A PARAMETRIC FAMILY OF BAYESIAN ESTIMATORS FOR NON-STANDARD LOSS FUNCTIONS

Stefan Uhlich and Bin Yang

Chair of System Theory and Signal Processing, Universität Stuttgart, Pfaffenwaldring 47, 70550 Stuttgart, Germany
{stefan.uhlich,bin.yang}@lss.uni-stuttgart.de, http://www.lss.uni-stuttgart.de

ABSTRACT

This paper introduces a new parametric family of Bayesian estimators. As the estimation with non-standard loss functions can often only be stated as an optimization problem which has to be solved for each new observation, it is advantageous to use such a parametric family. We proof that many well known estimators are included in our family. Among them are the MMSE and MAP estimator as well as the optimal Bayesian estimator (OBE) under LINEX loss. By restricting the estimator to lie in this family, we split the estimation problem into two parts: In a first step, we have to find the best estimator with respect to the Bayes risk for a given loss function, which has to be done only once. The second step then calculates the estimate for a given observation. We demonstrate the usefulness of the proposed parametric family in an example.

1. INTRODUCTION

Most often in Bayesian estimation, the MAP or MMSE estimators are used to estimate an unknown parameter $\underline{\theta} \in \Theta \subset \mathbb{R}^M$ from the observation $\underline{x} \in \mathbb{R}^N$. It is well known that the underlying loss functions $L(\underline{\theta}, \hat{\underline{\theta}})$ are the hit-or-miss loss and the quadratic loss [1]

$$L_{\text{MAP}}(\underline{\theta}, \hat{\underline{\theta}}) = \begin{cases} 1 & \|\underline{\theta} - \hat{\underline{\theta}}\| > \delta \\ 0 & \|\underline{\theta} - \hat{\underline{\theta}}\| < \delta \end{cases} \quad \text{and} \tag{1a}$$

$$L_{\text{MMSE}}(\underline{\theta}, \hat{\underline{\theta}}) = (\underline{\theta} - \hat{\underline{\theta}})^T \mathbf{W} (\underline{\theta} - \hat{\underline{\theta}}), \ \mathbf{W} \text{ pos. def.} \tag{1b}$$

The reason that they are used so widely is often not their suitability to the problem at hand but that the corresponding optimal Bayesian estimators (OBE) are well known and, at least for the MAP estimator, are often computable. They are the maximum and mean of the a posteriori density $p(\underline{\theta}|\underline{x})$. Powerful methods are available to calculate the estimate $\hat{\underline{\theta}}$ from an observation $\underline{x}$, ranging from optimization algorithms [1] and the Expectation-Maximization (EM) algorithm [2] to sampling techniques including Markov chain Monte Carlo methods [3].

In this paper, we will consider Bayesian estimation with other than those loss functions given in (1). This problem is very important for practical applications as the loss function should reflect the cost that is connected with a certain estimation error, see. e.g. [4,5]. The following two examples illustrate this more clearly:

- Consider the problem of constructing a dam [6]. Underestimating the peak water level from older measurements is clearly more serious than overestimating it and this fact should be reflected in the choice of the loss function $L(\underline{\theta}, \hat{\underline{\theta}})$. This example motivates the use of an asymmetric loss function and it is obvious that the two loss functions in (1) are not suited for such an estimation problem.
- Another example that gives rise to other loss functions than those given in (1) can be found in the field of image processing. Traditionally, the mean squared error is used to compare images and therefore many algorithms are optimized for this loss function [7]. The problem with the MSE is that it does not well represent the human perception. Images which have a small mean squared error may still look very different and therefore in [7] it is suggested to use other distance measures. One is the structural similarity (SSIM) index, which was introduced by Wang in [8] and e.g. used in [9] for the design of linear equalizers. Another related example that discusses the design of loss functions for the reconstruction of images is given by Rue in [10].

However, calculating the OBE for many non-standard loss functions is not trivial and it can often only be stated in terms of an optimization problem which has to be solved for each new observation $\underline{x}$. Therefore, we propose in this paper a parametric family $\mathcal{F}$ of estimators which are suited for a large variety of loss functions but still have a computationally complexity comparable to the MMSE estimator for the same problem. Thus, using the best estimator in $\mathcal{F}$ that has the smallest Bayes risk for a given loss function will be a good approximation of the OBE. Our parametric family of estimators can be viewed as a compromise between the perfect OBE at the one side and a (nonlinear) regression approach on the other. It trades off performance against computational complexity as it will have a larger Bayes risk than the OBE but will be easier to learn due to the small and fixed number of parameters compared to a regression approach.

This paper is organized as follows: First, we review in Sec. 2 the Bayesian estimation problem and introduce the OBE which minimizes the Bayes risk. As the OBE can often not be computed in closed form, we propose in Sec. 3 and 4 two new parametric families of estimators. We start in Sec. 3 by considering a basic family $\mathcal{F}_{\mathcal{B}}$ of estimators, which includes the MMSE and the MAP estimator. This family, however, has the disadvantage that the underlying loss functions are always symmetric. Therefore, we generalize the estimator family in Sec. 4. This generalized family $\mathcal{F}$ also includes the OBE under LINEX loss and is thus more versatile. In Sec. 5 we consider the general approach how to use the estimator family and discuss its computational complexity. We show that we can use importance sampling to efficiently compute an estimate. Finally, an example in Sec. 6 demonstrates the usefulness of our parametric family to approximate the OBE.

Following notations are used throughout this paper: $\underline{x}$ denotes a column vector, $\mathbf{X}$ a matrix and in particular $\mathbf{I}$ the identity matrix. The trace operator, determinant, matrix transpose and euclidean norm are denoted by $\text{tr}\{.\}$, $\det\{.\}$, $(.)^T$ and $\|.\|$, respectively. $\text{diag}\{\underline{x}\}$ returns a squared matrix which has the elements of $\underline{x}$ on its diagonal. Finally, $\mathbf{X} \circ \mathbf{Y}$ denotes the elementwise product, also known as Hadamard product.

2. REVIEW OF BAYESIAN ESTIMATION

Suppose we have an estimator $\hat{\underline{\theta}}(\underline{x})$ that estimates the unknown, random parameter $\underline{\theta} \in \mathbb{R}^M$ from the observation $\underline{x} \in \mathbb{R}^N$. To evaluate the quality of the estimator, we assign a loss $L(\underline{\theta}, \hat{\underline{\theta}}) \geq 0$ to the error of estimating $\hat{\underline{\theta}}(\underline{x})$ although the true value is $\underline{\theta}$. Following different types of loss functions can be distinguished:

Definition. *A loss function $L(\underline{\theta}, \hat{\underline{\theta}})$ is called*
(i) symmetric, if $L(\underline{\theta}, \hat{\underline{\theta}}) = L(-\underline{\theta}, -\hat{\underline{\theta}})$;
(ii) spherical, if $L(\underline{\theta}, \hat{\underline{\theta}}) = \tilde{L}(\|\underline{\theta} - \hat{\underline{\theta}}\|)$.

Note, that a spherical loss function is also symmetric but the converse is in general not true. An example is the loss (1b) which is symmetric but not spherical for $\mathbf{W} \neq \alpha \mathbf{I}$. Besides these two properties, scale invariance [11, 12] and boundedness [13, 14] are other

 ISSN 2076-1465

characteristics of the loss function that may be desired for practical applications.

Averaging this loss with respect to the joint probability density function (PDF) $p(\underline{\theta},\underline{x})$ yields an important characteristic value for an estimator. It is called the Bayes risk (BR) and given by [15]

$$\mathrm{BR} = \iint L(\underline{\theta},\underline{\hat{\theta}}(\underline{x}))p(\underline{\theta},\underline{x})d\underline{\theta}d\underline{x}. \tag{2}$$

The optimal Bayesian estimator (OBE) is now that estimator that minimizes the Bayes risk, i.e.

$$\begin{aligned}\underline{\hat{\theta}}_{\mathrm{OBE}}(\underline{x}) &= \arg\min_{\underline{\hat{\theta}}(\underline{x})} \mathrm{BR} = \arg\min_{\underline{\hat{\theta}}(\underline{x})} \iint L(\underline{\theta},\underline{\hat{\theta}}(\underline{x}))p(\underline{\theta},\underline{x})d\underline{\theta}d\underline{x} \\ &= \arg\min_{\underline{\hat{\theta}}(\underline{x})} \int L(\underline{\theta},\underline{\hat{\theta}}(\underline{x}))p(\underline{\theta}|\underline{x})d\underline{\theta} \end{aligned} \tag{3}$$

where we used in the last line of (3) the fact that $\underline{\hat{\theta}}$ is a function of $\underline{x}$ and thus $\arg\min_{\underline{\hat{\theta}}}\mathrm{BR}$ is equivalent to minimizing the loss averaged over the a posteriori distribution. Therefore, we immediately see that all information to find the OBE is included in the a posteriori density $p(\underline{\theta}|\underline{x})$.

Assuming that the loss function $L(\underline{\theta},\underline{\hat{\theta}})$ is differentiable, we can calculate the first derivative with respect to the estimate and equate it to zero to obtain a necessary condition to find the OBE, i.e.

$$\frac{\partial}{\partial\underline{\hat{\theta}}}\int L(\underline{\theta},\underline{\hat{\theta}}(\underline{x}))p(\underline{\theta}|\underline{x})d\underline{\theta} = \int \frac{\partial L(\underline{\theta},\underline{\hat{\theta}})}{\partial\underline{\hat{\theta}}}p(\underline{\theta}|\underline{x})d\underline{\theta} \stackrel{!}{=} \underline{0}. \tag{4}$$

Solving (4) can often not be done analytically and therefore Bayesian estimation with most loss functions is difficult. We will thus introduce in the next section a parametric family $\mathcal{F}_{\mathcal{B}}$ of estimators that will transform (3) into an optimization problem to find one parameter. This family is then extended in Sec. 4 to asymmetric loss functions.

3. BASIC FAMILY OF ESTIMATORS

3.1 Definition

Let $\mathcal{F}_{\mathcal{B}}$ be the set of estimators that have the form

$$\underline{\hat{\theta}}(\underline{x};\lambda) = \frac{\int \underline{\theta}\, p(\underline{\theta},\underline{x})^{\lambda}d\underline{\theta}}{\int p(\underline{\theta},\underline{x})^{\lambda}d\underline{\theta}} \tag{5}$$

and are parameterized by λ. We call $\mathcal{F}_{\mathcal{B}}$ the *basic family of estimators*. Thinking of $p(\underline{\theta},\underline{x})^{\lambda}$ as a new (unnormalized) density, we see that (5) calculates the mean of the conditional density $p(\underline{\theta},\underline{x})^{\lambda}/\int p(\underline{\theta},\underline{x})^{\lambda}d\underline{\theta}$ and therefore looks similar to the MMSE estimator except for the modified PDF.

Note that it is reasonable to restrict λ to positive values, i.e. $\lambda \in [0,\infty)$. Otherwise we average over a new density $p(\underline{\theta},\underline{x})^{\lambda}/\int p(\underline{\theta},\underline{x})^{\lambda}d\underline{\theta}$ which is inverted in the sense that it has large values at positions where $p(\underline{\theta}|\underline{x})$ is small, i.e. it emphasizes points $(\underline{\theta},\underline{x}) \in \mathbb{R}^{M+N}$ that are not likely to occur and we can expect therefore a poor performance for $\lambda < 0$.[1]

3.2 Relationship to other Estimators

In this section, we will show that $\mathcal{F}_{\mathcal{B}}$ includes three important estimators, namely the uniform a priori MMSE estimator, the MMSE estimator and the MAP estimator. By uniform a priori MMSE estimator, we refer to the estimator where we have no observation $\underline{x}$ about $\underline{\theta} \in \Theta \subset \mathbb{R}^M$ and the a priori distribution $p(\underline{\theta})$ is assumed to be uniform in Θ. The estimator with the minimum MSE is then the "center of gravity" of Θ, i.e. $\underline{\hat{\theta}} = E[\underline{\theta}] = \int \underline{\theta}p(\underline{\theta})d\underline{\theta} = \int_{\Theta}\underline{\theta}d\underline{\theta}/\int_{\Theta}1d\underline{\theta}$ which is well defined if Θ is bounded. The following theorem proofs that all three estimators are in $\mathcal{F}_{\mathcal{B}}$.

Theorem 1. *The estimator family $\mathcal{F}_{\mathcal{B}}$ defined in* (5) *includes the following special cases:*

(a) If $\Theta \subset \mathbb{R}^M$ is bounded and $p(\underline{\theta},\underline{x}) \neq 0$ then $\underline{\hat{\theta}}(\underline{x};\lambda)$ for $\lambda = 0$ exists and is equivalent to the uniform a priori MMSE estimator.
(b) The case $\lambda = 1$ corresponds to the MMSE estimator.
(c) The case $\lambda \to \infty$ corresponds to the MAP estimator.

Proof. (a) Assuming Θ to be a bounded set on $\mathbb{R}^M$, we immediately see that $\lim_{\lambda\to 0} p(\underline{\theta},\underline{x})^{\lambda}/\int p(\underline{\theta},\underline{x})^{\lambda}d\underline{\theta} = \text{const.}$, i.e. it converges pointwise to a uniform distribution on Θ. Therefore, $\underline{\hat{\theta}}(\underline{x};0)$ calculates the center of gravity of Θ which is equivalent to the a priori MMSE estimator.

(b) Setting $\lambda = 1$ in (5), we obtain $p(\underline{\theta},\underline{x})/\int p(\underline{\theta},\underline{x})d\underline{\theta} = p(\underline{\theta}|\underline{x})$ and thus $\underline{\hat{\theta}}(\underline{x};1) = \int \underline{\theta}\, p(\underline{\theta}|\underline{x})d\underline{\theta} = \mathrm{E}[\underline{\theta}|\underline{x}]$, which is the MMSE estimator.

(c) To proof this theorem, we use a result from Pincus [16]: Given a continuous function $f(\underline{\theta})$, which attains a global maximum at exactly one point in Θ, then Pincus showed

$$\arg\max_{\underline{\theta}} f(\underline{\theta}) = \lim_{\lambda\to\infty} \frac{\int_{\Theta}\underline{\theta}\, f(\underline{\theta})^{\lambda}d\underline{\theta}}{\int_{\Theta} f(\underline{\theta})^{\lambda}d\underline{\theta}}. \tag{6}$$

Using this theorem, we conclude that $\lim_{\lambda\to\infty}\underline{\hat{\theta}}(\underline{x};\lambda)$ is the MAP estimator. ■

3.3 Corresponding Loss Functions

Although it is interesting to see the relationship of this basic family of estimators to other estimators, we also see that the loss functions associated with $\lambda = \{0,1,\infty\}$ are all symmetric as they are the hit-or-miss error (1a) and the squared error (1b). In the following, we will proof in Theorem 2 that if there is a continuously differentiable loss function that results in $\underline{\hat{\theta}}(\underline{x};\lambda)$, then the loss function has to be symmetric.[2] For the proof of Theorem 2, we need the following Lemma. The proofs can be found in the appendix A.

Lemma. *The estimator $\underline{\hat{\theta}}(\underline{x};\lambda)$ for the PDFs $p(\underline{\theta},\underline{x}) = \delta(\underline{\theta}-\underline{\theta}_0)$ and $p(\underline{\theta},\underline{x}) = P\delta(\underline{\theta}-\underline{\theta}_0)+(1-P)\delta(\underline{\theta}-\underline{\theta}_1)$ is given by $\underline{\hat{\theta}}(\underline{x};\lambda) = \underline{\theta}_0$ and $\underline{\hat{\theta}}(\underline{x};\lambda) = (P^{\lambda}\underline{\theta}_0+(1-P)^{\lambda}\underline{\theta}_1)/(P^{\lambda}+(1-P)^{\lambda})$, respectively.*

Theorem 2. *Let $L(\underline{\theta},\underline{\hat{\theta}})$ be a continuously differentiable loss function that results in the optimal Bayesian estimator $\underline{\hat{\theta}}(\underline{x};\lambda)$ for an arbitrary PDF $p(\underline{\theta},\underline{x})$. Then $L(\underline{\theta},\underline{\hat{\theta}})$ is symmetric, i.e. $L(\underline{\theta},\underline{\hat{\theta}}) = L(-\underline{\theta},-\underline{\hat{\theta}})$.*

From this Theorem, we see that no estimator resulting from an asymmetric, continuously differentiable loss function is included in $\mathcal{F}_{\mathcal{B}}$. However, we would like to cover such estimation problems due to their practical relevance and hence we have to extend $\mathcal{F}_{\mathcal{B}}$. This is done in the next section.

4. GENERALIZATION TO ASYMMETRIC LOSS FUNCTIONS

Let $\mathcal{F}$ be the set of estimators where each estimator has the form

$$\underline{\hat{\theta}}(\underline{x};\mathcal{P}) = \underline{f}_1\left(\frac{\int \underline{f}_2(\underline{\theta};\mathcal{P}_2)\, p(\underline{\theta},\underline{x})^{\lambda}d\underline{\theta}}{\int p(\underline{\theta},\underline{x})^{\lambda}d\underline{\theta}};\mathcal{P}_1\right) \tag{7a}$$

and depends on the $2M+4$ parameters $\mathcal{P} = \{\lambda,\mathcal{P}_1,\mathcal{P}_2\}$ with $\mathcal{P}_1 = \{\xi_1,\underline{\phi},\ldots,\phi_M\}$ and $\mathcal{P}_2 = \{\xi_2,\xi_3,\underline{\psi},\ldots,\psi_M\}$. $\underline{f}_1$ and $\underline{f}_2$ are defined as

$$\underline{f}_1(\underline{z};\mathcal{P}_1) = \xi_1\underline{z} + \underline{\phi}\circ\ln|\underline{z}|, \tag{7b}$$

$$\underline{f}_2(\underline{z};\mathcal{P}_2) = \xi_2\underline{z} + \xi_3 e^{\underline{\psi}\circ\underline{z}} \tag{7c}$$

with $\underline{\phi} = [\phi_1,\ldots,\phi_M]^T$ and $\underline{\psi} = [\psi_1,\ldots,\psi_M]^T$. Note that $e^{\underline{z}}$, $\ln\underline{z}$ and $|\underline{z}|$ are understood elementwise. λ is again chosen such that $\lambda \in [0,\infty)$ as discussed in 3.1.

[1] For example the loss $L(\underline{\theta},\underline{\hat{\theta}}) = 1 - L_{\mathrm{MAP}}(\underline{\theta},\underline{\hat{\theta}})$ results in seeking the minimum of $p(\underline{\theta}|\underline{x})$ which is related (but in general not identical) to $\underline{\hat{\theta}}(\underline{x};\lambda)$ for $\lambda \to -\infty$.

[2] Note that it is difficult to proof the existence of such a loss function for an arbitrary λ and the corresponding estimator $\underline{\hat{\theta}}(\underline{x};\lambda)$.

Note that $\mathcal{F}_{\mathcal{B}} \subset \mathcal{F}$ as all estimators $\hat{\underline{\theta}}(\underline{x};\lambda)$ from (5) are included in $\hat{\underline{\theta}}(\underline{x};\mathcal{P})$ for $\xi_1 = \xi_2 = 1, \xi_3 = 0$ and $\phi_1 = \cdots = \phi_M = 0$. Therefore, we already know from Theorem 1 that the uniform a priori MMSE, the MMSE and the MAP estimator are included in this family.

In the following, we will show that the LINEX loss is also included in (7). The LINEX loss is frequently used in Bayesian estimation, see e.g. [6, 17]. It rises approximately linear on one side and exponential on the other. The univariate LINEX loss function is given by $L_{\text{LINEX}}(\theta,\hat{\theta}) = b\left(e^{a\Delta} - a\Delta - 1\right)$ where $\Delta = \hat{\theta} - \theta$, $a \neq 0$ and $b > 0$. The multivariate LINEX loss is defined as a straightforward extension and given by [6]

$$L(\underline{\theta},\hat{\underline{\theta}}) = \sum_{m=1}^{M} b_m \left(e^{a_m \Delta_m} - a_m \Delta_m - 1\right) \tag{8}$$

where $\Delta_m = \hat{\theta}_m - \theta_m$, $a_m \neq 0$ and $b_m > 0$. To calculate the OBE, we use (4) with $\partial L(\underline{\theta},\hat{\underline{\theta}})/\partial\hat{\theta}_m = b_m a_m \left(e^{a_m\Delta_m} - 1\right)$ and finally obtain

$$\hat{\theta}_m = -\frac{1}{a_m} \ln \int e^{-a_m\theta_m} p(\theta_m|\underline{x}) d\theta_m, \quad m = 1,\ldots,M. \tag{9}$$

The next theorem shows that this estimator is included in the family $\mathcal{F}$.

Theorem 3. *The OBE for the multivariate LINEX loss function* (8) *is included in the estimator family* $\mathcal{F}$.

Proof. Plugging in the values $\xi_1 = \xi_2 = 0, \xi_3 = 1, \lambda = 1$ and $\psi_m = 1/\phi_m = -a_m$ for $m = 1,\ldots,M$ into (7) proofs the theorem. ■

From Theorem 3 we conclude that the new estimator family $\mathcal{F}$ is more general than $\mathcal{F}_{\mathcal{B}}$ and also includes estimators with asymmetric loss functions. Actually, the parametric family of estimators in (7) is designed as a kind of "superposition" of both $\mathcal{F}_{\mathcal{B}}$ and the OBE resulting from the LINEX loss.

5. PRACTICAL CONSIDERATIONS

This section explains the general approach how to obtain the estimator for a given signal model and loss function and also shows how the estimate $\hat{\underline{\theta}}(\underline{x};\mathcal{P})$ can be calculated efficiently for a given observation $\underline{x}$. In the sequel, we will make the following two assumptions:

- The generation of samples $(\underline{\theta}_k, \underline{x}_k) \sim p(\underline{\theta},\underline{x})$ is manageable, where $p(\underline{\theta},\underline{x})$ is the joint PDF of $\underline{\theta}$ and $\underline{x}$. This is often the case as $p(\underline{\theta},\underline{x})$ can be written as $p(\underline{\theta},\underline{x}) = p(\underline{x}|\underline{\theta})p(\underline{\theta})$, where $p(\underline{\theta})$ is the a priori PDF of $\underline{\theta}$ and $p(\underline{x}|\underline{\theta})$ is the likelihood PDF. Very often, both are known: $p(\underline{\theta})$ from expert knowledge and $p(\underline{x}|\underline{\theta})$ through the signal model.
- The generation of samples $\underline{\theta}_k \sim p(\underline{\theta}|\underline{x})$ is manageable. This is not a hard restriction as the MMSE estimator is often calculated using Markov chain Monte Carlo methods (MCMC) [3, 18]. MCMC allows the generation of samples from the a posteriori distribution and the MMSE estimator is then simply the average over all samples. Here, we will use importance sampling where the conditional distribution $p(\underline{\theta}|\underline{x})$ is the importance function.

Given the loss function and the signal model, the use of our estimator family for a general estimation problem consists of two steps:

Step 1 – Find the optimal estimator in $\mathcal{F}$

In a first step, we have to find the estimator $\hat{\underline{\theta}}(\underline{x};\mathcal{P}_0) \in \mathcal{F}$ that has the smallest Bayes risk for the particular loss function and joint PDF $p(\underline{\theta},\underline{x})$, i.e. we have to solve the optimization problem

$$\mathcal{P}_0 = \arg\min_{\mathcal{P}} \iint L(\underline{\theta},\hat{\underline{\theta}}(\underline{x};\mathcal{P}))p(\underline{\theta},\underline{x})d\underline{\theta}d\underline{x}. \tag{10}$$

This optimization has only to be carried out once to learn the optimal values of the parameters $\mathcal{P}$. In the appendix B, we give the gradient vector of the Bayes risk in (10) with respect to the parameters in $\mathcal{P}$. The knowledge of the gradient vector allows to use a gradient descent method to find the optimal parameter values. Note that the integration with respect to $\underline{\theta}$ and $\underline{x}$ can be carried out by a plain Monte Carlo (MC) integration using samples $(\underline{\theta}_k, \underline{x}_k) \sim p(\underline{\theta},\underline{x})$. The optimization problem (10) becomes then

$$\mathcal{P}_0 = \arg\min_{\mathcal{P}} \frac{1}{K_1} \sum_{k=1}^{K_1} L(\underline{\theta}_k, \hat{\underline{\theta}}(\underline{x}_k;\mathcal{P})). \tag{11}$$

If the generation of samples from $p(\underline{\theta},\underline{x})$ is not directly possible then importance sampling as discussed below is another possibility to obtain an accurate approximation of the integral.

Step 2 – Calculate the estimate $\hat{\underline{\theta}}(\underline{x};\mathcal{P}_0)$

In a second step, we calculate the estimate for a given observation $\underline{x}$. Therefore, we need an efficient method to compute both integrals in (7a). Note that (7a) can be written as

$$\begin{aligned}\hat{\underline{\theta}}(\underline{x};\mathcal{P}) &= \underline{f}_1\left(\int \underline{f}_2(\underline{\theta};\mathcal{P}_2)\frac{p(\underline{\theta},\underline{x})^\lambda}{\int p(\underline{\theta},\underline{x})^\lambda d\underline{\theta}} d\underline{\theta};\mathcal{P}_1\right) \\ &= \underline{f}_1\left(E_{p_\lambda}\left[\underline{f}_2(\underline{\theta};\mathcal{P}_2)\right];\mathcal{P}_1\right).\end{aligned} \tag{12}$$

We see that we can write the integrals as the expectation of $\underline{f}_2(\underline{\theta};\mathcal{P})$ with respect to a new conditional density $p_\lambda(\underline{\theta}|\underline{x}) = p(\underline{\theta},\underline{x})^\lambda / \int p(\underline{\theta},\underline{x})^\lambda d\underline{\theta}$. Assuming that we can generate samples from the a posteriori distribution $\underline{\theta}_k \sim p(\underline{\theta}|\underline{x}) = p(\underline{\theta},\underline{x})/\int p(\underline{\theta},\underline{x})d\underline{\theta}$, we can use importance sampling [3] for (12). The importance sampling algorithm is as follows: Suppose we want to calculate $\mathrm{E}[h(\underline{\theta})] = \int h(\underline{\theta})p(\underline{\theta})d\underline{\theta}$. Then we can use the approximation

$$\mathrm{E}[h(\underline{\theta})] \approx \sum_{k=1}^{K} w_k h(\underline{\theta}_k) \Big/ \sum_{k=1}^{K} w_k \tag{13}$$

where $\underline{\theta}_k$ are drawn from a trial distribution $\tilde{p}(\underline{\theta})$ and the importance weights w_k are defined as $w_k = p(\underline{\theta}_k)/\tilde{p}(\underline{\theta}_k)$. Note that w_k has only to be known up to a multiplicative constant. Using importance sampling for our problem, we finally obtain the approximation

$$\hat{\underline{\theta}}(\underline{x};\mathcal{P}) \approx \underline{f}_1\left(\sum_{k=1}^{K_2} w_k\, \underline{f}_2(\underline{\theta}_k;\mathcal{P}_2) \Big/ \sum_{k=1}^{K_2} w_k;\mathcal{P}_1\right) \tag{14}$$

with $\tilde{p}(\underline{\theta}) = p(\underline{\theta},\underline{x})$ and thus $w_k = p(\underline{\theta}_k,\underline{x})^{\lambda-1}$. The computational complexity is hence comparable to that of a MMSE estimation if the MMSE estimator also uses MC integration.

6. EXAMPLE

The example is as follows: Given the signal model $x = \theta + z$, estimate θ which is uniformly distributed in $[0,1]$ from the observation x where we know that the observation is disturbed by additive Gaussian noise $z \sim \mathcal{N}(0,\sigma^2)$. Furthermore, z and θ are independently distributed. The considered loss function is the bounded LINEX (BLINEX) loss introduced in [14]. The univariate BLINEX loss function is given by

$$L_{\text{BLINEX}}(\theta,\hat{\theta}) = \frac{L_{\text{LINEX}}(\theta,\hat{\theta})}{1 + \rho L_{\text{LINEX}}(\theta,\hat{\theta})},\ \rho > 0. \tag{15}$$

Plugging $L_{\text{LINEX}}(\theta,\hat{\theta})$ from (8) into (15), we obtain

$$L_{\text{BLINEX}}(\theta,\hat{\theta}) = \frac{1}{\rho}\left(1 - \frac{1}{1 + c(e^{a(\hat{\theta}-\theta)} - a(\hat{\theta}-\theta) - 1)}\right) \tag{16}$$

with $c = \rho b$. It differs from the usually used loss functions (1) in two main properties, namely it is (a) asymmetric and (b) bounded:

(a) If $a > 0$ then the positive error $\hat{\theta} > \theta$ results in a larger loss than the corresponding negative error of the same magnitude. If $a < 0$ then negative errors $\hat{\theta} < \theta$ have a larger loss. A case where such an emphasis of negative errors is useful is the dam construction example given in Sec. 1 as underestimating the peak water level is more severe then overestimating it.

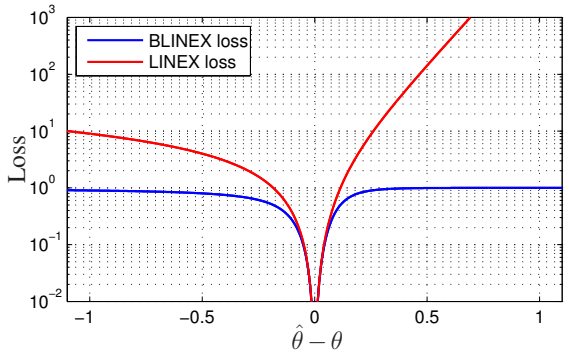

Figure 1: LINEX and BLINEX loss ($\rho=1$, $a=10$ and $b=1$)

(b) $L_{\text{BLINEX}}(\theta,\hat{\theta})$ is bounded by 0 and $1/\rho$. Such a requirement for a loss function may occur naturally out of the considered problem or may be artificially introduced to improve the robustness of the estimator in the case of outliers.

In our example, we choose $\rho=1$, $a=10$ and $b=1$. Fig. 1 shows the graph of the BLINEX loss function for this choice of parameters. Furthermore, the noise variance is $\sigma^2=0.5$. Note that both loss functions in Fig. 1 differ substantially for such a noise variance.

We compare the following five estimators with respect to the squared error loss (1b) and the BLINEX loss (16):

- *MAP estimator*: The MAP estimator is in general given by $\hat{\theta}=\arg\max_\theta p(\theta|x)$ with $p(\theta|x)\sim e^{-(x-\theta)^2/(2\sigma^2)}u_{[0,1]}(\theta)$ and $u_{[0,1]}(\theta)$ is the a priori PDF of θ which is uniformly distributed in $[0,1]$. This yields

$$\hat{\theta}_{\text{MAP}}=\begin{cases}0 & x<0\\ x & 0\le x\le 1\\ 1 & x>1\end{cases} \tag{17}$$

- *MMSE estimator*: The MMSE estimator is given by $\hat{\theta}_{\text{MMSE}}=E[\theta|x]$. For our signal model, the conditional mean can be calculated analytically and one obtains

$$\hat{\theta}_{\text{MMSE}}=x+\sqrt{\frac{2}{\pi}}\sigma\frac{e^{-\frac{x^2}{2\sigma^2}}-e^{-\frac{(x-1)^2}{2\sigma^2}}}{\operatorname{erf}\left(\frac{x}{\sqrt{2}\sigma}\right)-\operatorname{erf}\left(\frac{x-1}{\sqrt{2}\sigma}\right)} \tag{18}$$

- *OBE under LINEX loss*: The OBE for LINEX loss is given by (9) which can be calculated analytically.
- *OBE under BLINEX loss*: The optimization problem (3) for this example can not be carried out analytically and thus (3) has to be solved for each new observation x individually, either by Monte Carlo integration or numerical quadrature which we used here.
- *Estimator family* (7) *with optimal parameters*: The optimal parameters are found via the Matlab function `fmincon` and are $\xi_1\approx 6.32\times10^{-1}$, $\xi_2\approx 2.57\times10^{-1}$, $\xi_3=1.58\times10^{-1}$, $\lambda\approx 1.10\times10^{1}$, $\phi\approx -5.92\times10^{-5}$ and $\psi\approx 1.90$. $K_1=5\,000$ samples are used for the Monte Carlo approximation in (11) and $K_2=5\,000$ samples are drawn from the a posteriori density $p(\theta|x)$ for (14).

Table 1 shows the results averaged over 10 000 trials. Clearly, the MMSE estimator is optimal in terms of the squared error loss as expected. Similarly, the OBE under BLINEX loss gives the smallest Bayes risk if the BLINEX loss function is used. The optimal estimator $\underline{\hat{\theta}}(\underline{x};\mathcal{P}_0)$ from the set $\mathcal{F}$ is a good approximation of the OBE under BLINEX loss as it has a similar Bayes risk. Thus, although the OBE under BLINEX loss itself is not an element of $\mathcal{F}$, there is an estimator $\underline{\hat{\theta}}(\underline{x};\mathcal{P}_0)$ in $\mathcal{F}$ which gives nearly the same performance.

7. CONCLUSIONS AND FUTURE WORK

In this paper a family of estimators was proposed for the Bayesian estimation with non-standard loss functions. This family has the advantage that it is parameterized by a small number of variables which can be determined offline for a particular loss function. We proofed that the family includes many important estimators known from the literature, namely MMSE, MAP, and OBE under LINEX loss which shows that it is quite versatile. The computational complexity of our approach is comparable to that of an MMSE estimation for the same signal model if we assume that Monte Carlo integration is used for the calculation of the MMSE estimator.

Because of space limitations, we could only give a simple example in Sec. 6. We are currently working on a more sophisticated application for image denoising using the SSIM quality index [7] that we want to publish in a follow-up paper.

	Squared error loss	BLINEX loss
MAP estimator	1.67×10^{-1}	6.58×10^{-1}
MMSE estimator	7.09×10^{-2}	5.85×10^{-1}
OBE under LINEX loss	1.38×10^{-1}	5.78×10^{-1}
Optimal estimator $\in\mathcal{F}$	1.06×10^{-1}	5.45×10^{-1}
OBE under BLINEX loss	1.03×10^{-1}	5.43×10^{-1}

Table 1: Comparison of the Bayes risks

A. PROOFS

Proof of the Lemma. First of all, we would like to point out that the delta function can be expressed as a limit of the normal distribution, i.e.

$$g(\underline{\theta};a^2)=\frac{1}{a^M\pi^{M/2}}e^{-\|\underline{\theta}\|^2/a^2}\xrightarrow{a\to0}\delta(\underline{\theta}).$$

They are equivalent in the sense that $f(\underline{0})=\int f(\underline{\theta})\delta(\underline{\theta})d\underline{\theta}=\lim_{a\to0}\int f(\underline{\theta})g(\underline{\theta};a^2)d\underline{\theta}$.

(a) $p(\underline{\theta},\underline{x})=\delta(\underline{\theta}-\underline{\theta}_0)$:

$$\underline{\hat{\theta}}(\underline{x};\lambda)=\lim_{a\to0}\frac{\int\underline{\theta}\,g(\underline{\theta}-\underline{\theta}_0;a^2)^\lambda d\underline{\theta}}{\int g(\underline{\theta}-\underline{\theta}_0;a^2)^\lambda d\underline{\theta}}=\lim_{a\to0}\frac{\int\underline{\theta}\,g(\underline{\theta}-\underline{\theta}_0;\frac{a^2}{\lambda})d\underline{\theta}}{\int g(\underline{\theta}-\underline{\theta}_0;\frac{a^2}{\lambda})d\underline{\theta}}$$

$$=\lim_{a\to0}\int\underline{\theta}\,g(\underline{\theta}-\underline{\theta}_0;\frac{a^2}{\lambda})d\underline{\theta}=\underline{\theta}_0$$

(b) $p(\underline{\theta},\underline{x})=P\delta(\underline{\theta}-\underline{\theta}_0)+(1-P)\delta(\underline{\theta}-\underline{\theta}_1)$:

$$\underline{\hat{\theta}}(\underline{x};\lambda)=\lim_{a\to0}\frac{\int\underline{\theta}[Pg(\underline{\theta}-\underline{\theta}_0;a^2)+(1-P)g(\underline{\theta}-\underline{\theta}_1,a^2)]^\lambda d\underline{\theta}}{\int[Pg(\underline{\theta}-\underline{\theta}_0;a^2)+(1-P)g(\underline{\theta}-\underline{\theta}_1;a^2)]^\lambda d\underline{\theta}}$$

$$=\lim_{a\to0}\frac{P^\lambda}{P^\lambda+(1-P)^\lambda}\int\underline{\theta}\,g(\underline{\theta}-\underline{\theta}_0;\frac{a^2}{\lambda})d\underline{\theta}$$

$$+\lim_{a\to0}\frac{(1-P)^\lambda}{P^\lambda+(1-P)^\lambda}\int\underline{\theta}\,g(\underline{\theta}-\underline{\theta}_1;\frac{a^2}{\lambda})d\underline{\theta}$$

$$=\frac{P^\lambda\underline{\theta}_0+(1-P)^\lambda\underline{\theta}_1}{P^\lambda+(1-P)^\lambda}$$

where we used the fact that $[Pg(\underline{\theta}-\underline{\theta}_0;a^2)+(1-P)g(\underline{\theta}-\underline{\theta}_1;a^2)]^\lambda\to P^\lambda g(\underline{\theta}-\underline{\theta}_0;a^2)^\lambda+(1-P)^\lambda g(\underline{\theta}-\underline{\theta}_1;a^2)^\lambda$ for $a\to0$. ■

Proof of Theorem 2. We will proof this theorem by contradiction. Suppose $\underline{\hat{\theta}}(\underline{x};\lambda)$ has a corresponding loss function $L(\underline{\theta},\underline{\hat{\theta}})$ which is continuously differentiable but not symmetric. Then at least one of the following two cases has to be true:

(a) There is a $\underline{\theta}_0$ such that

$$\left|\frac{\partial L(\underline{\theta},\underline{\hat{\theta}})}{\partial\underline{\hat{\theta}}}\right|_{\substack{\underline{\theta}=\underline{\theta}_0\\ \underline{\hat{\theta}}=\underline{\theta}_0}}\neq\left|\frac{\partial L(\underline{\theta},\underline{\hat{\theta}})}{\partial\underline{\hat{\theta}}}\right|_{\substack{\underline{\theta}=-\underline{\theta}_0\\ \underline{\hat{\theta}}=-\underline{\theta}_0}} \tag{$\star$}$$

Now, consider the special PDF $p(\underline{\theta},\underline{x})=\delta(\underline{\theta}-\underline{\theta}_0)$. As $\underline{\hat{\theta}}(\underline{x};\lambda)$ from (5) holds for all densities, we can directly use the result of

the Lemma and obtain $\hat{\underline{\theta}}(\underline{x};\lambda) = \underline{\theta}_0$. A necessary condition that $\hat{\underline{\theta}}(\underline{x};\lambda)$ is the OBE under the loss function $L(\underline{\theta},\hat{\underline{\theta}})$ is (4)

$$\left.\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|_{\substack{\underline{\theta}=\underline{\theta}_0 \\ \hat{\underline{\theta}}=\underline{\theta}_0}} = \underline{0}.$$

Furthermore, consider the special PDF $p(\underline{\theta},\underline{x}) = \delta(\underline{\theta}+\underline{\theta}_0)$ which has the OBE $\hat{\underline{\theta}}(\underline{x};\lambda) = -\underline{\theta}_0$. Using again (4), we obtain the necessary condition

$$\left.\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|_{\substack{\underline{\theta}=-\underline{\theta}_0 \\ \hat{\underline{\theta}}=-\underline{\theta}_0}} = \underline{0}$$

which can not be true as we assumed ($\star$).

(b) There is a $\underline{\theta}_0$ and $\underline{\theta}_1$ such that

$$\left|\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|\Bigg|_{\substack{\underline{\theta}=\underline{\theta}_0 \\ \hat{\underline{\theta}}=\underline{\theta}_1}} \neq \left|\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|\Bigg|_{\substack{\underline{\theta}=-\underline{\theta}_0 \\ \hat{\underline{\theta}}=-\underline{\theta}_1}} \qquad (\star\star)$$

Now consider the special PDF $p(\underline{\theta},\underline{x}) = P\delta(\underline{\theta}-\underline{\theta}_0) + (1-P)\delta(\underline{\theta}-\underline{\theta}_1)$ which, according to the above Lemma, has the OBE $\underline{u} = \hat{\underline{\theta}}(\underline{x};\lambda) = (P^\lambda \underline{\theta}_0 + (1-P)^\lambda \underline{\theta}_1)/(P^\lambda + (1-P)^\lambda)$. A necessary condition that has to be fulfilled is (4) which yields

$$P\left.\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|_{\substack{\underline{\theta}=\underline{\theta}_0 \\ \hat{\underline{\theta}}=\underline{u}}} + (1-P)\left.\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|_{\substack{\underline{\theta}=\underline{\theta}_1 \\ \hat{\underline{\theta}}=\underline{u}}} = \underline{0}.$$

Furthermore, the PDF $p(\underline{\theta},\underline{x}) = P\delta(\underline{\theta}+\underline{\theta}_0) + (1-P)\delta(\underline{\theta}+\underline{\theta}_1)$ results in the OBE $-\underline{u}$ and the necessary condition (4) is

$$P\left.\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|_{\substack{\underline{\theta}=-\underline{\theta}_0 \\ \hat{\underline{\theta}}=-\underline{u}}} + (1-P)\left.\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|_{\substack{\underline{\theta}=-\underline{\theta}_1 \\ \hat{\underline{\theta}}=-\underline{u}}} = \underline{0}.$$

Without loss of generality, we can assume $\left|\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|\Big|_{\substack{\underline{\theta}=\underline{\theta}_1 \\ \hat{\underline{\theta}}=\underline{\theta}_1}} = \left|\frac{\partial L(\underline{\theta},\hat{\underline{\theta}})}{\partial \hat{\underline{\theta}}}\right|\Big|_{\substack{\underline{\theta}=-\underline{\theta}_1 \\ \hat{\underline{\theta}}=-\underline{\theta}_1}}$ as we can otherwise use (a) and show that the loss is asymmetric. Taking the limit $P \to 0$ ($P > 0$), we see that both necessary conditions contradict the assumption ($\star\star$). ■

B. GRADIENT OF THE BAYES RISK

In this section, we derive the gradient of the Bayes risk with respect to an element $\gamma \in \mathcal{P}$. Using the gradient is advantageous to solve the optimization problem (10) as gradient descent methods can be used. Taking the derivative of BR in (10) with respect to γ, we obtain for the first-order derivative

$$\frac{\partial \mathrm{BR}}{\partial \gamma} = \iint \left(\left.\frac{\partial L(\underline{\theta},\underline{u})}{\partial \underline{u}}\right|_{\underline{u}=\hat{\underline{\theta}}(\underline{x};\mathcal{P})}\right)^T \frac{\partial \hat{\underline{\theta}}(\underline{x};\mathcal{P})}{\partial \gamma} p(\underline{\theta},\underline{x}) d\underline{\theta} d\underline{x}.$$

Using the shorthand notations $p_\lambda(\underline{\theta}|\underline{x}) = p(\underline{\theta},\underline{x})^\lambda / \int p(\underline{\theta},\underline{x})^\lambda d\underline{\theta}$ and $\mathbf{D} = \frac{\partial \underline{f}}{\partial \underline{z}} = \xi_1 \mathbf{I} + \mathrm{diag}\{\phi_1/z_1, \ldots, \phi_M/z_M\}$ evaluated at $\underline{z} = \int \underline{f}_2(\underline{\theta},\mathcal{P}_2) p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta}$, we obtain

$$\frac{\partial \hat{\underline{\theta}}(\underline{x};\mathcal{P})}{\partial \xi_1} = \int \underline{f}_2(\underline{\theta},\mathcal{P}_2) p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta}$$

$$\frac{\partial \hat{\underline{\theta}}(\underline{x};\mathcal{P})}{\partial \xi_2} = \mathbf{D} \cdot \int \underline{\theta}\, p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta}$$

$$\frac{\partial \hat{\underline{\theta}}(\underline{x};\mathcal{P})}{\partial \xi_3} = \mathbf{D} \cdot \int e^{\underline{\psi} \circ \underline{\theta}} p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta}$$

$$\frac{\partial \hat{\underline{\theta}}(\underline{x};\mathcal{P})}{\partial \lambda} = \mathbf{D} \cdot \Bigg(\int \underline{f}_2(\underline{\theta};\mathcal{P}_2) \ln\left(p(\underline{\theta},\underline{x})\right) p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta} - \int \underline{f}_2(\underline{\theta};\mathcal{P}_2)\, p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta} \int \ln\left(p(\underline{\theta},\underline{x})\right) p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta} \Bigg)$$

$$\frac{\partial \hat{\underline{\theta}}(\underline{x};\mathcal{P})}{\partial \underline{\phi}} = \mathrm{diag}\left\{ \ln \left| \int \underline{f}_2(\underline{\theta};\mathcal{P}_2)\, p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta} \right| \right\}$$

$$\frac{\partial \hat{\underline{\theta}}(\underline{x};\mathcal{P})}{\partial \underline{\psi}} = \xi_3 \mathbf{D} \cdot \mathrm{diag}\left\{ \int \underline{\theta} \circ e^{\underline{\psi} \circ \underline{\theta}} p_\lambda(\underline{\theta}|\underline{x}) d\underline{\theta} \right\}$$

Note that all integrals can again be calculated using Monte Carlo integration, especially importance sampling as was shown in Sec. 5.

REFERENCES

[1] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 1: Estimation Theory*, Prentice-Hall, 1993.

[2] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood estimation from incomplete data via the EM algorithm," *J. Roy. Stat. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[3] J. S. Liu, *Monte Carlo Strategies in Scientific Computing*, Springer, 2008.

[4] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, Chapman & Hall, 2003.

[5] J. G. Norstrom, "The use of precautionary loss functions in risk analysis," *IEEE Trans. on Reliability*, vol. 45, no. 3, pp. 400–403, 1996.

[6] A. Zellner, "Bayesian estimation and prediction using asymmetric loss functions," *Journal of the American Statistical Association*, vol. 81, no. 394, pp. 446–451, 1986.

[7] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 4, 2004.

[9] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath Jr., "Design of linear equalizers optimized for the structural similarity index," *IEEE Trans. Image Processing*, vol. 17, no. 6, pp. 857–872, 2008.

[10] H. Rue, "New loss functions in Bayesian imaging," *Journal of the American Statistical Association*, vol. 90, no. 431, pp. 900–908, 1995.

[11] C. Stein, "Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean," *Annals of the Institute of Statistical Mathematics*, vol. 16, no. 1, pp. 155–160, 1964.

[12] L. D. Brown, "Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters," *Annals of Mathematical Statistics*, vol. 39, no. 1, pp. 29–48, 1968.

[13] A. Kaminska and Z. Porosinski, "On robust Bayesian estimation under some asymmetric and bounded loss function," *Statistics*, vol. 43, no. 3, pp. 253–265, 2009.

[14] D. Wen and M. S. Levy, "BLINEX: a bounded asymmetric loss function with application to Bayesian estimation," *Communications in Statistics - Theory and Methods*, vol. 30, no. 1, pp. 147–153, 2001.

[15] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation and Time Series Analysis*, Addison-Wesley, 1990.

[16] M. Pincus, "A closed form solution of certain programming problems," *Operations Research*, vol. 16, no. 3, pp. 690–694, 1968.

[17] H. R. Varian, "A Bayesian approach to real estate assessment," in *Studies in Bayesian Econometrics and Statistics in Honor of Leonard J. Savage*, S. E. Fienberg and A. Zellner, Eds., pp. 195–208. North Holland Press, 1974.

[18] C. P. Robert, *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer-Verlag, 2001.