Distributionally Robust State Estimation

Wang Shixiong
(s.wang@u.nus.edu)

Department of Industrial Systems Engineering and Management,
National University of Singapore

April 2022

Sources and Declaration

Contents excerpted from

1. **Shixiong Wang**, Zhongming Wu, and Andrew Lim, "Robust State Estimation for Linear Systems Under Distributional Uncertainty", *IEEE Transactions on Signal Processing*, vol. 69, pp. 5963–5978, 2021. DOI:10.1109/TSP.2021.3118540.
2. **Shixiong Wang** and Zhisheng Ye, "Distributionally Robust State Estimation for Linear Systems Subject to Uncertainty and Outlier", *IEEE Transactions on Signal Processing*, vol. 70, pp. 452-467, 2021. DOI:10.1109/TSP.2021.3136804.
3. **Shixiong Wang**, "Distributionally Robust State Estimation for Nonlinear Systems", Major Revision at *IEEE Transactions on Signal Processing*.

Content

Content

A motivating example

Research on state estimation is active in many fields, e.g., astronautics, robotics, reliability engineering, geodesy, power system.

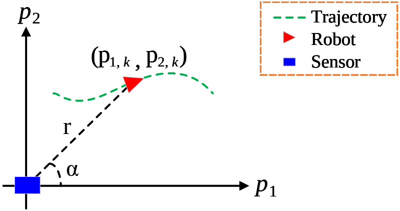

Figure: A 2-dimensional robot tracking problem.

- **Aim** to infer the real-time positions and velocities of the moving robot with observable information from a sensor.

A motivating example

Research on state estimation is active in many fields, e.g., astronautics, robotics, reliability engineering, geodesy, power system.

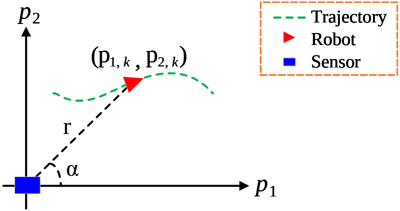

Figure: A 2-dimensional robot tracking problem.

- **Aim** to infer the real-time positions and velocities of the moving robot with observable information from a sensor.
- At time k, the sensor can capture the noisy value of $\mathbf{p}_k$, or the noisy values of some transforms of $\mathbf{p}_k$. Typical transforms include:

A motivating example

Research on state estimation is active in many fields, e.g., astronautics, robotics, reliability engineering, geodesy, power system.

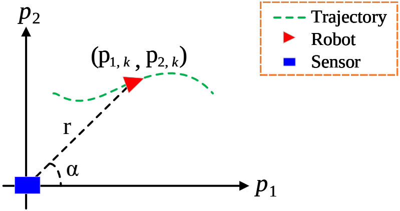

Figure: A 2-dimensional robot tracking problem.

- **Aim** to infer the real-time positions and velocities of the moving robot with observable information from a sensor.
- At time k, the sensor can capture the noisy value of $\mathbf{p}_k$, or the noisy values of some transforms of $\mathbf{p}_k$. Typical transforms include:
  - The range r.

A motivating example

Research on state estimation is active in many fields, e.g., astronautics, robotics, reliability engineering, geodesy, power system.

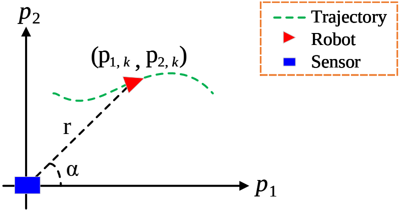

Figure: A 2-dimensional robot tracking problem.

- **Aim** to infer the real-time positions and velocities of the moving robot with observable information from a sensor.
- At time k, the sensor can capture the noisy value of $\mathbf{p}_k$, or the noisy values of some transforms of $\mathbf{p}_k$. Typical transforms include:
  - The range r.
  - The azimuth α.

A motivating example

According to basic kinematics, we have

$$\begin{bmatrix} \mathbf{p}_k \\ \boldsymbol{v}_k \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}_{k-1} \\ \boldsymbol{v}_{k-1} \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{a}_{k-1}.$$

- T denotes the sampling time between the time instant $k-1$ and the time instant k.

A motivating example

According to basic kinematics, we have

$$\begin{bmatrix} \mathbf{p}_k \\ \boldsymbol{v}_k \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}_{k-1} \\ \boldsymbol{v}_{k-1} \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{a}_{k-1}.$$

- T denotes the sampling time between the time instant $k-1$ and the time instant k.
- $\boldsymbol{v}_{k-1}$ is the average velocity in-between the time instants $k-1$ and k.

A motivating example

According to basic kinematics, we have

$$\begin{bmatrix} \mathbf{p}_k \\ \boldsymbol{v}_k \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}_{k-1} \\ \boldsymbol{v}_{k-1} \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{a}_{k-1}.$$

- T denotes the sampling time between the time instant $k-1$ and the time instant k.
- $\boldsymbol{v}_{k-1}$ is the average velocity in-between the time instants $k-1$ and k.
- $\mathbf{a}_{k-1}$ is the average acceleration during the same time slot.

A motivating example

According to basic kinematics, we have

$$\begin{bmatrix} \mathbf{p}_k \\ \boldsymbol{v}_k \end{bmatrix} = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{p}_{k-1} \\ \boldsymbol{v}_{k-1} \end{bmatrix} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{a}_{k-1}.$$

- T denotes the sampling time between the time instant $k-1$ and the time instant k.
- $\boldsymbol{v}_{k-1}$ is the average velocity in-between the time instants $k-1$ and k.
- $\mathbf{a}_{k-1}$ is the average acceleration during the same time slot.
- But $\forall k$, $\mathbf{p}_k$, $\mathbf{p}_{k-1}$, $\boldsymbol{v}_{k-1}$, and $\mathbf{a}_{k-1}$ are all unknown to us.

A motivating example

The **process dynamics equation** (also known as the state evolution equation or the state transition equation)

$$\mathbf{x}_k = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{w}_{k-1},$$

- $\mathbf{x}_k := [\mathbf{p}_k^\top, \boldsymbol{v}_k^\top]^\top$ and $\mathbf{x}_k$ is termed the **state** vector.

A motivating example

The **process dynamics equation** (also known as the state evolution equation or the state transition equation)

$$\mathbf{x}_k = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{w}_{k-1},$$

- $\mathbf{x}_k := [\mathbf{p}_k^\top, \boldsymbol{v}_k^\top]^\top$ and $\mathbf{x}_k$ is termed the **state** vector.
- We are using a random vector $\mathbf{w}_{k-1}$ to model the unknown acceleration $\mathbf{a}_{k-1}$.

A motivating example

The **process dynamics equation** (also known as the state evolution equation or the state transition equation)

$$\mathbf{x}_k = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{w}_{k-1},$$

- $\mathbf{x}_k := [\mathbf{p}_k^\top, \boldsymbol{v}_k^\top]^\top$ and $\mathbf{x}_k$ is termed the **state** vector.
- We are using a random vector $\mathbf{w}_{k-1}$ to model the unknown acceleration $\mathbf{a}_{k-1}$.
- $\mathbf{w}_{k-1}$ is the **process noise** vector.

A motivating example

The **measurement dynamics** equation (also known as the state measurement equation or the state observation equation) might be

$$\mathbf{y}_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}_k \\ \boldsymbol{v}_k \end{bmatrix} + \mathbf{v}_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}_k + \mathbf{v}_k.$$

or, if a different sensor is used, it might be

$$\mathbf{y}_k = \begin{bmatrix} \mathrm{r} \\ \alpha \end{bmatrix} + \mathbf{v}_k = \begin{bmatrix} \sqrt{\mathrm{p}_{1,k}^2 + \mathrm{p}_{2,k}^2} \\ \arctan\left(\frac{\mathrm{p}_{2,k}}{\mathrm{p}_{1,k}}\right) \end{bmatrix} + \mathbf{v}_k.$$

- $\mathbf{y}_k$ is the **measurement** vector

A motivating example

The **measurement dynamics** equation (also known as the state measurement equation or the state observation equation) might be

$$\mathbf{y}_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{p}_k \\ \boldsymbol{v}_k \end{bmatrix} + \mathbf{v}_k = \begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x}_k + \mathbf{v}_k.$$

or, if a different sensor is used, it might be

$$\mathbf{y}_k = \begin{bmatrix} \mathrm{r} \\ \alpha \end{bmatrix} + \mathbf{v}_k = \begin{bmatrix} \sqrt{\mathrm{p}_{1,k}^2 + \mathrm{p}_{2,k}^2} \\ \arctan\left(\frac{\mathrm{p}_{2,k}}{\mathrm{p}_{1,k}}\right) \end{bmatrix} + \mathbf{v}_k.$$

- $\mathbf{y}_k$ is the **measurement** vector
- $\mathbf{v}_k$ is the **measurement noise** vector.

Linear Systems

A **linear system** is given as

$$\begin{cases} \mathbf{x}_k & = \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k & = \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

In the contexts of the robot tracking problem above, we specifically have

$$\boldsymbol{F}_{k-1} := \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}, \quad \boldsymbol{G}_{k-1} := \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix}, \quad \text{and} \quad \boldsymbol{H}_k := \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

Nonlinear Systems

A **nonlinear system** is given as

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}), \\ \mathbf{y}_k &= \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k), \end{cases}$$

where $\boldsymbol{f}_k(\cdot,\cdot)$ and $\boldsymbol{h}_k(\cdot,\cdot)$ are termed the **process dynamics** function and the **measurement dynamics** function, respectively. In the contexts of the robot tracking problem above, we specifically have

$$\mathbf{x}_k = \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) := \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix} \mathbf{x}_{k-1} + \begin{bmatrix} \frac{T^2}{2} \\ T \end{bmatrix} \mathbf{w}_{k-1},$$

where $\boldsymbol{f}_k(\cdot,\cdot)$ degenerates to a linear form and

$$\mathbf{y}_k = \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k) := \begin{bmatrix} \sqrt{\mathrm{p}_{1,k}^2 + \mathrm{p}_{2,k}^2} \\ \arctan\left(\frac{\mathrm{p}_{2,k}}{\mathrm{p}_{1,k}}\right) \end{bmatrix} + \mathbf{v}_k,$$

where $\boldsymbol{h}_k(\cdot,\cdot)$ is of a nonlinear form.

Problem Statement

- Definition: State estimation is to estimate the unknown state $\mathbf{x}_k$ based on observable information $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$ and the linear/nonlinear system dynamics, for every discrete time index k.

Problem Statement

- Definition: State estimation is to estimate the unknown state $\mathbf{x}_k$ based on observable information $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$ and the linear/nonlinear system dynamics, for every discrete time index k.
- Issues: Possible model mismatches. **e.g., in the robot example:**

Problem Statement

- Definition: State estimation is to estimate the unknown state $\mathbf{x}_k$ based on observable information $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$ and the linear/nonlinear system dynamics, for every discrete time index k.
- Issues: Possible model mismatches. **e.g., in the robot example:**
  - $\forall k$, use random vector $\mathbf{w}_k$ with assumed-known distribution (usually a white Gaussian process) to model unknown acceleration $\mathbf{a}_k$ [13]. This is not always exact.

Problem Statement

- Definition: State estimation is to estimate the unknown state $\mathbf{x}_k$ based on observable information $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$ and the linear/nonlinear system dynamics, for every discrete time index k.
- Issues: Possible model mismatches. **e.g., in the robot example:**
  - $\forall k$, use random vector $\mathbf{w}_k$ with assumed-known distribution (usually a white Gaussian process) to model unknown acceleration $\mathbf{a}_k$ [13]. This is not always exact.
  - Measurement noise $\mathbf{v}_k$ is usually assumed to be Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{R}_k$ [16, 3]. However, the true distribution might be non-Gaussian, and/or the noise statistics are not exactly the same as $\mathbf{0}$ and $\boldsymbol{R}_k$ [6, 2].

Problem Statement

- Definition: State estimation is to estimate the unknown state $\mathbf{x}_k$ based on observable information $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$ and the linear/nonlinear system dynamics, for every discrete time index k.
- Issues: Possible model mismatches. **e.g., in the robot example:**
  - $\forall k$, use random vector $\mathbf{w}_k$ with assumed-known distribution (usually a white Gaussian process) to model unknown acceleration $\mathbf{a}_k$ [13]. This is not always exact.
  - Measurement noise $\mathbf{v}_k$ is usually assumed to be Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{R}_k$ [16, 3]. However, the true distribution might be non-Gaussian, and/or the noise statistics are not exactly the same as $\mathbf{0}$ and $\boldsymbol{R}_k$ [6, 2].
  - Due to clock error, the sensor's true sampling time might be different from the nominal value T; i.e., $\boldsymbol{F}_k = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$ might be uncertain.

Problem Statement

- Definition: State estimation is to estimate the unknown state $\mathbf{x}_k$ based on observable information $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$ and the linear/nonlinear system dynamics, for every discrete time index k.
- Issues: Possible model mismatches. **e.g., in the robot example:**
  - $\forall k$, use random vector $\mathbf{w}_k$ with assumed-known distribution (usually a white Gaussian process) to model unknown acceleration $\mathbf{a}_k$ [13]. This is not always exact.
  - Measurement noise $\mathbf{v}_k$ is usually assumed to be Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{R}_k$ [16, 3]. However, the true distribution might be non-Gaussian, and/or the noise statistics are not exactly the same as $\mathbf{0}$ and $\boldsymbol{R}_k$ [6, 2].
  - Due to clock error, the sensor's true sampling time might be different from the nominal value T; i.e., $\boldsymbol{F}_k = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$ might be uncertain.
- Consequences: Optimal state estimator designed for the nominal model degrades or even diverges.

Problem Statement

- Definition: State estimation is to estimate the unknown state $\mathbf{x}_k$ based on observable information $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$ and the linear/nonlinear system dynamics, for every discrete time index k.
- Issues: Possible model mismatches. **e.g., in the robot example:**
  - $\forall k$, use random vector $\mathbf{w}_k$ with assumed-known distribution (usually a white Gaussian process) to model unknown acceleration $\mathbf{a}_k$ [13]. This is not always exact.
  - Measurement noise $\mathbf{v}_k$ is usually assumed to be Gaussian with mean $\mathbf{0}$ and covariance $\boldsymbol{R}_k$ [16, 3]. However, the true distribution might be non-Gaussian, and/or the noise statistics are not exactly the same as $\mathbf{0}$ and $\boldsymbol{R}_k$ [6, 2].
  - Due to clock error, the sensor's true sampling time might be different from the nominal value T; i.e., $\boldsymbol{F}_k = \begin{bmatrix} 1 & T \\ 0 & 1 \end{bmatrix}$ might be uncertain.
- Consequences: Optimal state estimator designed for the nominal model degrades or even diverges.
- Philosophy: Robust solutions insensitive to model mismatches are expected.

Literature Review

Literature Review: For linear systems.

- Unknown-input filters, [8] etc.:

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{\Gamma}_{k-1}\mathbf{d}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

where $\mathbf{d}_{k-1} \in \mathbb{R}^q$ is the unknown input used to describe the parameter uncertainties. But need to elegantly specify $\boldsymbol{\Gamma}_{k-1}$.

Literature Review: For linear systems.

- Unknown-input filters, [8] etc.:

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{\Gamma}_{k-1}\mathbf{d}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

where $\mathbf{d}_{k-1} \in \mathbb{R}^q$ is the unknown input used to describe the parameter uncertainties. But need to elegantly specify $\boldsymbol{\Gamma}_{k-1}$.

- In [15, 18] etc.,

$$\begin{cases} \mathbf{x}_k &= (\boldsymbol{F}_{k-1} + \delta\boldsymbol{F}_{k-1})\mathbf{x}_{k-1} + (\boldsymbol{G}_{k-1} + \delta\boldsymbol{G}_{k-1})\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

where $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ are used to model the perturbations imposed on the nominal system matrices $\boldsymbol{F}_{k-1}$ and $\boldsymbol{G}_{k-1}$, respectively. But need to elegantly specify $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$.

Literature Review

Literature Review: For linear systems.

- Distributionally Robust Estimation Using Wasserstein Metric [1]: At each k, assuming true joint distribution of $(\mathbf{x}_k, \mathbf{y}_k) \sim \mathbb{P}$ is in

$$\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta) := \left\{\mathbb{P} \in \mathcal{N}_{n+m} \,\middle|\, \mathrm{W}(\mathbb{P}, \ \bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}) \leq \theta \right\}$$

  where W defines Wasserstein distance and $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$ is the nominal distribution (which is Gaussian). $\mathcal{N}_{n+m}$: Gaussian distributions. Only linear estimator studied, and cannot handle outliers in $\mathbf{y}_k$. Resulted Nonlinear Semi-Definite Program is hard to solve.

Literature Review

Literature Review: For linear systems.

- Distributionally Robust Estimation Using Wasserstein Metric [1]: At each k, assuming true joint distribution of $(\mathbf{x}_k, \mathbf{y}_k) \sim \mathbb{P}$ is in

$$\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta) := \left\{\mathbb{P} \in \mathcal{N}_{n+m} \,\middle|\, \mathrm{W}(\mathbb{P}, \ \bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}) \leq \theta \right\}$$

  where W defines Wasserstein distance and $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$ is the nominal distribution (which is Gaussian). $\mathcal{N}_{n+m}$: Gaussian distributions. Only linear estimator studied, and cannot handle outliers in $\mathbf{y}_k$. Resulted Nonlinear Semi-Definite Program is hard to solve.

- In [9, 17, 14] etc., for linear systems

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

  the distribution of $\mathbf{v}_k$ is assumed to be t-distributed, Laplacian, etc. But cannot handle parameter uncertainties.

Literature Review

Literature Review: For nonlinear systems.

- Particle filters are most popular.

Literature Review

Literature Review: For nonlinear systems.

- Particle filters are most popular.
- Over the years, efforts are in designing efficient sampling and resampling techniques [4, 12, 5, 11].

Literature Review: For nonlinear systems.

- Particle filters are most popular.
- Over the years, efforts are in designing efficient sampling and resampling techniques [4, 12, 5, 11].
- Virtually all of the past literature assume that the process dynamics and measurement dynamics are accurate.

Literature Review

Literature Review: For nonlinear systems.

- Particle filters are most popular.
- Over the years, efforts are in designing efficient sampling and resampling techniques [4, 12, 5, 11].
- Virtually all of the past literature assume that the process dynamics and measurement dynamics are accurate.
- There is no literature addressing model uncertainties and measurement outliers in particle filtering.

Research Aim and Methodology

- Research Aim:

Research Aim and Methodology

- Research Aim:
  - We study robust state estimators for both linear systems and nonlinear systems.

Research Aim and Methodology

- Research Aim:
  - We study robust state estimators for both linear systems and nonlinear systems.
  - For linear systems, the robust state estimators should be insensitive to both parameter uncertainties and measurement outliers.

Research Aim and Methodology

- Research Aim:
  - We study robust state estimators for both linear systems and nonlinear systems.
  - For linear systems, the robust state estimators should be insensitive to both parameter uncertainties and measurement outliers.
  - For nonlinear systems, we focus on robustifying the particle filter so that it can be insensitive to model uncertainties and measurement outliers.

Research Aim and Methodology

- Research Aim:
  - We study robust state estimators for both linear systems and nonlinear systems.
  - For linear systems, the robust state estimators should be insensitive to both parameter uncertainties and measurement outliers.
  - For nonlinear systems, we focus on robustifying the particle filter so that it can be insensitive to model uncertainties and measurement outliers.
- Methodology: We leverage Distributionally Robust Optimization Theories [7] and Robust Statistics Theories [10].

Content

Aim: to estimate the hidden state vector $\mathbf{x}_k$ of a linear Markov system given the measurement set $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$.

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

- where k is the discrete time index.

Aim: to estimate the hidden state vector $\mathbf{x}_k$ of a linear Markov system given the measurement set $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$.

$$\begin{cases} \mathbf{x}_k & = \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k & = \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

- where k is the discrete time index.
- $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{w}_k$, and $\mathbf{v}_k$ have finite second moments.

Aim: to estimate the hidden state vector $\mathbf{x}_k$ of a linear Markov system given the measurement set $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$.

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

- where k is the discrete time index.
- $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{w}_k$, and $\mathbf{v}_k$ have finite second moments.
- In literature, typical assumptions:

Aim: to estimate the hidden state vector $\mathbf{x}_k$ of a linear Markov system given the measurement set $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$.

$$\begin{cases} \mathbf{x}_k & = \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k & = \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

- where k is the discrete time index.
- $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{w}_k$, and $\mathbf{v}_k$ have finite second moments.
- In literature, typical assumptions:
  1. $\mathbf{w}_k \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{Q}_k)$, and $\mathbf{v}_k \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{R}_k)$; Hence, no measurement outliers modeled.

Linear System Case

Problem Background, Statements, and Motivations

Aim: to estimate the hidden state vector $\mathbf{x}_k$ of a linear Markov system given the measurement set $\{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$.

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

- where k is the discrete time index.
- $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{w}_k$, and $\mathbf{v}_k$ have finite second moments.
- In literature, typical assumptions:
  1. $\mathbf{w}_k \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{Q}_k)$, and $\mathbf{v}_k \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{R}_k)$; Hence, no measurement outliers modeled.
  2. $\boldsymbol{Q}_k$, $\boldsymbol{R}_k$, $\boldsymbol{F}_{k-1}$, $\boldsymbol{G}_{k-1}$, and $\boldsymbol{H}_k$ are exactly known.

For every discrete time index $k = 1, 2, ..., $ let

$$\mathcal{Y}_k := (\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k)$$

denote the measurement set.

Let

$$\mathcal{H}'_{\mathcal{Y}_k} := \left\{ \phi(\mathbf{y}_1, ..., \mathbf{y}_k) \;\middle|\; \begin{array}{l} \phi : \underbrace{\mathbb{R}^m \times \cdots \times \mathbb{R}^m}_{k} \to \mathbb{R}^n \\ \phi \text{ is Borel-measurable} \\ \int_{\mathbb{R}^{m\times k}} [\phi(\boldsymbol{Y}_k)]^\top [\phi(\boldsymbol{Y}_k)] \mathrm{d}\mathbb{P}_{\mathcal{Y}_k}(\boldsymbol{Y}_k) < \infty \end{array} \right\}.$$

Intuitively, $\mathcal{H}'_{\mathcal{Y}_k}$ contains all possible estimator of $\mathbf{x}_k$:

$$\hat{\mathbf{x}}_k = \phi(\mathbf{y}_1, ..., \mathbf{y}_k), \quad \forall k.$$

- Suppose the nominal joint state-measurement distribution at time k is $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$. We would like to solve the following optimization problem

$$\min_{\phi \in \mathcal{H}'_{\mathcal{Y}_k}} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \phi(\mathcal{Y}_k)][\mathbf{x}_k - \phi(\mathcal{Y}_k)]^\top,$$

- Suppose the nominal joint state-measurement distribution at time k is $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$. We would like to solve the following optimization problem

$$\min_{\phi \in \mathcal{H}'_{\mathcal{Y}_k}} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \phi(\mathcal{Y}_k)][\mathbf{x}_k - \phi(\mathcal{Y}_k)]^\top,$$

- Expectation is taken over $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$ and $\phi(\cdot)$ is referred to as an estimator (the optimal one is called the optimal estimator).

- Suppose the nominal joint state-measurement distribution at time k is $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$. We would like to solve the following optimization problem

$$\min_{\phi \in \mathcal{H}'_{\mathcal{Y}_k}} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \phi(\mathcal{Y}_k)][\mathbf{x}_k - \phi(\mathcal{Y}_k)]^\top,$$

- Expectation is taken over $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$ and $\phi(\cdot)$ is referred to as an estimator (the optimal one is called the optimal estimator).
- The optimal estimator of $\mathbf{x}_k$ in this minimum mean square error sense is $\mathbb{E}(\mathbf{x}_k|\mathcal{Y}_k)$.

- Question: What if the true joint state-measurement distribution at time k, i.e., $\mathbb{P}_{\mathbf{x}_k, \mathcal{Y}_k}$, deviates from the nominal $\bar{\mathbb{P}}_{\mathbf{x}_k, \mathcal{Y}_k}$?

- Question: What if the true joint state-measurement distribution at time k, i.e., $\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}$, deviates from the nominal $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$?
- Philosophy: We aim to find a robust state estimation solution that is insensitive to the model deviation.

- Question: What if the true joint state-measurement distribution at time k, i.e., $\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}$, deviates from the nominal $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$?
- Philosophy: We aim to find a robust state estimation solution that is insensitive to the model deviation.
- Motivation: We study the distributionally robust counterpart:

$$\min_{\phi \in \mathcal{H}'_{\mathcal{Y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta)} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \phi(\mathcal{Y}_k)][\mathbf{x}_k - \phi(\mathcal{Y}_k)]^\top,$$

Linear System Case

- Question: What if the true joint state-measurement distribution at time k, i.e., $\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}$, deviates from the nominal $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$?
- Philosophy: We aim to find a robust state estimation solution that is insensitive to the model deviation.
- Motivation: We study the distributionally robust counterpart:

$$\min_{\phi\in\mathcal{H}'_{\mathcal{Y}_k}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta)} \operatorname{Tr}\mathbb{E}[\mathbf{x}_k-\phi(\mathcal{Y}_k)][\mathbf{x}_k-\phi(\mathcal{Y}_k)]^\top,$$

- where

$$\mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta) := \left\{\mathbb{P}\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^{m\times k})\,\middle|\,D(\mathbb{P},\ \bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k})\le\theta\right\}$$

is **ambiguity set** constructed around the nominal distribution $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$.

- Question: What if the true joint state-measurement distribution at time k, i.e., $\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}$, deviates from the nominal $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$?
- Philosophy: We aim to find a robust state estimation solution that is insensitive to the model deviation.
- Motivation: We study the distributionally robust counterpart:

$$\min_{\phi \in \mathcal{H}'_{\mathcal{Y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta)} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \phi(\mathcal{Y}_k)][\mathbf{x}_k - \phi(\mathcal{Y}_k)]^\top,$$

- where

$$\mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta) := \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^{m \times k}) \,\middle|\, D(\mathbb{P},\ \ \bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}) \leq \theta \right\}$$

is **ambiguity set** constructed around the nominal distribution $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$.
  - $D(\cdot,\cdot)$ is a possible statistical similarity measure, e.g., Wasserstein distance, Kullback–Leibler divergence.

- Question: What if the true joint state-measurement distribution at time k, i.e., $\mathbb{P}_{\mathbf{x}_k,\mathcal{Y}_k}$, deviates from the nominal $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$?
- Philosophy: We aim to find a robust state estimation solution that is insensitive to the model deviation.
- Motivation: We study the distributionally robust counterpart:

$$\min_{\phi\in\mathcal{H}'_{\mathcal{Y}_k}} \max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta)} \operatorname{Tr}\mathbb{E}[\mathbf{x}_k-\phi(\mathcal{Y}_k)][\mathbf{x}_k-\phi(\mathcal{Y}_k)]^\top,$$

- where

$$\mathcal{F}_{\mathbf{x}_k,\mathcal{Y}_k}(\theta) := \left\{\mathbb{P}\in\mathcal{P}(\mathbb{R}^n\times\mathbb{R}^{m\times k})\,\middle|\,D(\mathbb{P},\ \ \bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k})\leq\theta\right\}$$

is **ambiguity set** constructed around the nominal distribution $\bar{\mathbb{P}}_{\mathbf{x}_k,\mathcal{Y}_k}$.
  - $D(\cdot,\cdot)$ is a possible statistical similarity measure, e.g., Wasserstein distance, Kullback–Leibler divergence.
  - We do not know the true distribution, but we assume that it lies in a ball centered at the nominal distribution.

Recap: $$\min_{\phi \in \mathcal{H}'_{\mathcal{Y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathcal{Y}_k}(\theta)} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \phi(\mathcal{Y}_k)][\mathbf{x}_k - \phi(\mathcal{Y}_k)]^\top,$$

- Issue: What if measurements $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k$ sequentially arrives along k?

Recap: $$\min_{\phi \in \mathcal{H}'_{\mathcal{Y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathcal{Y}_k}(\theta)} \operatorname{Tr} \mathbb{E}[\mathbf{x}_k - \phi(\mathcal{Y}_k)][\mathbf{x}_k - \phi(\mathcal{Y}_k)]^\top,$$

- Issue: What if measurements $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_k$ sequentially arrives along k?
- We study time-incremental (a.k.a. time-series, online) version:

$$\min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

where the ambiguity set is defined as

$$\mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta) := \left\{ \mathbb{P} \in \mathcal{P}(\mathbb{R}^n \times \mathbb{R}^m) \middle| D(\mathbb{P}, \ \ \bar{\mathbb{P}}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}) \leq \theta \right\}$$

Recap: $$\min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

- It can be proved that the min-max problem is equivalent to the max-min problem below

$$\max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

Recap: $$\min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

- It can be proved that the min-max problem is equivalent to the max-min problem below

$$\max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

  - Intuition: The objective is positive-definite quadratic (thus convex) in ϕ, and linear (thus concave) in $\mathbb{P}$.

Recap: $$\min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

- It can be proved that the min-max problem is equivalent to the max-min problem below

$$\max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

  - Intuition: The objective is positive-definite quadratic (thus convex) in ϕ, and linear (thus concave) in $\mathbb{P}$.
  - Fact: The saddle point exists.

Recap: $$\min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

- It can be proved that the min-max problem is equivalent to the max-min problem below

$$\max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

  - Intuition: The objective is positive-definite quadratic (thus convex) in ϕ, and linear (thus concave) in $\mathbb{P}$.
  - Fact: The saddle point exists.

- **Note**. The latter is easier to solve because for every $\mathbb{P}$, we can find the associated optimal estimator, i.e., associated conditional mean.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr}\mathbb{E}\left\{[\mathbf{x}_k-\phi(\mathbf{y}_k)][\mathbf{x}_k-\phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1}\right\},$$

- Therefore, at each k, we are inspired to **first** study a one-stage distributionally robust Bayesian estimation problem

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\phi(\mathbf{y})][\mathbf{x}-\phi(\mathbf{y})]^\top$$

subject to

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr}\mathbb{E}\left\{[\mathbf{x}_k-\phi(\mathbf{y}_k)][\mathbf{x}_k-\phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1}\right\},$$

- Therefore, at each k, we are inspired to **first** study a one-stage distributionally robust Bayesian estimation problem

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\phi(\mathbf{y})][\mathbf{x}-\phi(\mathbf{y})]^\top$$

subject to

- the nominal prior state distribution $\bar{\mathbb{P}}_{\mathbf{x}}$,

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr}\mathbb{E}\left\{[\mathbf{x}_k-\phi(\mathbf{y}_k)][\mathbf{x}_k-\phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1}\right\},$$

- Therefore, at each k, we are inspired to **first** study a one-stage distributionally robust Bayesian estimation problem

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\phi(\mathbf{y})][\mathbf{x}-\phi(\mathbf{y})]^\top$$

subject to

- the nominal prior state distribution $\bar{\mathbb{P}}_{\mathbf{x}}$,
- the nominal conditional measurement distribution $\bar{\mathbb{P}}_{\mathbf{y}|\mathbf{x}}$,

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr}\mathbb{E}\left\{[\mathbf{x}_k-\phi(\mathbf{y}_k)][\mathbf{x}_k-\phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1}\right\},$$

- Therefore, at each k, we are inspired to **first** study a one-stage distributionally robust Bayesian estimation problem

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\phi(\mathbf{y})][\mathbf{x}-\phi(\mathbf{y})]^\top$$

subject to

- the nominal prior state distribution $\bar{\mathbb{P}}_{\mathbf{x}}$,
- the nominal conditional measurement distribution $\bar{\mathbb{P}}_{\mathbf{y}|\mathbf{x}}$,
- a properly constructed ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ that contains fat-tailed (marginal) distributions for $\mathbf{y}$,

Recap: $$\max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x}_k, \mathbf{y}_k | \mathcal{Y}_{k-1}}(\theta)} \min_{\phi \in \mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr} \mathbb{E} \left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1} \right\},$$

- Therefore, at each k, we are inspired to **first** study a one-stage distributionally robust Bayesian estimation problem

$$\max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\phi \in \mathcal{H}'_{\mathbf{y}}} \operatorname{Tr} \mathbb{E}[\mathbf{x} - \phi(\mathbf{y})][\mathbf{x} - \phi(\mathbf{y})]^\top$$

subject to

- the nominal prior state distribution $\bar{\mathbb{P}}_{\mathbf{x}}$,
- the nominal conditional measurement distribution $\bar{\mathbb{P}}_{\mathbf{y}|\mathbf{x}}$,
- a properly constructed ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ that contains fat-tailed (marginal) distributions for $\mathbf{y}$,
- the linear measurement equation $\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{v}$.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}_k}} \operatorname{Tr}\mathbb{E}\left\{ [\mathbf{x}_k - \phi(\mathbf{y}_k)][\mathbf{x}_k - \phi(\mathbf{y}_k)]^\top \middle| \mathcal{Y}_{k-1}\right\},$$

- Therefore, at each k, we are inspired to **first** study a one-stage distributionally robust Bayesian estimation problem

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\phi\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x} - \phi(\mathbf{y})][\mathbf{x} - \phi(\mathbf{y})]^\top$$

subject to

- the nominal prior state distribution $\bar{\mathbb{P}}_{\mathbf{x}}$,
- the nominal conditional measurement distribution $\bar{\mathbb{P}}_{\mathbf{y}|\mathbf{x}}$,
- a properly constructed ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ that contains fat-tailed (marginal) distributions for $\mathbf{y}$,
- the linear measurement equation $\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{v}$.

- **Then**, by identifying $\mathbb{P}_{\mathbf{x}_k,\mathbf{y}_k|\mathcal{Y}_{k-1}}$, we can solve the distributionally robust state estimation problem.

Theorem

Consider the joint distribution $\mathbb{P}_{\mathbf{x},\mathbf{y}}$ *and* $\mathbf{y} = \boldsymbol{H}\mathbf{x} + \mathbf{v}$. *Suppose* $\mathbf{x} \sim \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$, *the mean of* $\mathbf{v}$ *is* $\boldsymbol{c}_{\boldsymbol{v}}$, *the covariance of* $\mathbf{v}$ *is* $\boldsymbol{\Sigma}_{\boldsymbol{v}}$, $\mathbf{x}$ *is independent of* $\mathbf{v}$, *all involved densities exist and twice continuously differentiable.*
Let $\mathbf{e} := \mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} - \boldsymbol{c}_{\boldsymbol{v}}$ *denote the innovation vector,* $\boldsymbol{S}$ *the covariance of* $\mathbf{e}$, *and* $\mathbf{u} := \boldsymbol{S}^{-1/2}\mathbf{e}$ *the diagonalized and normalized innovation.*
Then the optimal estimator $\hat{\mathbf{x}}$ *of* $\mathbf{x}$ *given* $\mathbf{y}$, *i.e.,* $\mathbb{E}(\mathbf{x}|\mathbf{y})$, *is*

$$\hat{\mathbf{x}} = \boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\left[-\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\mu}}\ln p_{\mathbf{u}}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathbf{u}}.$$

and the estimation error covariance, i.e., $\boldsymbol{P} := \mathbb{E}(\hat{\mathbf{x}} - \mathbf{x})(\hat{\mathbf{x}} - \mathbf{x})^\top$, *is*

$$\boldsymbol{P} = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top\boldsymbol{S}^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p_{\mathrm{u}}(\mu)\Big|_{\mu=\mathrm{u}}\right],$$

where $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}$, $p_{\mathbf{u}}(\boldsymbol{\mu}) = p_{\mathbf{y}}(\boldsymbol{S}^{1/2}\boldsymbol{\mu} + \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{c}_{\boldsymbol{v}}) \cdot \det(\boldsymbol{S}^{1/2})$ *is the density of* $\mathbf{u}$, $p_{\mathbf{y}}(\cdot)$ *is the density of* $\mathbf{y}$, *and* $\det(\cdot)$ *denotes the determinant of a matrix.*

In the Theorem above, we do not specify the type of the distribution of $\mathbf{v}$. It can be fat-tailed.

Example:

Suppose $\mathbf{v}$ is Gaussian: $\mathbf{v} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma_v})$.

- The innovation $\mathbf{e} := \mathbf{y} - \boldsymbol{H}\boldsymbol{c_x} = \boldsymbol{H}(\mathbf{x} - \boldsymbol{c_x}) + \mathbf{v}$ is also Gaussian with mean of $\mathbf{0}$ and covariance $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{\Sigma_x}\boldsymbol{H}^\top + \boldsymbol{\Sigma_v}$.

In the Theorem above, we do not specify the type of the distribution of $\mathbf{v}$. It can be fat-tailed.

Example:

Suppose $\mathbf{v}$ is Gaussian: $\mathbf{v} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.

- The innovation $\mathbf{e} := \mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} = \boldsymbol{H}(\mathbf{x} - \boldsymbol{c}_{\boldsymbol{x}}) + \mathbf{v}$ is also Gaussian with mean of $\mathbf{0}$ and covariance $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}$.
- The normalized innovation $\mathbf{u} := \boldsymbol{S}^{-1/2}\mathbf{e}$ is Gaussian with mean of $\mathbf{0}$ and covariance of $\boldsymbol{I}$.

In the Theorem above, we do not specify the type of the distribution of $\mathbf{v}$. It can be fat-tailed.

Example:

Suppose $\mathbf{v}$ is Gaussian: $\mathbf{v} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.

- The innovation $\mathbf{e} := \mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} = \boldsymbol{H}(\mathbf{x} - \boldsymbol{c}_{\boldsymbol{x}}) + \mathbf{v}$ is also Gaussian with mean of $\mathbf{0}$ and covariance $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}$.
- The normalized innovation $\mathbf{u} := \boldsymbol{S}^{-1/2}\mathbf{e}$ is Gaussian with mean of $\mathbf{0}$ and covariance of $\boldsymbol{I}$.
- Namely, the density of $\mathbf{u}$ is $p_{\mathbf{u}}(\boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^m}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\mu}\right)$. As a result, we have

$$-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}} = \frac{1}{2}\frac{\mathrm{d}\boldsymbol{\mu}^\top\boldsymbol{\mu}}{\mathrm{d}\boldsymbol{\mu}} = \boldsymbol{\mu},$$

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

In the Theorem above, we do not specify the type of the distribution of $\mathbf{v}$. It can be fat-tailed.

Example:

Suppose $\mathbf{v}$ is Gaussian: $\mathbf{v} \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.

- The innovation $\mathbf{e} := \mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} = \boldsymbol{H}(\mathbf{x} - \boldsymbol{c}_{\boldsymbol{x}}) + \mathbf{v}$ is also Gaussian with mean of $\mathbf{0}$ and covariance $\boldsymbol{S} = \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}$.
- The normalized innovation $\mathbf{u} := \boldsymbol{S}^{-1/2}\mathbf{e}$ is Gaussian with mean of $\mathbf{0}$ and covariance of $\boldsymbol{I}$.
- Namely, the density of $\mathbf{u}$ is $p_{\mathbf{u}}(\boldsymbol{\mu}) = \frac{1}{\sqrt{(2\pi)^m}} \exp\left(-\frac{1}{2}\boldsymbol{\mu}^\top\boldsymbol{\mu}\right)$. As a result, we have

$$-\frac{\mathrm{d}\ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}} = \frac{1}{2}\frac{\mathrm{d}\boldsymbol{\mu}^\top\boldsymbol{\mu}}{\mathrm{d}\boldsymbol{\mu}} = \boldsymbol{\mu},$$

- The optimal **estimator** is given as

$$\begin{aligned}
\hat{\mathbf{x}} &= \boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\left[-\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\mu}}\ln p_{\mathbf{u}}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathbf{u}} \\
&= \boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\boldsymbol{S}^{-1/2}\left[\mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}}\right] \\
&= \boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\left[\mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}}\right].
\end{aligned}$$

Example (cont'd):
Likewise,

$$-\frac{\mathrm{d}^2 \ln p_{\mathbf{u}}(\boldsymbol{\mu})}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top} = \boldsymbol{I},$$

and therefore,

$$\begin{aligned}
\boldsymbol{P} &= \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\mathbb{E}\left\{\left[-\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\mu}\mathrm{d}\boldsymbol{\mu}^\top}\ln p_{\mathbf{u}}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathbf{u}}\right\}\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top\boldsymbol{S}^{-1/2}\boldsymbol{I}\boldsymbol{S}^{-1/2}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top\boldsymbol{S}^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \\
&= \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}.
\end{aligned}$$

We end up with the standard Kalman formulas.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})][\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})]^\top$$

Corollary (Reducing to finding worst-case distribution)

The distributionally robust Bayesian estimation can be reformulated as

$$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr}\boldsymbol{P},$$

where

$$\boldsymbol{P} = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top+\boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\cdot\mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p_{\mathrm{u}}(\mu)\Big|_{\mu=\mathrm{u}}\right].$$

Because for every possible joint distribution $\mathbb{P}_{\mathbf{x},\mathbf{y}}$,

$$\hat{\mathrm{x}} = \boldsymbol{\phi}(\mathbf{y}) = \boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top+\boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1/2}\left[-\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\mu}}\ln p_{\mathbf{u}}(\boldsymbol{\mu})\right]_{\boldsymbol{\mu}=\mathbf{u}}$$

is uniquely determined.

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\left.\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\right|_{\mu=\mathrm{u}} \right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Recap: $$\max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E} \left[- \frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2} \bigg|_{\mu = \mathrm{u}} \right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top (\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}} \right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.
- 1. Nominal $\bar{\mathbb{P}}_{\mathbf{x}} := \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ and candidate $\mathbb{P}_{\mathbf{x}} := \mathcal{D}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[- \frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}} \right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.
-
  1. Nominal $\bar{\mathbb{P}}_{\mathbf{x}} := \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ and candidate $\mathbb{P}_{\mathbf{x}} := \mathcal{D}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$.
  2. Nominal $\bar{\mathbb{P}}_{\mathbf{v}} := \mathcal{N}_m(\mathbf{0}, \boldsymbol{R})$ and candidate $\mathbb{P}_{\mathbf{v}} := \mathcal{D}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top (\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}} \right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.
- 1. Nominal $\bar{\mathbb{P}}_{\mathbf{x}} := \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ and candidate $\mathbb{P}_{\mathbf{x}} := \mathcal{D}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$.
  2. Nominal $\bar{\mathbb{P}}_{\mathbf{v}} := \mathcal{N}_m(\mathbf{0}, \boldsymbol{R})$ and candidate $\mathbb{P}_{\mathbf{v}} := \mathcal{D}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.
  3. Nominal $\bar{\mathbb{P}}_{\mathrm{u}}$ is a Standard Gaussian and candidate $\mathbb{P}_{\mathrm{u}}$.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}} \right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.
- 1. Nominal $\bar{\mathbb{P}}_{\mathbf{x}} := \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ and candidate $\mathbb{P}_{\mathbf{x}} := \mathcal{D}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$.
  2. Nominal $\bar{\mathbb{P}}_{\mathbf{v}} := \mathcal{N}_m(\mathbf{0}, \boldsymbol{R})$ and candidate $\mathbb{P}_{\mathbf{v}} := \mathcal{D}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.
  3. Nominal $\bar{\mathbb{P}}_{\mathrm{u}}$ is a Standard Gaussian and candidate $\mathbb{P}_{\mathrm{u}}$.
- Recall $\mathbf{u} = \boldsymbol{S}^{-1/2}(\mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} - \boldsymbol{c}_{\boldsymbol{v}})$ and $\mathbf{y} = (\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{1/2}\mathbf{u} + \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{c}_{\boldsymbol{v}}$

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}} \right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.
-
  1. Nominal $\bar{\mathbb{P}}_{\mathbf{x}} := \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ and candidate $\mathbb{P}_{\mathbf{x}} := \mathcal{D}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$.
  2. Nominal $\bar{\mathbb{P}}_{\mathbf{v}} := \mathcal{N}_m(\mathbf{0}, \boldsymbol{R})$ and candidate $\mathbb{P}_{\mathbf{v}} := \mathcal{D}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.
  3. Nominal $\bar{\mathbb{P}}_{\mathrm{u}}$ is a Standard Gaussian and candidate $\mathbb{P}_{\mathrm{u}}$.
- Recall $\mathbf{u} = \boldsymbol{S}^{-1/2}(\mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} - \boldsymbol{c}_{\boldsymbol{v}})$ and $\mathbf{y} = (\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{1/2}\mathbf{u} + \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{c}_{\boldsymbol{v}}$
- $\mathbb{P}_{\mathrm{u}}$ must be fat-tailed. Because the type of the distribution of $\mathbf{y}$ only depends on the type of $\mathbf{u}$.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}}\right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.
- 1. Nominal $\bar{\mathbb{P}}_{\mathbf{x}} := \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ and candidate $\mathbb{P}_{\mathbf{x}} := \mathcal{D}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$.
  2. Nominal $\bar{\mathbb{P}}_{\mathbf{v}} := \mathcal{N}_m(\mathbf{0}, \boldsymbol{R})$ and candidate $\mathbb{P}_{\mathbf{v}} := \mathcal{D}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.
  3. Nominal $\bar{\mathbb{P}}_{\mathrm{u}}$ is a Standard Gaussian and candidate $\mathbb{P}_{\mathrm{u}}$.
- Recall $\mathbf{u} = \boldsymbol{S}^{-1/2}(\mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} - \boldsymbol{c}_{\boldsymbol{v}})$ and $\mathbf{y} = (\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{1/2}\mathbf{u} + \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{c}_{\boldsymbol{v}}$
- $\mathbb{P}_{\mathrm{u}}$ must be fat-tailed. Because the type of the distribution of $\mathbf{y}$ only depends on the type of $\mathbf{u}$.
- Assuming candidate $\mathbb{P}_{\mathbf{x}}$ and candidate $\mathbb{P}_{\mathbf{v}}$ to be Gaussian is sufficient to describe the uncertainties in parameters $\{\bar{\boldsymbol{x}}, \mathbf{0}, \boldsymbol{M}, \boldsymbol{R}\}$.

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}}\right]$$

- We need to define the ambiguity set $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$, which contains fat-tailed distributions for $\mathbf{y}$. We call every $\mathbb{P}$ in $\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)$ a **candidate**.
-
  1. Nominal $\bar{\mathbb{P}}_{\mathbf{x}} := \mathcal{N}_n(\bar{\boldsymbol{x}}, \boldsymbol{M})$ and candidate $\mathbb{P}_{\mathbf{x}} := \mathcal{D}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}})$.
  2. Nominal $\bar{\mathbb{P}}_{\mathbf{v}} := \mathcal{N}_m(\mathbf{0}, \boldsymbol{R})$ and candidate $\mathbb{P}_{\mathbf{v}} := \mathcal{D}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}})$.
  3. Nominal $\bar{\mathbb{P}}_{\mathrm{u}}$ is a Standard Gaussian and candidate $\mathbb{P}_{\mathrm{u}}$.
- Recall $\mathbf{u} = \boldsymbol{S}^{-1/2}(\mathbf{y} - \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} - \boldsymbol{c}_{\boldsymbol{v}})$ and $\mathbf{y} = (\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{1/2}\mathbf{u} + \boldsymbol{H}\boldsymbol{c}_{\boldsymbol{x}} + \boldsymbol{c}_{\boldsymbol{v}}$
- $\mathbb{P}_{\mathrm{u}}$ must be fat-tailed. Because the type of the distribution of $\mathbf{y}$ only depends on the type of $\mathbf{u}$.
- Assuming candidate $\mathbb{P}_{\mathbf{x}}$ and candidate $\mathbb{P}_{\mathbf{v}}$ to be Gaussian is sufficient to describe the uncertainties in parameters $\{\bar{\boldsymbol{x}}, \mathbf{0}, \boldsymbol{M}, \boldsymbol{R}\}$.
  1. The uncertain counterpart of $\{\bar{\boldsymbol{x}}, \mathbf{0}, \boldsymbol{M}, \boldsymbol{R}\}$ is $\{\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}\}$.

Ambiguity sets for $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{v}}$: mainly accounts for parameter uncertainties

- **1. Kullback–Leibler divergence** (KL divergence).

$$\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}}) = \left\{ \mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \middle| \ \mathrm{KL}(\mathbb{P}_{\mathbf{x}} \| \bar{\mathbb{P}}_{\mathbf{x}}) \leq \theta_{\boldsymbol{x}} \ \right\},$$

$$\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}}) = \left\{ \mathbb{P}_{\mathbf{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \middle| \ \mathrm{KL}(\mathbb{P}_{\mathbf{v}} \| \bar{\mathbb{P}}_{\mathbf{v}}) \leq \theta_{\boldsymbol{v}} \ \right\}.$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the KL divergence and under Gaussianity assumption, $\mathrm{KL}(\mathbb{P}_{\mathbf{x}} \| \bar{\mathbb{P}}_{\mathbf{x}}) = \frac{1}{2}[\|\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2_{\boldsymbol{M}^{-1}} + \mathrm{Tr}\left[\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{I}\right] - \ln\det\left(\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)]$.

Ambiguity sets for $\mathbb{P}_{\mathbf{x}}$ and $\mathbb{P}_{\mathbf{v}}$: mainly accounts for parameter uncertainties

- **1. Kullback–Leibler divergence** (KL divergence).

$$\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}}) = \left\{\mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \,\middle|\, \mathrm{KL}(\mathbb{P}_{\mathbf{x}} \| \bar{\mathbb{P}}_{\mathbf{x}}) \le \theta_{\boldsymbol{x}} \right\},$$

$$\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}}) = \left\{\mathbb{P}_{\mathbf{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \,\middle|\, \mathrm{KL}(\mathbb{P}_{\mathbf{v}} \| \bar{\mathbb{P}}_{\mathbf{v}}) \le \theta_{\boldsymbol{v}} \right\}.$$

where $\mathrm{KL}(\cdot\|\cdot)$ denotes the KL divergence and under Gaussianity assumption, $\mathrm{KL}(\mathbb{P}_{\mathbf{x}} \| \bar{\mathbb{P}}_{\mathbf{x}}) = \frac{1}{2}[\|\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2_{\boldsymbol{M}^{-1}} + \mathrm{Tr}\left[\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{I}\right] - \ln\det\left(\boldsymbol{M}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\right)]$.

- **2. Wasserstein distance**.

$$\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}}) = \left\{\mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \,\middle|\, \mathrm{W}(\mathbb{P}_{\mathbf{x}}, \bar{\mathbb{P}}_{\mathbf{x}}) \le \theta_{\boldsymbol{x}} \right\},$$

$$\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}}) = \left\{\mathbb{P}_{\mathbf{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \,\middle|\, \mathrm{W}(\mathbb{P}_{\mathbf{v}}, \bar{\mathbb{P}}_{\mathbf{v}}) \le \theta_{\boldsymbol{v}} \right\}.$$

where $\mathrm{W}(\cdot,\cdot)$ denotes the Wasserstein metric and under Gaussianity assumption, the type-2 Wasserstein distance is given as
$\mathrm{W}(\mathbb{P}_{\mathbf{x}}, \bar{\mathbb{P}}_{\mathbf{x}}) = \sqrt{\|\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}\|^2 + \mathrm{Tr}[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}})^{\frac{1}{2}}]}$.

3. Moment-based set.

$$\begin{aligned}
\mathcal{F}_{\mathbf{x}}(\theta_{1,\boldsymbol{x}}, \theta_{2,\boldsymbol{x}}, \theta_{3,\boldsymbol{x}}) &= \left\{ \mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \,\middle|\, \begin{array}{l} [\mathbb{E}\mathbf{x} - \bar{\boldsymbol{x}}]^\top \boldsymbol{M}^{-1} [\mathbb{E}\mathbf{x} - \bar{\boldsymbol{x}}] \le \theta_{3,\boldsymbol{x}} \\ \mathbb{E}(\mathbf{x} - \bar{\boldsymbol{x}})(\mathbf{x} - \bar{\boldsymbol{x}})^\top \preceq \theta_{2,\boldsymbol{x}} \boldsymbol{M} \\ \mathbb{E}(\mathbf{x} - \bar{\boldsymbol{x}})(\mathbf{x} - \bar{\boldsymbol{x}})^\top \succeq \theta_{1,\boldsymbol{x}} \boldsymbol{M} \end{array} \right\} \\
&= \left\{ \mathbb{P}_{\mathbf{x}} = \mathcal{N}_n(\boldsymbol{c}_{\boldsymbol{x}}, \boldsymbol{\Sigma}_{\boldsymbol{x}}) \,\middle|\, \begin{array}{l} [\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}]^\top \boldsymbol{M}^{-1} [\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}}] \le \theta_{3,\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} + (\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})(\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})^\top \preceq \theta_{2,\boldsymbol{x}} \boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} + (\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})(\boldsymbol{c}_{\boldsymbol{x}} - \bar{\boldsymbol{x}})^\top \succeq \theta_{1,\boldsymbol{x}} \boldsymbol{M} \end{array} \right\}.
\end{aligned}$$

$$\mathcal{F}_{\mathbf{v}}(\theta_{1,\boldsymbol{v}}, \theta_{2,\boldsymbol{v}}, \theta_{3,\boldsymbol{v}}) = \left\{ \mathbb{P}_{\mathbf{v}} = \mathcal{N}_m(\boldsymbol{c}_{\boldsymbol{v}}, \boldsymbol{\Sigma}_{\boldsymbol{v}}) \,\middle|\, \begin{array}{l} [\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0}]^\top \boldsymbol{R}^{-1} [\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0}] \le \theta_{3,\boldsymbol{v}} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} + (\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0})(\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0})^\top \preceq \theta_{2,\boldsymbol{v}} \boldsymbol{R} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} + (\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0})(\boldsymbol{c}_{\boldsymbol{v}} - \mathbf{0})^\top \succeq \theta_{1,\boldsymbol{v}} \boldsymbol{R} \end{array} \right\}.$$

Ambiguity sets for $\mathbb{P}_{\mathrm{u}}$: mainly accounts for measurement outliers.

- **1. ϵ-contamination set**.

$$\mathcal{F}_{\mathrm{u}}(\epsilon) = \left\{ \mathbb{P}_{\mathrm{u}} \in \mathcal{P}(\mathbb{R}) \;\middle|\; \begin{array}{l} F_{\mathrm{u}}(\mu) = \mathbb{P}_{\mathrm{u}}(\mathrm{u} \le \mu) \\ F_{\mathrm{u}}(\mu) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu) \\ H(\mu) = 1 - H(-\mu), \quad H(\mu) \text{ is a cumulative on } \mathbb{R} \end{array} \right\}.$$

Intuitively, it means with probability ϵ, say 5%, a transformed measurement u is from the contamination distribution $H(\cdot)$, and with probability $1-\epsilon$, u is from the Gaussian distribution $\Phi(\cdot)$.

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Ambiguity sets for $\mathbb{P}_{\mathrm{u}}$: mainly accounts for measurement outliers.

- **1. ϵ-contamination set**.

$$\mathcal{F}_{\mathrm{u}}(\epsilon) = \left\{ \mathbb{P}_{\mathrm{u}} \in \mathcal{P}(\mathbb{R}) \;\middle|\; \begin{array}{l} F_{\mathrm{u}}(\mu) = \mathbb{P}_{\mathrm{u}}(\mathrm{u} \le \mu) \\ F_{\mathrm{u}}(\mu) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu) \\ H(\mu) = 1 - H(-\mu), \quad H(\mu) \text{ is a cumulative on } \mathbb{R} \end{array} \right\}.$$

  Intuitively, it means with probability ϵ, say 5%, a transformed measurement u is from the contamination distribution $H(\cdot)$, and with probability $1-\epsilon$, u is from the Gaussian distribution $\Phi(\cdot)$.

- **2. ϵ-normal set**.

$$\mathcal{F}_{\mathrm{u}}(\epsilon) = \left\{ \mathbb{P}_{\mathrm{u}} \in \mathcal{P}(\mathbb{R}) \;\middle|\; \begin{array}{l} F_{\mathrm{u}}(\mu) = \mathbb{P}_{\mathrm{u}}(\mathrm{u} \le \mu) \\ \sup\limits_{\mu \in \mathbb{R}} \|F_{\mathrm{u}}(\mu) - \Phi(\mu)\| \le \epsilon \\ F_{\mathrm{u}}(\mu) = 1 - F_{\mathrm{u}}(-\mu) \end{array} \right\}.$$

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Ambiguity sets for $\mathbb{P}_{\mathrm{u}}$: mainly accounts for measurement outliers.

- **1. ϵ-contamination set.**

$$\mathcal{F}_{\mathrm{u}}(\epsilon) = \left\{ \mathbb{P}_{\mathrm{u}} \in \mathcal{P}(\mathbb{R}) \;\middle|\; \begin{array}{l} F_{\mathrm{u}}(\mu) = \mathbb{P}_{\mathrm{u}}(\mathrm{u} \le \mu) \\ F_{\mathrm{u}}(\mu) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu) \\ H(\mu) = 1 - H(-\mu), \quad H(\mu) \text{ is a cumulative on } \mathbb{R} \end{array} \right\}.$$

  Intuitively, it means with probability ϵ, say 5%, a transformed measurement u is from the contamination distribution $H(\cdot)$, and with probability $1-\epsilon$, u is from the Gaussian distribution $\Phi(\cdot)$.

- **2. ϵ-normal set.**

$$\mathcal{F}_{\mathrm{u}}(\epsilon) = \left\{ \mathbb{P}_{\mathrm{u}} \in \mathcal{P}(\mathbb{R}) \;\middle|\; \begin{array}{l} F_{\mathrm{u}}(\mu) = \mathbb{P}_{\mathrm{u}}(\mathrm{u} \le \mu) \\ \sup_{\mu \in \mathbb{R}} \|F_{\mathrm{u}}(\mu) - \Phi(\mu)\| \le \epsilon \\ F_{\mathrm{u}}(\mu) = 1 - F_{\mathrm{u}}(-\mu) \end{array} \right\}.$$

- **Note.** Given ϵ, a ϵ-contamination set is a subset of ϵ-normal set. Because

$$F_{\mathrm{u}}(\mu) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu) \quad \Longrightarrow \quad \sup_{\mu \in \mathbb{R}} \|F_{\mathrm{u}}(\mu) - \Phi(\mu)\| \le \epsilon$$

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\left.\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\right|_{\mu=\mathrm{u}}\right]$$

- Hence, we solve it independently and sequentially, i.e., solving the innermost first and the outermost last.

$$\max_{\mathbb{P}_{\mathbf{x}}\in\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})} \max_{\mathbb{P}_{\mathbf{v}}\in\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})} \max_{\mathbb{P}_{\mathrm{u}}\in\mathcal{F}_{\mathrm{u}}(\epsilon)} \operatorname{Tr} \boldsymbol{P}.$$

i.e., (due to parameterizations of distributions)

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \max_{i_\mu} \operatorname{Tr} \boldsymbol{P}$$

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}}\right]$$

- Hence, we solve it independently and sequentially, i.e., solving the innermost first and the outermost last.

$$\max_{\mathbb{P}_{\mathbf{x}}\in\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})} \max_{\mathbb{P}_{\mathbf{v}}\in\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})} \max_{\mathbb{P}_{\mathrm{u}}\in\mathcal{F}_{\mathrm{u}}(\epsilon)} \operatorname{Tr} \boldsymbol{P}.$$

i.e., (due to parameterizations of distributions)

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \max_{i_\mu} \operatorname{Tr} \boldsymbol{P}$$

- Define i_μ: Fisher information of $p(\mu)$, leading to

$$\boldsymbol{P} = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top}(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu.$$

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[-\frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2}\bigg|_{\mu=\mathrm{u}}\right]$$

- Hence, we solve it independently and sequentially, i.e., solving the innermost first and the outermost last.

$$\max_{\mathbb{P}_{\mathbf{x}}\in\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})} \max_{\mathbb{P}_{\mathbf{v}}\in\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})} \max_{\mathbb{P}_{\mathrm{u}}\in\mathcal{F}_{\mathrm{u}}(\epsilon)} \operatorname{Tr} \boldsymbol{P}.$$

i.e., (due to parameterizations of distributions)

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \max_{i_\mu} \operatorname{Tr} \boldsymbol{P}$$

- Define i_μ: Fisher information of $p(\mu)$, leading to

$$\boldsymbol{P} = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu.$$

- First, we solve the innermost sub-problem over i_μ. Note that $\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \mathbf{0}$ because $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathbb{S}^n_+$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} \in \mathbb{S}^m_{++}$.

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Recap: $$\max_{\mathbb{P} \in \mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \operatorname{Tr} \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot \mathbb{E}\left[- \frac{\mathrm{d}^2 \ln p_{\mathrm{u}}(\mu)}{\mathrm{d}\mu^2} \bigg|_{\mu = \mathrm{u}} \right]$$

- Hence, we solve it independently and sequentially, i.e., solving the innermost first and the outermost last.

$$\max_{\mathbb{P}_{\mathbf{x}} \in \mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})} \max_{\mathbb{P}_{\mathbf{v}} \in \mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})} \max_{\mathbb{P}_{\mathrm{u}} \in \mathcal{F}_{\mathrm{u}}(\epsilon)} \operatorname{Tr} \boldsymbol{P}.$$

i.e., (due to parameterizations of distributions)

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \max_{i_\mu} \operatorname{Tr} \boldsymbol{P}$$

- Define i_μ: Fisher information of $p(\mu)$, leading to

$$\boldsymbol{P} = \boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu.$$

- First, we solve the innermost sub-problem over i_μ. Note that $\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \mathbf{0}$ because $\boldsymbol{\Sigma}_{\boldsymbol{x}} \in \mathbb{S}^n_+$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}} \in \mathbb{S}^m_{++}$.
- The non-negative and minimal i_μ maximizes $\operatorname{Tr} \boldsymbol{P}$.

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Lemma

The functional optimization over the ϵ-contamination ambiguity set

$$\min_{p(\mu)} \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2} \ln p(\mu)\Big|_{\mu=\mathrm{u}} \right] \quad \textit{s.t.} \quad \begin{cases} p(\mu) = \dfrac{\mathrm{d}F_{\mathrm{u}}(\mu)}{\mathrm{d}\mu} \\ F_{\mathrm{u}}(\mu) = (1-\epsilon)\Phi(\mu) + \epsilon H(\mu) \\ H(\mu) = 1 - H(-\mu) \end{cases}$$

is solved by the Laplacian-tailed least-favorable distribution

$$p(\mu) = \begin{cases} (1-\epsilon)\frac{1}{\sqrt{2\pi}} e^{K\mu + \frac{1}{2}K^2}, & \mu \leq -K \\ (1-\epsilon)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}, & |\mu| \leq K \\ (1-\epsilon)\frac{1}{\sqrt{2\pi}} e^{-K\mu + \frac{1}{2}K^2}, & \mu \geq K, \end{cases}$$

where $K \in \mathbb{R}_+$ *is implicitly defined by* ϵ*:* $\int_{-K}^{K} p(\mu) dt + \frac{2p(K)}{K} = 1$.
Furthermore, $\min i_\mu = \min \mathbb{E}\left[-\frac{d^2}{d\mu^2} \ln p(\mu) \right] = (1-\epsilon)[1 - 2\Phi(-K)]$.

Lemma

Given $0 \leq \epsilon \lesssim 0.0303$, *the functional optimization over the* ϵ*-normal ambiguity set*

$$\min_{p(\mu)} \mathbb{E}\left[-\frac{\mathrm{d}^2}{\mathrm{d}\mu^2}\ln p(\mu)\Big|_{\mu=\mathrm{u}}\right] \qquad \textit{s.t.} \quad \begin{cases} p(\mu) = \dfrac{\mathrm{d}F_{\mathrm{u}}(\mu)}{\mathrm{d}\mu} \\ \sup\limits_{\mu\in\mathbb{R}} \|F_{\mathrm{u}}(\mu) - \Phi(\mu)\| \leq \epsilon \\ F_{\mathrm{u}}(\mu) = 1 - F_{\mathrm{u}}(-\mu), \end{cases}$$

is solved by the Laplacian-tailed least-favorable distribution

$$p(\mu) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}a^2} \cdot \cos^{-2}(\frac{1}{2}ca) \cdot \cos^2(\frac{1}{2}c\mu), & 0 \leq \mu \leq a \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mu^2}, & a \leq \mu \leq b \\ \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}b^2} \cdot e^{-b\mu + b^2}, & \mu \geq b \end{cases}$$

and $p(\mu) = p(-\mu)$, *where* a, b, *and* c *are implicitly defined by* ϵ *as: 1)* $c\tan(\frac{1}{2}ca) = a \quad (0 \leq ca < \pi)$, *2)* $\int_0^a p(\mu)d\mu = \int_0^a d\Phi(\mu) - \epsilon$, *and 3)* $\int_b^\infty p(\mu)d\mu = \int_b^\infty d\Phi(\mu) + \epsilon$.

Furthermore, $\min i_\mu = \min \mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right] = \frac{c^2 a}{\cos^2(\frac{1}{2}ca)} p(a) + 2\Phi(b) - 2\Phi(a)$.

Theorem

The distributionally robust Bayesian estimation is equivalent to

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \; \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}\right],$$

where $i_\mu^{\min} := \min i_\mu := \min \mathbb{E}\left[-\frac{d^2}{d\mu^2}\ln p(\mu)\right]$ *is a constant defined in two Lemmas above, whichever is adopted. Besides,* $0 \le i_\mu^{\min} \le 1$.

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

After solving the inner-most optimization over $p(\mu)$, we next we solve the outer sub-problems over $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}\right],$$

Under Wasserstein ambiguities of $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ and $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$, we have

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}\right],$$

subject to

$$\begin{cases} \sqrt{\operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} + \boldsymbol{M} - 2\left(\boldsymbol{M}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{M}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \leq \theta_{\boldsymbol{x}} \\ \sqrt{\operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{v}} + \boldsymbol{R} - 2\left(\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{v}}\boldsymbol{R}^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \leq \theta_{\boldsymbol{v}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \mathbf{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \mathbf{0}. \end{cases}$$

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Theorem (Under Wasserstein Ambiguity)

Suppose $\boldsymbol{R} \succ \mathbf{0}$. It can be reformulated as a linear SDP

$$\max_{\boldsymbol{\Sigma_x}, \boldsymbol{\Sigma_v}, \boldsymbol{V_x}, \boldsymbol{V_v}, \boldsymbol{U}} \operatorname{Tr}\left[\boldsymbol{\Sigma_x} - i_{\mu}^{\min} \cdot \boldsymbol{U}\right],$$

$$s.t. \quad \begin{cases} \begin{bmatrix} \boldsymbol{U} & \boldsymbol{\Sigma_x H}^\top \\ \boldsymbol{H\Sigma_x} & \boldsymbol{H\Sigma_x H}^\top + \boldsymbol{\Sigma_v} \end{bmatrix} \succeq \mathbf{0} \\ \operatorname{Tr}\left[\boldsymbol{\Sigma_x} + \boldsymbol{M} - 2\boldsymbol{V_x}\right] \le \theta_{\boldsymbol{x}}^2 \\ \begin{bmatrix} \boldsymbol{M}^{\frac{1}{2}} \boldsymbol{\Sigma_x} \boldsymbol{M}^{\frac{1}{2}} & \boldsymbol{V_x} \\ \boldsymbol{V_x} & \boldsymbol{I} \end{bmatrix} \succeq \mathbf{0} \\ \operatorname{Tr}\left[\boldsymbol{\Sigma_v} + \boldsymbol{R} - 2\boldsymbol{V_v}\right] \le \theta_{\boldsymbol{v}}^2 \\ \begin{bmatrix} \boldsymbol{R}^{\frac{1}{2}} \boldsymbol{\Sigma_v} \boldsymbol{R}^{\frac{1}{2}} & \boldsymbol{V_v} \\ \boldsymbol{V_v} & \boldsymbol{I} \end{bmatrix} \succeq \mathbf{0} \\ \boldsymbol{\Sigma_x} \succeq \mathbf{0}, \boldsymbol{\Sigma_v} \succ \mathbf{0}, \boldsymbol{V_x} \succeq \mathbf{0}, \boldsymbol{V_v} \succeq \mathbf{0}, \boldsymbol{U} \succeq \mathbf{0}. \end{cases}$$

Under moment-based ambiguities of $\mathcal{F}_{\mathbf{x}}(\theta_{\boldsymbol{x}})$ and $\mathcal{F}_{\mathbf{v}}(\theta_{\boldsymbol{v}})$, we have

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \mathrm{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}\right],$$

subject to

$$\begin{cases} \boldsymbol{\Sigma}_{\boldsymbol{x}} \preceq \theta_{2,\boldsymbol{x}}\boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \theta_{1,\boldsymbol{x}}\boldsymbol{M} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \preceq \theta_{2,\boldsymbol{v}}\boldsymbol{R} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succeq \theta_{1,\boldsymbol{v}}\boldsymbol{R} \succ \mathbf{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}} \succeq \mathbf{0} \\ \boldsymbol{\Sigma}_{\boldsymbol{v}} \succ \mathbf{0}. \end{cases}$$

Theorem (Under Moment-Based Ambiguity)

It is analytically solved by $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \theta_{2,\boldsymbol{x}}\boldsymbol{M}$ *and* $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \theta_{2,\boldsymbol{v}}\boldsymbol{R}$.

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Recap: $$\max_{\mathbb{P}\in\mathcal{F}_{\mathbf{x},\mathbf{y}}(\theta)} \min_{\boldsymbol{\phi}\in\mathcal{H}'_{\mathbf{y}}} \operatorname{Tr}\mathbb{E}[\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})][\mathbf{x}-\boldsymbol{\phi}(\mathbf{y})]^\top$$

Theorem (Solution to Distributionally Robust Bayesian Estimation)

Optimal Estimator.

$$\hat{\mathbf{x}} = \bar{\boldsymbol{x}} + \boldsymbol{\Sigma}_{\boldsymbol{x}}^{*}\boldsymbol{H}^\top \boldsymbol{S}^{*-1/2}\cdot\psi[\boldsymbol{S}^{*-1/2}(\mathbf{y}-\boldsymbol{H}\bar{\boldsymbol{x}})],$$

where $\boldsymbol{S}^{*} := \boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{*}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}^{*}$ where $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{*}$ and $\boldsymbol{\Sigma}_{\boldsymbol{v}}^{*}$ are optimal solutions of nonlinear SDPs associated with the Wasserstein metric or the moment-based set.

$\psi(\mu)$ *is entry-wise identical and for each entry*

$$\psi(\mu) = \begin{cases} -K, & \mu \le -K \\ \mu, & |\mu| \le K \\ K, & \mu \ge K, \end{cases}$$

if the ϵ-contamination ambiguity set is used, or

$$\psi(\mu) = -\psi(-\mu) = \begin{cases} c\tan(\frac{1}{2}c\mu), & 0 \le \mu \le a \\ \mu, & a \le \mu \le b \\ b, & \mu \ge b, \end{cases}$$

if ϵ-normal ambiguity set is used. Whenever a measurement $\mathbf{y} = \boldsymbol{y}$ is large, the value of $\psi(\cdot)$ is limited to $\pm K$ or $\pm b$.

Linear System Case

Solve Distributionally Robust Bayesian Estimation Problem

Recap:
$$\begin{cases} \mathbf{x}_k &= \boldsymbol{F}_{k-1}\mathbf{x}_{k-1} + \boldsymbol{G}_{k-1}\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

Theorem (Solution to Distributionally Robust State Estimation)

Optimal Recursive State Estimator.

$$\hat{\mathbf{x}}_{k|k} = \hat{\mathbf{x}}_{k|k-1} + \boldsymbol{\Sigma}^*_{\boldsymbol{x},k}\boldsymbol{H}_k^\top \boldsymbol{S}_k^{*-1/2} \cdot \psi[\boldsymbol{S}_k^{*-1/2}(\mathbf{y}_k - \boldsymbol{H}_k\hat{\mathbf{x}}_{k|k-1})],$$

where

$$\hat{\mathbf{x}}_{k|k-1} = \boldsymbol{F}_{k-1}\hat{\mathbf{x}}_{k-1|k-1},$$

and

$$\boldsymbol{S}_k^* := \boldsymbol{H}_k\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}\boldsymbol{H}_k^\top + \boldsymbol{\Sigma}^*_{\boldsymbol{v},k};$$

$\psi(\cdot)$, $\boldsymbol{\Sigma}^*_{\boldsymbol{x},k}$, *and* $\boldsymbol{\Sigma}^*_{\boldsymbol{v},k}$ *are defined in Theorem above.*

In the nominal case, the distributionally robust state estimator degenerates to the Kalman filter: e.g., $\psi(\mu) = \mu$ (i.e., there is no longer outlier treatment).

Content

Nonlinear System Case

Problem Background, Statements, and Motivations

We consider a state estimation problem for nonlinear systems.

$$\begin{cases} \mathbf{x}_k & = \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \\ \mathbf{y}_k & = \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k) \end{cases}$$

- $k = 1, 2, 3, ...$ denote discrete time index.

Nonlinear System Case

We consider a state estimation problem for nonlinear systems.

$$\begin{cases} \mathbf{x}_k & = \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \\ \mathbf{y}_k & = \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k) \end{cases}$$

- $k = 1, 2, 3, ...$ denote discrete time index.
- $\mathbf{x}_k \in \mathbb{R}^n$ is hidden state, $\mathbf{y}_k \in \mathbb{R}^m$ is measurement, $\mathbf{w}_{k-1} \in \mathbb{R}^p$ is process noise, $\mathbf{v}_k \in \mathbb{R}^q$ is measurement noise.

We consider a state estimation problem for nonlinear systems.

$$
\begin{cases}
\mathbf{x}_k &= \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \\
\mathbf{y}_k &= \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k)
\end{cases}
$$

- $k = 1, 2, 3, ...$ denote discrete time index.
- $\mathbf{x}_k \in \mathbb{R}^n$ is hidden state, $\mathbf{y}_k \in \mathbb{R}^m$ is measurement, $\mathbf{w}_{k-1} \in \mathbb{R}^p$ is process noise, $\mathbf{v}_k \in \mathbb{R}^q$ is measurement noise.
- $\boldsymbol{f}_k(\cdot, \cdot)$ is process dynamics function, and $\boldsymbol{h}_k(\cdot, \cdot)$ is measurement dynamics function.

We consider a state estimation problem for nonlinear systems.

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \\ \mathbf{y}_k &= \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k) \end{cases}$$

- $k = 1, 2, 3, \ldots$ denote discrete time index.
- $\mathbf{x}_k \in \mathbb{R}^n$ is hidden state, $\mathbf{y}_k \in \mathbb{R}^m$ is measurement, $\mathbf{w}_{k-1} \in \mathbb{R}^p$ is process noise, $\mathbf{v}_k \in \mathbb{R}^q$ is measurement noise.
- $\boldsymbol{f}_k(\cdot, \cdot)$ is process dynamics function, and $\boldsymbol{h}_k(\cdot, \cdot)$ is measurement dynamics function.
- **Assume**: $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{w}_k$, and $\mathbf{v}_k$ have finite second moments; $\mathbf{w}_k$ and $\mathbf{v}_k$ have known distributions.

We consider a state estimation problem for nonlinear systems.

$$\begin{cases} \mathbf{x}_k &= \boldsymbol{f}_k(\mathbf{x}_{k-1}, \mathbf{w}_{k-1}) \\ \mathbf{y}_k &= \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k) \end{cases}$$

- $k = 1, 2, 3, ...$ denote discrete time index.
- $\mathbf{x}_k \in \mathbb{R}^n$ is hidden state, $\mathbf{y}_k \in \mathbb{R}^m$ is measurement, $\mathbf{w}_{k-1} \in \mathbb{R}^p$ is process noise, $\mathbf{v}_k \in \mathbb{R}^q$ is measurement noise.
- $\boldsymbol{f}_k(\cdot, \cdot)$ is process dynamics function, and $\boldsymbol{h}_k(\cdot, \cdot)$ is measurement dynamics function.
- **Assume**: $\mathbf{x}_k$, $\mathbf{y}_k$, $\mathbf{w}_k$, and $\mathbf{v}_k$ have finite second moments; $\mathbf{w}_k$ and $\mathbf{v}_k$ have known distributions.
- **Task**: estimate/infer the hidden $\mathbf{x}_k$ based on measurement sequence $\mathcal{Y}_k := \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k\}$.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
   - Specifically, for given k, at least one of the following is inexact.

Nonlinear System Case

Problem Background, Statements, and Motivations

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
   - Specifically, for given k, at least one of the following is inexact.
     - Process dynamics $\boldsymbol{f}_k(\cdot)$.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
    - Specifically, for given k, at least one of the following is inexact.
        - Process dynamics $\boldsymbol{f}_k(\cdot)$.
        - Measurement dynamics $\boldsymbol{h}_k(\cdot)$.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
    - Specifically, for given k, at least one of the following is inexact.
        - Process dynamics $\boldsymbol{f}_k(\cdot)$.
        - Measurement dynamics $\boldsymbol{h}_k(\cdot)$.
        - Types and/or parameters of distributions of $\mathbf{w}_{k-1}$.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
   - Specifically, for given k, at least one of the following is inexact.
     - Process dynamics $\boldsymbol{f}_k(\cdot)$.
     - Measurement dynamics $\boldsymbol{h}_k(\cdot)$.
     - Types and/or parameters of distributions of $\mathbf{w}_{k-1}$.
     - Types and/or parameters of distributions of $\mathbf{v}_k$.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
   - Specifically, for given k, at least one of the following is inexact.
     - Process dynamics $\boldsymbol{f}_k(\cdot)$.
     - Measurement dynamics $\boldsymbol{h}_k(\cdot)$.
     - Types and/or parameters of distributions of $\mathbf{w}_{k-1}$.
     - Types and/or parameters of distributions of $\mathbf{v}_k$.
   - **Aim:** Hence, a robust state estimation scheme is expected.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
    - Specifically, for given k, at least one of the following is inexact.
        - Process dynamics $\boldsymbol{f}_k(\cdot)$.
        - Measurement dynamics $\boldsymbol{h}_k(\cdot)$.
        - Types and/or parameters of distributions of $\mathbf{w}_{k-1}$.
        - Types and/or parameters of distributions of $\mathbf{v}_k$.
    - **Aim:** Hence, a robust state estimation scheme is expected.
2. **Issue 2**: For a general nonlinear $\mathbf{y}_k = \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k)$, the likelihood of prior state $\boldsymbol{x}_k$ given measurement $\boldsymbol{y}_k$ is hard to be evaluated.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
   - Specifically, for given k, at least one of the following is inexact.
     - Process dynamics $\boldsymbol{f}_k(\cdot)$.
     - Measurement dynamics $\boldsymbol{h}_k(\cdot)$.
     - Types and/or parameters of distributions of $\mathbf{w}_{k-1}$.
     - Types and/or parameters of distributions of $\mathbf{v}_k$.
   - **Aim:** Hence, a robust state estimation scheme is expected.
2. **Issue 2**: For a general nonlinear $\mathbf{y}_k = \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k)$, the likelihood of prior state $\boldsymbol{x}_k$ given measurement $\boldsymbol{y}_k$ is hard to be evaluated.
   - Only easy for additive and multiplicative measurement noises.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
   - Specifically, for given k, at least one of the following is inexact.
     - Process dynamics $\boldsymbol{f}_k(\cdot)$.
     - Measurement dynamics $\boldsymbol{h}_k(\cdot)$.
     - Types and/or parameters of distributions of $\mathbf{w}_{k-1}$.
     - Types and/or parameters of distributions of $\mathbf{v}_k$.
   - **Aim:** Hence, a robust state estimation scheme is expected.
2. **Issue 2**: For a general nonlinear $\mathbf{y}_k = \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k)$, the likelihood of prior state $\boldsymbol{x}_k$ given measurement $\boldsymbol{y}_k$ is hard to be evaluated.
   - Only easy for additive and multiplicative measurement noises.
   - **Aim:** Therefore, a general likelihood evaluation method is expected.

Issues regarding the state estimation for the nonlinear system.

1. **Issue 1**: Typically, we assume **nominal** forms of nonlinear mappings $\boldsymbol{f}_k(\cdot)$ and $\boldsymbol{h}_k(\cdot)$, and **nominal** types and parameters of the distributions of $\mathbf{w}_{k-1}$ and $\mathbf{v}_k$ are exactly true. However, in practice, they might be uncertain.
   - Specifically, for given k, at least one of the following is inexact.
     - Process dynamics $\boldsymbol{f}_k(\cdot)$.
     - Measurement dynamics $\boldsymbol{h}_k(\cdot)$.
     - Types and/or parameters of distributions of $\mathbf{w}_{k-1}$.
     - Types and/or parameters of distributions of $\mathbf{v}_k$.
   - **Aim:** Hence, a robust state estimation scheme is expected.
2. **Issue 2**: For a general nonlinear $\mathbf{y}_k = \boldsymbol{h}_k(\mathbf{x}_k, \mathbf{v}_k)$, the likelihood of prior state $\boldsymbol{x}_k$ given measurement $\boldsymbol{y}_k$ is hard to be evaluated.
   - Only easy for additive and multiplicative measurement noises.
   - **Aim:** Therefore, a general likelihood evaluation method is expected.
3. **Issue 3**: What if measurement outliers exist? How to treat them?

Nonlinear System Case

Problem Background, Statements, and Motivations

Recall the Bayesian estimation procedure (n.b., $\boldsymbol{Y}_{k-1} := \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{k-1}\}$):

A-Priori Step: $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}) = \int_{\boldsymbol{x}_{k-1}} p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}) p(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}) d\boldsymbol{x}_{k-1}$

A-Posteriori Step: $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_k) \propto p(\boldsymbol{y}_k \mid \boldsymbol{x}_k) \cdot p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$

Handle Issue 1:

- Uncertain models render induced prior state distribution $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ and likelihood distribution $p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$ being uncertain as well.

Nonlinear System Case

Problem Background, Statements, and Motivations

Recall the Bayesian estimation procedure (n.b., $\boldsymbol{Y}_{k-1} := \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{k-1}\}$):

A-Priori Step: $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}) = \int_{\boldsymbol{x}_{k-1}} p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}) p(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}) d\boldsymbol{x}_{k-1}$

A-Posteriori Step: $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_k) \propto p(\boldsymbol{y}_k \mid \boldsymbol{x}_k) \cdot p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$

Handle Issue 1:

- Uncertain models render induced prior state distribution $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ and likelihood distribution $p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$ being uncertain as well.
- Recall Linear Case:

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top(\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1}\boldsymbol{H}\boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}\right]$$

  where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is variance of the prior, and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ is variance of the likelihood.

Recall the Bayesian estimation procedure (n.b., $\boldsymbol{Y}_{k-1} := \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{k-1}\}$):

A-Priori Step: $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}) = \int_{\boldsymbol{x}_{k-1}} p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}) p(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}) d\boldsymbol{x}_{k-1}$

A-Posteriori Step: $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_k) \propto p(\boldsymbol{y}_k \mid \boldsymbol{x}_k) \cdot p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$

Handle Issue 1:

- Uncertain models render induced prior state distribution $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ and likelihood distribution $p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$ being uncertain as well.
- Recall Linear Case:

$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^{\top} (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_{\mu}^{\min}\right]$$

  where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is variance of the prior, and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ is variance of the likelihood.
- If $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ is uncertain, we should make main use of $p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$, and let $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ be non-informative (i.e., flat).

Recall the Bayesian estimation procedure (n.b., $\boldsymbol{Y}_{k-1} := \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{k-1}\}$):

A-Priori Step: $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right) = \int_{\boldsymbol{x}_{k-1}} p\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}\right) p\left(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}\right) d\boldsymbol{x}_{k-1}$

A-Posteriori Step: $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_k\right) \propto p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right) \cdot p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$

Handle Issue 1:

- Uncertain models render induced prior state distribution $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and likelihood distribution $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ being uncertain as well.
- Recall Linear Case:
$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top \left(\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_\mu^{\min}\right]$$
where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is variance of the prior, and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ is variance of the likelihood.
- If $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ is uncertain, we should make main use of $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$, and let $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ be non-informative (i.e., flat).
- If $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ is uncertain, we should make main use of $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$, and let $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ be non-informative (i.e., flat).

Recall the Bayesian estimation procedure (n.b., $\boldsymbol{Y}_{k-1} := \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{k-1}\}$):

A-Priori Step: $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right) = \int_{\boldsymbol{x}_{k-1}} p\left(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1}\right) p\left(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}\right) d\boldsymbol{x}_{k-1}$

A-Posteriori Step: $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_k\right) \propto p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right) \cdot p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$

Handle Issue 1:

- Uncertain models render induced prior state distribution $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and likelihood distribution $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ being uncertain as well.
- Recall Linear Case:
$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top \left(\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^\top + \boldsymbol{\Sigma}_{\boldsymbol{v}}\right)^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_{\mu}^{\min}\right]$$
where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is variance of the prior, and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ is variance of the likelihood.
- If $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ is uncertain, we should make main use of $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$, and let $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ be non-informative (i.e., flat).
- If $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ is uncertain, we should make main use of $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$, and let $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ be non-informative (i.e., flat).
- Find worst-case prior state distribution $p^*\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and worst-case likelihood distribution $p^*\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$. "Worst-case" scenario defined by "entropy".

Nonlinear System Case

Problem Background, Statements, and Motivations

Recall the Bayesian estimation procedure (n.b., $\boldsymbol{Y}_{k-1} := \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_{k-1}\}$):

A-Priori Step: $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}) = \int_{\boldsymbol{x}_{k-1}} p(\boldsymbol{x}_k \mid \boldsymbol{x}_{k-1})\, p(\boldsymbol{x}_{k-1} \mid \boldsymbol{Y}_{k-1}) d\boldsymbol{x}_{k-1}$

A-Posteriori Step: $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_k) \propto p(\boldsymbol{y}_k \mid \boldsymbol{x}_k) \cdot p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$

Handle Issue 1:

- Uncertain models render induced prior state distribution $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ and likelihood distribution $p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$ being uncertain as well.
- Recall Linear Case:
$$\max_{\boldsymbol{\Sigma}_{\boldsymbol{x}}} \max_{\boldsymbol{\Sigma}_{\boldsymbol{v}}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{\boldsymbol{x}} - \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^{\top} (\boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{H}^{\top} + \boldsymbol{\Sigma}_{\boldsymbol{v}})^{-1} \boldsymbol{H} \boldsymbol{\Sigma}_{\boldsymbol{x}} \cdot i_{\mu}^{\min}\right]$$
where $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is variance of the prior, and $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ is variance of the likelihood.
- If $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ is uncertain, we should make main use of $p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$, and let $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ be non-informative (i.e., flat).
- If $p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$ is uncertain, we should make main use of $p(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$, and let $p(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$ be non-informative (i.e., flat).
- Find worst-case prior state distribution $p^*(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1})$ and worst-case likelihood distribution $p^*(\boldsymbol{y}_k \mid \boldsymbol{x}_k)$. "Worst-case" scenario defined by "entropy".
- **Note.** Why not directly consider $p(\boldsymbol{x}_k|\boldsymbol{Y}_k)$? Because it is overly conservative — it lacks flexibility (only process uncertainty or only measurement uncertainty).

Handle Issue 2: The maximum-entropy scheme can serve as a general likelihood evaluation method.

Handle Issue 3: Evaluating likelihoods of all prior state particles at the given measurement. If the **largest** likelihood (of all prior state particles) is smaller than a threshold (e.g., 5%), we treat this measurement as an outlier because none of these prior state particle can possibly generate this measurement.

- In particle filtering, $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ are presented by particles:

- In particle filtering, $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ are presented by particles:
    - For $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$, we have nominal prior state particles $\boldsymbol{x}^i_{k|k-1}$, $i \in [N]$.

- In particle filtering, $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ are presented by particles:
  - For $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$, we have nominal prior state particles $\boldsymbol{x}^i_{k|k-1}$, $i \in [N]$.
  - For $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$, we have nominal likelihood particles $\boldsymbol{y}^r_k|\boldsymbol{x}^i_{k|k-1}$, $r \in [R]$, $\forall i \in [N]$. (n.b., for every $\boldsymbol{x}^i_{k|k-1}$, we have R likelihood particles.)

- In particle filtering, $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ are presented by particles:
  - For $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$, we have nominal prior state particles $\boldsymbol{x}^i_{k|k-1}$, $i \in [N]$.
  - For $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$, we have nominal likelihood particles $\boldsymbol{y}^r_k|\boldsymbol{x}^i_{k|k-1}$, $r \in [R]$, $\forall i \in [N]$. (n.b., for every $\boldsymbol{x}^i_{k|k-1}$, we have R likelihood particles.)
- At each k, we have nominal prior state particles $\boldsymbol{x}^i$, $i \in [N]$ and their weights, and nominal likelihood particles $\boldsymbol{y}^r|\boldsymbol{x}^i$, $r \in [R]$, $\forall i \in [N]$ and their weights.

- In particle filtering, $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$ are presented by particles:
  - For $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$, we have nominal prior state particles $\boldsymbol{x}^i_{k|k-1}$, $i \in [N]$.
  - For $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$, we have nominal likelihood particles $\boldsymbol{y}^r_k | \boldsymbol{x}^i_{k|k-1}$, $r \in [R]$, $\forall i \in [N]$. (n.b., for every $\boldsymbol{x}^i_{k|k-1}$, we have R likelihood particles.)
- At each k, we have nominal prior state particles $\boldsymbol{x}^i$, $i \in [N]$ and their weights, and nominal likelihood particles $\boldsymbol{y}^r | \boldsymbol{x}^i$, $r \in [R]$, $\forall i \in [N]$ and their weights.
- For worst-case $p\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$ and $p\left(\boldsymbol{y}_k \mid \boldsymbol{x}_k\right)$, they can be either continuous or discrete.

1. **On Worst-Case Prior Distribution**: Given the particle-represented nominal $\hat{p}\left(\boldsymbol{x}_k \mid \boldsymbol{Y}_{k-1}\right)$, find a **max**imum-**ent**ropy distribution near it. Note that $\hat{p}(\cdot|\cdot)$ is supported on $\{\boldsymbol{x}^i\}_{i\in[N]}$.

Continuous: If the maxent is continous.

$$\begin{array}{cl} \max\limits_{p(\boldsymbol{x})\in L^1} & \displaystyle\int -p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} \\ s.t. & \left\{ \begin{array}{ll} D(p,\hat{p}) & \le \theta \\ \int p(\boldsymbol{x}) d\boldsymbol{x} & = 1 \end{array} \right. \end{array}$$

Discrete: If the maxent is discrete and supported on $\{\boldsymbol{x}^j\}_{j\in[M]}$ (not necessarily the same to $\{\boldsymbol{x}^i\}_{i\in[N]}$ but usually can be).

$$\begin{array}{cl} \max\limits_{p(\boldsymbol{x})\in l^1} & \displaystyle\sum_j -p(\boldsymbol{x}^j) \ln p(\boldsymbol{x}^j) \\ s.t. & \left\{ \begin{array}{ll} D(p,\hat{p}) & \le \theta \\ \sum_j p(\boldsymbol{x}^j) & = 1. \end{array} \right. \end{array}$$

2. **On Worst-Case Likelihood Distribution**: Given the particle-represented nominal $\hat{p}\left(\boldsymbol{y} \mid \boldsymbol{x}^j\right), \forall j \in [M]$, find a maxent distribution near it. Note that $\hat{p}(\cdot|\cdot)$ is supported on $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r\in[R]}, \forall j \in [M]$.

Continuous: If the maxent is continuous.

$$\max_{p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) \in L^1} \quad \int -p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) \ln p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) d\boldsymbol{y}$$
$$\text{s.t.} \quad \begin{cases} D(p, \hat{p}) & \leq \theta \\ \int p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) d\boldsymbol{y} & = 1 \end{cases}$$

Discrete: If the maxent is discrete and supported on $\{\boldsymbol{y}^t|\boldsymbol{x}^j\}_{t\in[T]}, \ \forall j \in [M]$ (not necessarily the same to $\{\boldsymbol{y}^r|\boldsymbol{x}^j\}_{r\in[R]}$ but usually can be).

$$\max_{p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}) \in l^1} \quad \sum_t -p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}^t) \ln p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}^t)$$
$$\text{s.t.} \quad \begin{cases} D(p, \hat{p}) & \leq \theta \\ \sum_t p_{\mathbf{y}|\boldsymbol{x}^j}(\boldsymbol{y}^t) & = 1. \end{cases}$$

where

1. $D(\cdot, \cdot)$ is a statistical similarity measures of two distributions. Not necessarily divergences or metrics/distances. We consider three cases.

where

1. $D(\cdot, \cdot)$ is a statistical similarity measures of two distributions. Not necessarily divergences or metrics/distances. We consider three cases.

    1. Moments-Based

where

1. $D(\cdot, \cdot)$ is a statistical similarity measures of two distributions. Not necessarily divergences or metrics/distances. We consider three cases.

    1. Moments-Based
    2. Wasserstein Metric

where

1. $D(\cdot, \cdot)$ is a statistical similarity measures of two distributions. Not necessarily divergences or metrics/distances. We consider three cases.

    1. Moments-Based
    2. Wasserstein Metric
    3. ϕ-Divergences (incl. especially KL-divergence)

where

1. $D(\cdot, \cdot)$ is a statistical similarity measures of two distributions. Not necessarily divergences or metrics/distances. We consider three cases.

   1. Moments-Based
   2. Wasserstein Metric
   3. ϕ-Divergences (incl. especially KL-divergence)

2. L^1 denote absolutely integrable function space. Because a continuous $p(\cdot)$ must be integrated to unit.

where

1. $D(\cdot, \cdot)$ is a statistical similarity measures of two distributions. Not necessarily divergences or metrics/distances. We consider three cases.
   1. Moments-Based
   2. Wasserstein Metric
   3. ϕ-Divergences (incl. especially KL-divergence)
2. L^1 denote absolutely integrable function space. Because a continuous $p(\cdot)$ must be integrated to unit.
3. l^1 denote absolutely summable sequence space. Because a discrete $p(\cdot)$ must be summed to unit.

One Example:

Theorem (Continuous Case Under Wasserstein)

The continuous maximum entropy distribution in Wasserstein ball is

$$p(\boldsymbol{x}) = \exp\left\{ -v_0 \min_{i \in [N]} \left\{ \|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i \right\} - v_1 - 1 \right\}$$

where $v_0 \in \mathbb{R}^1$, $v_1 \in \mathbb{R}^1$, and $\lambda_i \in \mathbb{R}^1, \forall i$ solve the following convex and smooth problem (n.b., almost-everywhere smooth in terms of λ_i; non-smooth only on zero-measure boundaries):

$$\begin{aligned} \min_{v_0, v_1, \boldsymbol{\lambda}} \quad & v_0 \cdot (\theta - \textstyle\sum_{i=1}^{N} \lambda_i q_i) + v_1 + \\ & \int \exp\left\{ -v_0 \min_{i \in [N]} \left\{ \|\boldsymbol{x} - \boldsymbol{x}^i\| - \lambda_i \right\} - v_1 - 1 \right\} d\boldsymbol{x} \\ \text{s.t.} \quad & v_0 \geq 0, \end{aligned}$$

where $\boldsymbol{\lambda} := [\lambda_1, \lambda_2, ..., \lambda_N]^\top$.

Projected Gradient Descent to solve the minimization sub-problem.

Another Example:

Theorem (Discrete Case Under KL-Divergence)

The discrete maximum entropy distribution in KL-Divergence ball is

$$p_i = \exp\left\{ \frac{-\lambda_0 \ln(q_i) + \lambda_1}{-(\lambda_0 + 1)} - 1 \right\}, \qquad \forall i \in [N],$$

where $\lambda_0 \in \mathbb{R}^1, \lambda_1 \in \mathbb{R}^1$ solve the following the convex and smooth problem:

$$\begin{aligned} \min_{\lambda_0, \lambda_1} \quad & \lambda_0 \theta + \lambda_1 + (\lambda_0 + 1) \sum_{i=1}^{N} p_i \\ \textit{s.t.} \quad & \lambda_0 \geq 0. \end{aligned}$$

Projected Gradient Descent to solve the minimization sub-problem.

After solving maximum entropy problems, distributionally robust state estimation for nonlinear systems is ready.

Recall the three proposed three steps to robustify particle filter.

1. Calculate worst-case prior state distribution.

After solving maximum entropy problems, distributionally robust state estimation for nonlinear systems is ready.

Recall the three proposed three steps to robustify particle filter.

1. Calculate worst-case prior state distribution.
2. Evaluate worst-case likelihood.

After solving maximum entropy problems, distributionally robust state estimation for nonlinear systems is ready.

Recall the three proposed three steps to robustify particle filter.

1. Calculate worst-case prior state distribution.
2. Evaluate worst-case likelihood.
3. Outlier identification and treatment.

Blue texts are modifications compared to standard particle filter.

Inputs: $\boldsymbol{x}^i_{k-1|k-1}, i \in [N]$ **Outputs**: $\boldsymbol{x}^i_{k|k}, i \in [N]$

Step 1 (Prior Estimation): $\boldsymbol{x}^i_{k|k-1} = \boldsymbol{f}_k(\boldsymbol{x}^i_{k-1|k-1}, \boldsymbol{w}^i_{k-1})$, $\forall i \in [N]$ where $\boldsymbol{w}^i_{k-1}$ are sampled from its distribution. Then, using this set of nominal prior state particles $\boldsymbol{x}^i_{k|k-1}$ to find worst-case prior state particles (and/or updating their weights). After which, nominal prior particles $\boldsymbol{x}^i_{k|k-1}$ are replaced by worst-case prior particles.

Step 2 (Likelihood Evaluation): For every (worst-case) $\boldsymbol{x}^i_{k|k-1}$, evaluate its (worst-case) likelihood $p(\boldsymbol{y}_k|\boldsymbol{x}^i_{k|k-1})$ at $\boldsymbol{y}_k$, during which outlier identification and treatment are applied.

Step 3 (Posterior Evaluation): Every (worst-case) prior $\boldsymbol{x}^i_{k|k-1}$ becomes (worst-case) posterior $\boldsymbol{x}^i_{k|k}$, weights update: $u_{\boldsymbol{x}^i_{k|k}} \leftarrow u_{\boldsymbol{x}^i_{k|k-1}} \cdot p(\boldsymbol{y}_k|\boldsymbol{x}^i_{k|k-1}), \forall i \in [N]$. Weights normalization is required.

Step 4 (Resampling): If most of weights are close to zero, "particle degeneracy" happens. Resampling $\boldsymbol{x}^i_{k|k}, i \in [N]$ according to the discrete distribution $\{u_{\boldsymbol{x}^i_{k|k}}\}_{i\in[N]}$.

Content

1. Robust filters are just remedial solutions.

1. Robust filters are just remedial solutions.
   - Reducing modeling uncertainties is always important.

1. Robust filters are just remedial solutions.
    - Reducing modeling uncertainties is always important.
    - Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).

1. Robust filters are just remedial solutions.
    - Reducing modeling uncertainties is always important.
    - Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).
    - Robust solutions are usable only when

1. Robust filters are just remedial solutions.
    - Reducing modeling uncertainties is always important.
    - Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).
    - Robust solutions are usable only when
        - Model uncertainties cannot be reduced, given fixed cost/budget.

1. Robust filters are just remedial solutions.
    - Reducing modeling uncertainties is always important.
    - Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).
    - Robust solutions are usable only when
        - Model uncertainties cannot be reduced, given fixed cost/budget.
        - Robust solutions are good enough for the problem of interest.

1. Robust filters are just remedial solutions.
    - Reducing modeling uncertainties is always important.
    - Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).
    - Robust solutions are usable only when
        - Model uncertainties cannot be reduced, given fixed cost/budget.
        - Robust solutions are good enough for the problem of interest.
    - It is better to take into consideration uncertainties from immediate sources where uncertainties occur. For example, for the model:

$$\begin{cases} \mathbf{x}_k &= (\boldsymbol{F}_{k-1} + \delta\boldsymbol{F}_{k-1})\mathbf{x}_{k-1} + (\boldsymbol{G}_{k-1} + \delta\boldsymbol{G}_{k-1})\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

1. Robust filters are just remedial solutions.
    - Reducing modeling uncertainties is always important.
    - Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).
    - Robust solutions are usable only when
        - Model uncertainties cannot be reduced, given fixed cost/budget.
        - Robust solutions are good enough for the problem of interest.
    - It is better to take into consideration uncertainties from immediate sources where uncertainties occur. For example, for the model:

$$\begin{cases} \mathbf{x}_k &= (\boldsymbol{F}_{k-1} + \delta\boldsymbol{F}_{k-1})\mathbf{x}_{k-1} + (\boldsymbol{G}_{k-1} + \delta\boldsymbol{G}_{k-1})\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

        - It is better if parameter uncertainties $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ can be directly modeled and constrained.

1. Robust filters are just remedial solutions.
    - Reducing modeling uncertainties is always important.
    - Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).
    - Robust solutions are usable only when
        - Model uncertainties cannot be reduced, given fixed cost/budget.
        - Robust solutions are good enough for the problem of interest.
    - It is better to take into consideration uncertainties from immediate sources where uncertainties occur. For example, for the model:

$$\begin{cases} \mathbf{x}_k & = (\boldsymbol{F}_{k-1} + \delta\boldsymbol{F}_{k-1})\mathbf{x}_{k-1} + (\boldsymbol{G}_{k-1} + \delta\boldsymbol{G}_{k-1})\mathbf{w}_{k-1}, \\ \mathbf{y}_k & = \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

        - It is better if parameter uncertainties $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ can be directly modeled and constrained.
        - Distributionally robust solutions are supplementary, not dominating. It is extremely useful when $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ cannot be directly modeled.

1. Robust filters are just remedial solutions.
   - Reducing modeling uncertainties is always important.
   - Readers should not expect that the proposed method is optimal or satisfactory in all scenarios, e.g., for a model with t-distributed measurement noises (which implies that the true model is known).
   - Robust solutions are usable only when
     - Model uncertainties cannot be reduced, given fixed cost/budget.
     - Robust solutions are good enough for the problem of interest.
   - It is better to take into consideration uncertainties from immediate sources where uncertainties occur. For example, for the model:

$$\begin{cases} \mathbf{x}_k &= (\boldsymbol{F}_{k-1} + \delta\boldsymbol{F}_{k-1})\mathbf{x}_{k-1} + (\boldsymbol{G}_{k-1} + \delta\boldsymbol{G}_{k-1})\mathbf{w}_{k-1}, \\ \mathbf{y}_k &= \boldsymbol{H}_k\mathbf{x}_k + \mathbf{v}_k, \end{cases}$$

     - It is better if parameter uncertainties $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ can be directly modeled and constrained.
     - Distributionally robust solutions are supplementary, not dominating. It is extremely useful when $\delta\boldsymbol{F}_{k-1}$ and $\delta\boldsymbol{G}_{k-1}$ cannot be directly modeled.
   - Why not directly constrain parameters?, e.g., $D(\boldsymbol{F}_k, \hat{\boldsymbol{F}}_k) \leq \theta$. It raises a matrix optimization problem that is even not a SDP. Tractability is a big issue!

2. The performance of the proposed method depends heavily on the size of uncertainty sets.

2. The performance of the proposed method depends heavily on the size of uncertainty sets.
   - If the size is too small, the algorithm has no sufficient robustness against the uncertainty. Contrarily, if the size is too large, the algorithm is too conservative to obtain a good performance as well.

2. The performance of the proposed method depends heavily on the size of uncertainty sets.
    - If the size is too small, the algorithm has no sufficient robustness against the uncertainty. Contrarily, if the size is too large, the algorithm is too conservative to obtain a good performance as well.
    - State estimation problems are not data-driven problems, and therefore, sizes cannot be determined from data (by, e.g., concentration inequalities).

2. The performance of the proposed method depends heavily on the size of uncertainty sets.
   - If the size is too small, the algorithm has no sufficient robustness against the uncertainty. Contrarily, if the size is too large, the algorithm is too conservative to obtain a good performance as well.
   - State estimation problems are not data-driven problems, and therefore, sizes cannot be determined from data (by, e.g., concentration inequalities).
   - One should carefully (and pragmatically) tune this parameter to achieve good performances for their specific real problems.

Content

Contributions

Part 1) For Linear System Case:

1. A distributionally robust state estimation framework against both parameter uncertainties and measurement outliers.

Contributions

Part 1) For Linear System Case:

1. A distributionally robust state estimation framework against both parameter uncertainties and measurement outliers.
    - It uses a family of distributions to describe the parameter uncertainties and measurement outliers, and the robust state estimation is made over the least-favorable distribution.

Contributions

Part 1) For Linear System Case:

1. A distributionally robust state estimation framework against both parameter uncertainties and measurement outliers.
   - It uses a family of distributions to describe the parameter uncertainties and measurement outliers, and the robust state estimation is made over the least-favorable distribution.
2. We show that the proposed framework generalizes several existing estimation methodologies.

Contributions

Part 1) For Linear System Case:

1. A distributionally robust state estimation framework against both parameter uncertainties and measurement outliers.
   - It uses a family of distributions to describe the parameter uncertainties and measurement outliers, and the robust state estimation is made over the least-favorable distribution.
2. We show that the proposed framework generalizes several existing estimation methodologies.
   - Including the fading Kalman filter, the Student's t Kalman filter, the risk-sensitive Kalman filter, the M-estimation-based Kalman filters, the relative-entropy Kalman filter, the τ-divergence Kalman filter, and the Wasserstein Kalman filter.

Contributions

Part 1) For Linear System Case:

1. A distributionally robust state estimation framework against both parameter uncertainties and measurement outliers.
   - It uses a family of distributions to describe the parameter uncertainties and measurement outliers, and the robust state estimation is made over the least-favorable distribution.
2. We show that the proposed framework generalizes several existing estimation methodologies.
   - Including the fading Kalman filter, the Student's t Kalman filter, the risk-sensitive Kalman filter, the M-estimation-based Kalman filters, the relative-entropy Kalman filter, the τ-divergence Kalman filter, and the Wasserstein Kalman filter.
3. We show that the proposed distributionally robust state estimation problem can be reformulated into a linear/nonlinear semi-definite program.

Contributions

Part 1) For Linear System Case:

1. A distributionally robust state estimation framework against both parameter uncertainties and measurement outliers.
   - It uses a family of distributions to describe the parameter uncertainties and measurement outliers, and the robust state estimation is made over the least-favorable distribution.
2. We show that the proposed framework generalizes several existing estimation methodologies.
   - Including the fading Kalman filter, the Student's t Kalman filter, the risk-sensitive Kalman filter, the M-estimation-based Kalman filters, the relative-entropy Kalman filter, the τ-divergence Kalman filter, and the Wasserstein Kalman filter.
3. We show that the proposed distributionally robust state estimation problem can be reformulated into a linear/nonlinear semi-definite program.
   - In some special cases it can be analytically (i.e., efficiently) solved.

Contributions

Part 2) For Nonlinear System Case:

1. We propose a robustification scheme for particle filters.

Contributions

Part 2) For Nonlinear System Case:

1. We propose a robustification scheme for particle filters.
    - We use worst-case distribution to generate new prior state particles and/or update their weights, and use worst-case distribution to evaluate the worst-case likelihoods.

Contributions

Part 2) For Nonlinear System Case:

1. We propose a robustification scheme for particle filters.
    - We use worst-case distribution to generate new prior state particles and/or update their weights, and use worst-case distribution to evaluate the worst-case likelihoods.
2. We derive maximum entropy distributions in balls centered at nominal distributions with radii defined by the Wasserstein distance and the Kullback-Leibler divergence.

Contributions

Part 2) For Nonlinear System Case:

1. We propose a robustification scheme for particle filters.
    - We use worst-case distribution to generate new prior state particles and/or update their weights, and use worst-case distribution to evaluate the worst-case likelihoods.
2. We derive maximum entropy distributions in balls centered at nominal distributions with radii defined by the Wasserstein distance and the Kullback-Leibler divergence.
3. We show that this robustification scheme serves yet a new resampling strategy against particle degeneracy.

Contributions

Part 2) For Nonlinear System Case:

1. We propose a robustification scheme for particle filters.
   - We use worst-case distribution to generate new prior state particles and/or update their weights, and use worst-case distribution to evaluate the worst-case likelihoods.
2. We derive maximum entropy distributions in balls centered at nominal distributions with radii defined by the Wasserstein distance and the Kullback-Leibler divergence.
3. We show that this robustification scheme serves yet a new resampling strategy against particle degeneracy.
   - Maximum entropy distributions tend to be flat, and therefore, in particle filter, particles tend to have equal weights.

Contributions

Part 2) For Nonlinear System Case:

1. We propose a robustification scheme for particle filters.
   - We use worst-case distribution to generate new prior state particles and/or update their weights, and use worst-case distribution to evaluate the worst-case likelihoods.
2. We derive maximum entropy distributions in balls centered at nominal distributions with radii defined by the Wasserstein distance and the Kullback-Leibler divergence.
3. We show that this robustification scheme serves yet a new resampling strategy against particle degeneracy.
   - Maximum entropy distributions tend to be flat, and therefore, in particle filter, particles tend to have equal weights.
4. We show that the proposed robustification scheme offers a universal likelihood evaluation method for prior state particles when measurement dynamics is driven by non-additive and non-multiplicative noises.

Contributions

Part 2) For Nonlinear System Case:

1. We propose a robustification scheme for particle filters.
   - We use worst-case distribution to generate new prior state particles and/or update their weights, and use worst-case distribution to evaluate the worst-case likelihoods.
2. We derive maximum entropy distributions in balls centered at nominal distributions with radii defined by the Wasserstein distance and the Kullback-Leibler divergence.
3. We show that this robustification scheme serves yet a new resampling strategy against particle degeneracy.
   - Maximum entropy distributions tend to be flat, and therefore, in particle filter, particles tend to have equal weights.
4. We show that the proposed robustification scheme offers a universal likelihood evaluation method for prior state particles when measurement dynamics is driven by non-additive and non-multiplicative noises.
5. We show: Gaussian approximation state estimators are distributionally robust.

Contributions

Part 2) For Nonlinear System Case:

1. We propose a robustification scheme for particle filters.
   - We use worst-case distribution to generate new prior state particles and/or update their weights, and use worst-case distribution to evaluate the worst-case likelihoods.
2. We derive maximum entropy distributions in balls centered at nominal distributions with radii defined by the Wasserstein distance and the Kullback-Leibler divergence.
3. We show that this robustification scheme serves yet a new resampling strategy against particle degeneracy.
   - Maximum entropy distributions tend to be flat, and therefore, in particle filter, particles tend to have equal weights.
4. We show that the proposed robustification scheme offers a universal likelihood evaluation method for prior state particles when measurement dynamics is driven by non-additive and non-multiplicative noises.
5. We show: Gaussian approximation state estimators are distributionally robust.
   - Fixing second moments, maximum entropy distributions are Gaussian.

Thank You

Thank You!

Content

References I

[1] Soroosh Shafieezadeh Abadeh et al. "Wasserstein distributionally robust Kalman filtering". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8474–8483.

[2] Gabriel Agamennoni, Juan I Nieto, and Eduardo M Nebot. "Approximate inference in state-space models with heavy-tailed noise". In: *IEEE Transactions on Signal Processing* 60.10 (2012), pp. 5024–5037.

[3] B. D. Anderson and John Barratt Moore. *Optimal Filtering*. Prentice-Hall, 1979.

[4] M Sanjeev Arulampalam et al. "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking". In: *IEEE Transactions on Signal Processing* 50.2 (2002), pp. 174–188.

[5] Monica F Bugallo et al. "Adaptive importance sampling: The past, the present, and the future". In: *IEEE Signal Processing Magazine* 34.4 (2017), pp. 60–79.

[6] Guanrong Chen. *Approximate Kalman Filtering*. Vol. 2. World Scientific, 1993.

[7] Ruidi Chen, Ioannis Ch Paschalidis, et al. "Distributionally Robust Learning". In: *Foundations and Trends® in Optimization* 4.1-2 (2020), pp. 1–243.

[8] Steven Gillijns and Bart De Moor. "Unbiased minimum-variance input and state estimation for linear discrete-time systems". In: *Automatica* 43.1 (2007), pp. 111–116.

[9] Yulong Huang et al. "A Novel Robust Gaussian-Student's t Mixture Distribution Based Kalman Filter". In: *IEEE Transactions on Signal Processing* 67.13 (2019), pp. 3606–3620.

References III

[10] Peter J Huber. *Robust Statistics (2nd edition)*. John Wiley & Sons, 2009.

[11] Yousef El-Laham, Victor Elvira, and Mónica F Bugallo. "Robust covariance adaptation in adaptive importance sampling". In: *IEEE Signal Processing Letters* 25.7 (2018), pp. 1049–1053.

[12] Tiancheng Li, Miodrag Bolic, and Petar M Djuric. "Resampling methods for particle filtering: classification, implementation, and strategies". In: *IEEE Signal Processing Magazine* 32.3 (2015), pp. 70–86.

[13] X Rong Li and Vesselin P Jilkov. "Survey of maneuvering target tracking. Part I. Dynamic models". In: *IEEE Transactions on Aerospace and Electronic Systems* 39.4 (2003), pp. 1333–1364.

References IV

[14] Julian Neri, Philippe Depalle, and Roland Badeau. “Approximate inference and learning of state space models with Laplace noise”. In: *IEEE Transactions on Signal Processing* 69 (2021), pp. 3176–3189.

[15] Ali H Sayed. “A framework for state-space estimation with uncertain models”. In: *IEEE Transactions on Automatic Control* 46.7 (2001), pp. 998–1013.

[16] Dan Simon. *Optimal State Estimation: Kalman, H_∞, and Nonlinear Approaches*. John Wiley & Sons, 2006.

[17] Lu Sun et al. “Recursive maximum likelihood estimation with t-distribution noise model”. In: *Automatica* 132 (2021), p. 109789.

[18] Fan Wang and Venkataramanan Balakrishnan. “Robust Kalman filters for linear time-varying systems with stochastic parametric uncertainties”. In: *IEEE Transactions on Signal Processing* 50.4 (2002), pp. 803–813.