E}[X|Z] = \mathbb{E}[X|\tilde{Z}] = \boldsymbol{\Psi}\mathbb{E}[A|\tilde{Z}]. \tag{22}$$

Now, for every $i = 1, \dots, M$, we have that $\tilde{Z}_i = \alpha A_i + \tilde{V}_i$, where $\tilde{V} = \boldsymbol{\Psi}^T V$ is distributed $\mathcal{N}(0, \sigma^2 \boldsymbol{I})$. Therefore, the set $\{\tilde{Z}_j\}_{j \neq i}$ is statistically independent of the pair $(A_i, \tilde{Z}_i)$ and consequently

$$\begin{aligned} \mathbb{E}[A_i|\tilde{Z}] &= \mathbb{E}[A_i|\tilde{Z}_i] \\ &= \mathbb{E}[A_i|\tilde{Z}_i, S_i = 0]\mathbb{P}(S_i = 0|\tilde{Z}_i) \\ &\quad + \mathbb{E}[A_i|\tilde{Z}_i, S_i = 1]\mathbb{P}(S_i = 1|\tilde{Z}_i). \end{aligned} \tag{23}$$

If $S_i = 0$ then also $A_i = 0$, so that the first term in this expression vanishes. Under the event $S_i = 1$, the RVs A_i and $\tilde{Z}_i$ are jointly normally distributed with mean zero, implying that

$$\mathbb{E}[A_i|\tilde{Z}_i, S_i = 1] = \frac{\mathbb{C}\mathrm{ov}(A_i, \tilde{Z}_i)}{\mathbb{C}\mathrm{ov}(\tilde{Z}_i)} = \frac{\alpha \sigma_{B_i}^2}{\alpha^2 \sigma_{B_i}^2 + \sigma_W^2} \tilde{Z}_i. \tag{24}$$

Finally, using Bayes rule, the term $\mathbb{P}(S_i = 1|\tilde{Z}_i)$ reduces to

$$\begin{aligned} &\frac{f_{\tilde{Z}_i|S_i}(\tilde{Z}_i|S_i = 1)\mathbb{P}(S_i = 1)}{f_{\tilde{Z}_i|S_i}(\tilde{Z}_i|S_i = 0)\mathbb{P}(S_i = 0) + f_{\tilde{Z}_i|S_i}(\tilde{Z}_i|S_i = 1)\mathbb{P}(S_i = 1)} \\ &\quad = \frac{\mathcal{N}(0, \tilde{Z}_i; 0, \alpha^2 \sigma_B^2 + \sigma^2) p}{\mathcal{N}(0, \tilde{Z}_i; 0, \sigma^2)(1 - p) + \mathcal{N}(0, \tilde{Z}_i; 0, \alpha^2 \sigma_B^2 + \sigma^2) p}, \end{aligned} \tag{25}$$

where $\mathcal{N}(\gamma; \mu, \sigma^2)$ denotes the Gaussian density function with mean μ and variance σ^2, evaluated at γ. Substituting (25) and (24) into (23) leads to the following observation.

Theorem 4 *The MMSE estimate of* X *of* (17) *given* Z *of* (20) *is*

$$\mathbb{E}[X|Z] = \boldsymbol{\Psi} \tilde{f}\left(\boldsymbol{\Psi}^T Z\right), \tag{26}$$

where $\tilde{f}(\tilde{z}) = (f(\tilde{z}_1), \dots, f(\tilde{z}_M))^T$, *with*

$$f(\tilde{z}_i) = \frac{\frac{\alpha \sigma_{B_i}^2}{\alpha^2 \sigma_{B_i}^2 + \sigma_W^2} \, p_i \, \mathcal{N}(\tilde{z}_i; 0, \alpha^2 \sigma_{B_i}^2 + \sigma^2) \, \tilde{z}_i}{p_i \, \mathcal{N}(\tilde{z}_i; 0, \alpha^2 \sigma_{B_i}^2 + \sigma^2) + (1 - p_i) \, \mathcal{N}(\tilde{z}_i; 0, \sigma^2)}. \tag{27}$$

B. PLMMSE Estimation of a Sparse Signal From Two Observations

Equipped with a closed form expression for $\mathbb{E}[X|Z]$, we can now obtain an expression for the PLMMSE estimator (13). Specifically, we have that

$$\boldsymbol{\Gamma}_{XX} = \boldsymbol{\Psi} \boldsymbol{\Gamma}_{AA} \boldsymbol{\Psi}^T, \tag{28}$$

where $\boldsymbol{\Gamma}_{AA}$ is a diagonal matrix with $(\boldsymbol{\Gamma}_{AA})_{i,i} = p_i \sigma_{B_i}^2$. Similarly,

$$\boldsymbol{\Gamma}_{\hat{X}_Z \hat{X}_Z} = \boldsymbol{\Psi} \mathbb{C}\mathrm{ov}(\tilde{f}(\tilde{Z})) \boldsymbol{\Psi}^T, \tag{29}$$

where $\mathbb{C}\mathrm{ov}(\tilde{f}(\tilde{Z}))$ is a diagonal matrix whose (i, i) element is $\beta_i = \mathbb{C}\mathrm{ov}(f(\tilde{Z}_i))$. This is due to the fact that the elements of $\tilde{Z}$ are statistically independent and the fact that the function $\tilde{f}(\cdot)$ operates element-wise on its argument. Therefore, the PLMMSE estimator is given in our setting by equation (13) with $\mathbb{E}[X|Z]$ of (26) and with the matrix

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{\Psi}(\boldsymbol{\Gamma}_{AA} - \mathbb{C}\mathrm{ov}(\tilde{f}(\tilde{Z}))) \boldsymbol{\Psi}^T \\ &\quad \times \boldsymbol{H}^T \left(\boldsymbol{H} \boldsymbol{\Psi}(\boldsymbol{\Gamma}_{AA} - \mathbb{C}\mathrm{ov}(\tilde{f}(\tilde{Z}))) \boldsymbol{\Psi}^T \boldsymbol{H}^T + \sigma \boldsymbol{I}\right)^{\dagger}. \end{aligned} \tag{30}$$

We note that if $p_i = p$ and $\sigma_{B_i}^2 = \sigma_B^2$ for every i, then also $\beta_i = \beta$ for every i. In this case,

$$\boldsymbol{\Gamma}_{XX} = \boldsymbol{\Psi} \left(p \sigma_B^2 \boldsymbol{I}\right) \boldsymbol{\Psi}^T = p \sigma_B^2 \boldsymbol{I} \tag{31}$$

and

$$\boldsymbol{\Gamma}_{\hat{X}_Z \hat{X}_Z} = \boldsymbol{\Psi}(\beta \boldsymbol{I}) \boldsymbol{\Psi}^T = \beta \boldsymbol{I}, \tag{32}$$

so that $\boldsymbol{A}$ is simplified to

$$\boldsymbol{A} = (p \sigma_B^2 - \beta) \boldsymbol{H}^T \left((p \sigma_B^2 - \beta) \boldsymbol{H} \boldsymbol{H}^T + \sigma \boldsymbol{I}\right)^{\dagger}. \tag{33}$$

Observe that there is generally no closed form expression for the scalars $\beta_i = \mathbb{C}\mathrm{ov}(f(\tilde{Z}_i))$, rendering it necessary to compute them numerically.

C. Numerical Study

Figure 2 compares the MSE attained by $\hat{X}_{\mathrm{PLMMSE}}$ to that attained by $\hat{X}_Z$, $\hat{X}_Y^{\mathrm{L}}$ and the approximation to $\mathbb{E}[X|Y, Z]$ produced by the FBMP method. In this experiment $\boldsymbol{\Psi} \in \mathbb{R}^{64 \times 64}$ was taken to be a Hadamard matrix with normalized columns. The matrix $\boldsymbol{H}$ corresponded to (circular) convolution with the sequence $h[n] = \exp\{-|n|/12.8\}$. To comply with the assumption made in [3] that the columns of the measurement matrix are normalized, we normalized the columns of $\boldsymbol{H}$ to be of norm 0.99 and set the scalar α to be 0.01. Figure 2 depicts the MSE of all estimators as a function of the input SNR, which we define as $10 \log_{10}(p \sigma_B^2 / \sigma^2)$. As can be seen, the MSE of the PLMMSE method is significantly lower then that of $\hat{X}_Z$ and $\hat{X}_Y^{\mathrm{L}}$ and is very close to that attained by the FBMP method. At low SNR levels and low sparsity levels (high p) the performance of the PLMMSE method is even slightly better than the FBMP approach. Considering the fact that the PLMMSE method is also much faster than the FBMP method in our setting, it seems that there is a clear advantage to using it in scenarios of similar nature.

A word of caution is in place, though. In situations in which the SNR of the measurement Y is roughly the same as that of Z (or better), the FBMP method is advantageous in terms of performance. Therefore in this regime, decision on the use of the PLMMSE method boils down a performance-complexity tradeoff.

V. Conclusions

In this paper we derived the PLMMSE estimator, which is the method whose MSE is minimal among all functions that are linear in Y. We showed that the PLMMSE solution depends only on the joint second-order statistics of X and Y, which renders it applicable in a wide variety of situations. Furthermore, we showed that this estimator attains the lowest worst-case MSE over the set of distributions whose joint second-order moments of X and Y are fixed. Finally, we demonstrated our approach in the context of recovering a sparse vector from noisy measurements. In this application, the PLMMSE solution achieves an MSE very close to that attained by iterative approximation strategies, such as the FBMP method of [3], and is cheaper computationally.

Appendix A
Proof of Theorem 1

Using the smoothing property, the MSE of any estimator of the form (1) is given by

$$\mathbb{E}\left[\mathbb{E}\left[\|X - \boldsymbol{A}(Z)Y - b(Z)\|^2 \,|Z\right]\right]. \tag{34}$$

Thus, for every specific value z that Z can take, the optimal choice of $\boldsymbol{A}(z)$ and $b(z)$ is that minimizing the inner expectation. The solution to this minimization problem corresponds to the LMMSE estimate of X based on Y, under the the joint distribution of (X, Y) given Z, concluding the proof.

Appendix B
Proof of Theorem 2

We start by noting that the set $\mathcal{B}$ of RVs constituting candidate estimates is a closed linear subspace within the space of finite-second-order-moment RVs taking values in $\mathbb{R}^M$. Therefore, the MMSE estimate $\hat{X}$ within this subspace, which is the projection of the RV

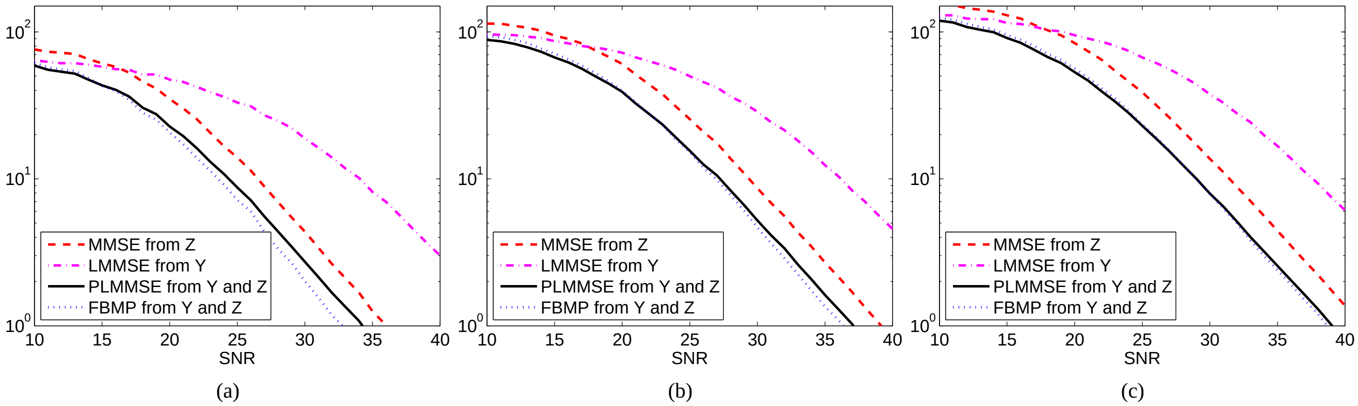

Fig. 2: The MSE attained by $\hat{X}_Z$, $\hat{X}_Y^{\mathrm{L}}$, $\hat{X}_{\mathrm{PLMMSE}}$ and the approximation of $\mathbb{E}[X|Y,Z]$ produced by the FBMP method [3]. (a) $p = 1/3$. (b) $p = 1/2$. (c) $p = 2/3$.

X onto $\mathcal{B}$, is the unique[1] RV whose estimation error $\hat{X} - X$ is orthogonal to every RV of the form $\boldsymbol{A}Y + b(Z)$. To demonstrate that $\hat{X}$ of (4) indeed satisfies this property, note that the inner product between $\hat{X} - X$ and $\boldsymbol{A}Y + b(Z)$ is given by

$$\mathbb{E}\left[(\hat{X} - X)^T(\boldsymbol{A}Y + b(Z))\right] = \mathrm{Tr}\left\{\mathbb{E}\left[(\hat{X} - X)Y^T\right]\boldsymbol{A}^T\right\} + \mathrm{Tr}\left\{\mathbb{E}\left[(\hat{X} - X)b(Z)^T\right]\right\}. \tag{35}$$

Substituting (4), the expectation within the second term becomes

$$\begin{aligned}&\mathbb{E}\left[\left(\boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}W + \mathbb{E}[X|Z] - X\right)b(Z)^T\right]\\ &\quad= \boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}\mathbb{E}\left[Wb(Z)^T\right] + \mathbb{E}\left[\left(\mathbb{E}[X|Z] - X\right)b(Z)^T\right]. \end{aligned}\tag{36}$$

Recall that $W = Y - \mathbb{E}[Y|Z]$ is the estimation error incurred in estimating Y from Z. Consequently, W and $X - \mathbb{E}[X|Z]$ are uncorrelated with every function of Z and, in particular, with $b(Z)$, so that this expression vanishes. Similarly, substituting (4) and expressing $Y = W + \mathbb{E}[Y|Z]$, the expectation within the first summand in (35) becomes

$$\begin{aligned}&\mathbb{E}\left[\left(\boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}W + \mathbb{E}[X|Z] - X\right)Y^T\right]\\ &\qquad= \boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}\mathbb{E}\left[W(W + \mathbb{E}[Y|Z])^T\right]\\ &\qquad\quad- \mathbb{E}\left[(X - \mathbb{E}[X|Z])(W + \mathbb{E}[Y|Z])^T\right]. \end{aligned}\tag{37}$$

Being a function of Z, the RV $\mathbb{E}[Y|Z]$ is uncorrelated with W and $X - \mathbb{E}[X|Z]$ so that this expression can be reduced to

$$\begin{aligned}&\boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}\mathbb{E}\left[WW^T\right] - \mathbb{E}\left[(X - \mathbb{E}[X|Z])W^T\right]\\ &\quad= \boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}\boldsymbol{\Gamma}_{WW} - \mathbb{E}\left[(X - \mu_X + \mu_X - \mathbb{E}[X|Z])W^T\right]\\ &\quad= \boldsymbol{\Gamma}_{XW} - \boldsymbol{\Gamma}_{XW} + \mathbb{E}\left[(\mathbb{E}[X|Z] - \mu_X)W^T\right]\\ &\quad= \boldsymbol{\Gamma}_{XW} - \boldsymbol{\Gamma}_{XW}\\ &\quad= 0, \end{aligned}\tag{38}$$

where we used the facts that $\mathbb{E}[W] = 0$, that $\boldsymbol{\Gamma}_{XW}\boldsymbol{\Gamma}_{WW}^{\dagger}\boldsymbol{\Gamma}_{WW} = \boldsymbol{\Gamma}_{XW}$ [8, Lemma 2], and that W is uncorrelated with $\mathbb{E}[X|Z]$ (due to the same argument as above). This completes the proof.

Appendix C
Derivation of Equation (6)

By definition,

$$\begin{aligned}\boldsymbol{\Gamma}_{XW} &= \mathbb{E}[(X - \mu_X)(Y - \mathbb{E}[Y|Z])^T]\\ &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y + \mu_Y - \mathbb{E}[Y|Z])^T]\\ &= \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^T] - \mathbb{E}[(X - \mu_X)(\mathbb{E}[Y|Z] - \mu_Y)^T]\\ &= \boldsymbol{\Gamma}_{XY} - \mathbb{E}[\mathbb{E}[(X - \mu_X)(\mathbb{E}[Y|Z] - \mu_Y)^T|Z]]\\ &= \boldsymbol{\Gamma}_{XY} - \mathbb{E}[(\mathbb{E}[X|Z] - \mu_X)(\mathbb{E}[Y|Z] - \mu_Y)^T]\\ &= \boldsymbol{\Gamma}_{XY} - \boldsymbol{\Gamma}_{\hat{X}_Z\hat{Y}_Z}, \end{aligned}\tag{39}$$

where the fourth equality is a result of the smoothing property and the last equality follows from the facts that $\mathbb{E}[\mathbb{E}[X|Z]] = \mu_X$ and $\mathbb{E}[\mathbb{E}[Y|Z]] = \mu_Y$. In a similar manner, it is easy to show that

$$\boldsymbol{\Gamma}_{YW} = \boldsymbol{\Gamma}_{YY} - \boldsymbol{\Gamma}_{\hat{Y}_Z\hat{Y}_Z}. \tag{40}$$

Using (40) and the fact that W is uncorrelated with $\mathbb{E}[Y|Z] - \mu_Y$, we obtain

$$\begin{aligned}\boldsymbol{\Gamma}_{WW} &= \mathbb{E}[WW^T]\\ &= \mathbb{E}[(Y - \mathbb{E}[Y|Z])W^T]\\ &= \mathbb{E}[(Y - \mu_Y)W^T] - \mathbb{E}[(\mathbb{E}[Y|Z] - \mu_Y)W^T]\\ &= \boldsymbol{\Gamma}_{YW}\\ &= \boldsymbol{\Gamma}_{YY} - \boldsymbol{\Gamma}_{\hat{Y}_Z\hat{Y}_Z}. \end{aligned}\tag{41}$$

Substituting (39) and (41) into (4) leads to (6).

Appendix D
Proof of Theorem 3

Let $\varepsilon(F_{XYZ}, \hat{X}) = \mathbb{E}_{F_{XYZ}}[\|\hat{X} - X\|^2]$ denote the MSE incurred by an estimator $\hat{X}$ of X based on Y and Z, when the joint distribution of X, Y and Z is $F_{XYZ}(x, y, z)$. It is easily verified that

$$\begin{aligned}&\varepsilon(F_{XYZ}, \hat{X}_{\mathrm{PLMMSE}}) = \mathrm{Tr}\{\boldsymbol{\Gamma}_{XX}\}\\ &\quad- \mathrm{Tr}\left\{(\boldsymbol{\Gamma}_{XY} - \boldsymbol{\Gamma}_{\hat{X}_Z\hat{Y}_Z})(\boldsymbol{\Gamma}_{YY} - \boldsymbol{\Gamma}_{\hat{Y}_Z\hat{Y}_Z})^{\dagger}(\boldsymbol{\Gamma}_{XY} - \boldsymbol{\Gamma}_{\hat{X}_Z\hat{Y}_Z})^T\right\} \end{aligned}\tag{42}$$

for all $F_{XYZ} \in \mathcal{A}$. Therefore, in particular, (42) is also the worst-case MSE of $\hat{X}_{\mathrm{PLMMSE}}$ over $\mathcal{A}$. We next make use of the following lemma.

[1] In an almost-sure sense.

Lemma 1 *There exists a distribution F^*_{XYZ} in the set $\mathcal{A}$ of distributions satisfying* (15)*, under which the PLMMSE estimate of X based on Y and Z coincides with the MMSE estimate $\mathbb{E}[X|Y,Z]$.*

Proof: See Appendix E. ■

Now, any estimator $\hat{X}$ that is a function of Y and Z satisfies

$$\begin{aligned}\sup_{F_{XYZ}\in\mathcal{A}} \varepsilon(F_{XYZ},\hat{X}) &\geq \varepsilon(F^*_{XYZ},\hat{X})\\ &\geq \min_{\hat{X}} \varepsilon(F^*_{XYZ},\hat{X})\\ &= \varepsilon(F^*_{XYZ},\mathbb{E}[X|Y,Z])\\ &= \varepsilon(F^*_{XYZ},\hat{X}_{\mathrm{PLMMSE}})\\ &= \max_{F_{XYZ}\in\mathcal{A}} \varepsilon(F_{XYZ},\hat{X}_{\mathrm{PLMMSE}}), \qquad (43)\end{aligned}$$

where the first line follows from the fact that $F^*_{XYZ}\in\mathcal{A}$, the third line is a result of the fact that the MMSE and PLMMSE estimators coincide under F^*_{XYZ}, and the last line is due to the fact that $\varepsilon(F_{XYZ},\hat{X}_{\mathrm{PLMMSE}})$ is constant as a function of F_{XYZ} over $\mathcal{A}$. We have thus established that the worst-case MSE of any estimator over $\mathcal{A}$ is greater or equal to the worst-case MSE of the PLMMSE solution over $\mathcal{A}$, proving that $\hat{X}_{\mathrm{PLMMSE}}$ is minimax optimal.

Appendix E
Proof of Lemma 1

We prove the statement by construction. Let Y and Z be two RVs distributed according to F_{YZ} and denote $h(Z)=\mathbb{E}[Y|Z]$ and $W=Y-h(Z)$. Let U be a zero-mean RV, statistically independent of the pair (W,Z), whose covariance matrix is

$$\mathbf{\Gamma}_{UU}=\mathbf{\Gamma}_{XX}-\mathbb{C}\mathrm{ov}(g(Z))-\mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}\mathbf{\Gamma}_{WX}, \qquad (44)$$

Consider the RV[2]

$$X=\mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}W+g(Z)+U. \qquad (45)$$

We will show that the so constructed X, Y and Z satisfy the constraints (15). Indeed, using the fact that U has zero mean and is statistically independent of Z, we find that the conditional expectation of X of (45) given Z is

$$\mathbb{E}[X|Z]=g(Z). \qquad (46)$$

Furthermore, since W, $g(Z)$ and U are pairwise uncorrelated, the covariance of X of (45) can be computed as

$$\begin{aligned}\mathbb{C}\mathrm{ov}(X) &= \mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}\mathbf{\Gamma}_{WW}\mathbf{\Gamma}^{\dagger}_{WW}\mathbf{\Gamma}_{WX}+\mathbb{C}\mathrm{ov}(g(Z))+\mathbf{\Gamma}_{UU}\\ &= \mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}\mathbf{\Gamma}_{WX}+\mathbb{C}\mathrm{ov}(g(Z))\\ &\quad +\mathbf{\Gamma}_{XX}-\mathbb{C}\mathrm{ov}(g(Z))-\mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}\mathbf{\Gamma}_{WX}\\ &= \mathbf{\Gamma}_{XX}, \qquad (47)\end{aligned}$$

where we substituted (44). Finally, the cross covariance of X of (45) and Y is given by

$$\begin{aligned}\mathbb{C}\mathrm{ov}(X,Y) &= \mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}\mathbf{\Gamma}_{WY}+\mathbb{C}\mathrm{ov}(g(Z),h(Z))\\ &= \mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}\mathbf{\Gamma}_{WW}+\mathbb{C}\mathrm{ov}(g(Z),h(Z))\\ &= \mathbf{\Gamma}_{XW}+\mathbb{C}\mathrm{ov}(g(Z),h(Z))\\ &= \mathbf{\Gamma}_{XY}-\mathbb{C}\mathrm{ov}(g(Z),h(Z))+\mathbb{C}\mathrm{ov}(g(Z),h(Z))\\ &= \mathbf{\Gamma}_{XY}, \qquad (48)\end{aligned}$$

where the second and fourth equalities follow from the third and fourth lines of (41) and the third equality follows from [8, Lemma 2]. Equations (46), (47) and (48) demonstrate that the distribution F^*_{XYZ} associated with X, Y and Z, belongs to the family of distributions $\mathcal{A}$ satisfying (15).

Next, we show that the PLMMSE and MMSE estimators coincide under F^*_{XYZ}. Indeed, since U is statistically independent of the pair (W,Z), we have that $\mathbb{E}[U|W,Z]=\mathbb{E}[U]=0$, so that

$$\begin{aligned}\mathbb{E}[X|Y,Z] &= \mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}\mathbb{E}[W|Y,Z]+\mathbb{E}[g(Z)+U|Y,Z]\\ &= \mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}(Y-h(Z))+g(Z)+\mathbb{E}[U|W,Z]\\ &= \mathbf{\Gamma}_{XW}\mathbf{\Gamma}^{\dagger}_{WW}(Y-h(Z))+g(Z), \qquad (49)\end{aligned}$$

where we used the fact that there is a one-to-one transformation between the pair (Y,Z) and the pair (W,Z). This expression is partially linear in Y, implying that this is also the PLMMSE estimator in this setting. Thus, for the distribution F^*_{XYZ}, the PLMMSE estimator is optimal not only among all partially linear functions, but also among *all* functions of Y and Z.

References

[1] H. A. P. Blom and Y. Bar-Shalom, "The interacting multiple model algorithm for systems with Markovian switching coefficients," *IEEE Trans. Autom. Control*, vol. 33, no. 8, pp. 780–783, 1988.

[2] O. L. V. Costa, "Linear minimum mean square error estimation for discrete-time Markovian jump linear systems," *IEEE Trans. Autom. Control*, vol. 39, no. 8, pp. 1685–1689, 1994.

[3] P. Schniter, L. C. Potter, and J. Ziniel, "Fast Bayesian matching pursuit," in *Information Theory and Applications Workshop (ITA'08)*, 2008, pp. 326–333.

[4] M. Girolami, "A Variational method for learning sparse and overcomplete representations," *Neural Computation*, vol. 13, no. 11, pp. 2517–2532, 2001.

[5] S. M. Kakade and D. P. Foster, "Multi-view regression via canonical correlation analysis," in *Proceedings of the 20th annual conference on learning theory*, 2007, pp. 82–96.

[6] W. Härdle and H. Liang, *Statistical Methods for Biostatistics and Related Fields*, 2007, ch. Partially linear models, pp. 87–103.

[7] T. Michaeli and Y. C. Eldar, "Hidden relationships: Bayesian estimation with partial knowledge," *IEEE Trans. Signal Process.*, vol. 59, no. 5, 2011, to appear.

[8] A. Torokhti and P. Howlett, "An optimal filter of the second order," *IEEE Trans. Signal Process.*, vol. 49, no. 5, pp. 1044–1048, 2002.

[2] Recall that $\mathbf{\Gamma}_{XW}$ and $\mathbf{\Gamma}_{WW}$ are functions of $\mathbb{C}\mathrm{ov}(X,Y)$ and F_{YZ}, which are given.