Akademie věd České republiky
Ústav teorie informace a automatizace, v.v.i.

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Áron Hoffmann, Anthony Quinn

Ockham's Razor from a Fully Probabilistic Design Perspective

No. 2391 2021

GAČR 18-15970S

ÚTIA AV ČR, P.O.Box 18, 182 08 Prague, Czech Republic
Tel: (+420)266052422, Fax: (+420)286890378, Url: http://www.utia.cas.cz, E-mail: utia@utia.cas.cz

Any opinions and conclusions expressed in this report are those of the authors and do not necessarily represent the views of the Institute.

Contents

Abstract

This research report investigates an approach to the design of an Ockham prior penalising parametric complexity in the Hierarchical Fully Probabilistic Design (HFPD) [1] setting. We identify a term which penalises the introduction of an additional parameter in the Wold decomposition. We also derive the objective Ockham Parameter Prior (OPI) in this context, based on earlier work [2], and we show that the two are, in fact, closely related. This confers validity on the HFPD Ockham term.

1 Ockham's Razor in Parametric Inference

The desire to model a piece of data better pushes modellers to increase the complexity of their architectures and introduce additional explanatory variables. Beyond a certain point, this complexity does not correspond to the data generating process and begins to model the non-systematic (unpredictable or noise) portion of the data, to the detriment of its predictive ability. This phenomenon, called overfitting, is a significant and pervasive problem in the context of machine learning, signal processing, and related fields [3].

To address the parsimony-vs-prediction trade-off [2, 4], Ockham's Razor (i.e. the *Desideratum of Simplicity*) must be followed, stating that randomness must not be fitted with determinism. Importantly geometric (Least Squares) and Maximum A Posteriori (MAP) estimation cannot inherently fulfil this desideratum as they are conditioned on the full acceptance of the model and provide point estimates that lack the necessary context of a measure. As alternatives, several approaches have been developed to quantify complexity and embrace Ockham's razor in the model-based inference. Common examples are the Minimum Message Length (MML) criterion [5], the Minimum Description Length (MDL) criterion [6], and the Akaike Infromation Criterion (AIC) [7]. These act as *regularizers* of the model by optimising an *ad hoc* statistic associated with the model's complexity. In contrast, Quinn [2, 4] derives the axiomatically justified Ockham Parameter Inference (OPI) simply by using the probability calculus (i.e. the calculus of belief in the Bayesian perspective). This approach has been shown [8] to accurately assess whether there is sufficient evidence for increased model complexity in very general signal processing settings. Furthermore, it facilitates the simultaneous estimation of model order and parameters.

More recently, Fully Probabilistic Design (FPD) [9] and its generalisation, Hierarchical Fully Probabilistic Design (HFPD) [1] have been developed. These extend the paradigm of stochastic modelling to scenarios where establishing a joint model of the entire system may not be desirable or feasible. They describe an axiomatically justified approach to optimally designing probability models of uncertainty. (H)FPD has been successfully used in various scenarios, such as optimal control system design [10] and transfer learning [11, 12]. To date, however, the Desideratum of Simplicity has not been investigated from a (H)FPD point of view.

In this research report, we examine Ockham's Razor in the context of HFPD. We do this by adopting a simple parametric model for inference, and we use HFPD to assess it against an Ockham model, chosen as the 'ideal'. A term penalising the increase of complexity in adopting the parametric model emerges. We show that this term is proportional to the one derived through OPI theory [1, 9]. The Ockham term is evaluated in a simple signal processing context.

This report is laid out as follows: Section 2 specifies the parametric model setup used. It also specifies the notational conventions. Section 3 briefly reviews the Ockham Parameter Inference (OPI), instantiates it in the context of this report's parametric modelling, and shows explicitly how this OPI is a function of the model's complexity. Section 4 similarly introduces Hierarchical Fully Probabilistic Design (HFPD) and applies it to the modelling context of this report. It provides the main results by detailing an approach to eliciting Ockham's Razor via HFPD, and deriving an explicit HFPD-optimal function of complexity for the parametric signal modelling example in this report. Furthermore, it compares the functions of model complexity derived through HFPD and OPI theory. Section 5 concludes the report and lays out future directions for this research.

2 Parametric signal modelling in the additive noise setting

Consider two observers, the systematic observer and the Ockham observer, denoted as $\mathcal{I}_1$ and $\mathcal{I}_0$, respectively. They both observe and model the same data source, x. $\mathcal{I}_1$ believes that the data are a realisation of a signal-plus-additive-white-Gaussian-noise (AWGN) process, with signal $a\psi$ and noise variance r. Here, ψ is a known regressor, specific to the hypothesised signal, while a (the amplitude) and r are unknown probabilistic parameters. In contrast, $\mathcal{I}_0$ believes that the data are simply white Gaussian noise, with the *same* variance r. Thus, $\mathcal{I}_0$ acts as the Ockham observer in the context of the a parameter of $\mathcal{I}_1$. $\mathcal{I}_1$'s beliefs in the parameters a and r are described through hyperparameters V and ν of their normal-inverse-Gamma (NiG) prior for $\{a, r\}$ [13], where V is a 2×2 symmetric positive-definite matrix and ν is a positive scalar. Both are recursively computable statistics of the data [13]. The two observers agree in respect of their inverse-Gamma prior for their shared parameter, r. $\mathcal{I}_1$'s belief in a is then described by a Gaussian distribution conditional on r, resulting in $\mathcal{I}_1$'s NiG joint prior distribution for $\{a, r\}$. To summarise:

$$\begin{array}{ll} \mathcal{I}_1 & \mathcal{I}_0 \\ (x \mid a, \psi, r, \mathcal{I}_1) \sim \mathcal{N}_x(a\psi, r) & (x \mid r, \mathcal{I}_0) \sim \mathcal{N}_x(0, r) \\ (a, r \mid V, \nu, \mathcal{I}_1) \sim \mathcal{N}i\mathcal{G}_{a,r}(V, \nu) & (r \mid v_{11}, \nu, \mathcal{I}_0) \sim i\mathcal{G}_r\left(\frac{\nu - 3}{2}, \frac{v_{11}}{2}\right) \end{array} \tag{1}$$

$\mathcal{I}_0$'s prior for r is therefore equal to $\mathcal{I}_1$'s marginal prior for r [14].

$\mathcal{I}_1$'s conjugate update rules for computing the hyperparameters, V and ν, given a length-$n \geq 1$ realisation, $\{x_1, \ldots, x_n\}$, of the data source, are

$$\begin{aligned} V_n &= V + \sum_{i=1}^{n} \begin{bmatrix} x_i \\ \psi_i \end{bmatrix} \begin{bmatrix} x_i & \psi_i \end{bmatrix} \qquad (2) \\ \nu_n &= \nu + n. \end{aligned}$$

2.1 Notation

The following notational conventions are used in this document:

- x_i, ψ_i, ν_i, and V_i are time series data/statistics. However, the time index, $_i$, is omitted where feasible.
- The 2×2 matrix, V_i, has elements $\begin{bmatrix} v_{i,11} & v_{i,1a} \\ v_{i,a1} & v_{i,aa} \end{bmatrix}$ where $v_{i,a1} = v_{i,1a}$
- $\mathcal{D}_x(P \parallel Q)$ denotes the Kullback-Leibler divergence from $P(x)$ to $Q(x)$ over x, where these are probability distribution functions with respect to the Lebesgue measure, and have common support in x:

$$\mathcal{D}_x(P \parallel Q) = \int_x P(x) \ln\left(\frac{P(x)}{Q(x)}\right) dx \tag{3}$$

3 Ockham Parameter Inference (OPI)

The Ockham parameter inference (OPI) [2, 4, 8] asserts an additive noise setting — as explained above —, with the same noise model for both the systematic and the Ockham observer. It yields $\mathcal{I}_0$'s non-heuristic distribution over $\mathcal{I}_1$'s hypothesis space and is a monotonically decreasing function of the complexity of the latter. Effectively, it is $\mathcal{I}_0$'s distribution of the signals expected by $\mathcal{I}_1$:

$$p_s(s|\mathcal{I}_0) \tag{4}$$

Being a distribution (4), the OPI can be focused, via marginalisation, thereby quantifying the evidence for inclusion of only a subset of $\mathcal{I}_1$'s signal parameters.

The observers of this document are defined as in Eq. (1), with $\mathcal{I}_1$'s model for additive noise, e, in their Wold model being adopted also by $\mathcal{I}_0$. Specifically:

$$\begin{array}{ll} \mathcal{I}_1 & \mathcal{I}_0 \\ (x \mid a, \psi, r, \mathcal{I}_1) \overset{\mathrm{d}}{=} s(a,\psi) + (e \mid r) & (x \mid r, \mathcal{I}_0) \overset{\mathrm{d}}{=} (e \mid r) \\ \equiv a\psi + (e \mid r) & \end{array} \tag{5}$$

$$\begin{aligned} (e \mid r) &\sim \mathcal{N}_e(0, r) \equiv p_e(e \mid r) \\ (r \mid v_{11}, \nu) &\sim i\mathcal{G}_r\left(\frac{\nu-3}{2}, \frac{v_{11}}{2}\right) \end{aligned} \tag{6}$$

In Eq. (5), $\overset{\mathrm{d}}{=}$ denotes equality in distribution.

If the noise variance, r, is assumed known by $\mathcal{I}_1$ and $\mathcal{I}_0$ — as in the original setting for the OPI in [4, 8] — then the OPI (complexity prior) is given by

$$\begin{aligned} p_s(s(a,\psi) \mid \psi, r, \mathcal{I}_0) &= p_e(a\psi \mid r, \psi) \\ &= \mathcal{N}_a\left(0, \frac{r}{\psi^2}\right) \\ &\propto \exp\left(\frac{(a\psi)^2}{r}\right) \end{aligned} \tag{7}$$

However, if the variance is only known probabilistically, as in Eq. (6), then $\mathcal{I}_0$'s the complexity prior for a — being the local, systematic parameter adopted by $\mathcal{I}_1$ but rejected by $\mathcal{I}_0$ — is deduced via

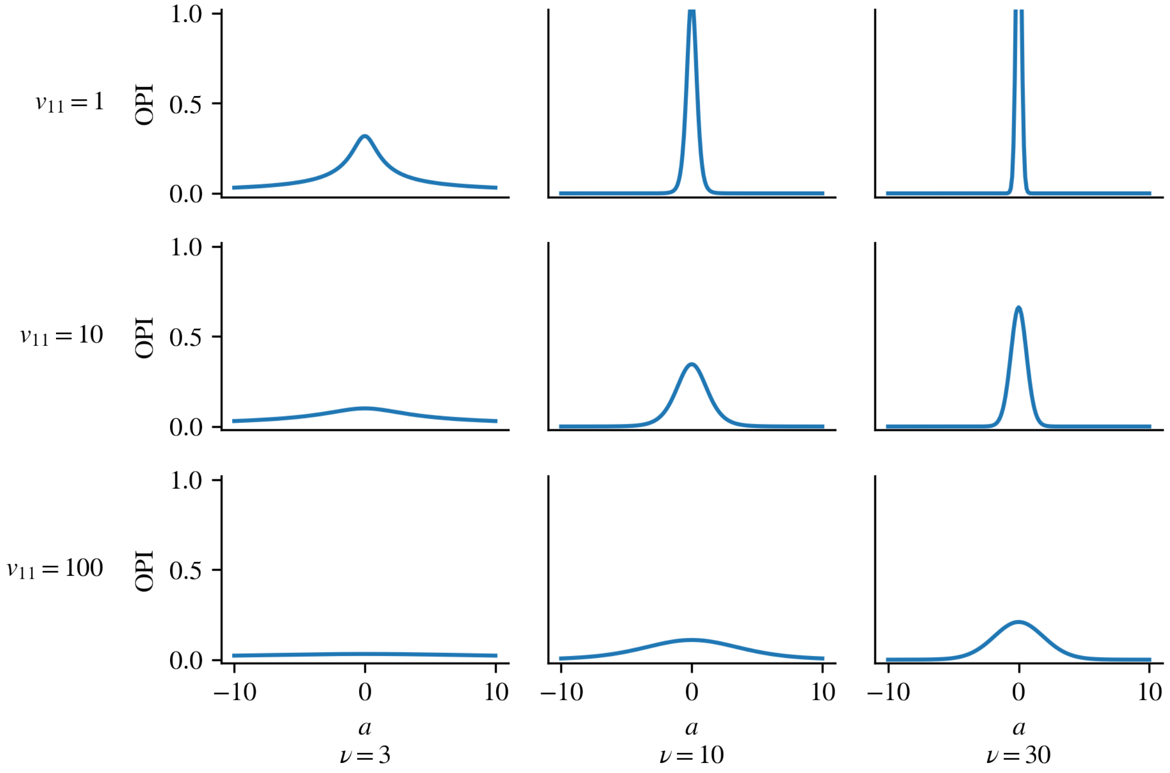

Figure 1: Plots of the Ockham Parameter Inference Eq. (8) at various combinations of v_{11} and ν and keeping the known regressor ψ constant (specifically, $v_{11} \in \{1, 10, 100\}$, $\nu \in \{3, 10, 30\}$, and $\psi = 1$).

marginalisation:

$$\begin{aligned} p_s(s(a,\psi) \mid \psi, v_{11}, \nu, \mathcal{I}_0) &= \int_r \mathcal{N}_a\left(0, \frac{r}{\psi^2}\right)\; i\mathcal{G}_r\left(\frac{\nu-3}{2}, \frac{v_{11}}{2}\right)\, dr \\ &= \mathcal{S}t\left(0, \frac{v_{11}}{\psi^2(\nu-3)}, \nu - 3\right) \text{ if } \nu > 3 \\ &= \frac{\Gamma\left(\frac{\nu-2}{2}\right)}{\sqrt{\pi \frac{v_{11}}{\psi^2}}\Gamma\left(\frac{\nu-3}{2}\right)} \left(\frac{v_{11}}{(a\psi)^2 + v_{11}}\right)^{\frac{\nu-2}{2}} \quad \text{if } \nu > 3 \\ &\propto \left(\frac{v_{11}}{(a\psi)^2 + v_{11}}\right)^{\frac{\nu-2}{2}} \quad \text{if } \nu > 3 \end{aligned} \tag{8}$$

Eqs. (7) and (8) are symmetric around their mode at $a = 0$. Additionally, Eq. (8), illustrated in Fig. 1 concentrates as ν increases and relaxes as $\frac{v_{11}}{\psi^2}$ increases. These distributions express $\mathcal{I}_0$'s preference for the rejection of complexity via choices close to $a = 0$, unless there is sufficient evidence otherwise.

4 Fully Probabilistic Design (FPD)

Fully probabilistic Design [9], in contrast with OPI, uses decision-theoretic arguments in settings where no joint model is posited. Through the works of [15] and [16], it asserts that knowledge processing is to be performed through the optimisation of a Bayes' *risk*, which it derives to be the Kullback-Leibler (KL) divergence 3. Hierarchical FPD (HFPD) [1] allows FPD-based knowledge processing in scenarios where the distributions themselves may not be known exactly, but only up to a distribution.

The full description of a model M is given by Eq. (9), where x is the modelled quantity, A is the probability distribution of x, and S is the hyper-distribution of A (or its parameters, if A is parametric). K represents the complete knowledge possessed by the model:

$$M(x, A|S, K) = A(x|K)S(A|K) \tag{9}$$

The FPD-optimal design (i.e. minimum-KLD choice) of S is given by S^O (10), where A_I and S_I represent the designer's *ideal* beliefs. The ideal distribution may be thought of as the stochastic model for (x, A) which the modeller would choose were it not for the belief constraints imposed by K:

$$S^O(A|K) \propto S_I(A|K) \exp\big(-\mathcal{D}_x(A \parallel A_I)\big) \tag{10}$$

Here, the constant of proportionality is given by

$$c_{S^O} = \int_{\boldsymbol{A}} S_I(A|K) \exp\big(-\mathcal{D}_x(A \parallel A_I)\big) dA \tag{11}$$

In the current context, we are assuming that $A(x)$ is finitely parameterised, that its optimal (hyper-) prior measure has density, $S^O(\cdot)$, with respect to Lebesgue measure. We instantiate the terms in (10) consistently with respect to the agreements in Section 2 (see Eq. (1)), as follows:

$$\begin{array}{ll} \text{FPD (variational) modeller, } \mathcal{I}_1: & \text{Ideal modeller, } \mathcal{I}_I \equiv \mathcal{I}_0: \\ A(x|a, r, \psi, \mathcal{I}_1) \equiv \mathcal{N}_x(a\psi, r) & A_I(x|r, \mathcal{I}_I) \equiv \mathcal{N}_x(0, r) \\ S(a, r|V, \nu, \mathcal{I}_1) \equiv \mathcal{N}i\mathcal{G}_{a,r}(V, \nu) & S_I(r|\nu_I, v_{I,11}, \mathcal{I}_I) \equiv i\mathcal{G}_r\left(\frac{\nu_I - 3}{2}, \frac{v_{I,11}}{2}\right) \end{array} \tag{12}$$

In this way, we posit the designer's ideal model to be the Ockham observer's $\mathcal{I}_0$ model of Sections 2 and 3. Furthermore, the designer's knowledge constraint, K, is precisely $\mathcal{I}_1$'s parametric (Wold) model (1) with respect to the known signal, ψ. Inserting these factors into (10):

$$S^O(r|\nu, \nu_{11}, a, \psi) \propto i\mathcal{G}_r\left(\frac{\nu - 3}{2}, \frac{v_{11}}{2}\right) \exp\big(-\mathcal{D}_x(\mathcal{N}_x(0, r) \parallel \mathcal{N}_x(a\psi, r)\big) \tag{13}$$

$$\tag{14}$$

$$= i\mathcal{G}_r\left(\frac{\nu - 3}{2}, \frac{v_{11}}{2}\right) \exp\left(-\frac{(a\psi)^2}{2r}\right) \tag{15}$$

The normalising constant is

$$c_{S^O} = \left(\frac{v_{11}}{(a\psi)^2 + v_{11}}\right)^{\frac{\nu - 2}{2}}, \text{ if } \quad \nu > 3. \tag{16}$$

Note that this normalising constant of the FPD-optimal model for the parameter of $\mathcal{I}_0$, i.e. r, is, itself, an unnormalised measure of complexity for the *additional* parameter of $\mathcal{I}_1$, i.e. a (12). It exhibits the same complexity-penalizing behaviour as the OPI (Section 3). The behaviour of c_{S^O} as a function of a is illustrated in Fig. 2, for various settings of ν and ν_{11}, holding the regressor constant at $\psi = 1$.

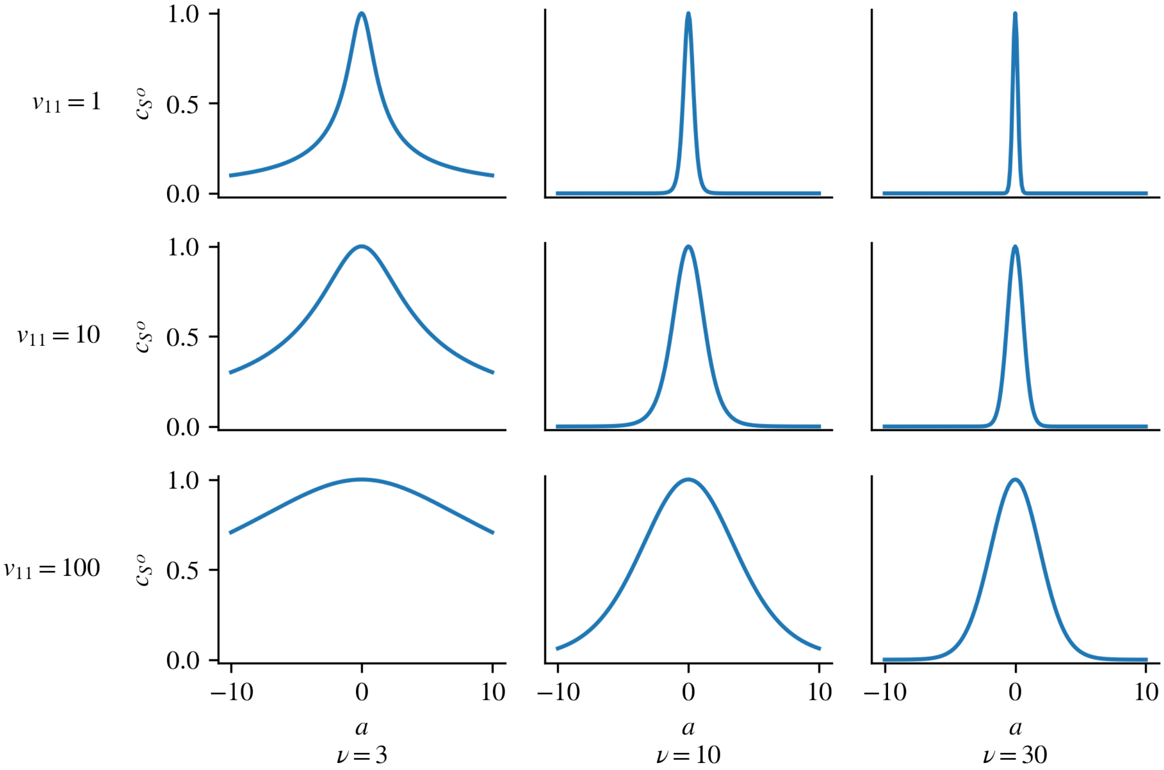

Figure 2: Plots of the HFPD Ockham factor c_{SO} at various combinations of v_{11} and ν and keeping the known regressor ψ constant (specifically, $v_{11} \in \{1, 10, 100\}$, $\nu \in \{3, 10, 30\}$, and $\psi = 1$).

4.1 Connection with OPI-based Ockham's Razor

The established OPI theory evaluates the systematic model through the viewpoint of the Ockham observer. Similarly in the FPD setting, the systematic modeller transfers knowledge from the ideal Ockham modeller. The normalising constant of the resulting distribution, c_{SO} (16) can be re-interpreted as a function of the additional parameter of the systematic modeller, a, where it acts as a penalty on model complexity (Fig. 2). Additionally, it is proportional to the marginalised OPI complexity measure Eq. (8), differing only by a normalising constant. This is further indication to utility of the approach outlined in Section 4 for penalising model complexity.

5 Conclusions and future work

This work is an initial exploration towards identifying and establishing an Ockham's razor in the context of Fully Probabilistic Design. We use a simple parametric signal processing context (the Wold representation of the observation process) and demonstrate how (H)FPD can elicit Ockham-sensitive priors when the ideal is chosen to express the designer's simplicity objective. We also draw a connection between this approach and Ockham Parameter Inference (OPI)-based complexity penalisation.

More work is left to be done to establish a thorough theory of simplicity in FPD. A rigorous forward-design of the knowledge constraints and the ideals needs to be performed. The scalar parametric additive noise (i.e. Wold) setup, which provides the context for the current paper, should be expanded to include multivariate and multi-parameter scenarios, beyond additive noise. In particular, it should be investigated how the 'temperature' parameter of HFPD influences the elicitation of Ockham's razor. Additionally, the opportunity to relax $A(x)$ nonparametrically, yield an optimal Ockham-regularised, nonparametric prior, $S^O(A \mid K)$, should be explored further via the hierarchical FPD (HFPD) framework [1]. Finally, an algorithm to use the (H)FPD-optimal Ockham prior in simultaneous model order selection and parameter estimation must be developed. Crucially, its ability to obviate overfitting and complexity must be validated with simulations and real-world data.

References

[1] A. Quinn, M. Kárný, and T. V. Guy, "Fully probabilistic design of hierarchical bayesian models," *Information Sciences*, vol. 369, pp. 532–547, 2016.

[2] A. Quinn, "A new objective measure of signal complexity using bayesian inference," in *IEEE Seventh SP Workshop on Statistical Signal and Array Processing*, pp. 79–82, IEEE, 1994.

[3] J. M. Zhang, M. Harman, L. Ma, and Y. Liu, "Machine learning testing: Survey, landscapes and horizons," *IEEE Transactions on Software Engineering*, 2020.

[4] A. P. Quinn, *Bayesian point inference in signal processing.* PhD thesis, University of Cambridge, 1992.

[5] C. S. Wallace and D. M. Boulton, "An information measure for classification," *The Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.

[6] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of statistics*, pp. 416–431, 1983.

[7] H. Akaike *et al.*, "On entropy maximization principle.," 1977.

[8] A. Quinn, "Censored marginal a posteriori bayesian inference for signal models," in *IEEE Winter Workshop on Nonlinear Digital Signal Processing*, pp. 6–3, IEEE, 1993.

[9] M. Kárný, "Axiomatisation of fully probabilistic design revisited," *Systems & Control Letters*, vol. 141, p. 104719, 2020.

[10] M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař, *Optimized Bayesian dynamic advising.* Springer London, 2005.

[11] A. Barber and A. Quinn, "Robust bayesian transfer learning between autoregressive inference tasks," in *2021 32st Irish Signals and Systems Conference (ISSC)*, 2021.

[12] 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)*, pp. 1–6, IEEE, 2018.

[13] V. Smidl and A. Quinn, "Mixture-based extension of the ar model and its recursive bayesian identification," *IEEE Transactions on Signal Processing*, vol. 53, no. 9, pp. 3530–3542, 2005.

[14] J. M. Bernardo and A. F. Smith, *Bayesian theory*, vol. 405. John Wiley & Sons, 2009.

[15] L. J. Savage, *The foundations of statistics.* Courier Corporation, 1972.

[16] J. M. Bernardo, "Expected information as expected utility," *the Annals of Statistics*, pp. 686–690, 1979.