

On the Robustness of the Bayes and Wiener Estimators under Model Uncertainty *

Mattia Zorzi †

Abstract

This paper deals with the robust estimation problem of a signal given noisy observations. We assume that the actual statistics of the signal and observations belong to a ball about the nominal statistics. This ball is formed by placing a bound on a suitable divergence (or distance) between the actual and the nominal statistics. Then, the robust estimator is obtained by minimizing the mean square error according to the least favorable statistics in that ball. Therefore, we obtain a divergence-based minimax approach to robust estimation. Choosing a set of divergences, called Tau divergence family, we show that the Bayes estimator based on the nominal statistics is the optimal solution. Moreover, in the dynamic case, the optimal offline estimator is the noncausal Wiener filter based on the nominal statistics.

Keywords: Robust filtering, minimax problem, Tau divergence family, risk-sensitive estimation problem, minimum entropy problem.

1 Introduction

Consider the problem of estimating a signal from noisy observations. Typically the actual statistics of the signal and observations are known only imprecisely that is only the nominal statistics are known. In this situation, one would wonder how to construct an estimator which is robust to this model uncertainty. According to the minimax approach, the statistics of the signal and observations are assumed to belong to a neighborhood of the nominal ones. Then, the robust estimator is characterized by a minimax problem consisting of finding the estimator which minimizes the mean square error for the least favorable statistics in the neighborhood. The latter can be specified in different ways, e.g. it can be based on a ϵ -contamination model, a total variation model or a spectral band model [8]. [5] propose to specify this neighborhood through an uncertainty ball which is formed by placing a bound (i.e. tolerance) c on the *Kullback-Leibler* divergence between the actual and the nominal statistics.

*This work has been partially supported by the FIRB project “Learning meets time” (RBF12M3AC) funded by MIUR.

†Dipartimento di Ingegneria dell’Informazione, Università degli studi di Padova, via Gradenigo 6/B, 35131 Padova, Italy. Email: zorzimat@dei.unipd.it

This characterization of the uncertainty is supported by the fact that nominal models are identified from data according to the maximum-likelihood principle which turns out to be equivalent to the minimization of the *Kullback-Leibler* divergence over a suitable parametric model class. Accordingly, the maximum-likelihood approach provides the nominal model and the bound c . [9] showed that if the nominal statistics is Gaussian then the least favorable statistics is Gaussian and the robust estimator coincides with the nominal *Bayes* estimator, i.e. the *Bayes* estimator based on the nominal statistics.

In this paper we consider a family of uncertainty balls which are formed by placing a bound on a set of divergences, called τ divergence family, between Gaussian statistics. Each divergence of this set is characterized by parameter τ which is a real number belonging to the interval $[0, 1]$. In particular, for $\tau = 0$ we obtain the *Kullback-Leibler* divergence. This characterization of the uncertainty is supported by the fact that recently it has been proposed a system identification procedure which finds the model from data by minimizing the τ divergence family over a suitable parametric model class [19]. Accordingly, this system identification procedure provides the nominal model and the bound c for the proposed uncertainty ball. It turns out that the (family of) robust estimators, solution to the minimax problem with this family of uncertainty balls, coincide with the *Bayes* estimator based on the nominal statistics. Accordingly, the nominal *Bayes* estimator is robust in this wide family of (Gaussian) uncertainty classes. Our result also gives the analytical form of the least favorable statistics of the estimation error. This result allows to define a new family of robust Kalman filters obtained by iterating the *Bayes* estimator with the least favorable statistics [21, 10]. Our minimax approach can be also relaxed expressing the τ divergence constraint as a soft one. We will show that this relaxation corresponds to a new family of risk-sensitive problems (in the sense of [1]) which is also linked to a new family of minimum entropy problems (in the sense of [13]). These results can be extended in the dynamic case: We will prove that the nominal noncausal *Wiener* filter is the optimal robust offline estimator according to the minimax approach based on the τ divergence family. Finally, through a simulation study we analyze the features of the least favorable statistics of the estimation error for these uncertainty classes. Simulations show that the parameter τ tunes how the uncertainty is spread among the components (components and frequencies for the dynamic case) of the least favorable statistics.

The outline of the paper is as follows. In Section 2 we define the τ divergence family for Gaussian random vectors and stationary stochastic processes. Section 3 deals with the static case: We show that the nominal *Bayes* estimator is the solution to the family of divergence-based minimax approaches. Section 4 deals with the dynamic case: We show that the nominal noncausal *Wiener* estimator is the solution to the family of divergence-based minimax approaches. In Section 5, we present the simulation study. Finally, in Section 6 we draw the conclusions. In order to streamline the presentation all the proofs are deferred to the Appendix.

In the paper, we will use the following conventions. \mathbb{Z}, \mathbb{N} and \mathbb{R} denote the set of integer, natural and real numbers, respectively. \mathbb{R}_+ denotes the set of

positive real numbers. Given $x \in \mathbb{R}^q$, $\|x\|$ denotes its *Euclidean* vector norm. Moreover, $\|x\|_K = \sqrt{x^T K x}$. I_q denotes the identity matrix of dimension q . $\text{tr}(P)$ denotes the trace of matrix P . \mathcal{Q}^q denotes the vector space of symmetric matrices of dimension $q \times q$. The i -th singular value of $P \in \mathcal{Q}^q$ is denoted by $\sigma_i(P)$ and we assume that $\sigma_1(P) \geq \sigma_2(P) \geq \dots \geq \sigma_q(P)$. $\|P\|$ denotes the spectral norm of P , that is $\|P\| = \sigma_1(P)$. \mathcal{Q}_+^q is the cone of positive definite symmetric matrices of dimension $q \times q$. $\log(P)$ denotes the logarithm of matrix P and P^τ is the τ -th power of matrix P . Matrix functions defined over the unit circle $\mathbb{T} = \{e^{j\vartheta} : \vartheta \in [0, 2\pi]\}$ are denoted by capital Greek letters and the dependence upon ϑ is sometime dropped to simplify the notation, i.e. Σ instead of $\Sigma(\vartheta)$. A star denotes transposition plus conjugation, that is $\Sigma(\vartheta)^* = \Sigma(-\vartheta)^T$. $\|\Sigma\|_\infty$ is the infinity matrix norm of Σ , that is $\|\Sigma\|_\infty = \sup_{\vartheta} \sigma_1(\Sigma(\vartheta))$. $\mathcal{Q}^q(\mathbb{T})$ denotes the vector space of para-symmetric matrix functions of dimension $q \times q$, i.e. if $\Sigma \in \mathcal{Q}^q(\mathbb{T})$ then $\Sigma = \Sigma^*$. $\mathcal{Q}_+^q(\mathbb{T})$ denotes the cone of para-symmetric matrix functions of dimension $q \times q$ which are positive definite over \mathbb{T} . Given $\Sigma \in \mathcal{Q}^q(\mathbb{T})$, the shorthand notation $\int \Sigma$ means the integration over the unit circle with respect to the normalized *Lebesgue* measure, that is $(2\pi)^{-1} \int_0^{2\pi} \Sigma(\vartheta) d\vartheta$. $\delta(\vartheta)$ denotes the *Dirac delta function*.

2 τ Divergence for Gaussian Vectors and Processes

Let z be a Gaussian random vector of dimension q with probability density

$$\tilde{f}(z) = \frac{1}{\sqrt{(2\pi)^q \det \tilde{K}_z}} \exp\left(-\frac{1}{2}(z - \tilde{m}_z)^T \tilde{K}_z^{-1}(z - \tilde{m}_z)\right), \quad (1)$$

where $\tilde{m}_z \in \mathbb{R}^q$ and $\tilde{K}_z \in \mathcal{Q}_+^q$. Let \hat{z} denote the minimum variance predictor of z based on the nominal probability density f

$$f(z) = \frac{1}{\sqrt{(2\pi)^q \det K_z}} \exp\left(-\frac{1}{2}(z - m_z)^T K_z^{-1}(z - m_z)\right) \quad (2)$$

with $m_z \in \mathbb{R}^q$ and $K_z \in \mathcal{Q}_+^q$. Thus, $\hat{z} = m_z$. Let $e = z - \hat{z}$ be the corresponding innovation vector. Accordingly, $e^N = L_z^{-1}(z - \hat{z})$ is the normalized innovation vector with L_z a square root of K_z , i.e. $K_z = L_z L_z^T$. It is not difficult to see that e^N is Gaussian with mean $m = L_z^{-1} \Delta m_z$, $\Delta m_z = \tilde{m}_z - m_z$, and covariance matrix $K = L_z^{-1} \tilde{K}_z L_z^{-T}$. If \tilde{f} coincides with f then we have that e^N is with zero mean and identically independently distributed components, i.e. $m = 0$ and $K = I$. Therefore, (m, K) represents a mismatch criterium which naturally occurs in prediction error estimation [12]. This leads us to measure

the mismatch between \tilde{f} and f by quantifying the mismatch between (m, K) and $(0, I)$:

$$\mathcal{D}(\tilde{f}\|f) = \ell(m, K), \quad (3)$$

where $\ell : \mathbb{R}^q \times \mathcal{Q}_+^q \rightarrow \mathbb{R} \cup \{\infty\}$ is a function such that $\ell \geq 0$ and equality holds if and only if $m = 0$ and $K = I$. We consider the following function parametrized by $\tau \in [0, 1]$:

$$\ell(m, K) = \begin{cases} \|m\|^2 + \text{tr}(-\log(K) + K - I_q), & \tau = 0 \\ \frac{1}{1-\tau}\|m\|^2 + \text{tr}\left(\frac{1}{\tau(\tau-1)}K^\tau + \frac{1}{1-\tau}K + \frac{1}{\tau}I_q\right), & 0 < \tau < 1 \\ \delta(m) + \text{tr}(K \log(K) - K + I_q), & \tau = 1. \end{cases} \quad (4)$$

There are several ways to construct ℓ . We will motivate our choice in Remark 3.2. Substituting (4) in (3) we obtain the following family of divergences indexed by τ :

$$\mathcal{D}_\tau(\tilde{f}\|f) = \begin{cases} \|\Delta m_z\|_{K_z^{-1}}^2 + \text{tr}\left(-\log(\tilde{K}_z K_z^{-1}) + \tilde{K}_z K_z^{-1} - I_q\right), & \tau = 0 \\ \frac{1}{1-\tau}\|\Delta m_z\|_{K_z^{-1}}^2 + \text{tr}\left(\frac{1}{\tau(\tau-1)}(L_z^{-1}\tilde{K}_z L_z^{-T})^\tau + \frac{1}{1-\tau}\tilde{K}_z K_z^{-1} + \frac{1}{\tau}I_q\right), & 0 < \tau < 1 \\ \delta(\Delta m_z) + \text{tr}\left(L_z^{-1}\tilde{K}_z L_z^{-T} \log(L_z^{-1}\tilde{K}_z L_z^{-T}) - \tilde{K}_z K_z^{-1} + I_q\right), & \tau = 1. \end{cases} \quad (5)$$

It is worth noting that (5) coincides with the τ divergence between covariance matrices [20] when \tilde{f} and f have the same mean. Moreover, (5) coincides with the *Kullback-Leibler* divergence for $\tau = 0$.

Proposition 2.1 $\mathcal{D}_\tau(\tilde{f}\|f) \geq 0$ and equality holds if and only if $\tilde{f} = f$.

The mismatch criterium above can be extended to the dynamic case. Let $z(t)$ be a stationary Gaussian process defined over $t \in \mathbb{Z}$ of dimension q with probability measure \tilde{f} . The latter is completely characterized by its power spectral density $\tilde{S}_z(\vartheta) = 2\pi\tilde{m}_z\tilde{m}_z^T\delta(\vartheta) + \tilde{\Sigma}_z(\vartheta)$, where $\tilde{m}_z \in \mathbb{R}^q$ is the mean and $\tilde{\Sigma}_z \in \mathcal{Q}_+^q(\mathbb{T})$ is the discrete time *Fourier* transform of the covariance matrix function

$$K_z(s) = \mathbb{E}_{\tilde{f}}[(z(t) - \tilde{m}_z)(z(t-s) - \tilde{m}_z)^T], \quad s \in \mathbb{Z}.$$

Let $\hat{z}(t)$ be the minimum variance linear one-step-ahead predictor based on the nominal probability measure f with power spectral density $S_z(\vartheta) = 2\pi m_z m_z^T \delta(\vartheta) + \Sigma_z(\vartheta)$. It is not difficult to see that corresponding normalized innovation process

$e^N(t)$ is stationary Gaussian with power spectral density $S(\vartheta) = mm^T \delta(\vartheta) + \Sigma(\vartheta)$, where $\Sigma_z = \Gamma_z^{-1} \tilde{\Sigma}_z \Gamma_z^{-*}$, Γ_z is a left squared spectral factor of Σ_z , i.e. $\Sigma_z = \Gamma_z \Gamma_z^*$, $m = \Gamma_z(0)^{-1} \Delta m_z$, $\Delta m_z = \tilde{m}_z - m_z$. Clearly, the more $e^N(t)$ is similar to white Gaussian noise, i.e. $S \approx I$, the closer \tilde{f} and f are. Accordingly, we measure their mismatch as follows $\mathcal{S}(\tilde{f}||f) = \int \ell(2\pi m, \Sigma)$. Choosing ℓ as in (4), we obtain the following family of divergences indexed by τ :

$$\mathcal{S}_\tau(\tilde{f}||f) = \begin{cases} \|\Delta m_z\|_{\Sigma_z(0)^{-1}}^2 + \int \text{tr} \left(-\log(\tilde{\Sigma}_z \Sigma_z^{-1}) \right. \\ \quad \left. + \tilde{\Sigma}_z \Sigma_z^{-1} - I_q \right), & \tau = 0 \\ \frac{1}{1-\tau} \|\Delta m_z\|_{\Sigma_z(0)^{-1}}^2 + \int \text{tr} \left(\frac{1}{\tau(\tau-1)} (\Gamma_z^{-1} \tilde{\Sigma}_z \Gamma_z^{-*})^\tau \right. \\ \quad \left. + \frac{1}{1-\tau} \tilde{\Sigma}_z \Sigma_z^{-1} + \frac{1}{\tau} I_q \right), & 0 < \tau < 1 \\ \delta(\Delta m_z) + \int \text{tr} \left(\Gamma_z^{-1} \tilde{\Sigma}_z \Gamma_z^{-*} \log(\Gamma_z^{-1} \tilde{\Sigma}_z \Gamma_z^{-*}) \right. \\ \quad \left. - \tilde{\Sigma}_z \Sigma_z^{-1} + I_q \right), & \tau = 1. \end{cases} \quad (6)$$

Note that, for $\tau = 0$ we obtain the *Itakura-Saito* distance, [7]. In the case that \tilde{f} and f have the same mean, we obtain the τ divergence defined in [20].

Proposition 2.2 $\mathcal{S}_\tau(\tilde{f}||f) \geq 0$ and equality holds if and only if $\tilde{f} = f$.

3 Robust Static Estimation

We consider a static estimation problem where we seek to estimate a random vector $x \in \mathbb{R}^n$ given an observation $y \in \mathbb{R}^p$. Assume the joint vector $z := [x^T; y^T]^T$ is Gaussian with nominal probability density f defined in (2) where $q = n + p$. We conformably partition the mean vector and the covariance matrix of z according to x and y :

$$m_z = \begin{bmatrix} m_x \\ m_y \end{bmatrix}, \quad K_z = \begin{bmatrix} K_x & K_{xy} \\ K_{yx} & K_y \end{bmatrix}.$$

Let $\tilde{f}(z)$ be the actual probability density defined in (1) where $q = n + p$. We consider the closed ball centered on f :

$$\mathcal{B}_\tau := \{\tilde{f} \text{ s.t. } \mathcal{D}_\tau(\tilde{f}||f) \leq c\}, \quad (7)$$

where \mathcal{D}_τ has been defined in (5), $c \in \mathbb{R}_+$ is a fixed tolerance which accounts for the maximum allowable deviance. Therefore, the hope is that \mathcal{B}_τ contains the actual (unknown) probability density \tilde{f} . Note that, \mathcal{B}_τ depends on $\tau \in [0, 1]$. Accordingly, by changing τ the set of all possible probability densities changes. In this way, we have a family of uncertainty classes parametrized by τ . It is worth noting that \mathcal{B}_τ also depends on c . However, we dropped this dependence to ease the notation.

Remark 3.1 *The nominal probability density function can be identified from data by solving $f = \operatorname{argmin}_{\tilde{f} \in \mathcal{C}} \mathcal{D}_\tau(f_S \| \tilde{f})$ where \mathcal{C} is a suitable parametric family of probability densities and f_S is Gaussian with mean and covariance matrix, respectively, the sample mean and the sample covariance matrix computed from the data [19]. Since f_S is the best probability density fitting the data, it is then realistic to specify the uncertainty with \mathcal{B}_τ with $c = \mathcal{D}_\tau(f_S \| f)$.*

We shall use the minimax viewpoint to design our robust estimator of x [9, 5]. More precisely, whenever we seek to design an estimator minimizing a suitable loss function, an hostile player, say “nature”, conspires to select the worst possible probability density in \mathcal{B}_τ . Let $g(y)$ denote an estimator of x based on the observation y . We evaluate its performance through the mean square error

$$J(\tilde{f}, g) = \mathbb{E}_{\tilde{f}}[\|x - g(y)\|^2] = \int_{\mathbb{R}^{n+p}} \|x - g(y)\|^2 \tilde{f}(z) dz.$$

Let \mathcal{G} denote the set of estimators $g(y)$ such that $\mathbb{E}_{\tilde{f}}[\|g(y)\|^2]$ is finite for any $\tilde{f} \in \mathcal{B}_\tau$. Our optimal robust estimator is the solution to the following minimax problem

$$\min_{g \in \mathcal{G}} \max_{\tilde{f} \in \mathcal{B}_\tau} J(\tilde{f}, g). \quad (8)$$

Theorem 3.1 *The optimal robust estimator, according to (8), is the Bayes estimator based on f*

$$g^\circ(y) = G^\circ(y - m_y) + m_x \quad (9)$$

with $G^\circ = K_{xy} K_y^{-1}$. The least favorable probability density \tilde{f}° has mean vector and covariance matrix

$$\tilde{m}_z^\circ = m_z, \quad \tilde{K}_z^\circ = \begin{bmatrix} \tilde{K}_x & K_{xy} \\ K_{yx} & K_y \end{bmatrix}, \quad (10)$$

wherein only the covariance of x is perturbed with respect to the nominal covariance matrix. The nominal and the least favorable estimation error have zero mean and covariance matrix, respectively,

$$P = K_x - K_{xy} K_y^{-1} K_{yx}, \quad \tilde{P} = \tilde{K}_x - K_{xy} K_y^{-1} K_{yx}.$$

Moreover,

$$\tilde{P} = \begin{cases} L_P (I_n - \frac{1-\tau}{\lambda} L_P^T L_P)^{\frac{1}{\tau-1}} L_P^T, & 0 \leq \tau < 1 \\ L_P \exp(\frac{1}{\lambda} L_P^T L_P) L_P^T, & \tau = 1, \end{cases} \quad (11)$$

where $P = L_P L_P^T$ and λ , with $\lambda > (1-\tau)\|P\|$, is the unique Lagrange multiplier such that $\mathcal{D}_\tau(\tilde{f} \| f) = c$.

Theorem 3.1 shows that the *Bayes* estimator based on the nominal statistics f is robust with respect to the τ divergence constraint. The worst situation occurs when all the mismodeling budget is allocated in a perturbation of the covariance matrix K_x . In Problem (8) several divergence families can be used to characterize \mathcal{B}_τ , such as the α divergence, [18], and the β divergence family [2]. Although the existence of the solution to (8) with those uncertainty classes is guaranteed, the optimal Bayes estimator is not necessarily based on the nominal statistic and such a solution does not admit a closed form. The mean square error (MSE) corresponding to the nominal probability density is $\text{MSE} = \mathbb{E}_f[\|e\|^2] = \text{tr}(P)$, while the MSE corresponding to the least favorable probability density \tilde{f}° is $\widetilde{\text{MSE}} = \mathbb{E}_{\tilde{f}^\circ}[\|e\|^2] = \text{tr}(\tilde{P})$. In view of (11), it follows that $\tilde{P} - P \in \mathcal{Q}_+^n$ therefore $\widetilde{\text{MSE}} > \text{MSE}$ and the additional MSE occasioned by the least favorable model perturbation is $\Delta\text{MSE} = \text{tr}(\tilde{P} - P)$.

Remark 3.2 *One would wonder why in (5) we consider $\frac{1}{1-\tau}\|\Delta m_z\|_{K_z^{-1}}^2$ instead of the simpler term $\|\Delta m_z\|_{K_z^{-1}}^2$. Indeed, it is not difficult to see that Theorem 3.1 still holds with $\frac{1}{1-\tau}\|\Delta m_z\|_{K_z^{-1}}^2$ replaced by $\|\Delta m_z\|_{K_z^{-1}}^2$ in (5). The unique difference is that λ must be such that $\lambda > \|P\|$ for any $\tau \in [0, 1]$. Therefore, with the choice $\|\Delta m_z\|_{K_z^{-1}}^2$ we restrict the allowable values for the Lagrange multiplier λ and thus the allowable least favorable covariance matrices \tilde{K}_z . Accordingly, taking the term $\|\Delta m_z\|_{K_z}^2$, the freedom of the nature (i.e. the hostile player) is restricted.*

It is worth comparing our result with the one in [9]. Theorem 3.1 generalizes the case $\tau = 0$ analyzed in Theorem 1 in [9]. On the other hand, Theorem 1 shows the least square estimator is robust over the more general allowable set $\{\tilde{f} : \mathbb{D}_{KL}(\tilde{f}\|f) \leq c\}$, where $\mathbb{D}_{KL}(\tilde{f}\|f)$ is the *Kullback-Leibler* divergence among probability densities, and \tilde{f} is not necessarily Gaussian. However, our result cannot be extended to such general case because \mathcal{D}_τ is a divergence family which only measures the deviation among Gaussian vectors.

In order to understand the influence of parameter τ on the uncertainty ball \mathcal{B}_τ , we consider the case in which z is a Gaussian random variable, i.e. $q = 1$, with nominal mean $m_z = 0.5$ and nominal variance $K_z = 0.03$. We consider \mathcal{B}_τ with $\tau = 0$ and $c = 0$, left panel of Figure 1, and \mathcal{B}_τ with $\tau = 0.8$ and $c = 0.454$, right panel of Figure 1. The tolerance c for the two balls is chosen in such a way that their measure is the same. As we can see, increasing τ the uncertainty increases for the variance while it decreases for the mean. This observation holds also with $q > 1$, indeed the first term in (5) measures the deviance between the actual and the nominal mean, and, increasing τ , this term becomes preponderant than the second one when $m_z \neq \tilde{m}_z$. In other words through parameter τ we tune how to allocate the mismodeling budget between the mean and the covariance matrix.

Problem (8) can be relaxed in the following way

$$\min_{g \in \mathcal{G}} \max_{\tilde{f} \in \mathcal{B}_\tau^\infty} J(\tilde{f}, g) + \lambda(c - \mathcal{D}_\tau(\tilde{f}\|f)), \quad (12)$$

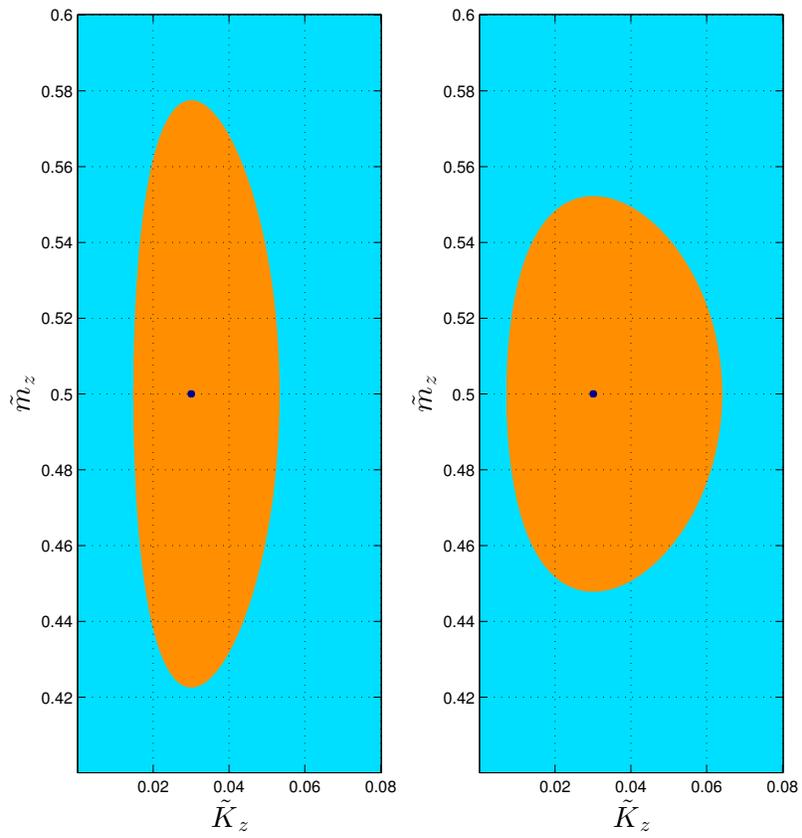


Figure 1: In orange the uncertainty ball with $\tau = 0$, $c = 0.2$ (left panel) and $\tau = 0.8$ and $c = 0.454$ (right panel); the blue point is the nominal statistics.

where $\mathcal{B}_\tau^\infty = \{\tilde{f} \text{ s.t. } \mathcal{D}_\tau(\tilde{f}\|f) < \infty\}$ and \mathcal{G} is the set of all estimators such that $\mathbb{E}_{\tilde{f}}[\|g(y)\|^2]$ is finite for any $\tilde{f} \in \mathcal{B}_\tau^\infty$. Here, $\lambda > 0$ is *a priori* fixed and such that $\lambda > (1 - \tau)\|P\|$. In this way the mismodeling tolerance is expressed as a soft constraint adding the penalty term $\lambda(c - \mathcal{D}_\tau(\tilde{f}\|f))$.

Corollary 3.1 *The optimal estimator, according to (12), is still the nominal Bayes estimator. The least favorable probability density \tilde{f}° has mean vector $\tilde{m}_z = m_z$ and covariance matrix \tilde{K}_z° as in (10). The least favorable estimation error has zero mean and covariance matrix \tilde{P} as in (11) where λ now has been chosen a priori.*

In the perspective presented in [1, 3, 4, 5, 9], Problem (12) represents a generalization of the risk-sensitive static estimation problem [15, 14, 6, 16, 22, 11]. Thus, the nominal *Bayes* estimator is also optimal for the τ risk-sensitive static estimation problem.

Now we show that Problem (12) is the solution to a new minimum entropy problem in the sense of [13]. Let $g(y) = Gy + h$ be an estimator of x . Under the nominal model, the estimation error $e = x - g(y)$ is Gaussian with

$$m_e = \mathbb{E}_f[e] = \begin{bmatrix} I_n & -G \end{bmatrix} m_z - h$$

$$K_e = \begin{bmatrix} I_n & -G \end{bmatrix} K_z \begin{bmatrix} I_n \\ -G^T \end{bmatrix}.$$

Note that, the mean and covariance matrix of e depends on $g(y)$. The idea is to characterize the robust estimator through the following minimum entropy problem

$$\min_{g \in \mathcal{G}} \mathcal{H}_\tau(e, \lambda), \quad (13)$$

where \mathcal{H}_τ is an entropy-like function which guarantees that the mean and the covariance matrix of e are bounded in some sense. Such boundedness is tuned by parameter $\lambda \in \mathbb{R}_+$. Next, we characterize \mathcal{H}_τ .

Definition 3.1 *The τ entropy family of e is defined as*

$$\mathcal{H}_\tau(e, \lambda) = \quad (14)$$

$$\begin{cases} m_e^T (I_n - \frac{1}{\lambda} K_e)^{-1} m_e - \lambda \log \det (I_n - \frac{1}{\lambda} K_e), & \tau = 0 \\ m_e^T (I_n - \frac{1-\tau}{\lambda} K_e)^{-1} m_e \\ \quad + \frac{\lambda}{\tau} \text{tr} \left((I_n - \frac{1-\tau}{\lambda} K_e)^{\frac{\tau}{\tau-1}} - I_n \right), & 0 < \tau < 1 \\ m_e m_e^T + \lambda \text{tr} \left(\exp \left(\frac{1}{\lambda} K_e \right) - I_n \right), & \tau = 1 \end{cases}$$

for $\lambda > (1 - \tau)\|K_e\|$ otherwise $\mathcal{H}_\tau(e, \lambda) = \infty$.

It is not difficult to see that \mathcal{H}_τ is continuous with respect to $\tau \in [0, 1]$. Note that, \mathcal{H}_0 is the usual entropy, see [13], used in H_∞ control.

Proposition 3.1 *The following properties hold:*

- $\mathcal{H}_\tau(e, \lambda) \geq 0$ and equality holds if and only if $e = 0$
- $\mathcal{H}_\tau(e, \cdot)$ is a monotone decreasing function.

The Theorem below states the connection between our minimax approach for robust estimation and the minimum entropy estimation.

Theorem 3.2 *Problem (12) and Problem (13) are equivalent.*

4 Noncausal Robust Filtering

Let $x(t)$ and $y(t)$ be two jointly stationary Gaussian processes defined over $t \in \mathbb{Z}$ of dimension n and p , respectively. We consider the noncausal filtering problem, that is to estimate $x(t)$ given the observations $\{y(s), s \in \mathbb{Z}\}$. We define $z(t) = [x(t)^T \ y(t)^T]^T$. The nominal probability measure of $z(t)$, say f , has power spectral density $S_z(\vartheta) = 2\pi m_z m_z^T \delta(\vartheta) + \Sigma_z(\vartheta)$, where

$$m_z = \begin{bmatrix} m_x \\ m_y \end{bmatrix}, \quad \Sigma_z(\vartheta) = \begin{bmatrix} \Sigma_x(\vartheta) & \Sigma_{xy}(\vartheta) \\ \Sigma_{yx}(\vartheta) & \Sigma_y(\vartheta) \end{bmatrix}.$$

The actual one, say \tilde{f} , has power spectral density $\tilde{S}_z(\vartheta) = 2\pi \tilde{m}_z \tilde{m}_z^T \delta(\vartheta) + \tilde{\Sigma}_z(\vartheta)$, where

$$\tilde{m}_z = \begin{bmatrix} \tilde{m}_x \\ \tilde{m}_y \end{bmatrix}, \quad \tilde{\Sigma}_z(\vartheta) = \begin{bmatrix} \tilde{\Sigma}_x(\vartheta) & \tilde{\Sigma}_{xy}(\vartheta) \\ \tilde{\Sigma}_{yx}(\vartheta) & \tilde{\Sigma}_y(\vartheta) \end{bmatrix}.$$

Suppose that the actual power spectral density belongs to the following closed ball parametrized by τ

$$\mathcal{B}_\tau = \{\tilde{f} \text{ s.t. } \mathcal{S}_\tau(\tilde{f} \| f) \leq c\}, \quad (15)$$

where \mathcal{S}_τ has been defined in (6) and $c \in \mathbb{R}_+$ is a fixed tolerance. Similarly to the static case, we design the robust noncausal filter according to the minimax point of view. Let \mathcal{G} denote class of estimators of $x(t)$ having the following structure

$$g(y, t) = \sum_{k=-\infty}^{\infty} G_k y(t-k) + h, \quad (16)$$

where the filter $\Lambda(\vartheta) = \sum_{k=-\infty}^{\infty} G_k e^{-j\vartheta k}$ is Bounded Input Bounded Output (BIBO) stable. Our robust noncausal filter is the solution to the following minimax problem

$$\min_{g \in \mathcal{G}} \max_{\tilde{f} \in \mathcal{B}_\tau} J(\tilde{f}, g), \quad (17)$$

where $J(\tilde{f}, g) = \mathbb{E}_{\tilde{f}}[\|x(t) - g(y, t)\|^2]$.

Theorem 4.1 Let τ be such that $\frac{1}{1-\tau} \in \mathbb{N}$ and Σ_z have bounded McMillan degree. The robust estimator $g^\circ(y, t)$, according to (17), is the noncausal Wiener filter based on f , that is

$$h = [1 \quad -\Lambda(0)] m_z, \quad \Lambda = \Sigma_{xy} \Sigma_y^{-1}.$$

The least favorable probability measure \tilde{f}° has power spectral density $\tilde{S}_z^\circ(\vartheta) = 2\pi m_z m_z^T \delta(\vartheta) + \tilde{\Sigma}_z^\circ(\vartheta)$ with

$$\tilde{\Sigma}_z^\circ = \begin{bmatrix} \tilde{\Sigma}_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix},$$

where only $\tilde{\Sigma}_x$ is perturbed. The power spectral density of the estimation error with respect to the nominal and the least favorable probability measure, respectively, are

$$\Sigma_e = \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx}, \quad \tilde{\Sigma}_e = \tilde{\Sigma}_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx}.$$

Then, $\tilde{\Sigma}_e$ can be expressed in terms of Σ_e as

$$\tilde{\Sigma}_e = \begin{cases} \Gamma_e \left(I_n - \frac{1-\tau}{\lambda} \Gamma_e^* \Gamma_e \right)^{\frac{1}{\tau-1}} \Gamma_e^*, & 0 \leq \tau < 1 \\ \Gamma_e \exp\left(\frac{1}{\lambda} \Gamma_e^* \Gamma_e\right) \Gamma_e^*, & \tau = 1, \end{cases}$$

where $\Sigma_e = \Gamma_e \Gamma_e^*$. Here, λ , with $\lambda > (1-\tau) \|\Sigma_e\|_\infty$, is the unique Lagrange multiplier such that $\mathcal{S}_\tau(f\|f) = c$.

Therefore, the noncausal Wiener filter is robust with respect to the τ divergence constraint. Similarly to the static case, the worst situation occurs when all the mismatch budget is allocated in a perturbation of the covariance matrix function Σ_x . The connection with the results in [9] is analogous to the static case. Moreover, the additional MSE occasioned by the least favorable model perturbation is $\Delta \text{MSE} = \text{tr} \int (\tilde{\Sigma}_e - \Sigma_e)$.

Problem (17) can be relaxed in the following way

$$\min_{g \in \mathcal{G}} \max_{\tilde{f} \in \mathcal{B}_\tau^\infty} J(\tilde{f}, g) + \lambda(c - \mathcal{S}_\tau(\tilde{f}\|f)), \quad (18)$$

where $\mathcal{B}_\tau^\infty = \{\tilde{f} \text{ s.t. } \mathcal{S}_\tau(\tilde{f}\|f) < \infty\}$. Here, λ is fixed *a priori* and such that $\lambda > (1-\tau) \|\Sigma_e\|_\infty$.

Remark 4.1 Also in this case, it is possible to show that Problem (18) is a minimum entropy problem in the sense of [13].

5 Simulation study

We analyze the impact of parameters c and τ on the least favorable statistics of the estimation error corresponding to the least favorable model in \mathcal{B}_τ . In the static case, given τ and c , λ is given by solving equation (30)-(32) in Appendix. The latter computation can be efficiently performed using the bisection method. Then, the least favorable statistics of the estimation error is given by Theorem 3.1. The same strategy is applied in the dynamic case.

5.1 Static Estimator

We consider a bidimensional Gaussian random vector x , that is $n = 2$. We assume that the nominal covariance matrix of the estimation error is

$$P = \begin{bmatrix} 0.15 & 0.05 \\ 0.05 & 0.1 \end{bmatrix}.$$

We consider the least favorable statistics $\tilde{f}^\circ \in \mathcal{B}_\tau$ where \mathcal{B}_τ has been defined in (7) with $\tau = 0$, $\tau = 0.5$ and $\tau = 1$. In Figure 2 we show the additional MSE occurred when $c \in [0.001, 0.1]$. We see that, for each value of τ , the larger c

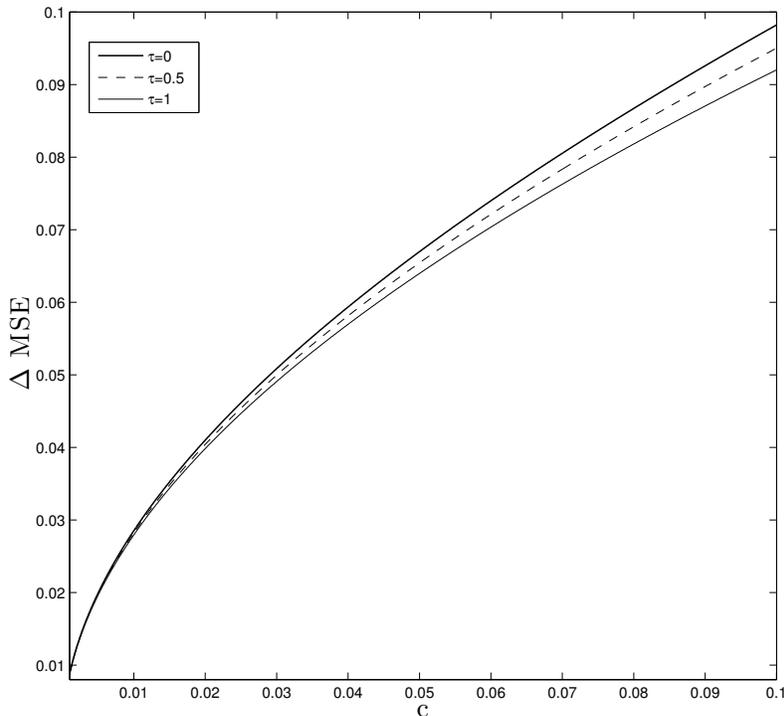


Figure 2: Additional MSE occasioned by the least favorable perturbation in \mathcal{B}_τ defined in (7) with $\tau = 0$, $\tau = 0.5$ and $\tau = 1$.

is, the larger ΔMSE is, as expected. Moreover, for c fixed, the larger τ is, the smaller ΔMSE is. Therefore, smaller values of τ corresponds to least favorable statistics with a larger ΔMSE . In order to compare the features of the three different balls \mathcal{B}_τ , with $\tau = 0$, $\tau = 0.5$ and $\tau = 1$, we fix the tolerance c for each

ball in such a way that $\Delta\text{MSE} = 0.08$ for the least favorable statistics. The tolerances, respectively, are $c_{\tau=0} = 0.0692$, $c_{\tau=0.5} = 0.0728$ and $c_{\tau=1} = 0.0767$. The corresponding covariance matrices of the least favorable estimation error are

$$\begin{aligned} \tilde{P}_{\tau=0} &= \begin{bmatrix} 0.2041 & 0.0783 \\ 0.0783 & 0.1259 \end{bmatrix}, \tilde{P}_{\tau=0.5} = \begin{bmatrix} 0.2039 & 0.0779 \\ 0.0779 & 0.1261 \end{bmatrix}, \\ \tilde{P}_{\tau=1} &= \begin{bmatrix} 0.2037 & 0.0775 \\ 0.0775 & 0.1263 \end{bmatrix}. \end{aligned}$$

One can see that the least favorable statistics in $\mathcal{B}_{\tau=0}$ tends to concentrate the perturbation on the component with larger nominal variance. On the contrary, the least favorable statistics in $\mathcal{B}_{\tau=1}$ tends to spread such perturbation among the two components. Finally, the least favorable statistics in $\mathcal{B}_{\tau=0.5}$ mitigates those two features.

5.2 Noncausal Filtering

We consider a Gaussian scalar process $x(t)$, that is $n = 1$. The spectral density of its nominal estimation error $e(t)$ is depicted in Figure 3. We consider three different balls \mathcal{B}_τ , see (15), with $\tau = 0$, $\tau = 0.5$ and $\tau = 1$. Also in this case we noticed that smaller values of τ corresponds to least favorable probability measures with a larger ΔMSE . Similarly to the static case, we fix the tolerance c for each ball in such a way that $\Delta\text{MSE} = 0.2$ for the least favorable statistics. The tolerances, respectively, are $c_{\tau=0} = 0.022$, $c_{\tau=0.5} = 0.025$ and $c_{\tau=1} = 0.028$. The corresponding least favorable spectral densities of $e(t)$ are depicted in Figure 3. One can see that the least favorable spectral density in $\mathcal{B}_{\tau=0}$ tends to concentrate the perturbation on the frequency band where the spectral density takes larger values. On the contrary, the least favorable spectral density in $\mathcal{B}_{\tau=1}$ tends to spread such perturbation over the entire frequency band. Also in this case, the least favorable spectral density in $\mathcal{B}_{\tau=0.5}$ mitigates those two features. Next, we consider a bidimensional Gaussian process $x(t)$, i.e. $n = 2$. The nominal spectral density of $e(t)$ is depicted in Figure 4. As before, we fix c for each ball in such a way that the least favorable statistics is such that $\Delta\text{MSE} = 1.7$. We found $c_{\tau=0} = 0.076$, $c_{\tau=0.5} = 0.088$ and $c_{\tau=1} = 0.1$. The corresponding least favorable spectral density in \mathcal{B}_τ , with $\tau = 0$, $\tau = 0.5$ and $\tau = 1$, is depicted in Figure 4. One can see that the least favorable spectral density in $\mathcal{B}_{\tau=0}$ allocates most of the perturbation in the second component, more precisely in the frequency band where the nominal spectral density takes the largest values (also compared with respect to the first component). On the contrary, the least favorable spectral density in $\mathcal{B}_{\tau=1}$ allocates more perturbation in the first component than the former; again, the latter tends to spread the perturbation among the two components and on the entire frequency band.

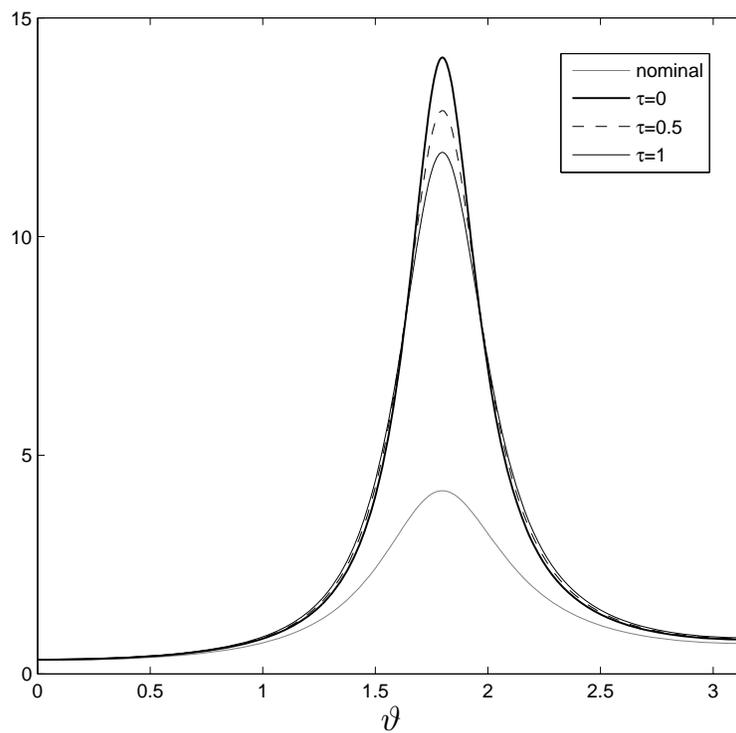


Figure 3: Spectral density of $e(t)$ with respect to the nominal statistics and with respect to the least favorable statistics in \mathcal{B}_τ defined in (15) with $\tau = 0$, $\tau = 0.5$ and $\tau = 1$.

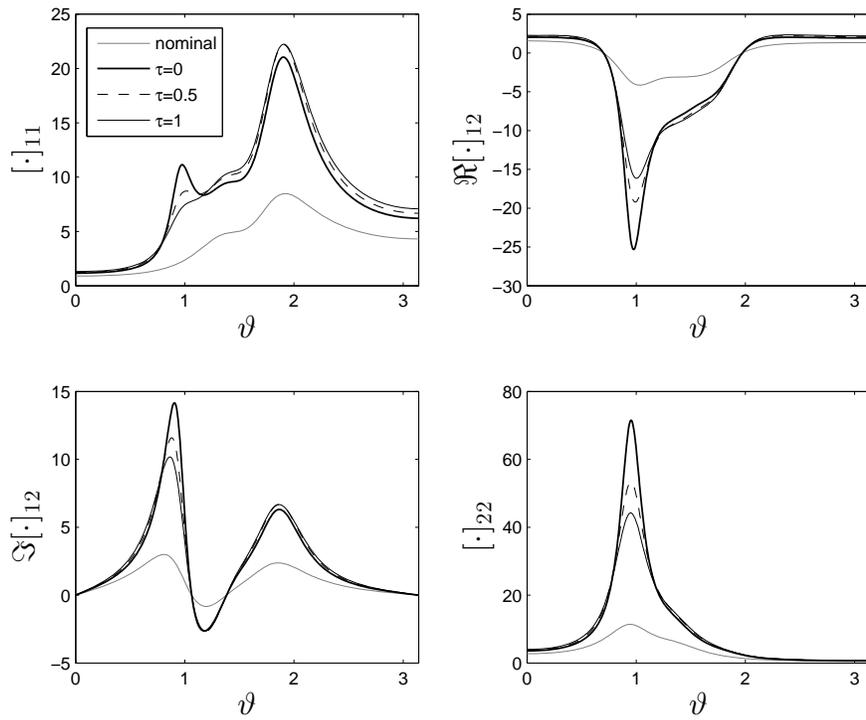


Figure 4: Spectral density of $e(t)$ with respect to the nominal statistics and with respect to the least favorable statistics in \mathcal{B}_τ defined in (15) with $\tau = 0$, $\tau = 0.5$ and $\tau = 1$.

6 Conclusions

In this paper, we showed that the *Bayes* estimator and the noncausal *Wiener* filter based on the nominal statistics are robust according to the minimax approach where the uncertainty is specified by a ball formed by placing a bound on the τ divergence family between the actual and the nominal statistics. Interestingly, the relaxation of this minimax problem can be understood as a family of risk-sensitive estimation problems which is also linked to a family of minimum entropy problems. Finally, through a simulation study, we have analyzed the features of this family of uncertainty classes. These results represent the starting point to derive a new family of robust *Kalman* filters characterized by the τ divergence family [21].

Appendix

A Proof of Proposition 2.1 and Proposition 2.2

To prove Proposition 2.1 it is sufficient to note that

$$\mathcal{D}_\tau(\tilde{f}||f) = \frac{1}{1-\tau} \|\Delta m_z\|_{K_z^{-1}}^2 + \overline{\mathcal{D}}_\tau(\tilde{K}_z||K_z), \quad (19)$$

where $\overline{\mathcal{D}}_\tau$ is the τ divergence between \tilde{K}_z and K_z [20]. The latter is nonnegative and is equal to zero if and only if $\tilde{K}_z = K_z$. Since $K_z \in \mathcal{Q}_+^q$, the first term in (19) is nonnegative and is equal to zero if and only if $\Delta m_z = 0$, that is $\tilde{m}_z = m_z$. Therefore, \mathcal{D}_τ is nonnegative and equality holds if and only if $\tilde{m}_z = m_z$ and $\tilde{K}_z = K_z$, that is $\tilde{f} = f$. The proof for cases $\tau = 0$ and $\tau = 1$ is similar. Finally, Proposition 2.2 can be proved along the same lines. ■

B Proof of Theorem 3.1

The proof is divided in three cases.

Case $\tau = 0$. \mathcal{D}_τ is equivalent to the *Kullback-Leibler* divergence and the statement has been proved in [9, Theorem 1].

Case $0 < \tau < 1$. We have to show that $J(\tilde{f}, g^\circ) \leq J(\tilde{f}^\circ, g^\circ) \leq J(\tilde{f}^\circ, g)$ for any $(\tilde{f}, g) \in \mathcal{B}_\tau \times \mathcal{G}$. Since $\tilde{f}^\circ \sim \mathcal{N}(\tilde{m}_z^\circ, \tilde{K}_z^\circ)$, the inequality $J(\tilde{f}^\circ, g^\circ) \leq J(\tilde{f}^\circ, g)$ implies that g° is the Bayesian estimator (9). Next, we show the inequality $J(\tilde{f}, g^\circ) \leq J(\tilde{f}^\circ, g^\circ)$ holds. Therefore, it is sufficient to show

$$\tilde{f}^\circ = \arg \max_{\tilde{f} \in \mathcal{B}_\tau} J(\tilde{f}, g^\circ). \quad (20)$$

Let $e = x - g^\circ(y) = [I_n \quad -G^\circ](z - m_z)$. Therefore, $\tilde{m}_e := \mathbb{E}_{\tilde{f}}[e] =$

$\begin{bmatrix} I_n & -G^\circ \end{bmatrix} \Delta m_z$ and

$$\begin{aligned} \tilde{K}_e &:= \mathbb{E}_{\tilde{f}}[(e - \tilde{m}_e)(e - \tilde{m}_e)^T] \\ &= \left(\begin{bmatrix} I_n & -G^\circ \end{bmatrix} \tilde{K}_z \begin{bmatrix} I_n \\ -(G^\circ)^T \end{bmatrix} \right). \end{aligned}$$

Moreover,

$$\begin{aligned} J(\tilde{f}, g^\circ) &= \text{tr}(\tilde{m}_e \tilde{m}_e^T + \tilde{K}_e) \\ &= \text{tr} \left(\begin{bmatrix} I_n & -G^\circ \end{bmatrix} (\tilde{K}_z + \Delta m_z \Delta m_z^T) \begin{bmatrix} I_n \\ -(G^\circ)^T \end{bmatrix} \right). \end{aligned}$$

In order to characterize \tilde{f} , we exploit the duality theory. The Lagrangian is

$$\begin{aligned} \mathcal{L}(\tilde{m}_z, \tilde{K}_z, \lambda) &= J(\tilde{f}, g^\circ) + \lambda(c - \mathcal{D}_\tau(\tilde{f} \| f)) \\ &= \Delta m_z^T W_\lambda \Delta m_z + \text{tr} \left(\begin{bmatrix} I_n & -G^\circ \end{bmatrix} \tilde{K}_z \begin{bmatrix} I_n \\ -(G^\circ)^T \end{bmatrix} \right) \\ &\quad + \lambda \left(c + \text{tr} \left(\frac{1}{\tau(1-\tau)} (L_z^{-1} \tilde{K}_z L_z^{-T})^\tau \right. \right. \\ &\quad \left. \left. - \frac{1}{1-\tau} \tilde{K}_z K_z^{-1} - \frac{1}{\tau} I_{n+p} \right) \right), \end{aligned} \quad (21)$$

where

$$W_\lambda = \begin{bmatrix} I_n \\ -(G^\circ)^T \end{bmatrix} \begin{bmatrix} I_n & -G^\circ \end{bmatrix} - \frac{\lambda}{1-\tau} K_z^{-1}$$

and $\lambda \geq 0$ is the *Lagrange* multiplier. Note that, \mathcal{L} is bounded above and strictly concave in \tilde{K}_z when $\lambda > 0$. Moreover, \mathcal{L} is bounded above and strictly concave in \tilde{m}_z if W_λ is negative definite. Define $M = \begin{bmatrix} I_n & -G^\circ \end{bmatrix} L_z$. Since $G^\circ = K_{xy} K_y^{-1}$ and $K_z = L_z L_z^T$, it is easy to check that

$$\begin{aligned} M M^T &= \begin{bmatrix} I_n & -G^\circ \end{bmatrix} K_z \begin{bmatrix} I_n \\ -(G^\circ)^T \end{bmatrix} \\ &= K_x - K_{xy} K_y^{-1} K_{yx} = P. \end{aligned}$$

Moreover, W_λ is congruent to $M^T M - \frac{\lambda}{1-\tau} I_{n+p}$ which is negative definite when

$$\lambda > (1-\tau) \|M^T M\| = (1-\tau) \|M M^T\| = (1-\tau) \|P\|. \quad (22)$$

Therefore, under assumption (22), \mathcal{L} is bounded above and strictly concave in $(\tilde{m}_z, \tilde{K}_z)$, and it is maximized by the points $(\tilde{m}_z^\circ, \tilde{K}_z^\circ)$ annihilating its first variations

$$\begin{aligned} \nabla_{\tilde{m}_z, u} \mathcal{L} &= 2 \Delta m_z^T W_\lambda u \\ \nabla_{\tilde{K}_z, V} \mathcal{L} &= \text{tr} \left(\left(\begin{bmatrix} I_n \\ -(G^\circ)^T \end{bmatrix} \begin{bmatrix} I_n & -G^\circ \end{bmatrix} \right. \right. \\ &\quad \left. \left. + \frac{\lambda}{1-\tau} (L_z^{-T} (L_z^{-1} \tilde{K}_z L_z^{-T})^{\tau-1} L_z^{-1} - K_z^{-1}) \right) V \right) \end{aligned}$$

for any direction $u \in \mathbb{R}^{n+p}$ and $V \in \mathcal{Q}^{n+p}$, respectively. To compute $\nabla_{\tilde{K}_z, V} \mathcal{L}$ we exploited the formula for the first variation of the exponentiation of a matrix given in [17]. Since W_λ is negative definite, it follows that $\Delta m_z = 0$, and thus $\tilde{m}_z^\circ = m_z$. Regarding \tilde{K}_z° , we have

$$\begin{aligned} & \frac{\lambda}{1-\tau} L_z^{-T} (L_z^{-1} \tilde{K}_z^\circ L_z^{-T})^{\tau-1} L_z^{-1} \\ &= \frac{\lambda}{1-\tau} K_z^{-1} - \begin{bmatrix} I_n & \\ & -(G^\circ)^T \end{bmatrix} \begin{bmatrix} I_n & -G^\circ \end{bmatrix} \end{aligned}$$

moreover,

$$\begin{aligned} & (L_z^{-1} \tilde{K}_z^\circ L_z^{-T})^{\tau-1} \\ &= I_{n+p} - \frac{1-\tau}{\lambda} L_z^T \begin{bmatrix} I_n & \\ & -(G^\circ)^T \end{bmatrix} \begin{bmatrix} I_n & -G^\circ \end{bmatrix} L_z, \end{aligned} \quad (23)$$

where the right hand side of (23) is positive definite by (22). Accordingly, under assumption (22), we have

$$\begin{aligned} & \tilde{K}_z^\circ = \\ & L_z \left(I_{n+p} - \frac{1-\tau}{\lambda} L_z^T \begin{bmatrix} I_n & \\ & -(G^\circ)^T \end{bmatrix} \begin{bmatrix} I_n & -G^\circ \end{bmatrix} L_z \right)^{\frac{1}{\tau-1}} L_z^T. \end{aligned} \quad (24)$$

Note that, K_z admits the following block upper diagonal lower (UDL) factorization

$$K_z = \begin{bmatrix} I_n & G^\circ \\ 0 & I_p \end{bmatrix} \begin{bmatrix} P & 0 \\ 0 & K_y \end{bmatrix} \begin{bmatrix} I_n & 0 \\ (G^\circ)^T & I_p \end{bmatrix} \quad (25)$$

and we choose L_z as

$$L_z = \begin{bmatrix} I_n & G^\circ \\ 0 & I_p \end{bmatrix} \begin{bmatrix} L_P & 0 \\ 0 & L_y \end{bmatrix}, \quad (26)$$

where $K_y = L_y L_y^T$ and $P = L_P L_P^T$. Accordingly,

$$\begin{aligned} & \left(I_{n+p} - \frac{1-\tau}{\lambda} L_z^T \begin{bmatrix} I_n & \\ & -(G^\circ)^T \end{bmatrix} \begin{bmatrix} I_n & -G^\circ \end{bmatrix} L_z \right)^{\frac{1}{\tau-1}} \\ &= \left(I_{n+p} - \frac{1-\tau}{\lambda} \begin{bmatrix} L_P^T & \\ & 0 \end{bmatrix} \begin{bmatrix} L_P & 0 \end{bmatrix} \right)^{\frac{1}{\tau-1}} \\ &= \left(\begin{bmatrix} I_n - \frac{1-\tau}{\lambda} L_P^T L_P & 0 \\ 0 & I_p \end{bmatrix} \right)^{\frac{1}{\tau-1}} \\ &= \begin{bmatrix} (I_n - \frac{1-\tau}{\lambda} L_P^T L_P)^{\frac{1}{\tau-1}} & 0 \\ 0 & I_p \end{bmatrix} \end{aligned} \quad (27)$$

and by (24), (26) and (27) we have

$$\tilde{K}_z^\circ = \begin{bmatrix} I_n & G^\circ \\ 0 & I_p \end{bmatrix} \begin{bmatrix} \tilde{P} & 0 \\ 0 & K_y \end{bmatrix} \begin{bmatrix} I_n & 0 \\ (G^\circ)^T & I_p \end{bmatrix}, \quad (28)$$

where

$$\tilde{P} = L_P \left(I_n - \frac{1-\tau}{\lambda} L_P^T L_P \right)^{\frac{1}{\tau-1}} L_P^T. \quad (29)$$

Moreover,

$$\tilde{K}_z^\circ = \begin{bmatrix} \tilde{P} + K_{xy} K_y^{-1} K_{yx} & K_{xy} \\ K_{yx} & K_y \end{bmatrix}$$

accordingly, $\tilde{K}_x = \tilde{P} + K_{xy} K_y^{-1} K_{yx}$. Since $\lambda > 0$, the duality gap between the primal and the dual is zero if and only if $\mathcal{D}_\tau(\tilde{f}^\circ \| f) = c$. It remains to be shown that there exists $\lambda > (1-\tau)\|P\|$ such that $\mathcal{D}_\tau(\tilde{f}^\circ \| f) = c$. By considering the factorizations (25), (28) and (29), we obtain

$$\begin{aligned} & \mathcal{D}_\tau(\tilde{f}^\circ \| f) \\ &= \text{tr} \left(\frac{1}{\tau(\tau-1)} (L_z^{-1} \tilde{K}_z^\circ L_z^{-T})^\tau + \frac{1}{1-\tau} \tilde{K}_z^\circ K_z^{-1} + \frac{1}{\tau} I_{n+p} \right) \\ &= \text{tr} \left(\frac{1}{\tau(\tau-1)} (L_P^{-1} \tilde{P} L_P^{-T})^\tau + \frac{1}{1-\tau} L_P^{-1} \tilde{P} L_P^{-T} + \frac{1}{\tau} I_n \right) \\ &= \text{tr} \left(\frac{1}{\tau(\tau-1)} \left(I_n - \frac{1-\tau}{\lambda} L_P^T L_P \right)^{\frac{\tau}{\tau-1}} \right. \\ & \quad \left. + \frac{1}{1-\tau} \left(I_n - \frac{1-\tau}{\lambda} L_P^T L_P \right)^{\frac{1}{\tau-1}} + \frac{1}{\tau} I_n \right). \end{aligned} \quad (30)$$

Let $L_P^T L_P = U D U^T$ be the eigenvalue decomposition of $L_P^T L_P$, where $U U^T = I_n$ and $D = \text{diag}(d_1, d_2, \dots, d_n)$ with $d_i = \sigma_i(L_P^T L_P)$. Therefore, we get

$$\begin{aligned} & \mathcal{D}_\tau(\tilde{f}^\circ \| f) \\ &= \text{tr} \left(\frac{1}{\tau(\tau-1)} \left(I_n - \frac{1-\tau}{\lambda} D \right)^{\frac{\tau}{\tau-1}} \right. \\ & \quad \left. + \frac{1}{1-\tau} \left(I_n - \frac{1-\tau}{\lambda} D \right)^{\frac{1}{\tau-1}} + \frac{1}{\tau} I_n \right) \\ &= \sum_{i=1}^n \gamma(\lambda, d_i) \end{aligned}$$

with

$$\begin{aligned}\gamma(\lambda, d_i) &= \frac{1}{\tau(\tau-1)} \left(1 - \frac{1-\tau}{\lambda} d_i\right)^{\frac{\tau}{\tau-1}} \\ &\quad + \frac{1}{1-\tau} \left(1 - \frac{1-\tau}{\lambda} d_i\right)^{\frac{1}{\tau-1}} + \frac{1}{\tau}.\end{aligned}$$

For $\lambda > (1-\tau)d_i > 0$, we have

$$\frac{d}{d\lambda} \gamma(\lambda, d_i) = -\frac{d_i^2}{\lambda^3} \left(1 - \frac{1-\tau}{\lambda} d_i\right)^{\frac{2-\tau}{\tau-1}} < 0$$

and

$$\lim_{\lambda \rightarrow \infty} \gamma(\lambda, d_i) = 0, \quad \lim_{\lambda \rightarrow ((1-\tau)d_i)^+} \gamma(\lambda, d_i) = \infty$$

so that $\gamma(\lambda, d_i)$ is a monotone decreasing function of λ . Since $\mathcal{D}_\tau(\tilde{f}^\circ \| f)$ is the sum of $\gamma(\lambda, d_i)$ with $i = 1 \dots n$, it is also a monotone decreasing function of λ over $((1-\tau)\|P\|, \infty)$ and

$$\lim_{\lambda \rightarrow \infty} \mathcal{D}_\tau(\tilde{f}^\circ \| f) = 0, \quad \lim_{\lambda \rightarrow ((1-\tau)\|P\|)^+} \mathcal{D}_\tau(\tilde{f}^\circ \| f) = \infty. \quad (31)$$

We conclude, for any $c > 0$ there exists a unique $\lambda > (1-\tau)\|P\|$ such that $\mathcal{D}_\tau(\tilde{f}^\circ \| f) = c$.

Case $\tau = 1$. The Lagrangian \mathcal{L} can be formed as in (21). Then, one can see that \mathcal{L} is bounded below if and only if $\tilde{m}_z = m_z$. Accordingly, the least favorable density \tilde{f}° has mean m_z . Then, the least favorable \tilde{K}_z° can be characterized using the duality theory similarly to the case $0 < \tau < 1$. In particular, it is not difficult to see that

$$\begin{aligned}\mathcal{D}_1(\tilde{f}^\circ \| f) &= \text{tr} \left(\exp \left(\frac{1}{\lambda} L_P^T L_P \right) \left(\frac{1}{\lambda} L_P^T L_P - I_n \right) + I_n \right).\end{aligned} \quad (32)$$

■

C Proof of Proposition 3.1

The statement can be proved along the same lines of Proposition 2.3.1 and Proposition 2.3.2 in [13]. ■

D Proof of Theorem 3.2

Exploiting similar argumentations used in the proof of Proposition 1 in [9], we have that

$$\max_{\tilde{f} \in \mathcal{B}^\infty} \mathcal{L}(\tilde{f}, \lambda, g) = \mathcal{H}_\tau(e, \lambda) + \lambda c,$$

where $\mathcal{L}(\tilde{f}, \lambda, g) = J(\tilde{f}, g) + \lambda(c - \mathcal{D}_\tau(\tilde{f} \| f))$. Since g does not depend on λ , we conclude that

$$\min_{g \in \mathcal{G}} \max_{\tilde{f} \in \mathcal{B}^\infty} \mathcal{L}(\tilde{f}, \lambda, g) = \min_{g \in \mathcal{G}} \mathcal{H}_\tau(e, \lambda). \quad \blacksquare$$

E Proof of Theorem 4.1

The proof is similar to the one of Theorem 3.1. The unique peculiarity follows. Since Σ_z has bounded *McMillan* degree and $\frac{1}{1-\tau} \in \mathbb{N}$, then the integrand function in $\mathcal{S}_\tau(\tilde{f}^\circ \| f)$ is rational. Moreover, as $\lambda \rightarrow ((1-\tau)\|\Sigma_e\|_\infty)^+$ this integrand function has at least one pole tending to \mathbb{T} . Accordingly, these assumptions allow to conclude that

$$\lim_{\lambda \rightarrow ((1-\tau)\|\Sigma_e\|_\infty)^+} \mathcal{S}_\tau(\tilde{f}^\circ \| f) = \infty. \quad \blacksquare$$

References

- [1] R. Boel, M. James, and I. Petersen. Robustness and risk-sensitive filtering. *IEEE Trans. Automat. Control*, 47(3):451–461, 2002.
- [2] Andrzej Cichocki and Shun-ichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.
- [3] L. Hansen and T. Sargent. Robust estimation and control under commitment. *Journal of Economic Theory*, 14:258–301, 2005.
- [4] L. Hansen and T. Sargent. Recursive robust estimation and control without commitment. *Journal of Economic Theory*, 136(1):1–27, 2007.
- [5] L. Hansen and T. Sargent. *Robustness*. Princeton University Press, Princeton, NJ, 2008.
- [6] B. Hassibi, A. Sayed, and T. Kailath. *Indefinite-Quadratic Estimation and Control - A Unified Approach to H^2 and H^∞ Theories*. Soc. Indust. Appl. Math., Philadelphia, 1999.
- [7] Fumitada Itakura and Shuzo Saito. Analysis synthesis telephony based on the maximum likelihood method. In *Proc. of the 6th Int. Congr. on Acoust.*, Tokyo, Japan, 1968.

- [8] S. Kassam and H. Poor. Robust techniques for signal processing: A survey. *Proc. IEEE*, 73(3):433–481, Mar. 1985.
- [9] B. Levy and R. Nikoukhah. Robust least-squares estimation with a relative entropy constraint. *IEEE Trans. Informat. Theory*, 50(1):89–104, Jan. 2004.
- [10] B. Levy and R. Nikoukhah. Robust state-space filtering under incremental model perturbations subject to a relative entropy tolerance. *IEEE Trans. Automat. Control*, 58:682–695, Mar. 2013.
- [11] B. Levy and M. Zorzi. A contraction analysis of the convergence of risk-sensitive filters. *SIAM J. Optimization Control*, 54:2154–2173, 2016.
- [12] A. Lindquist and G. Picci. *Linear Stochastic Systems: A Geometric Approach to Modeling, Estimation and Identification*. Springer-Verlag, Berlin Heidelberg, 2015.
- [13] D. Mustafa and K. Glover. *Minimum entropy H^∞ control*. Springer-Verlag, Berlin Heidelberg, 1990.
- [14] J. Speyer and W. Chung. *Stochastic Processes, Estimation, and Control*. SIAM, Philadelphia, 2008.
- [15] P. Whittle. *Risk-sensitive Optimal Control*. J. Wiley, Chichester, England, 1980.
- [16] K. Zhou, J. Doyle, and K. Glover. *Robust and optimal control*. Prentice hall, New Jersey, 1996.
- [17] M. Zorzi. A new family of high-resolution multivariate spectral estimators. *IEEE Trans. Automat. Control*, 59(4):892–904, Apr. 2014.
- [18] M. Zorzi. Rational approximations of spectral densities based on the alpha divergence. *Mathematics of Control, Signals, and Systems*, 26(2):259–278, 2014.
- [19] M. Zorzi. An interpretation of the dual problem of the THREE-like approaches. *Automatica*, 62:87–92, 2015.
- [20] M. Zorzi. Multivariate Spectral Estimation based on the concept of Optimal Prediction. *IEEE Trans. Automat. Control*, 60:1647–1652, Jun. 2015.
- [21] M. Zorzi. Robust Kalman filtering under model perturbations. *IEEE Trans. Automat. Control*, 62:to appear, 2017.
- [22] M. Zorzi and B. C. Levy. On the convergence of a risk sensitive like filter. In *54th IEEE Conference on Decision and Control*, Osaka, Japan, 2015.