Robust Bayesian Transfer Learning between Autoregressive Inference Tasks

Alec Barber [a]
barberalec2@gmail.com

Anthony Quinn [a,b]
aquinn@tcd.ie

[a] *Institute of Information Theory and Automation*
Czech Academy of Sciences
Prague, Czech Republic

[b] *Department of Electronic and Electrical Engineering*
Trinity College Dublin, the University of Dublin
Dublin, Ireland

2021 32nd Irish Signals and Systems Conference (ISSC) | 978-1-6654-3429-4/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSC52156.2021.9467857

***Abstract*—Bayesian transfer learning typically relies on a complete stochastic dependence specification between source and target learners. We advocate a solution to the Bayesian transfer learning paradigm which adopts Fully Probabilistic Design (FPD) to search for an optimal choice of distribution constrained by probabilistic source knowledge. Using this optimal decision-making strategy, an algorithm for accepting source knowledge is identified but is found to be effectively insensitive to source uncertainty. Therefore, we propose an adaptation of the FPD framework which results in a robust transfer learning algorithm.**

Experimental evidence gathered via synthetic data shows enhanced performance when employing both optimal algorithms in a low source data predictor variance regime. In a high source data predictor variance setting, only our adapted FPD-optimal algorithm achieves robustness.

***Index Terms*—Autoregressive (AR) model, Bayesian transfer learning, data-predictive transfer, FPD, robust transfer.**

I. Introduction

Transfer learning (multi-task learning) [1] [2] is the process of adopting knowledge from a set of source learning tasks to increase the learning rate of a given target task [3]. This paper is concerned with transfer learning in a Bayesian context and proposes two algorithms for knowledge transfer between a source and target autoregressive (AR) inference task via the sharing of a one-step-ahead data predictor or a conditional one-step-ahead mean data predictor (probability distributions). The objective of Bayesian Transfer Learning (BTL) is to identify an optimal target distribution M°, describing the targets beliefs in its generative parameters $\boldsymbol{\theta}$, conditioned on provided source knowledge $\mathsf{F_S}$ and n locally observed data $\mathbf{z_n}$. We are further concerned with the robustness property, i.e. positive knowledge transfer when quality source data is available and mitigating negative transfer otherwise.

BTL in literature is typically approached by the design of a prior for the target via knowledge provided by source learners [4], or by the assumption/design of a joint model between the target and source learners [5] which provides the setting for Bayesian conditioning. Prior design for knowledge transfer can prove effective but is restrictive due to its inability to incorporate synchronous, on-line learning. We advocate that any requirement for the design or assumption of a full model between target and sources is a restrictive form of transfer learning.

The research has been supported by GAČR grant 18-15970S

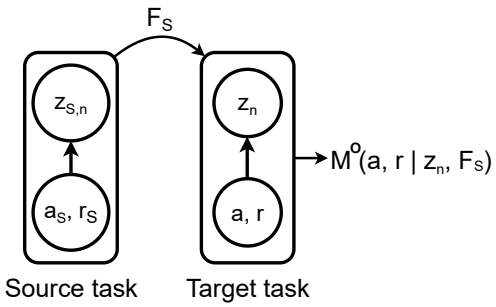

Fig. 1. Source and target inference tasks process their local data $\mathsf{z_{S,n}}$ and $\mathsf{z_n}$ to learn AR parameters $\mathsf{a_S}$, $\mathsf{r_S}$ and a, r respectively. At each time step the source makes available a distribution $\mathsf{F_S}$ to the target. The target elicits an FPD-optimal, $\mathsf{F_S}$-conditioned, posterior predictive distribution which improves its parameter inference performance on a and r.

This paper avoids the specification of a complete model between target and source by adopting principles of Fully Probabilistic Design (FPD) [6] [7] to elicit a source knowledge conditioned distribution on the target's parameters. FPD is as an axiomatically justified [8] optimal distribution design framework which is rooted in the minimum cross-entropy principle. FPD has previously been applied to transfer learning applications such as Gaussian process regression [9], Kalman filtering [10] and Student-t filtering [11].

This paper extends the application of FPD-based transfer learning to an AR setting. The AR model is favoured for its ability to describe relationships between lagged realisations of temporal data. AR models are frequently adopted in the domains of speech analysis [12] and finance [13].

The remainder of this paper is organised as follows. Section II gives an overview of Bayesian parameter inference for an isolated AR setting. Sections III and IV describe the principles of FPD and applies them to the AR model to achieve knowledge transfer between a source and target inference task. Section V identifies an alternative FPD-optimal robust algorithm for the AR transfer learning paradigm. Sections VI and VII outline the experimental setting for validations of the

978-1-6654-3429-4/21/$31.00 ©2021 IEEE

algorithms and provides a discussion of the findings. Section VIII concludes the paper.

II. Review of Bayesian Parameter Inference for the Autoregression Model

The objective when adopting an AR model of order p ($\mathsf{AR}(\mathsf{p})$) to describe n signal data $\mathbf{z}_\mathsf{n}$, is the inference of unknown static generative parameters $\boldsymbol{\theta} = \{\mathbf{a}, \mathsf{r}\}$. Adopting inferences on the parameters, any desired moments or marginals on objects of interest can be derived (e.g. data predictors).

Data is typically made available to the model at discrete steps indexed by $\mathsf{i} = 1, 2, 3, ...$ Equations (1) describe a univariate time-invariant $\mathsf{AR}(\mathsf{p})$ model where e_n is the Gaussian innovation at step n, $\mathsf{z}_\mathsf{n} \in \mathbb{R}$ is the scalar observation and $\boldsymbol{\psi}_\mathsf{n} \in \mathbb{R}^\mathsf{p}$ is the autoregression vector (p previous observations). This paper adopts $^\mathsf{T}$ to denote the transposition operator.

$$\begin{aligned} \mathsf{z}_\mathsf{n} = \mu_{\mathsf{z}_\mathsf{n}} + \mathsf{re}_\mathsf{n} = \boldsymbol{\psi}_\mathsf{n}^\mathsf{T}\mathbf{a} + \mathsf{re}_\mathsf{n}, \\ \mathbf{a} \in \mathbb{R}^\mathsf{p}, \mathsf{r} \in \mathbb{R}^+, \\ \mathsf{e}_\mathsf{n} \sim \mathcal{N}_\mathsf{e}(0, 1), \\ \boldsymbol{\psi}_\mathsf{n} = \left[\mathsf{z}_{\mathsf{n}-1} \dots \mathsf{z}_{\mathsf{n}-\mathsf{p}}\right]^\mathsf{T} \end{aligned} \tag{1}$$

$\mathsf{AR}(\mathsf{p})$ parameter estimation is classically solved via the normal equations which adopt the Wiener criterion [15]. In the Bayesian setting, we adopt a distribution to describe stochastic belief in $\{\mathbf{a}, \mathsf{r}\}$ which can be subsequently updated after each datum observation. A key advantage of the Bayesian method is that parameter (or other moments) estimates come equipped with a first principled measure of uncertainty.

We will now identify a solution for the processing of data $\mathbf{z}_\mathsf{n}$ in both an online and offline inference setting with respect to maintaining computational tractability (i.e. adopting conjugate priors on the model parameters).

$$\mathsf{F}(\mathsf{z}_\mathsf{n} \mid \mathbf{a}, \mathsf{r}, \mathbf{z}_{\mathsf{n}-1}) \equiv \mathcal{N}_{\mathsf{z}_\mathsf{n}}(\mu_{\mathsf{z}_\mathsf{n}}, \mathsf{r}) = \mathcal{N}_{\mathsf{z}_\mathsf{n}}(\boldsymbol{\psi}_\mathsf{n}^\mathsf{T}\mathbf{a}, \mathsf{r}) \tag{2}$$

$$\begin{aligned} \mathsf{F}(\mathbf{a}, \mathsf{r} \mid \mathbf{z}_\mathsf{n}) = \mathsf{F}(\mathbf{a}, \mathsf{r} \mid \mathbf{V}_\mathsf{n}, \nu_\mathsf{n}) = \mathcal{N}\mathrm{iG}_{\mathbf{a},\mathsf{r}}(\mathbf{V}_\mathsf{n}, \nu_\mathsf{n}), \\ \mathbf{V}_\mathsf{n} = \begin{bmatrix} \mathsf{v}_{11,\mathsf{n}} & \mathbf{v}_{\mathsf{a}1,\mathsf{n}}^\mathsf{T} \\ \mathbf{v}_{\mathsf{a}1,\mathsf{n}} & \mathbf{V}_{\mathsf{aa},\mathsf{n}} \end{bmatrix} \end{aligned} \tag{3}$$

The Gaussian AR generative model (2) is a member of the exponential family [14] and therefore is guaranteed to have a conjugate prior distribution and sufficient statistics. The conjugate prior for (2) is identified as the Normal inverse-Gamma ($\mathcal{N}\mathrm{iG}$) (4) distribution [14]. Conjugacy is important in signal processing applications as it facilitates a fully tractable, recursive, and online computational inference flow [18].

$$\begin{aligned} \mathcal{N}\mathrm{iG}_{\mathbf{a},\mathsf{r}}(\mathbf{V}_\mathsf{n}, \nu_\mathsf{n}) = \mathrm{K}_\mathsf{n}^{-1}\mathsf{r}^{-\frac{\nu_\mathsf{n}}{2}} \exp\left[-\frac{1}{2\mathsf{r}}\begin{bmatrix}-1 & \mathbf{a}^T\end{bmatrix}\mathbf{V}_\mathsf{n}\begin{bmatrix}-1 \\ \mathbf{a}\end{bmatrix}\right], \\ \lambda_\mathsf{n} = \mathsf{v}_{11,\mathsf{n}} - \mathbf{v}_{\mathsf{a}1,\mathsf{n}}^\mathsf{T}\mathbf{V}_{\mathsf{aa},\mathsf{n}}^{-1}\mathbf{v}_{\mathsf{a}1,\mathsf{n}} \\ \mathrm{K}_\mathsf{n} = \Gamma\left(\frac{\nu_\mathsf{n}}{2}\right)\lambda_\mathsf{n}^{-\frac{\nu_\mathsf{n}}{2}}|\mathbf{V}_{\mathsf{aa},\mathsf{n}}|^{-\frac{1}{2}}(2\pi)^{\frac{1}{2}}. \end{aligned} \tag{4}$$

Equation (4) describes the sufficient statistic parameterisation of $\mathcal{N}\mathrm{iG}$ which is adopted in this paper. $\mathbf{V}_\mathsf{n} \in \mathbb{R}^{(\mathsf{p}+1)\times(\mathsf{p}+1)}$ is the Extended Information Matrix (EIM), $\nu_\mathsf{n} \in \mathbb{R}^+$ is the degrees of freedom parameter, $\Gamma(\cdot)$ denotes the gamma function and K_n is the normalisation constant. A requirement of the EIM is that it must be symmetric and positive semi-definite. Equation (3) shows the partitioning of the EIM into its sub-components, isolating the scalar $\mathsf{v}_{11,\mathsf{n}}$, the vector $\mathbf{v}_{\mathsf{a}1,\mathsf{n}} \in \mathbb{R}^\mathsf{p}$ and the matrix $\mathbf{V}_{\mathsf{aa},\mathsf{n}} \in \mathbb{R}^{\mathsf{p}\times\mathsf{p}}$. The degrees of freedom parameter ν_n can be interpreted informally as a counter of observed data (note ν_n can be any positive, non-zero, real value). λ_n is the total Bayesian squared error. Provided in (5) below are some important moments of $\mathcal{N}\mathrm{iG}$.

$$\begin{aligned} \mathsf{E}[\mathbf{a} \mid \mathbf{z}_\mathsf{n}] = \mathsf{E}[\mathbf{a} \mid \mathsf{r}, \mathbf{z}_\mathsf{n}] = \mathbf{V}_{\mathsf{aa},\mathsf{n}}^{-1}\mathbf{v}_{\mathsf{a}1,\mathsf{n}}, \\ \mathsf{E}[\mathsf{r} \mid \mathbf{z}_\mathsf{n}] = \frac{\lambda_\mathsf{n}}{\nu_\mathsf{n} - \mathsf{p} + 2}, \qquad \nu_\mathsf{n} > \mathsf{p} - 2 \end{aligned} \tag{5}$$

It can be shown that the following are the fixed and finite dimensional updates of the Bayesian sufficient statistics prescribed by (2) and (3) after observing $\mathbf{z}_\mathsf{n}$ [15].

$$\begin{aligned} \mathbf{V}_\mathsf{n} = \mathbf{V}_{\mathsf{n}-1} + \begin{bmatrix}\mathsf{z}_\mathsf{n} \\ \boldsymbol{\psi}_n\end{bmatrix}\begin{bmatrix}\mathsf{z}_\mathsf{n} & \boldsymbol{\psi}_\mathsf{n}^\mathsf{T}\end{bmatrix} = \mathbf{V}_0 + \sum_{\mathsf{i}=1}^{\mathsf{n}}\begin{bmatrix}\mathsf{z}_\mathsf{i} \\ \boldsymbol{\psi}_\mathsf{i}\end{bmatrix}\begin{bmatrix}\mathsf{z}_\mathsf{i} & \boldsymbol{\psi}_\mathsf{i}^\mathsf{T}\end{bmatrix} \\ \nu_\mathsf{n} = \nu_{\mathsf{n}-1} + 1 = \nu_0 + \mathsf{n} \end{aligned} \tag{6}$$

The EIM $\mathbf{V}_{\mathsf{n}-1}$ is updated by the accumulation of the outer product of the vector z_n extended by $\boldsymbol{\psi}_\mathsf{n}$. Of particular note is that this isolated update procedure reappears in similar forms in both sections IV and V when conducting transfer learning. $\mathbf{V}_0$ and ν_0 in (6) represent the prior belief in the parameters before any data is considered. If no preference is provided, a diffuse distribution is typically adopted (i.e. $\mathbf{V}_0 = \epsilon I_{\mathsf{p}+1}, \nu_0 = \epsilon$, where ϵ is a small positive real number).

The one-step-ahead data predictor for the $\mathcal{N}\mathrm{iG}$ model is identified in (7) as the 3 parameter Student-t distribution.

$$\begin{aligned} \mathsf{F}(\mathsf{z}_{\mathsf{n}+1} \mid \mathbf{z}_\mathsf{n}) = \int_{\mathbf{a},\mathsf{r}} \mathsf{F}(\mathsf{z}_{\mathsf{n}+1} \mid \mathbf{a}, \mathsf{r}, \boldsymbol{\psi}_{\mathsf{n}+1})\mathsf{F}(\mathbf{a}, \mathsf{r} \mid \mathbf{z}_\mathsf{n})\mathsf{d}\mathbf{a}\mathsf{d}\mathsf{r} \\ = \mathsf{St}\text{-}\mathsf{t}_{\mathsf{z}_{\mathsf{n}+1}}\left(\boldsymbol{\psi}_{\mathsf{n}+1}^\mathsf{T}\mathbf{V}_{\mathsf{aa},\mathsf{n}}^{-1}\mathbf{v}_{\mathsf{a}1,\mathsf{n}}, \lambda_\mathsf{n}\frac{1 + \boldsymbol{\psi}_{\mathsf{n}+1}^\mathsf{T}\mathbf{V}_{\mathsf{aa},\mathsf{n}}^{-1}\boldsymbol{\psi}_{\mathsf{n}+1}}{\nu_\mathsf{n}}, \nu_\mathsf{n}\right) \end{aligned} \tag{7}$$

We will now provide a description of the FPD distribution design framework in the context of transfer learning which we will adopt to optimally choose a source knowledge conditioned object.

III. FPD-Optimal Bayesian Transfer Learning

This paper is concerned with BTL in an incompletely modeled setting (i.e. where no explicit relationship between source and target is assumed or designed). This incompletely modeled setting is shown in figure 1 where both source and target inference machines conduct independent modelling of local parameters with no explicit association. We desire a distribution on the target's parameters $\mathsf{M} \in \mathbf{M}$ that is conditioned on source knowledge F_S. The choice of M is not uniquely defined and is a decision making task. In this paper we adopt an optimal choice of distribution M° via the Bayesian minimum risk decision making strategy known as FPD [7].

Adoption of FPD design principles require the specifications of two key elements: 1) a user-specified choice of

an ideal, zero-loss choice distribution M_I, 2) a description of how knowledge constrains the set of possible candidate distributions $\mathsf{M} \in \mathbf{M}$. Note that this paper adopts M and F as variational (i.e. to be designed) and fixed-form distributions respectively.

$$\mathsf{M}^\circ \equiv \underset{\mathsf{M}\in\mathbf{M}}{\arg\min}\, \mathcal{D}\left(\mathsf{M}||\mathsf{M}_\mathsf{I}\right) \tag{8}$$

FPD dictates that the optimal distribution choice M° should be chosen as the candidate $\mathsf{M} \in \mathbf{M}$ which minimises the Kullback-Leibler divergence to the user-specified ideal M_I. When no restrictions are placed on $\mathbf{M}$, the optimal zero-loss choice in distribution is always M_I. The Kullback-Leibler divergence is defined in (9) as the expected log-odds between the candidate and ideal distributions [18].

$$\mathcal{D}\left(\mathsf{M}||\mathsf{M}_\mathsf{I}\right) = \mathsf{E}_\mathsf{M}\left[\log\left(\frac{\mathsf{M}}{\mathsf{M}_\mathsf{I}}\right)\right] \tag{9}$$

We will now instantiate equations (8) and (9) for the $\mathcal{N}$iG-AR tasks (2) and (3) where the object being transferred is the source's one-step-ahead data predictor transfer (7).

IV. FPD-Optimal BTL Between AR Inference Tasks

Equation (10) outlines the choice in ideal distribution as the joint model of $\mathbf{a}, \mathsf{r}$ and the next unknown datum $\mathsf{z}_{\mathsf{n}+1}$. This joint model expressed in terms of the target's fixed-form generative model (2) and the posterior predictive distribution of the parameters (3). This choice in ideal distribution is motivated by its complete description of the state of the inference task in the absence of external knowledge.

$$\mathsf{M}_\mathsf{I}(\mathbf{a}, \mathsf{r}, \mathsf{z}_{\mathsf{n}+1} \mid \mathbf{z}_\mathsf{n}) \equiv \mathsf{F}(\mathsf{z}_{\mathsf{n}+1} \mid \mathbf{a}, \mathsf{r}, \boldsymbol{\psi}_{\mathsf{n}+1})\mathsf{F}\left(\mathbf{a}, \mathsf{r} \mid \mathbf{z}_\mathsf{n}\right) \tag{10}$$

$$\mathsf{M}(\mathbf{a}, \mathsf{r}, \mathsf{z}_{\mathsf{n}+1} \mid \mathsf{F}_\mathsf{S}, \mathbf{z}_\mathsf{n}) \equiv \mathsf{F}_\mathsf{S}(\mathsf{z}_{\mathsf{n}+1} \mid \mathbf{z}_{\mathsf{S},\mathsf{n}})\mathsf{M}\left(\mathbf{a}, \mathsf{r} \mid \mathsf{F}_\mathsf{S}, \mathbf{z}_\mathsf{n}\right) \tag{11}$$

The variational candidate distribution (11) is a conditional on the source's one-step-ahead data predictor F_S (described in (7)). (11) is presumed unavailable, i.e. unknown functional form. We desire to find an optimal choice of this variational object which has been constrained by source knowledge F_S.

We assert in (11) that F_S is descriptive of the target's next datum $\mathsf{z}_{\mathsf{n}+1}$. Moreover, a conditional independence assumption between $\{\mathbf{a}, \mathsf{r}\}$ and $\mathsf{z}_{\mathsf{S},\mathsf{n}+1}$ given F_S is assumed. Equation (12) illustrates the conditional independence assumption made in (11).

$$\begin{aligned}\mathsf{M}(\mathsf{z}_{\mathsf{n}+1} \mid \mathsf{F}_\mathsf{S}, \mathbf{a}, \mathsf{r}, \mathbf{z}_\mathsf{n}) \equiv \mathsf{M}(\mathsf{z}_{\mathsf{n}+1} \mid \mathsf{F}_\mathsf{S}, \boldsymbol{\psi}_{\mathsf{n}+1}) \equiv \\ \mathsf{F}_\mathsf{S}(\mathsf{z}_{\mathsf{n}+1} \mid \boldsymbol{\psi}_{\mathsf{n}+1}, \mathbf{z}_{\mathsf{S},\mathsf{n}})\end{aligned} \tag{12}$$

$$\begin{aligned}\mathsf{M} \in \mathbf{M} \equiv \{&\text{models } (11) \text{ with } \mathsf{F}_\mathsf{S}(\mathsf{z}_{\mathsf{n}+1}) \\ &\text{fixed and } \mathsf{M}\left(\mathbf{a}, \mathsf{r} \mid \mathsf{F}_\mathsf{S}, \mathbf{z}_\mathsf{n}\right) \text{variational}\}\end{aligned} \tag{13}$$

With regards to brevity, $\boldsymbol{\psi}_{\mathsf{n}+1}$ will be omitted for the rest of this paper. We now wish to solve the FPD optimisation task (8) instantiated with (10), (11) and (13).

Proposition 1. *The unknown model belongs to the knowledge constrained set,* $\mathsf{M} \in \mathbf{M}$ *(13), and the ideal model* M_I *is (10). Then the FPD-optimal model is*

$$\mathsf{M}^\circ(\mathbf{a}, \mathsf{r}, \mathsf{z}_{\mathsf{n}+1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_\mathsf{n}) \propto \mathsf{F}_\mathsf{S}(\mathsf{z}_{\mathsf{n}+1})\mathsf{M}^\circ(\mathbf{a}, \mathsf{r}|\mathsf{F}_\mathsf{S}, \mathbf{z}_\mathsf{n}) \tag{14}$$

where

$$\mathsf{M}^\circ(\mathbf{a}, \mathsf{r} \mid \mathsf{F}_\mathsf{S}, \mathbf{z}_\mathsf{n}) \propto \mathsf{F}(\mathbf{a}, \mathsf{r}|\mathbf{z}_\mathsf{n}) \exp\left[\int_{\mathsf{z}_{\mathsf{n}+1}} \ln(\mathsf{F}(\mathsf{z}_{\mathsf{n}+1}|\mathbf{a}, \mathsf{r}))\mathsf{F}_\mathsf{S}(\mathsf{z}_{\mathsf{n}+1})\mathrm{d}\mathsf{z}_{\mathsf{n}+1}\right] \tag{15}$$

Proof. See Appendix A. □

The result achieved in proposition 1 is the canonical solution as reported in literature. A full description of a recursive and computationally tractable algorithm for the AR(p) model is identified in proposition 2.

Proposition 2. *At step* n*, the FPD-optimal source knowledge conditioned model is (15) and the conjugate prior is of the form (4). Then the appropriate recursive hyperparameter update is*

$$\begin{aligned}\mathbf{V}_\mathsf{n}^\circ = &\mathbf{V}_\mathsf{n} + \begin{bmatrix} \mathsf{m}_{\mathsf{S},\mathsf{n}+1} \\ \boldsymbol{\psi}_{\mathsf{n}+1} \end{bmatrix} \begin{bmatrix} \mathsf{m}_{\mathsf{S},\mathsf{n}+1} & \boldsymbol{\psi}_{\mathsf{n}+1}^\mathsf{T} \end{bmatrix} \\ &+ \begin{bmatrix} \mathsf{w}_{\mathsf{S},\mathsf{n}+1} & 0 & \cdots & 0 \\ 0 & 0 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}\end{aligned} \tag{16}$$

$$\nu_\mathsf{n}^\circ = \nu_\mathsf{n} + 1 \tag{17}$$

Where conjugacy between the prior and posterior is conserved and, the first and seconds moments of F_S*,* $\mathsf{m}_{\mathsf{S},\mathsf{n}+1}$ *and* $\mathsf{w}_{\mathsf{S},\mathsf{n}+1}$ *respectively, are given in (7).*

Proof. See Appendix B. □

The resulting computational flow adopts equations (6) for the data-step update procedure with the replacement of $\mathbf{V}_{\mathsf{n}-1}$ and $\nu_{\mathsf{n}-1}$ with the optimal posterior $\mathbf{V}_{\mathsf{n}-1}^\circ$ and $\nu_{\mathsf{n}-1}^\circ$ at step $\mathsf{n}-1$. The source knowledge transfer step adopts equations (16) and (17).

Remark 1. *The source distribution acceptance procedure outlined in proposition 2 processes both the first and second moments of* F_S*. This is an unexpected result as the second order moment transfer does not happen in similar FPD instantiations in literature [10]. Despite this, the expected values of the parameters* $\mathbf{a}$ *and future data realisations are invariant to the source predictor's uncertainty* $\mathsf{w}_{\mathsf{S},\mathsf{n}+1}$*, although it does have an effect on the target's confidence in these objects.*

We will now investigate an alternative algorithm for the transfer of knowledge with motivations to ensure robustness to source knowledge uncertainty.

V. Robust FPD-Optimal BTL

A new specification of ideal distribution M_I is defined in (18). Similar to (10), it is a joint model on the parameters $\{\mathbf{a}, \mathsf{r}\}$ and the target's next datum $\mathsf{z}_{\mathsf{n}+1}$. In contrast to (10), we choose to design our ideal model M_I as a conditional on the source provided object F_S where we define F_S as the source's conditional one-step ahead mean data predictor $\mathsf{F}_\mathsf{S}(\mu_{\mathsf{z}_{\mathsf{n}+1}}|\mathbf{z}_{\mathsf{S},\mathsf{n}}, \mathsf{r}_{\mathsf{CE},\mathsf{n}})$ (20). $\mu_{\mathsf{z}_{\mathsf{n}+1}}$ is the first parameter (mean) of

the AR(p) generative model in (1). $\mathsf{F_S}$ should not be confused with the one step ahead data predictor described in (7) and adopted in section IV.

$$\mathsf{M_I}(\mathbf{a}, \mathsf{r}, \mathsf{z}_{n+1} \mid \mathsf{F_S}, \mathbf{z}_n) \equiv \mathsf{F_S}(\mathsf{z}_{n+1})\mathsf{F}(\mathbf{a}, \mathsf{r} \mid \mathbf{z}_n) \tag{18}$$

$$\mathsf{M}(\mathbf{a}, \mathsf{r}, \mathsf{z}_{n+1} \mid \mathsf{F_S}, \mathbf{z}_n) \equiv \mathsf{F}(\mathsf{z}_{n+1} \mid \mathbf{a}, \mathsf{r})\mathsf{M}(\mathbf{a}, \mathsf{r} \mid \mathsf{F_S}, \mathbf{z}_n) \tag{19}$$

$$\mathsf{F_S}(\mu_{\mathsf{z}_{n+1}} \mid \mathbf{z}_{\mathsf{S},n}, \mathsf{r}_{\mathsf{CE},n}) = \mathcal{N}_{\mu_{\mathsf{z}_{n+1}}}(\boldsymbol{\psi}_{n+1}^{\mathsf{T}}\mathbf{V}_{\mathbf{aa},n}^{-1}\mathbf{v}_{\mathbf{a}1,n}, \frac{\lambda_n}{\nu_n - p + 2}\boldsymbol{\psi}_{n+1}^{\mathsf{T}}\mathbf{V}_{\mathbf{aa}}^{-1}\boldsymbol{\psi}_{n+1}) \tag{20}$$

We constrain our beliefs in M by asserting that z_{n+1} is described by the target's generative model only. When comparing (18) and (19) to (10) and (11), the effective difference is swapping the role of $\mathsf{F_S}$ and the generative model respectively.

$$\mathsf{M} \in \mathbf{M} \equiv \{\text{models (19) with } \mathsf{F}(\mathsf{z}_{n+1} \mid \mathbf{a}, \mathsf{r}) \text{ fixed and } \mathsf{M}(\mathbf{a}, \mathsf{r} \mid \mathsf{F_S}, \mathbf{z}_n) \text{ variational}\} \tag{21}$$

The objective is to once again optimally choose M° from the set of distributions (21) via the minimisation of (8) with the user-specified ideal (18).

Proposition 3. *The unknown model belongs to the knowledge constrained set,* $\mathsf{M} \in \mathbf{M}$ *(21), and the ideal model* $\mathsf{M_I}$ *is (18). Then the FPD-optimal model is*

$$\mathsf{M}^\circ(\mathbf{a}, \mathsf{r}, \mathsf{z}_{n+1} \mid \mathsf{F_S}, \mathbf{z}_n) \propto \mathsf{F_S}(\mathsf{z}_{n+1})\mathsf{M}^\circ(\mathbf{a}, \mathsf{r} \mid \mathsf{F_S}, \mathbf{z}_n) \tag{22}$$

where

$$\mathsf{M}^\circ(\mathbf{a}, \mathsf{r} \mid \mathsf{F_S}, \mathbf{z}_n) \propto \mathsf{F}(\mathbf{a}, \mathsf{r} \mid \mathbf{z}_n) \exp\left[-\mathcal{D}\Big(\mathsf{F}(\mathsf{z}_{n+1} \mid \mathbf{a}, \mathsf{r}) \Big\| \mathsf{F_S}(\mathsf{z}_{n+1})\Big)\right] \tag{23}$$

Proof. See Appendix C. □

Proposition 3 outlines a conical solution for transfer learning problem where both the ideal and candidate distributions are conditioned on the source knowledge $\mathsf{F_S}$. The next task is to identify a computationally tractable inference flow for the updating of belief at each data step. Unfortunately, proposition 3 does not result in a conjugate updating procedure when using the $\mathcal{N}$iG distribution. We now propose the use of an extended $\mathcal{N}$iG distribution to model belief in the parameters as follows.

$$\mathsf{M}(\mathbf{a}, \mathsf{r} \mid \mathbf{V}, \mathbf{B}, \nu, \mathsf{k}) \propto \mathsf{r}^{-\frac{\nu}{2}} \exp\left[-\frac{1}{2\mathsf{r}}\begin{bmatrix}-1 & \mathbf{a}^{\mathsf{T}}\end{bmatrix}\mathbf{V}\begin{bmatrix}-1 \\ \mathbf{a}\end{bmatrix} - \frac{1}{2}\begin{bmatrix}-1\mathbf{a}^{\mathsf{T}}\end{bmatrix}\mathbf{B}\begin{bmatrix}-1 \\ \mathbf{a}\end{bmatrix} - \mathsf{rk}\right] \tag{24}$$

A full investigation of (24) is out of scope of this paper. Regardless, it possess a number of key properties.

1) When diffuse choices of hyper-parameters $\mathbf{B}$ and k are adopted (ϵI_{p} and ϵ respectively, where ϵ is a small positive real number), the extended $\mathcal{N}$iG is equivalent to a standard $\mathcal{N}$iG with hyper parameters $\mathbf{V}, \nu$.
2) When conditioning on a known r, the extended $\mathcal{N}$iG reduces to a multivariate Normal distribution.
3) The hyper-parameter update procedure when processing local data $\mathbf{z}_n$ is unchanged (see (6)).

Considering Property 1 and Property 3, if the extended $\mathcal{N}$iG is adopted in sequential updating in an isolated learner when diffuse priors are adopted, the inference task is identical to traditional inference described in section II.

For the purposes of deriving a robust transfer learning algorithm, we are predominately concerned with producing a marginal distribution on $\mathbf{a}$. Unfortunately, this object is believed to be analytically intractable. This issue is addressed via the adoption of a certainty equivalent choice of r (i.e. $\mathsf{r_{CE}}$) which adopts $\mathsf{E}[\mathsf{r}]$ (5) as a known parameter. Considering Property 2, we can derive a $\mathsf{r_{CE}}$ conditional on $\mathbf{a}$ and is provided in (25) where $\bar{\mathbf{V}}_n = \mathbf{V}_n + \mathsf{r_{CE}}\mathbf{B}_n$.

$$\begin{aligned}\mathsf{M}(\mathbf{a} \mid \mathsf{r_{CE}}, \bar{\mathbf{V}}_n, \mathbf{B}_n, \nu_n, \mathsf{k}_n) &= \mathcal{N}(\bar{\mathbf{V}}_{\mathbf{aa},n}^{-1}\mathbf{v}_{\mathbf{a}1,n}, \mathsf{r_{CE}}\bar{\mathbf{V}}_{\mathbf{aa},n}^{-1}) \\ \mathsf{E}[\mathsf{z}_{n+1} \mid \mathsf{r_{CE}}] &= \boldsymbol{\psi}_{n+1}^{\mathsf{T}}\bar{\mathbf{V}}_{\mathbf{aa},n}^{-1}\mathbf{v}_{\mathbf{a}1,n} \\ \mathsf{VAR}[\mathsf{z}_{n+1} \mid \mathsf{r_{CE}}] &= \mathsf{r_{CE}}\boldsymbol{\psi}_{n+1}^{\mathsf{T}}\bar{\mathbf{V}}_{\mathbf{aa},n}^{-1}\boldsymbol{\psi}_{n+1}\end{aligned} \tag{25}$$

Note that $\mathsf{r_{CE}}$ must be generated at each step using isolated target knowledge only. Equation (25) is not a complete solution as it does not provide a description on r but is sufficient for making inferences on the parameters $\mathbf{a}$ and generating the expected values of future data. We now propose a recursive and tractable update procedure for the extended $\mathcal{N}$iG distribution.

Proposition 4. *At step* n*, the optimal source knowledge conditioned model is (23) and the conjugate prior is (24). Then the appropriate recursive hyper-parameter update is*

$$\begin{aligned}\nu_n^\circ &= \nu_n + 2 \\ \mathsf{k}_n^\circ &= \mathsf{k}_n + \frac{1}{\mathsf{w}_{\mathsf{S},n+1}} \\ \mathbf{V}_n^\circ &= \mathbf{V}_n \\ \mathbf{B}_n^\circ &= \mathbf{B}_n + \frac{1}{\mathsf{w}_{\mathsf{S},n+1}}\begin{bmatrix}\mathsf{m}_{\mathsf{S},n+1} \\ \boldsymbol{\psi}_{n+1}\end{bmatrix}\begin{bmatrix}\mathsf{m}_{\mathsf{S},n+1} & \boldsymbol{\psi}_{n+1}^{\mathsf{T}}\end{bmatrix}\end{aligned} \tag{26}$$

Where $\mathsf{m}_{\mathsf{S},n+1}$ *and* $\mathsf{w}_{\mathsf{S},n+1}$ *are the first and second moments of (20) respectively.*

Proof. See Appendix D. □

Adopting Property 1 of the extended $\mathcal{N}$iG model, the resulting computational flow is defined in (6) for the data-step update with the replacement of $\mathbf{V}_{n-1}$ and ν_{n-1} with the optimal posterior $\mathbf{V}_{n-1}^\circ$ and ν_{n-1}°. Equation (26) describes the source knowledge update. (25) is adopted for deducing moments.

VI. Experimentation

To illustrate the performance of the hitherto described FPD-optimal BTL algorithms, we design an experimental setting with a target AR(8) inference machine and a single source making available a distribution on the targets next datum at every step. We compare the following methods: No Transfer (NT); FPD-Optimal BTL (FPD-BTL); Robust FPD-Optimal BTL (RFPD-BTL). $\mathsf{n} = 30$ local target data are produced using the AR(8) parameters below with a driving Gaussian innovation with variance $\mathsf{r} = 5$.

$$\mathbf{a} = [0.175, \quad -0.126, \quad 0.067, \quad -0.035, \quad ...0.014, \quad -0.007, \quad 0.003, \quad -0.0001]$$

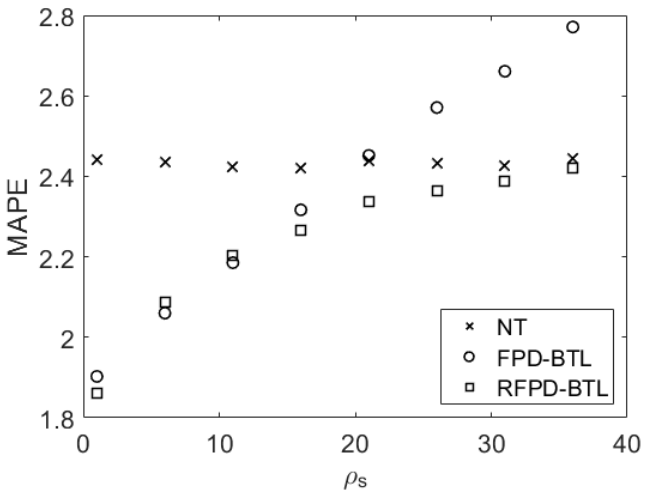

Fig. 2. Mean Absolute Prediction Error of the target AR(8) inference machine over the last 10 data with varying source prediction accuracy ρ_S. Methods compared: No Transfer (NT), FPD optimal BTL (FPD-BTL) and Robust FPD optimal BTL (RFPD-BTL). 30 target data considered with variance $r = 5$.

At each step the target processes its local datum, then accepts a distribution $F_S = \mathcal{N}(\mu_{z_{n+1}}, \rho_S)$ from the source (simulating the one-step-ahead data predictor and conditional one-step-ahead mean data predictor as appropriate). $\mu_{z_{n+1}}$ is sampled at each step from $\mathcal{N}(\psi_{n+1}^T \mathbf{a}, \rho_S)$. ρ_S is the operating condition which determines the source predictor variance and a substitute for the second parameter of the transferred distribution (20) and the second moment of (7). ρ_S dictates the quality of the source knowledge being transferred.

For each value of ρ_S, we plot in figure 1 the Mean Absolute Prediction Error (MAPE) of the target's one-step-ahead data predictor on the last 10 data points, over 5000 Monte-Carlo trials. The MAPE for a single experiment is defined in equation (27) where $\hat{z}_{i+1}$ is the predicted value of the next datum by the model at step i and z_{i+1} is the true realisation.

$$\text{MAPE} = \frac{1}{10} \sum_{i=n-10}^{n} |\hat{z}_{i+1} - z_{i+1}| \tag{27}$$

VII. Discussion

The isolated NT AR task acts a baseline for accessing the performance of the BTL algorithms. Unsurprisingly, its MAPE is invariant with ρ_S. In the low source data predictor variance regime ($\rho_S < 20$), both FPD algorithms perform well and provide positive knowledge transfer. For ($\rho_S < r$), RFPD-BTL outperforms FPD-BTL due to its ability to accept source knowledge weighted above locally sourced data. For $\rho_S \approx r$, FPD-BTL marginally outperforms RFPD-BTL. This is because the two algorithms are effectively equivalent when $\rho_S = r$ (with respect to expected data prediction) except RBTL-FPD has the added overhead of estimating r_{CE} (directly proportional to source knowledge acceptance weight). In the stressful regime of $\rho_S > 20$, RFPD-BTL achieves good levels of robustness while FPD-BTL does not.

The RFPD-BTL algorithm accepts the expected value of the source mean data predictor as a locally observed datum weighted by $\frac{r_{CE}}{w_S}$. When $r_{CE} \ll w_S$, we assign little weight to the source knowledge. When $r_{CE} > w_S$, we accept the source knowledge with a weight greater than 1 (i.e. weighting it higher than a locally observed datum). A key disadvantage to RFPD-BTL is the believed analytical intractability of the extended $\mathcal{N}$iG which omits the possibility for a full analytical Bayesian description of the source conditioned posterior.

The BTL technique presented in this paper has potentially widespread application in contexts of distributed and/or multi-output filtering and prediction, assuming that individual observation channels are amenable to all-pole modelling [16]. This includes microphone arrays and other networks of sensors in (low-noise) acoustic and vibrational fields. Recall that our framework involves knowledge transfer between local AR tasks where the interaction model does not need to be specified. It can be replace and robustify complete-modelling approaches to multi-output AR processing, such as the conventional vector-AR [17] process. More work will be required to assess the performance of our BTL framework for transfer from multiple AR sources. It will be interesting to test whether BTL can avoid sensitivity to (spatial) model mismatch.

VIII. Conclusion

This paper has presented an investigation of Bayesian Transfer Learning (BTL) in an incompletely modelled scenario. BTL is typically approached via the assumption or design of a complete model, allowing Bayesian conditioning between learners. We focus on the incomplete model scenario where this complete model is not designed or assumed. Two algorithms are derived for the AR BTL task via FPD principles, for the transfer of source data predictive knowledge. FPD-BTL adopts a standard FPD setting but is found to be effectively invariant to source uncertainty. We propose the RFPD-BTL instantiation which results in a robust algorithm and is validated in Monte-Carlo simulations. In future work, a formalisation of the RFPD-BTL algorithm will be investigated which does not require the adoption of a certainty equivalent, r_{CE}, in (25).

APPENDIX

A. Proof Proposition 1

$$\begin{aligned}
&\mathcal{D}(M \| M_I) = \\
&\quad \int F_S(z_{n+1} \mid \mathbf{z}_{S,n})\, M(\mathbf{a}, r \mid F_S, \mathbf{z}_n) \\
&\qquad \times \ln\left(\frac{F_S(z_{n+1} \mid \mathbf{z}_{S,n})\, M(\mathbf{a}, r \mid F_S, \mathbf{z}_n)}{F(z_{n+1} \mid \mathbf{a}, r)\, F(\mathbf{a}, r \mid \mathbf{z}_n)}\right) dz_{n+1} d\mathbf{a} dr \\
&= \int M(\mathbf{a}, r \mid F_S, \mathbf{z}_n) \\
&\qquad \times \ln\left(\frac{M(\mathbf{a}, r \mid F_S, \mathbf{z}_n)}{M^\circ(\mathbf{a}, r \mid F_S, \mathbf{z}_n)}\right) d\mathbf{a} dr - \mathcal{H}_{F_S} - \ln c_{M^\circ}
\end{aligned}$$

with differential entropy of F_S and normalising constant c_{M°

$$\mathcal{H}_{F_S} = -\int F_S(z_{n+1} \mid \mathbf{z}_{S,n}) \ln F_S(z_{n+1} \mid \mathbf{z}_{S,n})\, dz_{n+1}$$

$$c_{M^\circ} = \int F(\mathbf{a}, r \mid \mathbf{z}_n) \times \exp\left\{ \int \ln F(z_{n+1} \mid \mathbf{a}, r)\, F_S(z_{n+1} \mid \mathbf{z}_{S,n})\, dz_{n+1} \right\} d\mathbf{a} dr$$

B. Proof Proposition 2

$$M^\circ(\mathbf{a}, r F_S, \mathbf{z}_n) \propto \mathcal{N}\mathrm{iG}_{\mathbf{a},r}(\mathbf{V}_n, \nu_n) \exp\left[\int_{z_{n+1}} \ln \mathcal{N}_{z_{n+1}}(\psi_{n+1}^T \mathbf{a}, r) \mathrm{St}_{z_{n+1}}(\cdot) dz_{n+1} \right] \quad (28)$$

Considering the Boltzmann modulation term.

$$\begin{aligned} &\int_{z_{n+1}} \Big(\ln \frac{1}{\sqrt{2\pi r}} - \frac{1}{2r}(z_{n+1} - \psi_{n+1}^T \mathbf{a})^2 \Big) \mathrm{St}_{z_{n+1}}(\cdot) dz_{n+1} \\ &= \ln \frac{1}{\sqrt{2\pi r}} - \frac{1}{2r}\Big(\psi_{n+1}^T \mathbf{a}\mathbf{a}^T \psi_{n+1} - 2\psi_{n+1}^T \mathbf{a} m_{S,n+1} \\ &\quad + \int_{z_{n+1}} z_{n+1}^2 \mathrm{St}_{z_{n+1}}(\cdot) dz_{n+1} \Big) \\ &= \ln \frac{1}{\sqrt{2\pi r}} - \frac{1}{2r}\Big((\psi_{n+1}^T \mathbf{a} - m_{S,n+1})^2 + w_{S,n+1} \Big) \end{aligned} \quad (29)$$

Adopting (28), (29) and (4) where $m_{S,n+1}$ and $w_{S,n+1}$ are moments of $\mathrm{St}_{z_{n+1}}$ and can be generated via (7).

$$M^\circ(\mathbf{a}, r \mid F_S, \mathbf{z}_n) \propto r^{-\frac{\nu_n+1}{2}} \exp\left(-\frac{1}{2r} \begin{bmatrix} -1 & \mathbf{a}^T \end{bmatrix} \mathbf{V}_n \begin{bmatrix} -1 \\ \mathbf{a} \end{bmatrix} - \frac{1}{2r}\Big((\psi_{n+1}^T \mathbf{a} - m_{S,n})^2 + w_{S,n+1} \Big) \right)$$

C. Proof of Proposition 3

$$\begin{aligned} \mathcal{D}(M \| M_I) &= \int F(z_{n+1} \mid \mathbf{a}, r)\, M(\mathbf{a}, r \mid F_S, \mathbf{z}_n) \times \ln\left(\frac{F(z_{n+1} \mid \mathbf{a}, r)\, M(\mathbf{a}, r \mid F_S, \mathbf{z}_n)}{F_S(z_{n+1} \mid \mathbf{z}_{S,n})\, F(\mathbf{a}, r \mid \mathbf{z}_n)} \right) dz_{n+1} d\mathbf{a} dr \\ &= \int M(\mathbf{a}, r \mid F_S, \mathbf{z}_n) \left[\ln\left(\frac{M(\mathbf{a}, r \mid F_S, \mathbf{z}_n)}{F(\mathbf{a}, r \mid \mathbf{z}_n)} \right) + \int_{z_{n+1}} F(z_{n+1} \mid \mathbf{a}, r) \ln \frac{F(z_{n+1} \mid \mathbf{a}, r)}{F_S(z_{n+1} \mid \mathbf{z}_{S,n})} dz_{n+1} \right] d\mathbf{a} dr \\ &= \int M(\mathbf{a}, r \mid F_S, \mathbf{z}_n) \ln\left(\frac{M(\mathbf{a}, r \mid F_S, \mathbf{z}_n)}{F(\mathbf{a}, r \mid \mathbf{z}_n) \exp\{-\mathcal{D}(F \| F_S)\}} \right) d\mathbf{a} dr \end{aligned}$$

D. Proof of Proposition 4

$$M^\circ(\mathbf{a}, r \mid F_S, \mathbf{z}_n) \propto F(\mathbf{a}, r \mid \mathbf{z}_n) \exp\left[-\mathcal{D}\big(F(z_{n+1} \mid \mathbf{a}, r) \| F_S(z_{n+1}) \big) \right] \quad (30)$$

Considering the exponential modulation term only.

$$\begin{aligned} &\exp\left[-\frac{1}{2w_{S,n+1}} \Big(r + (\psi_{n+1}^T \mathbf{a} - m_{S,n+1})^2 \Big) - \ln(r) \right] \\ &= r^{-1} \exp\left[-\frac{1}{2}\Big(r \frac{1}{w_{S,n+1}} + \begin{bmatrix} -1 & \mathbf{a}^T \end{bmatrix} \frac{1}{w_{S,n+1}} \begin{bmatrix} m_{S,n+1} \\ \psi_{n+1} \end{bmatrix} \begin{bmatrix} m_{S,n+1} & \psi_{n+1}^T \end{bmatrix} \begin{bmatrix} -1 \\ \mathbf{a} \end{bmatrix} \Big) \right] \end{aligned}$$

Adopting (24) and (30).

$$\begin{aligned} M^\circ(\mathbf{a}, r \mid F_S, \mathbf{z}_n) \propto\ & r^{-\frac{\nu+2}{2}} \exp\left[-\frac{1}{2r} \begin{bmatrix} -1 & \mathbf{a}^T \end{bmatrix} \mathbf{V}_n \begin{bmatrix} -1 \\ \mathbf{a} \end{bmatrix} - \frac{1}{2} \begin{bmatrix} -1 & \mathbf{a}^T \end{bmatrix} \mathbf{B}_n \begin{bmatrix} -1 \\ \mathbf{a} \end{bmatrix} \right. \\ & -\frac{1}{2} \begin{bmatrix} -1 & \mathbf{a}^T \end{bmatrix} \frac{1}{w_{S,n+1}} \begin{bmatrix} m_{S,n+1} \\ \psi_{n+1} \end{bmatrix} \begin{bmatrix} m_{S,n+1} & \psi_{n+1}^T \end{bmatrix} \begin{bmatrix} -1 \\ \mathbf{a} \end{bmatrix} \\ & \left. -\frac{1}{w_{S,n+1}} r \right] \end{aligned}$$

REFERENCES

[1] Torrey, L., & Shavlik, J. (2010). Transfer learning. In Handbook of research on machine learning applications and trends: algorithms, methods, and techniques (pp. 242-264). IGI global.

[2] Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), 1345-1359.

[3] Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. Journal of Big data, 3(1), 9.

[4] Marx, Z., Rosenstein, M. T., Kaelbling, L. P., & Dietterich, T. G. (2005). Transfer learning with an ensemble of background tasks. Inductive Transfer, 10.

[5] Karbalayghareh, A., Qian, X., & Dougherty, E. R. (2018). Optimal Bayesian transfer learning. IEEE Transactions on Signal Processing, 66(14), 3724-3739.

[6] Kárný, M., & Guy, T. V. (2006). Fully probabilistic control design. Systems & Control Letters, 55(4), 259-265.

[7] Kárný, M. (1996). Towards fully probabilistic control design. Automatica, 32(12), 1719-1722.

[8] Kárný, M., & Kroupa, T. (2012). Axiomatisation of fully probabilistic design. Information Sciences, 186(1), 105-113.

[9] Papež, M., & Quinn, A. (2019, December). Bayesian transfer learning between Gaussian process regression tasks. In 2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) (pp. 1-6). IEEE.

[10] Foley, C., & Quinn, A. (2017). Fully probabilistic design for knowledge transfer in a pair of Kalman filters. IEEE Signal Processing Letters, 25(4), 487-490.

[11] Papež, M., & Quinn, A. (2020). Bayesian transfer learning between Student-t filters. Signal Processing, 107624.

[12] Ephraim, Y., & Roberts, W. J. (2005). Revisiting autoregressive hidden Markov modeling of speech signals. IEEE Signal processing letters, 12(2), 166-169.

[13] Pacurar, M. (2008). Autoregressive conditional duration models in finance: a survey of the theoretical and empirical literature. Journal of economic surveys, 22(4), 711-751.

[14] Bernardo, J. M., & Smith, A. F. (2009). Bayesian theory (Vol. 405). John Wiley & Sons.

[15] Kay, S. M. (1993). Fundamentals of statistical signal processing. Prentice Hall PTR.

[16] Tiao, G. C., & Box, G. E. (1981). Modeling multiple time series with applications. journal of the American Statistical Association, 76(376), 802-816.

[17] Corander, J., & Villani, M. (2006). A Bayesian approach to modelling graphical vector autoregressions. Journal of Time Series Analysis, 27(1), 141-156.

[18] Šmídl, V., & Quinn, A. (2006). The variational Bayes method in signal processing. Springer Science & Business Media.