Bayesian transfer learning between Student-t filters[⋆]

Milan Papež[a,*], Anthony Quinn[a,b]

[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*

Abstract

The problem of sequentially transferring a data-predictive probability distribution from a source to a target Bayesian filter is addressed in this paper. In many practical settings, this transfer is incompletely modelled, since the stochastic dependence structure between the filters typically cannot be fully specified. We therefore adopt fully probabilistic design to select the optimal transfer mechanism. We relax the target observation model via a scale-mixing parameter, which proves vital in successfully transferring the first and second moments of the source data predictor. This sensitivity to the transferred second moment ensures that imprecise predictors are rejected, achieving robust transfer. Indeed, Student-t state and observation models are adopted for *both* learning processes, in order to handle outliers in *all* hidden and observed variables. A recursive outlier-robust Bayesian transfer learning algorithm is recovered via a local variational Bayes approximation. The outlier rejection and positive transfer properties of the resulting algorithm are clearly demonstrated in a simulated planar position-velocity system, as is the key property of imprecise knowledge rejection (robust transfer), unavailable in current Bayesian transfer algorithms. Performance comparison with particle filter variants demonstrate the successful convergence of our robust variational Bayes transfer learning algorithm in sequential processing.

Keywords: Bayesian transfer learning, Student-t filtering, Incomplete modelling, Fully probabilistic design, Variational Bayes, Robust transfer

1. Introduction

Transfer learning [1] is one of the fundamental paradigms of artificial intelligence, addressing knowledge transfer between two (or more) learning tasks, known as the source task(s) and the target task(s), respectively [2, 3]. This research direction is of substantial interest in the statistical machine learning community [4, 5], and applications have been reported in protein folding [6], self-driving cars [7], natural language processing [8], biomedical image analysis [9], etc. This paper is specifically interested in Bayesian transfer learning—the transfer of knowledge expressed as probability distributions—and in the development of a consistent algorithm for networks of Bayesian filtering nodes.

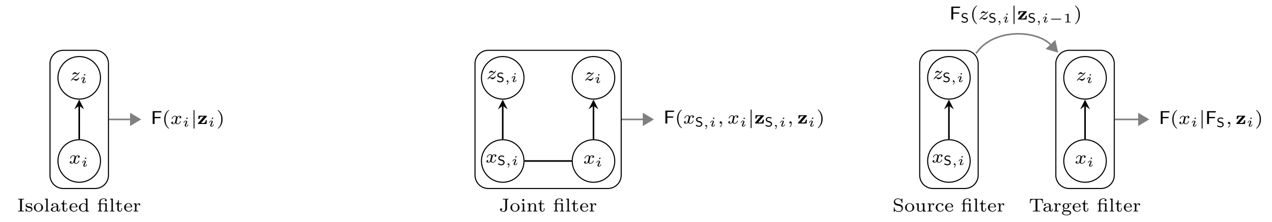
In Bayesian transfer learning [10], the challenge is to update the pre-prior distribution, prescribed via Bayesian foundations [11], by conditioning on a probability distribution made available by the source learning task [12, 13] (Fig. 1c). Standard Bayesian calculus relies on a complete specification of the stochastic dependence between the quantities of the target and source tasks, which we refer to as *complete modelling.* This may be practicable if the source knowledge takes the form of raw, stochastically modelled, data (i.e. a random process realization). Recently, an axiomatically justified approach based on fully probabilistic design (FPD) [14, 15]—which is rooted in the minimum cross-entropy principle for optimal prior design [16]—has emerged. FPD provides a principled and optimal way for designing a probability distribution that conditions on another probability distribution. This approach—in contrast to complete modelling—facilitates transfer in the form of a probability distribution and thus admits more general expressions of source knowledge. The main advantage lies in the fact that there is no longer the need to specify dependence assumptions between the target and source tasks (Fig. 1b). We refer to this evolved setting as *incomplete modelling.*

Recent work on FPD-based Bayesian transfer learning has been concerned with static [17, 18] and dynamic [19] knowledge transfer between a pair of Kalman filters. However, the fragile assumptions of Gaussianity adopted by the Kalman filter are rarely met in practical applications. We want to consider scenarios where outliers are present (i.e. outlierness or heavy-tailedness), so that the nominal noise values (inliers) of the state and observation processes are additionally contaminated by large, impulsive, and occasional disturbances. This happens, for example, in unreliable sensors or when tracking quickly manoeuvring targets. The performance and stability of the Kalman filter can be severely undermined in such situations. This has led to an increased interest in robustifying the Kalman filter against outliers [20]. Most of the recently proposed approaches rely on the heavy-tailed properties of the Student-t distribution to model only the observation process, involving optimization techniques that utilize expectation maximization [21] and variational approximations [22–27]. The design of filtering algorithms that adopt the Student-t distribution also to model the state process leads to tractability problems. However, ignoring such heavy-tailed state behaviours can have a significantly negative impact on estimation performance, in applications such as those mentioned above. To address this challenge, maximum likelihood-based techniques were proposed

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

[*] Corresponding author

Email addresses: papez@utia.cas.cz (Milan Papež), aquinn@tcd.ie (Anthony Quinn)


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

(a) No interaction

(b) Complete interaction model

(c) Incomplete interaction model

Figure 1: (a) *No interaction*: an isolated Bayesian filter, with a complete dependence structure between the variables, (z_i, x_i); (b) *complete interaction model*: a joint Bayesian filter, with complete modelling of all variables, $(z_{\mathsf{S},i}, x_{\mathsf{S},i}, z_i, x_i)$; (c) *incomplete interaction model*: source and target Bayesian filters, with isolated models of variables, $(z_{\mathsf{S},i}, x_{\mathsf{S},i})$ and (z_i, x_i), respectively. No stochastic interaction model is specified. The source filter provides its observation predictive distribution, F_{S}. The target filter utilizes this source knowledge to improve its performance.

in [28, 29]. More recently, heavy-tailed state process assumptions were imposed indirectly by modelling the one-step-ahead state predictor as Student-t, rather than directly modelling the state transitions as Student-t [30–32]. Alternatives to Student-t modelling of heavy-tailed state and observation processes have recently been proposed in [33].

Therefore, in this paper, we provide the following contributions:

1. We develop an online FPD-based static Bayesian transfer learning algorithm that accepts knowledge in the form of an observation predictor provided by a source filter.
2. Both the source and target filters are susceptible to outliers in both their state and observation processes. We propose a novel robust Student-t filter that is based on (i) modelling the heavy-tailed nature of both the state and observation processes with infinite Gaussian scale mixtures (consistent with Student-t modelling), and (ii) performing approximate inference using the coordinate ascent mean-field variational approach [34, 35], in order to recover a recursive algorithm.
3. We show that the introduction of a suitable auxiliary variable overcomes previous problems in achieving *robust transfer*, i.e. in rejecting imprecise source information. This variable augmentation now successfully transfers the second moment information of the source.
4. We provide extensive simulation results in the context of a planar position-velocity system, demonstrating that the reported Student-t transfer learning algorithm is more resistant to outliers than its Gaussian counterparts.
5. We also implement particle filtering variants of our variational Student-t algorithm, as well as of the classical measurement vector fusion (MF) algorithm [36]. This allows us to demonstrate the close tracking of our variational algorithm and these expensive stochastic variants, supporting the claims for convergence of our algorithm.

In reference to contribution 2 above, note that our method shares similarities with [31, 33] but adopts a novel second-order extension in order to avoid informal model adaptations necessitated in that previous work. This leads to a new computational flow for suppressing outliers in the state process. In our previous work [17, 19], we designed FPD-based static and dynamic Bayesian transfer learning strategies between a pair of Kalman filters. However, as stated in contribution 3, they could not achieve robust transfer, and it is a key contribution of this current work to design a robust transfer scheme.

The rest of this paper is organized as follows: Section 2 specifies the Bayesian transfer learning problem, and its general solution via the FPD-based framework, which transfers the source observation predictive distribution to the target Bayesian filter in incompletely modelled scenarios. Section 3 instantiates Section 2 in the Student-t filtering context, introducing a novel solution for handling outliers in the state process, and the essential scale-mixture relaxation which ensures robust transfer learning. Tractable and recursive processing is recovered via a local variational Bayes approximation at each step. Section 4 studies the key aspects of the proposed approach via a simulated planar position-velocity system, focusing on the robustness of the transfer, and its rejection of outliers. Detailed experimental comparisons with particle filter variants reveal the convergence properties of our sequential variational Bayesian transfer learning algorithm. Section 5 discusses the mechanism behind robust transfer learning and provides more comments on the newly developed Student-t filter. Section 6 offers concluding remarks.

2. Static FPD transfer of an observation predictor between a pair of Bayesian filters

We consider a state-space model of 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 the state variable $x_i \in \boldsymbol{x} \subseteq \mathbb{R}^{n_x}$ is indirectly (noisily) measured through the observation variable $z_i \in \boldsymbol{z} \subseteq \mathbb{R}^{n_z}$, with $i = 1, \ldots, n$ being the discrete-time index (and $x_0 \equiv \varnothing$ in the condition). The model (1) is specified by the state transition and observation probability distributions (1a) and (1b), respectively. The initial state variable is distributed according to $x_1 \sim \mathsf{F}(\cdot)$. All probability models are assumed to be expressed by distributions in this work. We use F to denote fixed-form (specified) distributions, and M and Q to denote variational (unspecified) distributions.

The fundamental and complete inferential object required to devise inference algorithms for the state-space model (1) is the joint model

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

where $\mathsf{F}(x_{i-1}|\mathbf{z}_{i-1})$ is the posterior distribution at the previous time step and $\mathbf{z}_{i-1} = (z_1, \ldots, z_{i-1})$ is the past observation record with $\mathbf{z}_0 \equiv \varnothing$. The central aim of this paper is to design an algorithm for transferring knowledge from a source to a target Bayesian filter, see Fig. 1c. In line with Bayesian principles, we assume that the source filter provides knowledge in the form of a probability distribution, F_S. Therefore, the target filter does not have access to the source observations, $z_{\mathsf{S},i}$, themselves. The inferential objective is to extend the basic setting (2) of the (isolated) target filter to condition also on this source distribution, F_S, i.e. to elicit the distribution,

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

The approach (3) provides notable benefits, including the following: (i) there is no need to specify explicit—and hard to elicit—dependence assumptions between the source $(z_{\mathsf{S},i}, x_{\mathsf{S},i})$ and target (z_i, x_i) quantities; (ii) F_S facilitates more general expressions of the source knowledge beyond crisp realizations, $\mathbf{z}_{\mathsf{S},n}$; and (iii) the degrees-of-freedom (dofs) of F_S—i.e. its sufficient statistics—are independent of n in dimension (and typically low-dimensional) in conjugate Bayesian systems [11]. The form of (3) is unknown in the absence of a complete model, and we therefore need to adopt a mechanism for conditioning (2) on F_S in this case.

In this paper, we transfer the observation predictor, F_S, of the source filter. This is achieved by restricting the functional form of the unknown joint model (3) according to

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

where "≡" denotes "is defined to be equal to". More specifically, we constrain the F_S-conditioned model of the target observations to be the observation predictor of the source filter evaluated at (target) $z_i \in \boldsymbol{z}$:

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

Fixing the transferred $\mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1})$ in (4), and admitting $\mathsf{M}(x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1})$ as the only variational quantity, the knowledge-constrained set of admissible models is defined to be

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

The joint model (2) is the complete knowledge specification for the selected (target) filter in the absence of knowledge transfer. Therefore, we choose it as the ideal (reference) model,

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

FPD chooses the optimal model, M°, for an unknown model, M, by searching for it within the knowledge-constrained set, $\mathsf{M} \in \boldsymbol{\mathsf{M}}$ (5), and expressing preferences about M via the (pre-specified) ideal model, M_I (6). Specifically, the FPD-optimal design, $\mathsf{M}^\circ \in \boldsymbol{\mathsf{M}}$, is chosen as the distribution that is closest to M_I in the minimum Kullback-Leibler divergence (KLD) sense:

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

where the KLD from M to M_I is given by

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

and E_M denotes the expected value under M.

Proposition 1. *If the unknown augmented model is a member of the knowledge constrained set,* $\mathsf{M} \in \boldsymbol{\mathsf{M}}$ *(5), and the ideal augmented model,* M_I*, is given by (6), then the FPD-optimal augmented model—and the solution of (7)—is*

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

where

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

Proof. See Appendix A. □

The FPD-optimal second-order state prior (9) processes the source observation predictor in the incompletely modelled case. It is the optimal update from the pre-prior $\mathsf{F}(x_i, x_{i-1}|\mathbf{z}_{i-1})$ to the prior $\mathsf{M}^\circ(x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1})$, and is subsequently adopted by the target filter in (2) by assigning

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

The FPD-optimal source-knowledge-constrained variant of (2) then becomes

$$\begin{aligned}\mathsf{F}(z_i, x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) \equiv \\ \mathsf{F}(z_i|x_i)\mathsf{M}^\circ(x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}).\end{aligned} \tag{10}$$

The joint augmented model (10) is now sufficient for designing the target filter.

3. Static FPD transfer of an observation predictor between a pair of Student-*t* filters

Outliers are defined as sudden disturbances or anomalies that are inconsistent with the assumed process model. Outlierness can be modelled via a linear mixture of the (clean) process model and another component with large variance. Examples include the ϵ-contamination model [30] and the spike-and-slab model [23]. In this paper, we opt to use the Student-t distribution, $\mathrm{St}(\cdot; \mu, \Sigma, \eta)$, which models the outliers via tails that are heavier than those of the standard Gaussian distribution, $\mathcal{N}(\cdot; \mu, \Sigma)$. Here, μ is the mean vector, Σ is the scale matrix (or covariance matrix for the Gaussian distribution), and $\eta > 0$ is the dof parameter. The tails of the Student-t distribution are tuned by the degrees-of-freedom (dof) parameter, η. If $\eta = 1$, the tails are heavy, thus modelling large outliers. For increasing η, the tails become lighter. Indeed, $\lim_{\eta\to\infty} \mathrm{St}(\cdot; \mu, \Sigma, \eta) = \mathcal{N}(\cdot; \mu, \Sigma)$, in which case outlierness is not explicitly modelled.

The conventional Kalman filter adopts a state-space model (1) in Gaussian form [37], and thus fails to model outlierness, significantly undermining its performance in outlier-present environments. In order to increase robustness to outliers, we relax $\mathcal{N}(\cdot; \mu, \Sigma)$ to its Student-t generalization, $\mathrm{St}(\cdot; \mu, \Sigma, \eta)$, via the following specification of the state-space model (1):

$$\mathsf{F}(x_i|x_{i-1}) \equiv \mathrm{St}(x_i; Ax_{i-1}, Q, \omega), \tag{11a}$$

$$\mathsf{F}(z_i|x_i) \equiv \mathrm{St}(z_i; Cx_i, R, \nu), \tag{11b}$$

where A, Q, and ω are, respectively, the state transition matrix, state noise scale matrix, and state dof parameter; and C, R, and ν are, respectively, the observation matrix, observation noise scale matrix, and observation dof parameter. It is typical for observation processes to contain outliers when using poor-quality sensors, sensors with sudden short-time failures, or sensors that interfere with a nearby device. Specific outlier contexts include multipath fading in satellite positioning applications [22] and electromagnetic wave reflections in radar applications [38]. Separately, the state process is susceptible to outliers in contexts such as the tracking a rapidly manoeuvring target or the processing corrupted observations from an inertial measurement unit [27, 39].

We ensure, by construction, that the posterior state distribution—i.e. the filtering distribution—at time $i-1$ is

$$\mathsf{F}(x_{i-1}|\mathbf{z}_{i-1}) \equiv \mathcal{N}(x_{i-1}; x_{i-1|i-1}, P_{i-1|i-1}), \tag{12}$$

where $\mathcal{N}(\cdot;\mu,\Sigma)$ denotes the Gaussian distribution with mean vector, μ, and covariance matrix, Σ. Specifically, $x_{i-1|i-1}$ and $P_{i-1|i-1}$ are the state estimate and covariance matrix at step $i-1$, respectively. It was shown in [31] that (12) enhances estimation performance in the context of (11), when compared to adoption of a Student-t distribution.

The conditional and marginal distributions of the joint model (10) are intractable under (11) and (12), as there is no closed-form expression for a joint distribution constructed either from Student-t distributions with different dof parameters [40], or from a combination of Student-t and Gaussian distributions. To simplify the subsequent design of the (approximate) inference algorithm, we use the fact that (11) can alternatively be expressed as infinite Gaussian scale mixtures:

$$\mathsf{F}(x_i|x_{i-1}) = \int_0^\infty \mathcal{N}\big(x_i; Ax_{i-1}, \xi Q\big) i\mathcal{G}\big(\xi; \tfrac{\omega}{2}, \tfrac{\omega}{2}\big) d\xi, \tag{13a}$$

$$\mathsf{F}(z_i|x_i) = \int_0^\infty \mathcal{N}\big(z_i; Cx_i, \lambda R\big) i\mathcal{G}\big(\lambda; \tfrac{\nu}{2}, \tfrac{\nu}{2}\big) d\lambda, \tag{13b}$$

where $i\mathcal{G}\big(\cdot; a, b\big)$ denotes the inverse-Gamma distribution with shape and scale parameters, a and b, receptively; and ξ and λ are scalar mixing variables. This allows us to reformulate the state-space model (11) hierarchically as follows:

$$\mathsf{F}(x_i|\xi, x_{i-1}) \equiv \mathcal{N}\big(x_i; Ax_{i-1}, \xi Q\big), \tag{14a}$$

$$\mathsf{F}(\xi) \equiv i\mathcal{G}\big(\xi; \tfrac{\omega}{2}, \tfrac{\omega}{2}\big), \tag{14b}$$

$$\mathsf{F}(z_i|\lambda, x_i) \equiv \mathcal{N}\big(z_i; Cx_i, \lambda R\big), \tag{14c}$$

$$\mathsf{F}(\lambda) \equiv i\mathcal{G}\big(\lambda; \tfrac{\nu}{2}, \tfrac{\nu}{2}\big). \tag{14d}$$

For the purposes of robust knowledge transfer, we augment the conditional observation model (14c) with the further auxiliary variable, κ, as follows:

$$\begin{aligned} \mathsf{F}(z_i, \lambda, \kappa|x_i) &\equiv \mathsf{F}(z_i|\kappa, x_i)\mathsf{F}(\lambda)\mathsf{F}(\kappa), \\ \mathsf{F}(z_i|\kappa, x_i) &\equiv \mathcal{N}\big(z_i; Cx_i, \kappa R\big), \\ \mathsf{F}(\kappa) &\equiv i\mathcal{G}\big(\kappa; \tfrac{\alpha}{2}, \tfrac{\beta}{2}\big), \end{aligned} \tag{15}$$

and $\mathsf{F}(\lambda)$ is given by (14d). Consequently, the joint model (10) is augmented in the following way:

$$\begin{aligned} &\mathsf{F}(z_i, \lambda, \kappa, \xi, x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) \equiv \\ &\qquad \mathsf{F}(z_i|\lambda, x_i)\mathsf{M}^\circ(\lambda, \kappa, \xi, x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}), \end{aligned} \tag{16}$$

where

$$\begin{aligned} \mathsf{M}^\circ(\lambda, \kappa, \xi, x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) &\propto \mathsf{F}(\lambda, \kappa, \xi, x_i, x_{i-1}|\mathbf{z}_{i-1}) \\ &\times \exp\left\{\int \log \mathsf{F}(z_i|\kappa, x_i)\mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1})dz_i\right\}. \end{aligned} \tag{17}$$

The central inference objective of this paper is to compute the joint augmented posterior model

$$\mathsf{F}(\lambda, \kappa, \xi, x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_i). \tag{18}$$

The normalizing constant of (18) is analytically intractable, which prevents us from finding an exact closed-form expression for computing (18). Therefore, we are forced to rely on approximate inference techniques. We adopt coordinate ascent mean-field variational inference (variational Bayes) [35] as a local approximation in each step of Bayesian filtering, since—as we shall see—it recovers a computationally efficient recursive filtering algorithm with good performance. In our current approach, we opt to approximate the second-order model,

$$\mathsf{F}(\lambda, \kappa, \xi, x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_i), \tag{19}$$

which is proportional to (16). This extension will engender second-order interactions in the resulting approximate distribution, which will prove vital in outlier suppression. We will discuss this point further in Section 5. Specifically, we seek an optimal posterior distribution from the mean-field variational class, as follows:

$$\begin{aligned} \mathsf{Q}(\lambda, \kappa, \xi, x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) &\equiv \mathsf{Q}(\kappa|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) \\ &\times \mathsf{Q}(\lambda|\mathsf{F}_\mathsf{S}, \mathbf{z}_i)\mathsf{Q}(\xi|\mathsf{F}_\mathsf{S}, \mathbf{z}_i)\mathsf{Q}(x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_i). \end{aligned} \tag{20}$$

The joint factor, $\mathsf{Q}(x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_i)$, is the key element in correctly handling the outliers in the state process. After marginalizing x_{i-1}, it yields the required F_S-conditioned state filtering factor, $\mathsf{Q}(x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_i)$.

Coordinate ascent mean-field variational inference seeks a local optimum of the variational objective function—the evidence lower bound [35]—by iteratively optimizing every independent (free) factor,

$$\mathsf{Q}^\circ(\theta_j) \propto \exp\left\{\mathsf{E}_{\text{-}\theta_j}[\log \mathsf{F}(\Theta)]\right\}, \tag{21}$$

while keeping the complementary ones,

$$\mathsf{Q}^\circ_{\text{-}\theta_j} = \prod_{l \neq j} \mathsf{Q}^\circ(\theta_l), \tag{22}$$

fixed. Here, $\Theta \equiv (\theta_1, \ldots, \theta_m)$, and $\mathsf{E}_{\text{-}\theta_j}$ denotes the expected value with respect to the complementary factors (22).

Proposition 2. *If the joint augmented model* (16) *is specified by* (14) *and* (15), *and the source observation predictor is*

$$\mathsf{F}_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1}) \equiv \mathsf{Q}^\circ_\mathsf{S}(z_i|\mathbf{z}_{\mathsf{S},i-1}) = \mathcal{N}(z_i; z_{\mathsf{S},i|i-1}, R_{\mathsf{S},i|i-1}), \tag{23}$$

then the optimal variational factors of (20) *are*

$$\mathsf{Q}^\circ(\xi|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) = i\mathcal{G}\big(\xi; \tfrac{a_\xi}{2}, \tfrac{b_\xi}{2}\big), \tag{24a}$$

$$\mathsf{Q}^\circ(\lambda|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) = i\mathcal{G}\big(\lambda; \tfrac{a_\lambda}{2}, \tfrac{b_\lambda}{2}\big), \tag{24b}$$

$$\mathsf{Q}^\circ(\kappa|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) = i\mathcal{G}\big(\kappa; \tfrac{a_\kappa}{2}, \tfrac{b_\kappa}{2}\big), \tag{24c}$$

$$\mathsf{Q}^\circ(x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) =$$

$$\mathcal{N}\left(\begin{bmatrix} x_i \\ x_{i-1}\end{bmatrix}; \begin{bmatrix} x_{i|i} \\ x_{i-1|i}\end{bmatrix}, \begin{bmatrix} P_{i|i} & P_{i|i}L^\top \\ LP_{i|i} & LP_{i|i}L^\top + P_{i-1|i}\end{bmatrix}\right), \tag{24d}$$

and, specifically, the FPD-optimal state predictor is

$$\mathsf{Q}^{o}(x_i|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) = \mathcal{N}(x_i; \bar{x}_{i|i-1}, \bar{P}_{i|i-1}). \tag{25}$$

The shape and scale hyperparameters of (24a-24c) are:

$$\begin{aligned} a_\xi &= \omega + n_x, \\ b_\xi &= \omega + \operatorname{tr}\left\{\mathsf{E}\left[(x_i - Ax_{i-1})(x_i - Ax_{i-1})^\top\right]Q^{-1}\right\}, \end{aligned} \tag{26}$$

$$\begin{aligned} a_\lambda &= \nu + n_z, \\ b_\lambda &= \nu + \operatorname{tr}\left\{\mathsf{E}\left[(z_i - Cx_i)(z_i - Cx_i)^\top\right]R^{-1}\right\}, \end{aligned} \tag{27}$$

$$\begin{aligned} a_\kappa &= \alpha + n_z, \\ b_\kappa &= \beta + \operatorname{tr}\Big\{\big(R_{\mathsf{S},i|i-1} + \mathsf{E}\big[(z_{\mathsf{S},i|i-1} - Cx_i) \\ &\qquad \times (z_{\mathsf{S},i|i-1} - Cx_i)^\top\big]\big)R^{-1}\Big\}. \end{aligned} \tag{28}$$

The shaping parameters of (24d) *and* (25) *are given by*

$$\begin{aligned} x_{i-1|i} &= x_{i-1|i-1} + L(x_{i|i} - x_{i|i-1}), \\ P_{i-1|i} &= (I_{n_x} - LA)P_{i-1|i-1}, \end{aligned} \tag{29}$$

$$\begin{aligned} x_{i|i} &= \bar{x}_{i|i-1} + K(z_i - \bar{z}_{i|i-1}), \\ P_{i|i} &= (I_{n_x} - KC)\bar{P}_{i|i-1}, \end{aligned} \tag{30}$$

$$\begin{aligned} \bar{x}_{i|i-1} &= x_{i|i-1} + M(z_{\mathsf{S},i|i-1} - z_{i|i-1}), \\ \bar{P}_{i|i-1} &= (I_{n_x} - MC)P_{i|i-1}. \end{aligned} \tag{31}$$

Furthermore, the gain terms are

$$\begin{aligned} K &= \bar{P}_{i|i-1}C^\top \bar{R}_{i|i-1}^{-1}, \\ L &= P_{i-|i-1}A^\top P_{i|i-1}^{-1}, \\ M &= P_{i|i-1}C^\top R_{i|i-1}^{-1}. \end{aligned} \tag{32}$$

The remaining statistics are

$$\begin{aligned} x_{i|i-1} &= Ax_{i-1|i-1}, \\ P_{i|i-1} &= AP_{i-1|i-1}A^\top + \mathsf{E}[\xi^{-1}]^{-1}Q, \\ \bar{z}_{i|i-1} &= C\bar{x}_{i|i-1}, \\ \bar{R}_{i|i-1} &= C\bar{P}_{i|i-1}C^\top + \mathsf{E}[\lambda^{-1}]^{-1}R, \\ z_{i|i-1} &= Cx_{i|i-1}, \\ R_{i|i-1} &= CP_{i|i-1}C^\top + \mathsf{E}[\kappa^{-1}]^{-1}R. \end{aligned} \tag{33}$$

Proof. See Appendix B. □

The shaping parameters of the variational factors (24) are coupled, and so there is no closed-form algebraic solution to update them directly. Instead, the computation of these factors is performed using a fixed-point iterative approach with N consecutive iterations per filtering step. Having expressed the required expected values in Proposition 2 in terms of the induced statistics, we obtain the iterative variational Bayes scheme summarized in Algorithm 1. Notice that (26, 29), (27, 30), and (28, 31) are represented by the same algebraic structure, which we therefore encode via the subroutine **B**. The latter is invoked three times per iteration for each step of the algorithm, namely the time-and-smoothing step, the transfer learning step, and the data step.

Algorithm 1: Student-t static variational Bayesian transfer learning

```
Input: $x_{i-1|i-1}$, $P_{i-1|i-1}$, $z_i$, $z_{\mathsf{S},i|i-1}$, $R_{\mathsf{S},i|i-1}$, $A$, $C$, $Q$, $R$, $\omega$, $\nu$, $\alpha$, $\beta$, $N$
Initialize $x_{i|i}^{(0)} = x_{i-1|i-1}$, $P_{i|i}^{(0)} = P_{i-1|i-1}$, $x_{i-1|i}^{(0)} = x_{i-2|i-1}$, $P_{i-1|i}^{(0)} = x_{i-2|i-1}$
for $k = 0, \dots, N-1$ do
    Time-and-smoothing step:
    $\Sigma = (I_{n_x} - AL)P_{i|i}^{(k)}(I_{n_x} - AL)^\top + AP_{i-1|i}^{(k)}A^\top$
    $(x_{i-1|i}^{(k+1)}, P_{i-1|i}^{(k+1)}, x_{i|i-1}, P_{i|i-1})=\mathbf{B}(\omega, \omega, x_{i|i}^{(k)}, A, Q, x_{i-1|i}^{(k)}, \Sigma, x_{i-1|i-1}, P_{i-1|i-1}, n_x)$
    Transfer learning step:
    $\Sigma = CP_{i|i}^{(k)}C^\top + R_{\mathsf{S},i|i-1}$
    $(\bar{x}_{i|i-1}, \bar{P}_{i|i-1}, \bar{z}_{i|i-1}, \bar{R}_{i|i-1})=\mathbf{B}(\alpha, \beta, z_{\mathsf{S},i|i-1}, C, R, x_{i|i}^{(k)}, \Sigma, x_{i|i-1}, P_{i|i-1}, n_z)$
    Data step:
    $\Sigma = CP_{i|i}^{(k)}C^\top$
    $(x_{i|i}^{(k+1)}, P_{i|i}^{(k+1)}, z_{i|i-1}, R_{i|i-1})=\mathbf{B}(\nu, \nu, z_i, C, R, x_{i|i}^{(k)}, \Sigma, \bar{x}_{i|i-1}, \bar{P}_{i|i-1}, n_z)$
Set $x_{i|i}=x_{i|i}^{(N)}$, $P_{i|i}=P_{i|i}^{(N)}$, $x_{i-1|i}=x_{i-1|i}^{(N)}$, $P_{i-1|i}=P_{i-1|i}^{(N)}$
Output: $x_{i|i}$, $P_{i|i}$
```

```
$(\widehat{x}, \widehat{P}, \widehat{y}, \widehat{S})=\mathbf{B}(c, d, y, H, S, \mu, \Sigma, x, P, n)$
$a = c + n$
$b = d + \operatorname{tr}\left\{[(y - H\mu)(y - H\mu)^\top + \Sigma]S^{-1}\right\}$
$\widehat{y} = Hx$
$\widehat{S} = HPH^\top + \frac{b}{a}S$
$N = PH^\top \widehat{S}^{-1}$
$\widehat{x} = x + N(y - \widehat{y})$
$\widehat{P} = P - N\widehat{S}N^\top$
```

Remark 1. *The isolated Student-t filter is obtained by approximating the second-order model,* $\mathsf{F}(\lambda, \xi, x_i, x_{i-1}|\mathbf{z}_i)$, *c.f.* (19), *via an optimal posterior distribution from the mean-field variational class,*

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

This is accomplished by a simple adaptation of the proof of Proposition 2 (which we do not present here for brevity). The isolated source Student-t filter (Fig. 1c) is designed in exactly the same way. Consequently, the source filter provides the observation predictor, F_S, *in the Gaussian form* (23).

4. Experiments

This section presents an extended simulation context in order to demonstrate the main features of the proposed method: (i) convergence of the local—variational Bayes—approximate inference scheme, (ii) robust and versatile transfer learning properties in the outlier-free setting, (iii) resistance to outliers of varying intensity, and (iv) estimation performance for various qualities of source knowledge in outlier-present settings. In all our experiments, we consider the linear state-space model with the following structure:

$$x_{i+1} = Ax_i + w_i, \qquad w_i \sim \mathrm{St}(w_i; \mathbf{0}, Q, \omega),$$

Algorithm		Description
*S*tudent-*t* filter with *No* *T*ransfer	(*SNT*)	Remark 2
*S*tudent-*t* *S*tatic Bayesian *T*ransfer learning	(*SST*)	Remark 3
*S*tudent-*t* *M*easurement vector *F*usion	(*SMF*)	Remark 4
*G*aussian filter with *No* *T*ransfer	(*GNT*)	[41]
*G*aussian *S*tatic Bayesian *T*ransfer learning	(*GST*)	[17]
*G*aussian *M*easurement vector *F*usion	(*GMF*)	[36] adapted with [17]
*P*article filter with *No* *T*ransfer	(*PNT*)	[42]
*P*article *S*tatic Bayesian *T*ransfer learning	(*PST*)	Remark 5
*P*article *M*easurement vector *F*usion	(*PMF*)	Remark 6

Table 1: The list of algorithms compared in the simulation study.

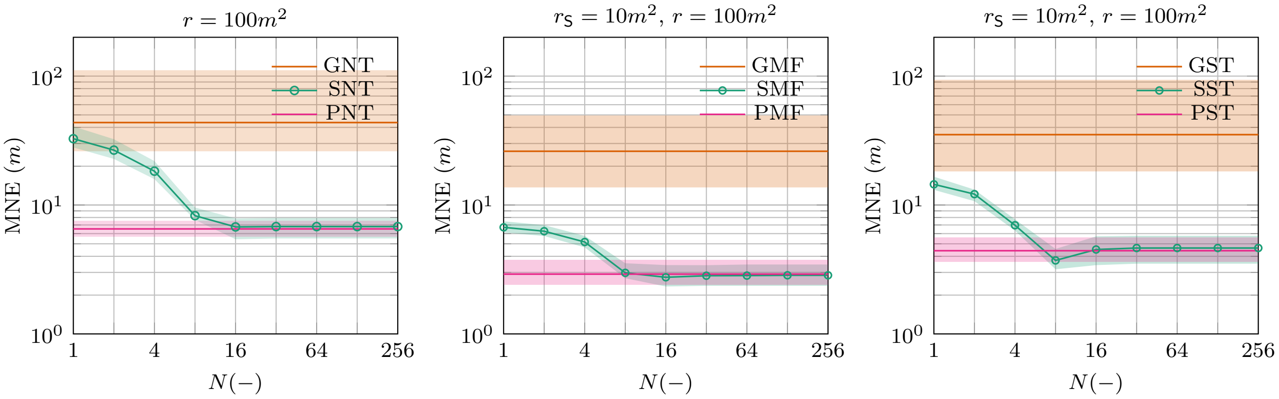


Figure 2: The position MNE versus the number of iterations, N, for the NT filters (left), MF filters (middle), and ST filters (right). The results are averaged over 100 independent simulation runs, with the solid line being the median and the shaded area delineating the interquartile range.

$$\begin{aligned} z_i &= Cx_i + v_i, & v_i &\sim \mathrm{St}(v_i; \mathbf{0}, R, \nu), \\ z_{\mathsf{S},i} &= Cx_i + v_{\mathsf{S},i}, & v_{\mathsf{S},i} &\sim \mathrm{St}(v_{\mathsf{S},i}; \mathbf{0}, R_{\mathsf{S}}, \nu_{\mathsf{S}}), \end{aligned} \tag{34}$$

where $w_i \in \boldsymbol{x}$, $v_i \in \boldsymbol{z}$, and $v_{\mathsf{S},i} \in \boldsymbol{z}$ are the target state, target observation, and source observation noise variables, respectively. We study the position-velocity model for tracking a highly manoeuvring target in the plane ($\mathbb{R}^2$). The parameters of (34) are therefore specified as [43]:

$$A = \begin{bmatrix} 1 & \Delta \\ 0 & 1 \end{bmatrix} \otimes I_2, \quad Q = q \begin{bmatrix} \frac{\Delta^3}{3} & \frac{\Delta^2}{2} \\ \frac{\Delta^2}{2} & \Delta \end{bmatrix} \otimes I_2,$$

$$C = \begin{bmatrix} I_2 & O_2 \end{bmatrix}, \quad R = rI_2, \quad R_{\mathsf{S}} = r_{\mathsf{S}} I_2.$$

Here, the state vector is $x_i \equiv (p_{x,i}, p_{y,i}, v_{x,i}, v_{y,i})$, where $p_{x,i}$ and $p_{y,i}$ are position coordinates in the x and y axes, respectively, and $v_{x,i}$ and $v_{y,i}$ are the velocities in the x and y axes, respectively. Only the positional states are (noisily) observed. The matrices of this model result from the discretization of the standard kinematic equations (see, for example, Section 6.2 [44]). We set the sampling period, state noise power spectral density, source observation variance, and target observation variance as $\Delta = 0.1$s, $q = 1\mathrm{m}^2/\mathrm{s}^3$, $r_{\mathsf{S}} = 10\mathrm{m}^2$ and $r = 100\mathrm{m}^2$, respectively. The initial posterior state estimate and covariance matrix are $x_{1|0} = \mathbf{0}$ and $P_{1|0} = I_4$, respectively. The dof parameters of the target filter (14) are taken as $\omega = 4$, $\nu = 1$, and the associated parameters of the inverse gamma prior (15) as $\alpha \to 0$, $\beta \to 0$. The dof parameters of the source filter are (also) taken as $\omega_{\mathsf{S}} = 4$ and $\nu_{\mathsf{S}} = 1$. Note that the correct parameter values are adopted in the generative model (34), and so we do not allow any model misspecification in these simulations. (34) implies that the common state process, x_i, is observed (with outlierness) via the conditionally independent source and target observation processes, $z_{\mathsf{S},i}$ and z_i, respectively. The state estimation performance is evaluated via the mean-norm error (MNE) between the true state and its posterior estimate, i.e. $\mathrm{MNE} = \frac{1}{n}\sum_{i=1}^{n} ||x_i - x_{i|i}||$, where $||\cdot||$ is the Euclidean norm and $n = 100$. We compare the algorithms listed in Table 1.

Remark 2. *The (target) SNT filter is the isolated Student-t filter (Fig. 1a) without any source information (Remark 1). It can readily be obtained from Algorithm 1 by omitting the transfer learning step (lines 6-8) and setting $\bar{x}_{i|i-1} \equiv x_{i|i-1}$ and $\bar{P}_{i|i-1} \equiv P_{i|i-1}$ in the data step. This filter acts as the datum for all the transfer learning algorithms.*

Remark 3. *The SST filter (Algorithm 1) receives the source observation predictor, F_{S} (23), of the isolated source SNT filter (Fig. 1c). F_{S} is an inference from learning the hidden state process from the history, $\mathbf{z}_{\mathsf{S},i}$, of the source observation process, and so validly constitutes 'transfer learning', i.e. we learn about the state process via the source task and use this knowledge to enhance the target task.*

Remark 4. *The SMF filter is the implied Student-t version of the measurement vector fusion (MF) algorithm [36], being a specific case of Fig. 1b. It is obtained by applying (the product of) the observation models* (34) *in the target SNT filter (Remark 2). The algorithm then follows from a simple adaptation of the proof of Proposition 2. The main disadvantage of this classical MF approach is the requirement for complete specification of the explicit dependence assumptions between the source and target tasks (i.e. complete modelling), which is—importantly—not required by our SST filter (see Section 1).*

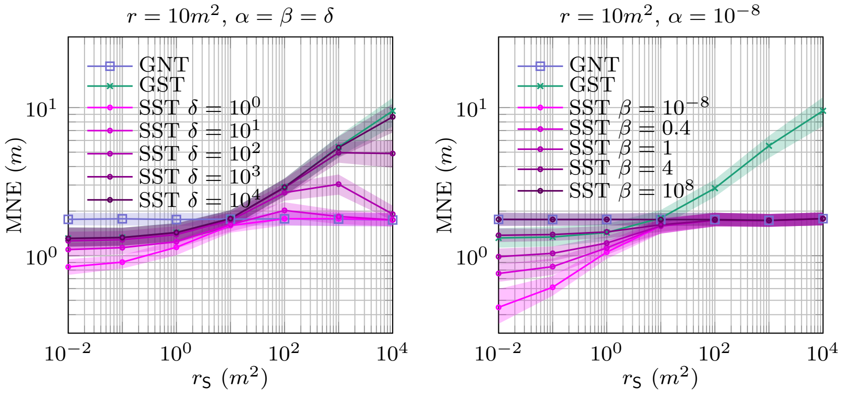


Figure 3: The position MNE versus the source observation variance, r_S for α and β varying (left), and α fixed and β varying (right). The results are averaged over 1000 independent simulation runs, with the solid line being the median and the shaded area delineating the interquartile range.

Remark 5. *The PST filter follows from an application of sequential importance sampling and resampling [42] in the context of Proposition 1. We directly use the Student-t state-space model* (11) *without its scale mixture decomposition. To implement the exponential term in* (9)*, the PST filter receives the source observation predictor,* F_S*, in the form of an empirical distribution provided by the isolated source PNT filter.*

Remark 6. *The PMF filter follows from application of sequential importance sampling and resampling [42] in the MF context described in Remark 4.*

4.1. Convergence properties

Although rigorous treatment of the convergence properties of *non-sequential* variational Bayes methods has recently been proposed [45], similar results on *sequential* variational Bayesian filtering are still elusive. On the other hand, particle filters [42] constitute a theoretically well supported stochastic approximation for the sequential Bayesian filtering problem. The essential feature of particle filters is that—with certain regularity assumptions—the approximation of the expected value of an unbounded function under the filtering distribution converges in the $\mathcal{L}^p$-norm (for $p \geq 2$) to the exact solution as the number of particles approaches infinity [46]. We adopt particle filters in our current simulation study in order to analyse the convergence properties of our proposed algorithm, whose tractability has been arranged via sequential local variational Bayes approximation at each step.

In Fig. 2, we present the position MNE of the SNT, SMF, and SST filters as the function of the number of (variational Bayes) iterations, N. The PNT, PMF, and PST filters run with 500 particles, the bootstrap proposal distribution, and multinomial resampling [42], delineating a sufficient lower MNE level in the present example. Increasing the number of particles extends this level only insignificantly, since the state-space model does not contain nonlinearities. The GNT, GMF, and GST filters provide an upper MNE level, which we seek to outperform. We see that the proposed Student-t filters indeed have lower MNEs than the upper MNE of the Gaussian filters, for as little as $N = 1$. Moreover, when increasing the number of iterations, the Student-t filters converge close to the lower MNE of the particle filters with no further improvements for $N > 16$. This experiment illustrates that our variational Student-t algorithm converges close to the *stable* solution provided by the particle filters, as the number of iterations, N, increases.

4.2. Robust transfer learning

The key feature of any transfer learning algorithm is its ability to reject poor-quality source knowledge, i.e. to achieve robust transfer. We demonstrate that the proposed method not only provides robust transfer but also allows us to tune the amount of transferred knowledge when processing high-quality source knowledge. We show this in the important special case of transfer between *Kalman* filters, and so we consider the outlier-free regime with the dof parameters of the source and target filters set to $\omega \to \infty$ and $\nu \to \infty$ (14b,14d). Under this setting, the tails of the Student-t distributions (34) correspond to the tails of the Gaussian distributions. Our earlier treatments of this situation—without the κ augmentation—failed to reject high-variance source knowledge [17, 19].

Fig. 3 illustrates how the MNE of the target filter depends on the quality of the source knowledge, by fixing the target observation variance, r, and varying the source observation variance, r_S. The GNT filter is *isolated* (i.e. it does not accept any source knowledge), and therefore it delineates the baseline MNE performance. The GST and SST filters accept source knowledge, and their MNE thus depends on the ratio of r to r_S. We say that these filters deliver *positive* or *negative* knowledge transfer if their MNE is below or above the performance level of the GNT filter, respectively. We see that the GST filter developed in [17] yields positive transfer for $r_\mathsf{S} < 10$, but negative transfer for $r_\mathsf{S} > 10$. Accordingly, since the MNE of the GST filter does not saturate at the baseline MNE level of the GNT filter, we say that the GST filter is not *robust*, i.e. it does not reject poor-quality source knowledge.

Fig. 3 (left) shows how different values of $\alpha = \beta = \delta$ (15) influence the transfer learning properties of the developed SST filter for poor-quality source knowledge ($r_\mathsf{S} > 10$). We observe that, for $\delta \to \infty$, the SST filter recovers the performance of the GST filter. A key result is that—as δ decreases—the ability to reject poor-quality source knowledge improves. Specifically, for $\delta = 1$, the SST filter provides positive transfer for $r_\mathsf{S} < 10$ and robust transfer for $r_\mathsf{S} > 10$. This investigation demonstrates that δ can be set by the modeller to enable any amount of

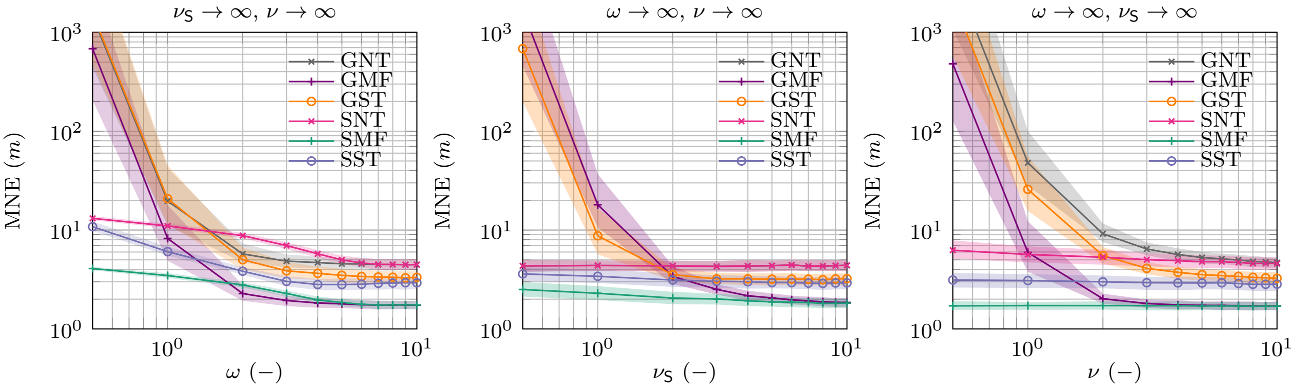


Figure 4: The position MNE versus the outlier intensity in the common states (left), source observations (middle), and target observations (right). The results are averaged over 1000 independent simulation runs, with the solid line being the median and the shaded area delineating the interquartile range.

rejection of poor-quality source knowledge, via $\delta \to \infty$ (no rejection) to $\delta = 1$ (complete rejection), respectively.

Fig. 3 (right) demonstrates how the parameters α and β of (15) influence the transfer learning properties of the developed SST filter in the interval of high-quality source knowledge, $r_\mathsf{S} < 10$. Specifically, we set $\alpha = 10^{-8}$ and change β from 10^{-8} to 10^8. This allows us to utilize all, or no, available high-quality source information, for $\beta = 10^{-8}$, or $\beta = 10^8$, respectively. For $r_\mathsf{S} > 10$, and for any setting of β, the SST filter achieves robust transfer. Importantly, for $\beta \to 0$, the proposed SST filter surpasses the GST filter. We offer more comments on these regimes in Section 5.

4.3. Robustness to state and observation outliers

Another principal purpose of the developed Student-t-based transfer learning algorithm—apart from robust transfer (above)—is to provide improved estimation performance in applications that suffer from outliers. Therefore, we compare the Student-t filters with the Gaussian filters when changing the outlier intensity (the dof parameters) in the state, source observation, and target observation noise variables (34).

Fig. 4 (left) shows the position MNE versus the state dof parameter, ω. The source and target observation dof parameters are set to $\nu_\mathsf{S} \to \infty$ and $\nu \to \infty$. In this case, the Student-t source and target observation distributions approach the Gaussian distribution (Section 3). For $\omega = 1$—corresponding to substantial outlier intensity—the difference between the Student-t and Gaussian filters is significant. When increasing ω (i.e. approaching Gaussianity), the MNE of the Student-t filters approaches the MNE of the Gaussian filters, eventually reaching the same values as $\omega \to \infty$.

To assess how the outliers affect the transfer learning properties of the proposed SST filter, we present Fig. 4 (middle) which depicts the position MNE while changing the source observation dof parameter, ν_S. The target state and observation dof parameters are set to $\omega \to \infty$ and $\nu \to \infty$. In this case, there is obviously no difference between the GNT and SNT filters, since they are not influenced by the source knowledge, and—as before—the Student-t noise distributions coincide with the Gaussian noise distributions. For large outliers in the source observations, $\nu_\mathsf{S} = 1$, we see that the Student-t filters offer increased performance compared to the Gaussian filters. Again, as the source observation dof parameter approaches infinity, $\nu_\mathsf{S} \to \infty$, we obtain a performance equivalent to the Gaussian filters.

The results in Fig. 4 (right)—where $\omega \to \infty$ and $\nu_\mathsf{S} \to \infty$—demonstrate that all methods behave in a similar way compared to Fig. 4 (left). Overall, the results in Fig. 4 confirm that the Student-t filters provide a lower MNE than the Gaussian filters when there are (even small) departures from the Gaussian modelling assumptions. Since the source and target observation variances, r and r_S, are set differently, we can notice the performance differences between the filters with and without transfer learning abilities.

4.4. Influence of source observation variance, r_S, on transfer

Fig. 5 illustrates the contrast between filters with and without the heavy-tailed assumptions on the state and observation processes while considering the presence of outliers and changing the quality of the source knowledge, controlled by r_S. Similarly as before, the GNT and SNT filters do not, of course, receive source knowledge, but they do provide a reference MNE level against which the transfer-based Gaussian and Student-t filters, respectively, can be compared. The clear difference between these two MNE levels demonstrates that the SNT filter provides increased resistance to outliers. The performance of the remaining filters depends on the ratio of r to r_S. The filters deliver *positive* or *negative* knowledge transfer whenever their MNE falls below or rises above the reference level, respectively. In particular, the GST filter offers positive transfer for $r_\mathsf{S} < 10$, but negative transfer for $r_\mathsf{S} > 10$ since the MNE does not saturate at the reference level of the GNT filter and thus does not reject imprecise source knowledge (i.e. it is not robust). The SST filter, on the other hand, provides positive transfer for $r_\mathsf{S} < 10$ and successfully rejects imprecise source knowledge by staying at the reference level of the SNT filter for $r_\mathsf{S} > 10$ (i.e. it is robust). Similarly, both the GMF and SMF filters—which imply a completely specified stochastic dependence structure between source and target processes—provide positive transfer and reject imprecise source knowledge (again, robust transfer). An overall look at the MNE and the associated interquartile ranges in Fig. 5 shows that the Gaussian filters are significantly more prone to outliers than the Student-t filters.

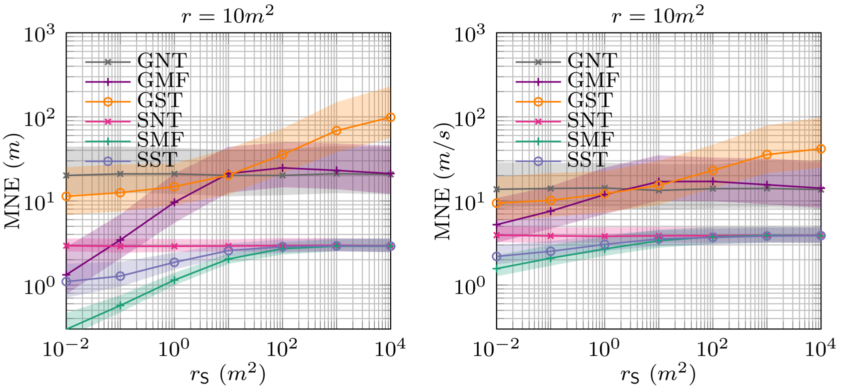


Figure 5: The MNE of the target filter versus the observation variance r_{S} of the source filter for position (left) and velocity (right). The results are averaged over 1000 independent simulation runs, with the solid line being the median and the shaded area delineating the interquartile range.

5. Discussion

Recall that the principal objective of the current paper is to achieve robust Bayesian knowledge transfer, i.e. the rejection of imprecise external knowledge. This was previously done only by an informal adaptation of FPD-optimal algorithms in the context of Kalman filters [17, 19]. The non-robustness is seen in the GST filter performance in Fig. 5. The problem arises from the fact that the FPD-optimal transfer is insensitive to the second moment of the source observation predictor in the Gaussian case [47]. The informal adaptation which was necessitated in order to achieve robust Gaussian transfer is obviated in the formal approach of this paper. The key progression in the current work has been the introduction of the auxiliary variable, κ (15). This allows the successful transfer of the source predictive covariance, $R_{\mathsf{S},i|i-1}$, as seen in (28). Equivalently, in the resulting Algorithm 1, Σ in line 7 successfully processes this second-order source statistics, entering the subroutine $\mathbf{B}$ as the seventh input variable. The framework reported in the current paper generalizes the hyperparameter-based relaxation for FPD-optimal robust transfer between *Kalman* filters [47], by allowing for Student-t outlierness in all the involved processes. Technically, the framework in the current paper specializes to [47] by setting $\omega \to \infty$ and $\nu \to \infty$ in (26) and (27), respectively.

We have shown that the prior relaxation of κ via (15) is vital to the success of transferring higher-order moments of the source observation predictor, in that the transfer fails at high δ (Fig. 3). In this regime, b_κ approaches δ (28) and the sensitivity on $R_{\mathsf{S},i|i-1}$ is lost.

This augmentation—at the cost of tractability in the exact FPD-optimal transfer learning algorithm—fails to preserve fixed functional forms sequentially. We have shown that the variational Bayes approximation, introduced as a local approximation at each time i, achieves functional closure of the parametric classes proposed in (23,24,25), recovering the recursive Algorithm 1. Note, however, that there are no guarantees in respect of the distributional accuracy achieved after sequential application of a local approximation such as variational Bayes [48].

The local variational Bayes approximation has previously been applied in Bayesian filtering [48], and, specifically, in filtering with outlier robust Student-t models [31, 33]. In the latter, the variational Bayes approximation proves to be intractable, and the authors overcome this problem via an informal adaptation of the one-step-ahead state predictor. We have circumvented this requirement in the current paper by approximating the second-order model (19), rather than the first-order model (18) which was adopted in [31, 33]. The resulting benefit for our algorithm is best seen in (33), where the dependence on state auxiliary variable ξ (13a) is engendered in the second term on the right-hand side, effectively modulating the nominal state-noise scale matrix, Q, instead of the covariance matrix of the one-step-ahead state predictor, $P_{i|i-1}$, as in [31, 33].

6. Conclusion

The sequential FPD-optimal Bayesian transfer learning algorithm developed in this paper has provided an important advance beyond previously available variants. The scale-mixture relaxation of the target observation process has allowed the transfer of higher-order moments of the source distribution, and we have seen that this ensures robust transfer. The reported framework explicitly models outliers in the source and target processes via the heavy-tailed Student-t distribution. In this respect, we proposed a novel and formal approach for dealing with outliers in the state process, which was previously unavailable in the literature. The simulation results in Section 4.4 show clearly that the algorithm rejects state and observation outliers when the isolated Kalman filter cannot.

The comparisons with MF-based algorithms reveal that the latter can still outperform our FPD-optimal Bayesian transfer, but they require a complete model of the dependence between the source and target state processes, which ours does not. Real-process environments that depart from these assumptions—which, anyway, are hard to elicit in practice—will undermine the MF performance. These problems are resisted by our FPD-optimal approach, which is, intrinsically, an optimal model completion strategy, and so does not depend on these fragile assumptions.

Appendix

A. Proof of Proposition 1

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

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

where

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

is the differential entropy of F_S, and

$$\begin{aligned}
c_{\mathsf{M}^\circ} = \int &\mathsf{F}(x_i, x_{i-1}|\mathbf{z}_{i-1}) \\
&\times \exp\{\mathsf{E}_{\mathsf{F}_\mathsf{S}}[\log \mathsf{F}(z_i|x_i)]\} dx_i dx_{i-1}
\end{aligned}$$

is the normalizing constant.

B. Proof of Proposition 2

We start the proof by finding an expression for the logarithm of (16), which—under (14) and (15)—yields

$$\begin{aligned}
&\log \mathsf{F}(z_i, \lambda, \kappa, \xi, x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_{i-1}) = \\
&\quad -\frac{n_z}{2}\log\lambda - \frac{1}{2}(z_i - Cx_i)^\top \lambda^{-1} R^{-1}(z_i - Cx_i) \\
&\quad -\frac{\nu+2}{2}\log\lambda - \frac{\nu}{2\lambda} \\
&\quad -\frac{n_z}{2}\log\kappa - \frac{1}{2}(z_{\mathsf{S},i|i-1} - Cx_i)^\top \kappa^{-1} R^{-1}(z_{\mathsf{S},i|i-1} - Cx_i) \\
&\quad -\frac{\alpha+2}{2}\log\kappa - \frac{\beta}{2\kappa} - \frac{1}{2}\operatorname{tr}\{R_{\mathsf{S},i|i-1}\kappa^{-1}R^{-1}\} \\
&\quad -\frac{n_x}{2}\log\xi - \frac{1}{2}(x_i - Ax_{i-1})^\top \xi^{-1} Q^{-1}(x_i - Ax_{i-1}) \\
&\quad -\frac{\omega+2}{2}\log\xi - \frac{\omega}{2\xi} \\
&\quad -\frac{1}{2}(x_{i-1} - x_{i-1|i-1})^\top P_{i-1|i-1}^{-1}(x_{i-1} - x_{i-1|i-1}) + c, \quad \text{(B.1)}
\end{aligned}$$

where c contains the constant terms.

After using (21) with (B.1), we gather the ξ-dependent terms as

$$\begin{aligned}
&\log \mathsf{Q}^\circ(\xi|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) = \\
&\quad -\frac{n_x}{2}\log\xi - \frac{1}{2\xi}\operatorname{tr}\{\mathsf{E}_{-\xi}[(x_i - Ax_{i-1})(x_i - Ax_{i-1})^\top]Q^{-1}\} \\
&\quad -\frac{\omega+2}{2}\log\xi - \frac{\omega}{2\xi} + c_\xi,
\end{aligned}$$

with c_ξ being a ξ-independent constant. This can be rearranged as

$$\begin{aligned}
\log \mathsf{Q}^\circ(\xi|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) &= \frac{a_\xi + 2}{2}\log\xi - \frac{b_\xi}{2\xi} + c_\xi \\
&= \log i\mathcal{G}\big(\xi; \tfrac{a_\xi}{2}, \tfrac{b_\xi}{2}\big) + c_\xi,
\end{aligned}$$

where the shaping parameters are given by (26).

Applying (21) and (B.1) allows us to gather the λ-dependent terms as

$$\begin{aligned}
&\log \mathsf{Q}^\circ(\lambda|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) = \\
&\quad -\frac{n_z}{2}\log\lambda - \frac{1}{2\lambda}\operatorname{tr}\{\mathsf{E}_{-\lambda}[(z_i - Cx_i)(z_i - Cx_i)^\top]R^{-1}\} \\
&\quad -\frac{\nu+2}{2}\log\lambda - \frac{\nu}{2\lambda} + c_\lambda,
\end{aligned}$$

where c_λ is a λ-independent constant. This leads to

$$\begin{aligned}
\log \mathsf{Q}^\circ(\lambda|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) &= \frac{a_\lambda + 2}{2}\log\lambda - \frac{b_\lambda}{2\lambda} + c_\lambda \\
&= \log i\mathcal{G}\big(\lambda; \tfrac{a_\lambda}{2}, \tfrac{b_\lambda}{2}\big) + c_\lambda,
\end{aligned}$$

where the shaping parameters are presented in (27).

Utilizing (21) with (B.1) reveals the κ-dependent terms

$$\begin{aligned}
&\log \mathsf{Q}^\circ(\kappa|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) = \\
&\quad -\frac{n_z}{2}\log\kappa - \frac{1}{2\kappa}\operatorname{tr}\{\mathsf{E}_{-\kappa}[(z_{\mathsf{S},i|i-1} - Cx_i)(z_{\mathsf{S},i|i-1} - Cx_i)^\top]R^{-1}\} \\
&\quad -\frac{\alpha+2}{2}\log\kappa - \frac{\beta}{2\kappa} - \frac{1}{2\kappa}\operatorname{tr}\{R_{\mathsf{S},i|i-1}R^{-1}\} + c_\kappa,
\end{aligned}$$

with c_κ being a κ-independent constant. Hence:

$$\begin{aligned}
\log \mathsf{Q}^\circ(\kappa|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) &= \frac{a_\kappa + 2}{2}\log\kappa - \frac{b_\kappa}{2\kappa} + c_\kappa \\
&= \log i\mathcal{G}\big(\kappa; \tfrac{a_\kappa}{2}, \tfrac{b_\kappa}{2}\big) + c_\kappa,
\end{aligned}$$

where the shaping parameters are summarized by (28).

Adopting (21) and (B.1), the (x_i, x_{i-1})-dependent terms are

$$\begin{aligned}
&\log \mathsf{Q}^\circ(x_i, x_{i-1}|\mathsf{F}_\mathsf{S}, \mathbf{z}_i) = \\
&\quad -\frac{1}{2}(z_i - Cx_i)^\top \mathsf{E}_{-\mathbf{x}}[\lambda^{-1}]R^{-1}(z_i - Cx_i) \\
&\quad -\frac{1}{2}(z_{\mathsf{S},i|i-1} - Cx_i)^\top \mathsf{E}_{-\mathbf{x}}[\kappa^{-1}]R^{-1}(z_{\mathsf{S},i|i-1} - Cx_i) \\
&\quad -\frac{1}{2}(x_i - Ax_{i-1})^\top \mathsf{E}_{-\mathbf{x}}[\xi^{-1}]Q^{-1}(x_i - Ax_{i-1}) \\
&\quad -\frac{1}{2}(x_{i-1} - x_{i-1|i-1})^\top P_{i-1|i-1}^{-1}(x_{i-1} - x_{i-1|i-1}) + c_\mathbf{x},
\end{aligned} \tag{B.2}$$

where $\mathbf{x} \equiv (x_i, x_{i-1})$. It can be seen that (B.2) is—up to the additive constant $c_\mathbf{x}$—equivalent to the logarithm of the joint density of the variables (z_i, x_i, x_{i-1}). Therefore, we have

$$\begin{aligned}\mathsf{Q}^{\circ}(z_i, x_i, x_{i-1}|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}) &= \mathcal{N}(z_i; Cx_i, \mathsf{E}[\lambda^{-1}]^{-1}R) \\ &\times \mathcal{N}(z_{\mathsf{S},i|i-1}; Cx_i, \mathsf{E}[\kappa^{-1}]^{-1}R)\mathcal{N}(x_i; Ax_{i-1}, \mathsf{E}[\xi^{-1}]^{-1}Q) \\ &\times \mathcal{N}(x_{i-1}; x_{i-1|i-1}, P_{i-1|i-1}). \end{aligned} \tag{B.3}$$

To find a closed-form expression for the F_{S}-conditioned joint smoothing density, we invoke the chain rule:

$$\mathsf{Q}^{\circ}(x_i, x_{i-1}|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_i) = \mathsf{Q}^{\circ}(x_{i-1}|x_i, \mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1})\mathsf{Q}^{\circ}(x_i|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_i). \tag{B.4}$$

The filtering density in (B.4) is derived using

$$\mathsf{Q}^{\circ}(x_i|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_i) \propto \mathsf{Q}^{\circ}(z_i|x_i)\mathsf{Q}^{\circ}(x_i|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}). \tag{B.5}$$

After marginalizing x_{i-1} in (B.3), we choose

$$\begin{aligned}\mathsf{Q}^{\circ}(z_i|x_i) &= \mathcal{N}(z_i; Cx_i, \mathsf{E}[\lambda^{-1}]^{-1}R) \\ \mathsf{Q}^{\circ}(x_i|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}) &\equiv \mathcal{N}(z_{\mathsf{S},i|i-1}; Cx_i, \mathsf{E}[\kappa^{-1}]^{-1}R)\mathcal{N}(x_i; x_{i|i-1}, P_{i|i-1}) \\ &= \mathcal{N}(x_i; \bar{x}_{i|i-1}, \bar{P}_{i|i-1}),\end{aligned}$$

where $\{x_{i|i-1}, P_{i|i-1}\}$ and $\{\bar{x}_{i|i-1}, \bar{P}_{i|i-1}\}$ are given in (33) and (31), respectively. Using these distributions, (B.5) leads to

$$\mathsf{Q}^{\circ}(x_i|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_i) = \mathcal{N}(x_i; x_{i|i}, P_{i|i}), \tag{B.6}$$

where $\{x_{i|i}, P_{i|i}\}$ is given in (30).

The backward transition kernel in (B.4) is computed as

$$\mathsf{Q}^{\circ}(x_{i-1}|x_i, \mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}) \propto \mathsf{Q}^{\circ}(x_i|x_{i-1})\mathsf{Q}^{\circ}(x_{i-1}|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}).$$

From (B.3), taking

$$\begin{aligned}\mathsf{Q}^{\circ}(x_i|x_{i-1}) &= \mathcal{N}(x_i; Ax_{i-1}, \mathsf{E}[\xi^{-1}]^{-1}Q), \\ \mathsf{Q}^{\circ}(x_{i-1}|\mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}) &= \mathcal{N}(x_{i-1}; x_{i-1|i-1}, P_{i-1|i-1}),\end{aligned}$$

we can write

$$\mathsf{Q}^{\circ}(x_{i-1}|x_i, \mathsf{F}_{\mathsf{S}}, \mathbf{z}_{i-1}) = \mathcal{N}(x_{i-1}; \tilde{x}_{i-1|i}, P_{i-1|i}), \tag{B.7}$$

where

$$\begin{aligned}\tilde{x}_{i-1|i} &= x_{i-1|i-1} + L(x_i - x_{i|i-1}), \\ P_{i-1|i} &= P_{i-1|i-1} - LP_{i|i-1}L^{\top},\end{aligned}$$

with L given in (32). Finally, inserting (B.6) and (B.7) in (B.4) yields (24d). □

References

1. Pan SJ. Transfer learning. In: *Data Classification: Algorithms and Applications.* Chapman and Hall/CRC; 2015:537–58. doi:10.1201/b17320.
2. Pan SJ, Yang Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 2010;22(10):1345–59. doi:10.1109/TKDE.2009.191.
3. Weiss K, Khoshgoftaar TM, Wang D. A survey of transfer learning. *Journal of Big Data* 2016;3(1):9. doi:10.1186/s40537-016-0043-6.
4. Bishop CM. Pattern recognition and machine learning. Springer; 2006.
5. Murphy KP. Machine Learning: A Probabilistic Perspective. MIT Press; 2012.
6. Wang S, Li Z, Yu Y, Xu J. Folding membrane proteins by deep transfer learning. *Cell systems* 2017;5(3):202–11. doi:10.1016/j.cels.2017.09.001.
7. Choi D, An TH, Ahn K, Choi J. Driving experience transfer method for end-to-end control of self-driving cars. *arXiv preprint arXiv:180901822* 2018;.
8. Mou L, Meng Z, Yan R, Li G, Xu Y, Zhang L, Jin Z. How transferable are neural networks in NLP applications? In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* 2016:479–89. doi:10.18653/v1/D16-1046.
9. Samala RK, Chan HP, Hadjiiski LM, Helvie MA, Cha KH, Richter CD. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Physics in Medicine & Biology* 2017;62(23):8894. doi:10.1088/1361-6560/aa93d4.
10. Karbalayghareh A, Qian X, Dougherty ER. Optimal Bayesian transfer learning. *IEEE Transactions on Signal Processing* 2018;66(14):3724–39. doi:10.1109/TSP.2018.2839583.
11. Bernardo JM, Smith AFM. Bayesian Theory. Wiley; 1994.
12. Torrey L, Shavlik J. Transfer learning. In: *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques.* IGI Global; 2010:242–64. doi:10.4018/978-1-60566-766-9.ch011.
13. Taylor ME, Stone P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research* 2009;10:1633–85. doi:10.1145/1577069.1755839.
14. Kárný M. Towards fully probabilistic control design. *Automatica* 1996;32(12):1719–22. doi:10.1016/S0005-1098(96)80009-4.
15. Kárný M, Kroupa T. Axiomatisation of fully probabilistic design. *Information Sciences* 2012;186(1):105–13. doi:10.1016/j.ins.2011.09.018.
16. Shore J, Johnson R. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory* 1980;26(1):26–37. doi:10.1109/TIT.1980.1056144.
17. Foley C, Quinn A. Fully probabilistic design for knowledge transfer in a pair of Kalman filters. *IEEE Signal Processing Letters* 2018;25(4):487–90. doi:10.1109/LSP.2017.2776223.
18. Quinn A, Kárný M, Guy TV. Optimal design of priors constrained by external predictors. *International Journal of Approximate Reasoning* 2017;84:150–8. doi:10.1016/j.ijar.2017.02.001.
19. Papež M, Quinn A. Dynamic Bayesian knowledge transfer between a pair of Kalman filters. In: *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP).* 2018:doi:10.1109/mlsp.2018.8517020.
20. Meinhold RJ, Singpurwalla ND. Robustification of Kalman filter models. *Journal of the American Statistical Association* 1989;84(406):479–86. doi:10.1080/01621459.1989.10478794.
21. Ting JA, Theodorou E, Schaal S. Learning an outlier-robust Kalman filter. In: *European Conference on Machine Learning.* Springer; 2007:748–56. doi:10.1007/978-3-540-74958-5_76.
22. Agamennoni G, Nieto JI, Nebot EM. An outlier-robust Kalman filter. In: *2011 IEEE International Conference on Robotics and Automation.* IEEE; 2011:1551–8. doi:10.1109/ICRA.2011.5979605.
23. Piché R, Särkkä S, Hartikainen J. Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate Student-t distribution. In: *2012 IEEE International Workshop on Machine Learning for Signal Processing.* IEEE; 2012:1–6. doi:10.1109/MLSP.2012.6349794.
24. Zhu H, Leung H, He Z. A variational Bayesian approach to robust sensor fusion based on Student-t distribution. *Information Sciences* 2013;221:201–14. doi:10.1016/j.ins.2012.09.017.
25. Nurminen H, Ardeshiri T, Piche R, Gustafsson F. Robust inference for state-space models with skewed measurement noise. *IEEE Signal Processing Letters* 2015;22(11):1898–902. doi:10.1109/LSP.2015.2437456.
26. Huang Y, Zhang Y, Li N, Chambers J. A robust Gaussian approximate filter for nonlinear systems with heavy tailed measurement noises. In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE; 2016:4209–13. doi:10.1109/ICASSP.2016.7472470.

27. Wang Z, Zhou W. Robust linear filter with parameter estimation under Student-t measurement distribution. *Circuits, Systems, and Signal Processing* 2019;38(6):2445–70. doi:10.1007/s00034-018-0972-8.

28. Gandhi MA, Mili L. Robust Kalman filter based on a generalized maximum-likelihood-type estimator. *IEEE Transactions on Signal Processing* 2010;58(5):2509–20. doi:10.1109/TSP.2009.2039731.

29. Huang Y, Zhang Y, Li N, Naqvi SM, Chambers J. A robust and efficient system identification method for a state-space model with heavy-tailed process and measurement noises. In: *2016 19th International Conference on Information Fusion (FUSION)*. IEEE; 2016:441–8.

30. Roth M, Özkan E, Gustafsson F. A Student's t filter for heavy tailed process and measurement noise. In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE; 2013:5770–4. doi:10.1109/ICASSP.2013.6638770.

31. Huang Y, Zhang Y, Li N, Wu Z, Chambers JA. A novel robust Student's t-based Kalman filter. *IEEE Transactions on Aerospace and Electronic Systems* 2017;53(3):1545–54. doi:10.1109/TAES.2017.2651684.

32. Dong P, Jing Z, Leung H, Shen K, Wang J. Student-t mixture labeled multi-Bernoulli filter for multi-target tracking with heavy-tailed noise. *Signal Processing* 2018;152:331–9. doi:10.1016/j.sigpro.2018.06.014.

33. Huang Y, Zhang Y, Shi P, Wu Z, Qian J, Chambers JA. Robust Kalman filters based on Gaussian scale mixture distributions with application to target tracking. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 2019;49(10):2082–96. doi:10.1109/TSMC.2017.2778269.

34. Šmídl V, Quinn A. The variational Bayes method in signal processing. Springer; 2006.

35. Blei DM, Kucukelbir A, McAuliffe JD. Variational inference: A review for statisticians. *Journal of the American Statistical Association* 2017;112(518):859–77. doi:10.1080/01621459.2017.1285773.

36. Willner D, Chang CB, Dunn KP. Kalman filter algorithms for a multi-sensor system. In: *Decision and Control including the 15th Symposium on Adaptive Processes, 1976 IEEE Conference on*; vol. 15. IEEE; 1976:570–4. doi:10.1109/CDC.1976.267794.

37. Särkkä S. Bayesian Filtering and Smoothing. Cambridge University Press; 2013.

38. Kim J, Tandale M, Menon P, Ohlmeyer E. Particle filter for ballistic target tracking with glint noise. *Journal of guidance, control, and dynamics* 2010;33(6):1918–21. doi:10.2514/1.51000.

39. Huang Y, Zhang Y. A new process uncertainty robust Student's-t based Kalman filter for SINS/GPS integration. *IEEE Access* 2017;5:14391–404. doi:10.1109/ACCESS.2017.2726519.

40. Roth M. On the multivariate t distribution. Tech. Rep. 3059; Division of Automatic Control, Linköping University; 2013.

41. Kalman RE. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* 1960;82(1):35–45. doi:10.1115/1.3662552.

42. Doucet A, Johansen AM. A tutorial on particle filtering and smoothing: Fifteen years later. In: Crisan D, Rozovsky B, eds. *The Oxford Handbook of Nonlinear Filtering*. Oxford University Press; 2009:.

43. Li XR, Jilkov VP. Survey of maneuvering target tracking. Part I. Dynamic models. *IEEE Transactions on aerospace and electronic systems* 2003;39(4):1333–64. doi:10.1109/TAES.2003.1261132.

44. Bar-Shalom Y, Li XR, Kirubarajan T. Estimation with applications to tracking and navigation: theory algorithms and software. Wiley; 2004.

45. Wang Y, Blei DM. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association* 2019;114(527):1147–61. doi:10.1080/01621459.2018.1473776.

46. Hu XL, Schön TB, Ljung L. A general convergence result for particle filtering. *IEEE Transactions on Signal Processing* 2011;59(7):3424–9. doi:10.1109/TSP.2011.2135349.

47. Papež M, Quinn A. Robust Bayesian transfer learning between Kalman filters. In: *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE; 2019:doi:10.1109/MLSP.2019.8918783.

48. Šmidl V, Quinn A. Variational Bayesian filtering. *IEEE Transactions on Signal Processing* 2008;56(10):5020–30. doi:10.1109/TSP.2008.928969.