2019 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, OCT. 13–16, 2019, PITTSBURGH, PA, USA

ROBUST BAYESIAN TRANSFER LEARNING BETWEEN KALMAN FILTERS

Milan Papež [a] *and Anthony Quinn* [a,b]

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

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

ABSTRACT

Bayesian transfer learning typically requires complete specification of the stochastic dependence between source and target domains. Fully probabilistic design-based Bayesian transfer learning—which transfers source knowledge in the form of a probability distribution—obviates these restrictive assumptions. However, this approach has suffered from negative transfer when the source knowledge is imprecise. We propose a scale variable relaxation to transfer all source moments successfully, achieving robust transfer (i.e. rejection of imprecise source knowledge). A recursive algorithm is recovered via local variational Bayes approximation. The solution offers positive transfer of precise source knowledge, while rejecting it when imprecise. Experiments show that the technique is competitive with or equivalent to alternative methods.

Index Terms— Bayesian transfer learning, robust knowledge transfer, scalar relaxation, fully probabilistic design, Kalman filtering

1. INTRODUCTION

The aim of transfer learning is to utilize knowledge learned in a source domain in order to improve learning performance in a related target domain [1]. Transfer learning has mostly been used to enhance traditional machine learning [2, 3] and reinforcement [4] learning algorithms, and has been widely deployed in various statistical signal processing applications, including genomics [5], cross-language speech recognition [6], fault diagnosis [7], video analysis [8], etc. This paper focuses on Bayesian transfer learning (BTL) and develops a transfer learning strategy suitable for networks of Bayesian filters.

BTL has until now relied on the traditional Bayesian paradigm. It undertakes a target learning task by using a prior distribution that is additionally conditioned on knowledge provided by a source learning task [9]. The central assumption is that knowledge is expressed via a probabilistic model conditioned on raw source data, rather than the raw source data being available themselves. The required conditioning of the target task on this source knowledge conventionally requires specification of the stochastic dependence between quantities in the source and target domains. We call this setting *complete modelling*. This paper assumes the more realistic scenario—in which explicit dependence of the target on the stochastic source knowledge is *not* available—obviating conventional Bayesian conditioning as the learning mechanism. We refer to this setting as *incomplete modelling*. In this case, fully probabilistic design (FPD) [10, 11]—an extension of the minimum cross-entropy principle [12]—provides an axiomatically justified approach to elicit the target model conditioned on the source distribution [13]. The freedom to optimize the knowledge conditioned mechanism yields more flexible and robust framework than that provided by completely modelled BTL.

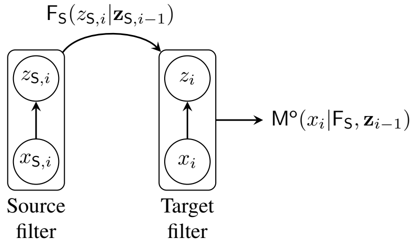
$\mathsf{F}_\mathsf{S}(z_{\mathsf{S},i}|\mathbf{z}_{\mathsf{S},i-1})$
$z_{\mathsf{S},i}$
$x_{\mathsf{S},i}$
z_i
x_i
$\mathsf{M}^\circ(x_i|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1})$
Source filter
Target filter

Fig. 1: Source and target Bayesian filters operating on source $(z_{\mathsf{S},i}, x_{\mathsf{S},i})$ and target (z_i, x_i) variables, respectively. The source filter transfers its observation predictor, F_S. The target filter improves its performance by conditioning on F_S in an FPD-optimal sense, yielding M°.

It has been found that FPD-optimal static BTL (Fig. 1) suffers from *negative* transfer—i.e., the inability to resist imprecise source knowledge—in the special case where the Bayesian filters implement the Kalman filter. This problem was addressed by informal adaptations in [14]. The dynamic transfer learning approach (i.e., transferring source knowledge over all time steps) was investigated in an effort to resolve this problem in [15], but again required informal adaptations. The current paper provides an important progression by proposing a scale-variable relaxation which avoids any such informalities. It achieves this by successfully transferring all source moments. The experiments show that this leads to *robust* transfer, i.e., successful rejection of imprecise source knowledge, and, moreover, offers fully knowledge-driven transfer learning capabilities.

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

978-1-7281-0824-7/19/$31.00 ©2019 IEEE

2. STATIC FPD TRANSFER BETWEEN BAYESIAN FILTERS

We consider a state-space model in the form

$$x_i \sim \mathsf{F}(x_i|x_{i-1}), \tag{1a}$$

$$z_i \sim \mathsf{F}(z_i|x_i), \tag{1b}$$

where $x_i \in \boldsymbol{x} \subseteq \mathbb{R}^{n_x}$ is the latent (hidden) state variable, $z_i \in \boldsymbol{z} \subseteq \mathbb{R}^{n_z}$ is the observation variable, and $i = 1, \ldots, n$ is the discrete-time index. The model (1) is characterized by the state transition and observation probabilities densities (1a) and (1b), respectively. The initial state variable is distributed according to $x_1 \sim \mathsf{F}(\cdot)$.

The basic object for developing state inference algorithms for the state-space model (1) is the joint predictive model,

$$\mathsf{F}(z_i, x_i|\mathbf{z}_{i-1}) = \mathsf{F}(z_i|x_i)\mathsf{F}(x_i|\mathbf{z}_{i-1}), \tag{2}$$

where $\mathsf{F}(x_i|\mathbf{z}_{i-1})$ is the state pre-prior density and $\mathbf{z}_{i-1} = (z_1, \ldots, z_{i-1})$ is the observation record. The importance of (2) is that it uniquely implies the conditional and marginal densities required in sequential Bayesian filtering [16].

The goal of this paper is to design an algorithm for transferring knowledge from a source to a target Bayesian filter. We assume that there is no explicit dependence assumption between the variables of the source and target domains. We also consider that the target filter has access only to the (probabilistic) observation predictor of the source filter, F_S, but not to any realized variable in the source domain, see Fig. 1. The inference objective is to extend the basic setting (2) of the target filter to condition additionally on the source density F_S,

$$\mathsf{M}(z_i, x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}). \tag{3}$$

The functional form of (3) is now unknown, and so we have to find a way to condition (2) on F_S. Throughout this paper, we use F to denote fixed-form (specified) densities, and M and Q to denote variational (unspecified) densities.

The transfer of the source observation predictor F_S is accomplished by restricting the functional form of the unknown joint model (3) via conditional independence:

$$\mathsf{M}(z_i, x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) \equiv \mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1})\mathsf{M}(x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}). \tag{4}$$

More specifically, we constrain the F_S-conditioned model of the target observations,

$$\mathsf{M}(z_i|x_i, \mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) \equiv \mathsf{F}_\mathsf{S}(z_{\mathsf{S},i}|\mathbf{z}_{\mathsf{S},i-1})\big|_{z_{\mathsf{S},i}=z_i},$$

to be the observation predictor of the source filter evaluated at $z_i \in \boldsymbol{z}$. Fixing $\mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1})$, and admitting $\mathsf{M}(x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1})$ as the only variational factor in (4), defines the knowledge-constrained set of admissible models

$$\mathsf{M} \in \mathbf{M} \equiv \{\text{models (4) with } \mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1}) \text{ fixed and } \mathsf{M}(x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) \text{ variational}\}. \tag{5}$$

The (prior) joint model (2) is valid in non-transfer contexts, and therefore, functions as the reference (ideal) design in the absence of source knowledge

$$\mathsf{M}_\mathsf{I}(z_i, x_i|\mathbf{z}_{i-1}) \equiv \mathsf{F}(z_i, x_i|\mathbf{z}_{i-1}). \tag{6}$$

FPD prescribes the optimal mechanism to condition on a probability density function. FPD achieves this by finding an optimal model, M^o, that incorporates the set-based knowledge constraints, $\mathsf{M} \in \mathbf{M}$ (5), and preferences about M expressed by an ideal model M_I (6). The FPD-optimal design, $\mathsf{M}^\mathsf{o} \in \mathbf{M}$, is the density that is closest to M_I in the minimum Kullback-Leibler divergence (KLD) [17] sense,

$$\mathsf{M}^\mathsf{o}(z_i, x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) \equiv \underset{\mathsf{M}\in\mathbf{M}}{\operatorname{argmin}}\, \mathcal{D}(\mathsf{M}||\mathsf{M}_\mathsf{I}), \tag{7}$$

where the KLD from M to M_I is

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

with E_M being the expected value under M.

Proposition 1. *The unknown model belongs to the knowledge constrained set,* $\mathsf{M} \in \mathbf{M}$ *(5), and the ideal model* M_I *is (6). Then, the FPD-optimal model—the solution of (7)—is*

$$\mathsf{M}^\mathsf{o}(z_i, x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) = \mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1})\mathsf{M}^\mathsf{o}(x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}), \tag{8}$$

where

$$\mathsf{M}^\mathsf{o}(x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) \propto \mathsf{F}(x_i|\mathbf{z}_{i-1}) \times \exp\left\{\int \log \mathsf{F}(z_i|x_i)\mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1})dz_i\right\}. \tag{9}$$

Proof. See Appendix 8.1. □

Proposition 1 shows that the source observation predictor is processed via the update from the pre-prior $\mathsf{F}(x_i|\mathbf{z}_{i-1})$ (2) to the FPD-optimal prior $\mathsf{M}^\mathsf{o}(x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1})$. This prior replaces the pre-prior in the Bayesian filtering equations, yielding a sequential Bayesian updating structure from the pre-prior to the prior and then to the posterior. Therefore, Proposition 1 introduces an additional step between the traditional *time* and *data steps* of the standard Bayesian filtering recursions, which we refer to as the *transfer learning step*.

3. STATIC FPD TRANSFER BETWEEN KALMAN FILTERS

We present the proposed transfer learning framework with (1) instantiated as the linear Gaussian state-space model:

$$\mathsf{F}(x_i|x_{i-1}) \equiv \mathcal{N}(x_i; Ax_{i-1}, Q), \tag{10a}$$

$$\mathsf{F}(z_i|\lambda, x_i) \equiv \mathcal{N}(z_i; Cx_i, \lambda R), \tag{10b}$$

where $\mathcal{N}(\cdot;\mu,\Sigma)$ denotes the Gaussian density with mean vector μ and covariance matrix Σ. The dynamic and stochastic properties of this model are defined by the state transition A, observation C, state noise Q, and observation noise R, matrices. To achieve robust transfer—as seen later in the paper—we augment (10b) with the scale variable λ, currently assumed known.

The source filtering task is assumed to be a Kalman filter (KF), which is the consistent solution of the Bayesian filtering equations for (10) (λ known), see Lemma 2 in Appendix 8.2. The target filter then processes the transferred source observation predictor via Proposition 1, which we now specify for the KF pair.

Lemma 1. *The observation model is specified by (10b), the state pre-prior is (22), and the source observation predictor is* $\mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1}) \equiv \mathcal{N}(z_i; z_{\mathsf{S},i|i-1}, R_{\mathsf{S},i|i-1})$*. Then, the FPD-optimal prior (9) becomes*

$$\begin{aligned}\mathsf{M}^\circ(x_i|\lambda,\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) \propto \mathcal{N}\big(x_i;\widehat{x}_{i|i-1}(\lambda),\widehat{P}_{i|i-1}(\lambda)\big) \qquad (11)\\ \exp\big\{-\tfrac{1}{2}\operatorname{tr}\big(\lambda^{-1}R^{-1}R_{\mathsf{S},i|i-1}\big)\big\},\end{aligned}$$

where the λ-conditional shaping parameters are

$$\begin{aligned}\widehat{x}_{i|i-1}(\lambda) &= x_{i|i-1} + L(z_{\mathsf{S},i|i-1} - z_{i|i-1}), \qquad (12)\\ \widehat{P}_{i|i-1}(\lambda) &= P_{i|i-1} - LR_{i|i-1}(\lambda)L^\top.\end{aligned}$$

Here, $L \equiv P_{i|i-1}C^\top R^{-1}_{i|i-1}(\lambda)$, and $\{z_{i|i-1}, R_{i|i-1}(\lambda)\}$ is given in Lemma 2.

Proof. The result follows from basic calculus with Gaussian densities, see, e.g., [14]. □

Note that for any known λ, say, $\lambda = 1$, Lemma 1 recovers the transfer learning step of [14]. In this case, the modulating exponential structure in (11)—which contains the second moment of the source observation predictor, $R_{\mathsf{S},i|i-1}$—is absorbed into the normalizing constant. Consequently, the algorithm successfully transfers only the first moment of the source observation predictor via (12) but fails to transfer the second one. This moment loss causes the target filter to perform non-robustly, as we will see in Section 5, i.e., the algorithm is then unable to reject imprecise source knowledge.

4. RELAXATION OF THE SCALE VARIABLE λ

To ensure robust knowledge transfer in the Gaussian setting of Section 3, we assume that the scale variable λ in (10b) is an unknown, and we assign a prior to it (relaxation). The λ-augmented FPD-optimal version of (9) is

$$\mathsf{M}^\circ(x_i,\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) \equiv \mathsf{M}^\circ(x_i|\lambda,\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1})\mathsf{F}(\lambda), \qquad (13)$$

where

$$\begin{aligned}\mathsf{M}^\circ(x_i|\lambda,\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) &\propto \mathsf{F}(x_i|\mathbf{z}_{i-1})\\ &\times \exp\left\{\int \log \mathsf{F}(z_i|\lambda,x_i)\mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1})dz_i\right\}.\end{aligned}$$

λ is a positive scale variable, and therefore, we adopt the inverse Gamma prior density for it,

$$\mathsf{F}(\lambda) \equiv i\mathcal{G}\big(\lambda;\tfrac{\alpha}{2},\tfrac{\beta}{2}\big), \qquad (14)$$

where $\frac{\alpha}{2} > 0$ and $\frac{\beta}{2} > 0$ are the shape and scale parameters. The FPD-optimal state predictor after transfer is then the infinite mixture,

$$\mathsf{M}^\circ(x_i|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) = \int \mathsf{M}^\circ(x_i,\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1})d\lambda. \qquad (15)$$

4.1. Local variational Bayesian approximation

The marginal density (15) is an infinite scale mixture under the present setting and does not admit a closed-form recursive updating formulae. Therefore, we use the coordinate ascent mean-filed variational inference [18] to find a local approximation of (13) at each step i. That is, we choose the variational density from the mean-field family,

$$\mathsf{Q}(x_i,\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) \equiv \mathsf{Q}(x_i|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1})\mathsf{Q}(\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}).$$

This allows us to find the KLD-optimal minimizer, which approximates the target density (13), as

$$\begin{aligned}\mathsf{M}^\circ(x_i,\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) &\approx \bar{\mathsf{Q}}(x_i,\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) \qquad (16)\\ &= \bar{\mathsf{Q}}(x_i|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1})\bar{\mathsf{Q}}(\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}),\end{aligned}$$

where

$$\begin{aligned}\bar{\mathsf{Q}}(x_i|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) &\propto \exp\big\{\mathsf{E}_\lambda[\log \mathsf{M}^\circ(x_i,\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1})]\big\},\\ \bar{\mathsf{Q}}(\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) &\propto \exp\big\{\mathsf{E}_x[\log \mathsf{M}^\circ(x_i,\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1})]\big\},\end{aligned}$$

with E_λ and E_x denoting the expected values under the KLD-optimal factors of λ and x_i, respectively. The approximate FPD-optimal state prior, $\bar{\mathsf{Q}}(x_i|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1})$, recovers a tractable recursive update.

Proposition 2. *The λ-augmented FPD-optimal prior (13) is given by the state pre-prior (22), observation model (10b), source observation predictor* $\mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1}) \equiv \mathcal{N}(z_i; z_{\mathsf{S},i|i-1}, R_{\mathsf{S},i|i-1})$*, and scale variable pre-prior (14). Then, the mean-field variational factors (16) are*

$$\begin{aligned}\bar{\mathsf{Q}}(\lambda|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) &= i\mathcal{G}\big(\lambda;\tfrac{\bar{\alpha}}{2},\tfrac{\bar{\beta}}{2}\big), \qquad (17a)\\ \bar{\mathsf{Q}}(x_i|\mathsf{F}_\mathsf{S},\mathbf{z}_{i-1}) &= \mathcal{N}(x_i;\bar{x}_{i|i-1},\bar{P}_{i|i-1}), \qquad (17b)\end{aligned}$$

where the shape and scale hyper-parameters of (17a) are

$$\begin{aligned}\bar{\alpha} &= \alpha + n_z,\\ \bar{\beta} &= \beta + \operatorname{tr}\big\{\big(R_{\mathsf{S},i|i-1} + \mathsf{E}_x\big[(z_{\mathsf{S},i|i-1} - Cx_i)\\ &\qquad \times (z_{\mathsf{S},i|i-1} - Cx_i)^\top\big]\big)R^{-1}\big\}, \qquad (18)\end{aligned}$$

and the shaping parameters of (17b) are

$$\begin{aligned}\bar{x}_{i|i-1} &= x_{i|i-1} + L(z_{\mathsf{S},i|i-1} - z_{i|i-1}),\\ \bar{P}_{i|i-1} &= (I_{n_x} - LC)P_{i|i-1}, \qquad (19)\end{aligned}$$

with

$$\begin{aligned} L &= P_{i|i-1}A^{\top}R_{i|i-1}^{-1} \qquad (20)\\ z_{i|i-1} &= Cx_{i|i-1},\\ R_{i|i-1} &= CP_{i|i-1}C^{\top} + \mathsf{E}_{\lambda}[\lambda^{-1}]^{-1}R. \end{aligned}$$

Here, $\mathrm{tr}(\cdot)$ *is the matrix trace operator.*

Proof. See Appendix 8.3. □

The shaping parameters of the variational factors (17) are coupled and cannot be updated under a direct closed-form solution. Therefore, we compute them with an iterative scheme at each step i. Evaluating the expected values in Proposition 2 and applying the FPD-optimal prior (17b) in the Bayesian recursions allows us to summarize the resulting procedure—at any time step i—in Algorithm 1.

Remark 1. *Let us focus on line 9 of Algorithm 1, i.e., on L which, in this transfer learning context, is analogous to the gain K of the conventional Kalman filter (Lemma 2). The relaxation via (14) engenders L, which involves the key quantity* $\frac{\bar{\beta}}{\bar{\alpha}}$*. This can be tuned via the hyper-parameters* α *and* β *in (14), which enter* $\bar{\alpha}$ *and* $\bar{\beta}$ *additively (lines 7 and 8, respectively). We consider three principal regimes:*

R1: $\alpha \to 0$ *and* $\beta \to \infty$*, so that* $\frac{\bar{\beta}}{\bar{\alpha}} \to \infty$*,* $L \to 0$*. This recovers the isolated Kalman filter, see lines 10 and 11.*

R2: $\alpha = \beta \to \infty$*, so that* $\frac{\bar{\beta}}{\bar{\alpha}} \to 1$*. In this case, L has the form reported in [14] (the static transfer (ST) filter).*

R3: $\alpha = \beta \to 0$*, so that* $\frac{\bar{\beta}}{\bar{\alpha}}$*, and therefore L, are influenced only by the transferred knowledge, without any influence from (14).*

5. EXPERIMENTS

This section compares the proposed approach with alternative strategies. In particular, we illustrate how different settings of the hyper-parameters in (14) influence the transfer learning properties of the proposed algorithm. The experiments consider a state-space model with a common state variable

$$\begin{aligned} x_i &= Ax_{i-1} + w_i,\\ z_i &= Cx_i + v_i,\\ z_{\mathsf{S},i} &= Cx_i + v_{\mathsf{S},i}, \qquad (21) \end{aligned}$$

where $w_i \sim \mathcal{N}(\cdot;\mathbf{0},Q)$, $v_i \sim \mathcal{N}(\cdot;\mathbf{0},R)$, and $v_{\mathsf{S},i} \sim \mathcal{N}(\cdot;\mathbf{0},R_{\mathsf{S}})$ are independent and identically distributed Gaussian noise variables associated with the common states, target observations, and source observations, respectively. We evaluate the state-estimation performance by computing the mean norm squared-error between the true state and its posterior estimate $\mathrm{MNSE} = \frac{1}{n}\sum_{i=1}^{n}||x_i - x_{i|i}||^2$, where $||\cdot||$ denotes the Euclidean norm and $n = 400$.

Algorithm 1: Static BTL filter with scale relaxation

```
   Input: $x_{i-1|i-1}$, $P_{i-1|i-1}$, $z_i$, $z_{\mathsf{S},i|i-1}$, $R_{\mathsf{S},i|i-1}$,
     $A$, $C$, $Q$, $R$, $\alpha$, $\beta$, $N$
1  Time step:
2  $x_{i|i-1} = Ax_{i-1|i-1}$
3  $P_{i|i-1} = AP_{i-1|i-1}A^{\top} + Q$
4  Transfer learning step:
5  Set $\bar{x}_{i|i-1}^{(0)} = x_{i|i-1}$ and $\bar{P}_{i|i-1}^{(0)} = P_{i|i-1}$.
6  for $k = 0, \dots, N-1$ do
7      $\bar{\alpha} = \alpha + n_z$
8      $\bar{\beta} = \beta + \mathrm{tr}\{[(z_{\mathsf{S},i|i-1} - C\bar{x}_{i|i-1}^{(k)})(z_{\mathsf{S},i|i-1} - C\bar{x}_{i|i-1}^{(k)})^{\top}$
                 $+ C\bar{P}_{i|i-1}^{(k)}C^{\top} + R_{\mathsf{S},i|i-1}]R^{-1}\}$
9      $L = P_{i|i-1}C^{\top}\left(CP_{i|i-1}C^{\top} + \frac{\bar{\beta}}{\bar{\alpha}}R\right)^{-1}$
10     $\bar{x}_{i|i-1}^{(k+1)} = x_{i|i-1} + L(z_{\mathsf{S},i|i-1} - Cx_{i|i-1})$
11     $\bar{P}_{i|i-1}^{(k+1)} = (I_{n_x} - LC)P_{i|i-1}$
12 Set $\bar{x}_{i|i-1} = \bar{x}_{i|i-1}^{(N)}$ and $\bar{P}_{i|i-1} = \bar{P}_{i|i-1}^{(N)}$.
13 Data step:
14 $K = \bar{P}_{i|i-1}C^{\top}\left(C\bar{P}_{i|i-1}C^{\top} + R\right)^{-1}$
15 $x_{i|i} = \bar{x}_{i|i-1} + K(z_i - C\bar{x}_{i|i-1})$
16 $P_{i|i} = (I_{n_x} - KC)\bar{P}_{i|i-1}$
   Output: $x_{i|i}$, $P_{i|i}$
```

We are concerned with a position-velocity model which specifies the matrices in (21) according to

$$A = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \otimes I_2, \quad G = \begin{bmatrix} \frac{\Delta^2}{2} \\ \Delta \end{bmatrix} \otimes I_2, \quad C = \begin{bmatrix} I_2 & O_2 \end{bmatrix},$$
$$Q = qGG^{\top}, \quad R = rI_2, \quad R_{\mathsf{S}} = r_{\mathsf{S}}I_2,$$

assuming that the state vector is $x_i = (p_{x,i}, p_{y,i}, v_{x,i}, v_{y,i})$ and that only the position can be observed. Here, we choose $\Delta = 0.1$, $q = 0.01$, and $r = 1$. The compared algorithms use $x_{1|0} = \mathbf{0}$, and $P_{1|0} = 10^{-5}I_4$. An increase in the number of iterations above $N = 5$ does not improve the estimation precision of the proposed method in the current example.

Fig. 2 demonstrates the influence of the source knowledge precision—as affected by the r_{S}-coefficient—on the MNSE. The NT filter does not depend on r_{S} and thus defines a reference MNSE level for comparing the remaining filters. If any filter has an MNSE that falls below or rises above this reference level, we say that the method delivers *positive* or *negative* transfer, respectively. If any filter saturates at this reference level when the source knowledge is imprecise, we say that the method achieves *robust* transfer (Section 1). The ST filter offers positive transfer for $r_{\mathsf{S}} < r$, but suffers negative transfer for $r_{\mathsf{S}} > r$. Therefore, as originally reported in [14], this method is not robust against imprecise source knowledge. For the purposes of reproducibility, note that we set the hyper-parameters of (14) at 10^{-10} and 10^{10}, respectively, to approximate limits at 0 and ∞, in Remark 1.

An interesting feature of the RST filter is that it possesses an extra degree of freedom, allowing the source knowledge

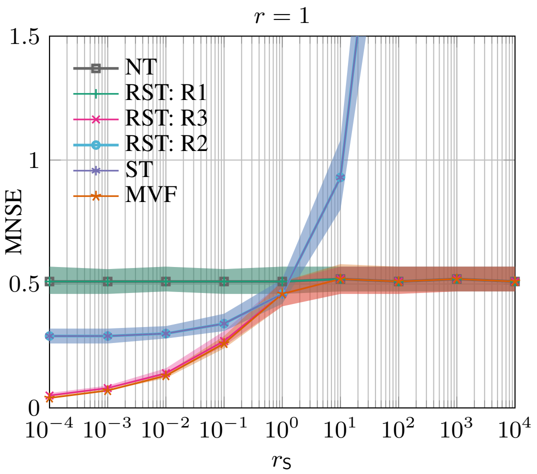

Fig. 2: The mean norm squared-error (MNSE) of the target filter versus the source observation variance, r_{S}. The results are averaged over 1000 independent simulation runs, with the solid line being the median and the shaded area delineating the interquartile range. The procedures that are compared are (i) the Kalman filter with *N*o *T*ransfer (*NT*) [20]; (ii) *S*tatic Bayesian *T*ransfer learning filter (*ST*) [14]; (iii) *S*tatic Bayesian *T*ransfer learning filter with scale *R*elaxation (*RST*) (Algorithm 1, this paper), under the three regimes in Remark 1; and (iv) *M*easurement *V*ector *F*usion (*MVF*) filter [19] as specified in the present context in [14].

to be accepted or rejected independently of its quality. This degree of freedom is set by the target filter via α and β (14). For example, under R1 source knowledge of any quality is rejected, recovering the isolated NT filter. Correspondingly, under R2 source knowledge of any quality is always accepted, recovering the ST filter. Interestingly, under R3, the source knowledge can be accepted or rejected, with a threshold set by $\frac{r_{\mathsf{S}}}{r}$. This threshold is at about $\frac{r_{\mathsf{S}}}{r} = 10$ for the current settings of α and β. Its positive and robust transfer properties are competitive with the MVF filter [19].

6. DISCUSSION

Recall that the transfer learning step in Lemma 1 does not incorporate the second moment of the source observation predictor, $R_{\mathsf{S},i|i-1}$, as long as λ is known. This corresponds to the ST filter—R2 in Remark 1—which fails to deliver robust transfer (Fig. 2). This was addressed in [14] by informally replacing R with $R_{\mathsf{S},i|i-1}$ in the expression for $R_{i|i-1}$ (23) used in Lemma 1. The key contribution of the current paper is to avoid this informal adjustment. The same idea can be adopted ins dynamic transfer, obviating the informal adaptation proposed in [15], thereby achieving robust dynamic transfer.

The experiments in Section 5 consider only those extreme settings of the hyper-parameters considered in Remark 1. Note, however, that there is a continuum of settings of α and β (14), allowing migration between these regimes. We will investigate these in future work.

As stated in Section 5, the transfer learning properties of the proposed RST filter can be controlled by the target modeller independently of the ratio $\frac{r}{r_{\mathsf{S}}}$ via their setting of the hyper-parameters, α and β, in (14). This represents a key advance over the previously developed ST filter since it allows the target modeller to switch on and off transfer learning from the source filter independently of the source knowledge quality. A question is how the modeller might set these hyper-parameters as a function—not only of the source and target knowledge—but also of the prior confidence of the target in respect of the source. This will be explored in future work.

Note that the λ-variable relaxation was neither applied in the source filter nor in the data step of the target filter. It is only needed in this work to ensure transfer of the second-order moments in the transfer learning step.

7. CONCLUSION

We have proposed the sequential FPD-optimal BTL method which resolves the difficulties in achieving robust transfer that were encountered with the previously developed algorithms [14, 15]. The central mechanism to deal with this issue is the successful transfer of higher-order moments of the source distribution via the scale variable augmentation of the FPD-optimal prior density. This optimal design framework does not require explicit dependence assumptions between source and target variables to be declared. Nevertheless, its performance is competitive with approaches that do rely on such assumptions, which are often hard to justify in practice. The proposed procedure offers operational adaptation of the transfer learning properties by tuning the hyper-parameters of the scale variable pre-prior. Future work will be focused on knowledge-driven tuning of these hyper-parameters.

8. APPENDIX

8.1. Proof of Proposition 1

Applying (4) and (6) in (7) leads to

$$\begin{aligned}\mathcal{D}(\mathsf{M}||\mathsf{M}_{\mathsf{I}}) &= \int \mathsf{F}_{\mathsf{S}}(z_i|\mathbf{z}_{\mathsf{S},i-1})\mathsf{M}(x_i|\mathsf{F}_{\mathsf{S}},\mathbf{z}_{i-1}) \\ &\quad \times \log\left(\frac{\mathsf{F}_{\mathsf{S}}(z_i|\mathbf{z}_{\mathsf{S},i-1})\mathsf{M}(x_i|\mathsf{F}_{\mathsf{S}},\mathbf{z}_{i-1})}{\mathsf{F}(z_i|x_i)\mathsf{F}(x_i|\mathbf{z}_{i-1})}\right) dz_i dx_i \\ &= \int \mathsf{M}(x_i|\mathsf{F}_{\mathsf{S}},\mathbf{z}_{i-1}) \\ &\quad \times \log\left(\frac{\mathsf{M}(x_i|\mathsf{F}_{\mathsf{S}},\mathbf{z}_{i-1})}{\mathsf{M}^{\mathsf{o}}(x_i|\mathsf{F}_{\mathsf{S}},\mathbf{z}_{i-1})}\right) dx_i - \mathcal{H}_{\mathsf{F}_{\mathsf{S}}} - \log c_{\mathsf{M}^{\mathsf{o}}},\end{aligned}$$

where the differential entropy of F_{S} is

$$\mathcal{H}_{\mathsf{F}_{\mathsf{S}}} = -\int \mathsf{F}_{\mathsf{S}}(z_i|\mathbf{z}_{\mathsf{S},i-1}) \log \mathsf{F}_{\mathsf{S}}(z_i|\mathbf{z}_{\mathsf{S},i-1}) dz_i,$$

and the normalizing constant is

$$c_{\mathsf{M}^{\circ}} = \int \mathsf{F}(x_i|\mathbf{z}_{i-1}) \\ \times \exp\left\{\int \log \mathsf{F}(z_i|x_i)\mathsf{F}_{\mathsf{S}}(z_i|\mathbf{z}_{\mathsf{S},i-1})dz_i\right\} dx_i. \qquad \square$$

8.2. The Kalman filter

Lemma 2. *The state space model is specified by (10) and the state pre-prior is* $\mathsf{F}(x_1) \equiv \mathcal{N}(x_1; x_{1|0}, P_{1|0})$. *Then, the joint model (2) yields the conditional and marginal densities of standard Bayesian filtering in the form*

$$\begin{aligned} \mathsf{F}(x_i|\mathbf{z}_{i-1}) &= \mathcal{N}(x_i; x_{i|i-1}, P_{i|i-1}), \\ \mathsf{F}(z_i|\mathbf{z}_{i-1}) &= \mathcal{N}(z_i; z_{i|i-1}, R_{i|i-1}), \\ \mathsf{F}(x_i|\mathbf{z}_i) &= \mathcal{N}(x_i; x_{i|i}, P_{i|i}), \end{aligned} \tag{22}$$

which are exactly computed under the following recursions:

$$\begin{aligned} x_{i|i-1} &= Ax_{i-1|i-1}, \\ P_{i|i-1} &= AP_{i-1|i-1}A^{\top} + Q, \\ z_{i|i-1} &= Cx_{i|i-1}, \\ R_{i|i-1} &= CP_{i|i-1}C^{\top} + \lambda R, \\ x_{i|i} &= x_{i|i-1} + K(z_i - z_{i|i-1}), \\ P_{i|i} &= P_{i|i-1} - KR_{i|i-1}K^{\top}, \end{aligned} \tag{23}$$

and $K \equiv P_{i|i-1}C^{\top}R^{-1}_{i|i-1}$, *where* $^{\top}$ *is the matrix transpose.*

Proof. See, e.g., [16]. □

8.3. Proof of Proposition 2

To find the variational factors (16), we first need to express the logarithm of the joint density (13),

$$\begin{aligned} &\log \mathsf{M}^{\circ}(x_i, \lambda|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}) = \\ &\quad - \tfrac{n_z}{2}\log\lambda - \tfrac{1}{2}(z_{\mathsf{S},i|i-1} - Cx_i)^{\top}\lambda^{-1}R^{-1}(z_{\mathsf{S},i|i-1} - Cx_i) \\ &\quad - \tfrac{1}{2\lambda}\operatorname{tr}(R_{\mathsf{S},i|i-1}R^{-1}) - \tfrac{\alpha+2}{2}\log\lambda - \tfrac{\beta}{2\lambda} \\ &\quad - \tfrac{1}{2}(x_i - x_{i|i-1})^{\top}P^{-1}_{i|i-1}(x_i - x_{i|i-1}) + c, \end{aligned} \tag{24}$$

where c is a constant.

Taking the expected value of (24) under $\bar{\mathsf{Q}}(x_i|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1})$ yields

$$\begin{aligned} &\log \bar{\mathsf{Q}}(\lambda|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}) = \\ &\qquad - \tfrac{\bar{\alpha}+2}{2}\log\lambda - \tfrac{\bar{\beta}}{2\lambda} + c_{\lambda} = \log i\mathcal{G}\big(\lambda; \tfrac{\bar{\alpha}}{2}, \tfrac{\bar{\beta}}{2}\big) + c_{\lambda}, \end{aligned}$$

where the shaping parameters are given by (18) and c_{λ} is a λ-independent constant.

Evaluating the expected value of (24) under $\bar{\mathsf{Q}}(\lambda|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1})$ leads to

$$\begin{aligned} \log \bar{\mathsf{Q}}(x_i|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}) &= \log\mathcal{N}(z_{\mathsf{S},i|i-1}; Cx_i, \mathsf{E}[\lambda^{-1}]^{-1}R) \\ &\qquad + \log\mathcal{N}(x_i; x_{i|i-1}, P_{i|i-1}) + a_x \\ &= \log\mathcal{N}(x_i; \bar{x}_{i|i-1}, \bar{P}_{i|i-1}) + b_x, \end{aligned} \tag{25}$$

with the shaping parameters given in (19) and a_x and b_x being x_i-independent constants. □

9. REFERENCES

[1] L. Torrey and J. Shavlik, “Transfer learning,” in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 242–264. IGI Global, 2010.

[2] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[3] K. Weiss, T. M. Khoshgoftaar, and D. D. Wang, “A survey of transfer learning,” *Journal of Big Data*, vol. 3, no. 1, pp. 9, 2016.

[4] M. E. Taylor and P. Stone, “Transfer learning for reinforcement learning domains: A survey,” *Journal of Machine Learning Research*, vol. 10, pp. 1633–1685, 2009.

[5] T. Yue and H. Wang, “Deep learning for genomics: A concise overview,” *arXiv preprint arXiv:1802.00810*, 2018.

[6] D. Wang and T. F. Zheng, “Transfer learning for speech and language processing,” in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2015, pp. 1225–1237.

[7] R. Zhang, H. Tao, L. Wu, and Y. Guan, “Transfer learning with neural networks for bearing fault diagnosis in changing working conditions,” *IEEE Access*, vol. 5, pp. 14347–14357, 2017.

[8] L. Duan, D. Xu, and S. Chang, “Exploiting web images for event recognition in consumer videos: A multiple source domain adaptation approach,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1338–1345.

[9] A. Karbalayghareh, X. Qian, and E. R. Dougherty, “Optimal Bayesian transfer learning,” *IEEE Transactions on Signal Processing*, vol. 66, no. 14, pp. 3724–3739.

[10] M. Kárný, “Towards fully probabilistic control design,” *Automatica*, vol. 32, no. 12, pp. 1719–1722, 1996.

[11] M. Kárný and T. Kroupa, “Axiomatisation of fully probabilistic design,” *Information Sciences*, vol. 186, no. 1, pp. 105–113, 2012.

[12] J. Shore and R. Johnson, “Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy,” *IEEE Transactions on Information Theory*, vol. 26, no. 1, pp. 26–37, 1980.

[13] A. Quinn, M. Kárný, and T. V. Guy, “Optimal design of priors constrained by external predictors,” *International Journal of Approximate Reasoning*, vol. 84, pp. 150–158, 2017.

[14] C. Foley and A. Quinn, “Fully probabilistic design for knowledge transfer in a pair of Kalman filters,” *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 487–490, 2018.

[15] M. Papež and A. Quinn, “Dynamic Bayesian knowledge transfer between a pair of Kalman filters,” in *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2018.

[16] S. Särkkä, *Bayesian filtering and smoothing*, vol. 3, Cambridge University Press, 2013.

[17] S. Kullback and R. A. Leibler, “On information and sufficiency,” *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[18] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, “Variational inference: A review for statisticians,” *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[19] D. Willner, C. B. Chang, and K. P. Dunn, “Kalman filter algorithms for a multi-sensor system,” in *1976 IEEE Conference on Decision and Control including the 15th Symposium on Adaptive Processes*. IEEE, 1976, pp. 570–574.

[20] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.