Supplementary Materials

Appendix H
Robustification Over Prior Model Probability

In this appendix, we discuss the case where we solve the genuine problem (24) with respect to $\boldsymbol{\omega}$, rather than the simplified case with respect to $\boldsymbol{\mu}$. In consideration of the high computational complexity, the genuine problem (24) with respect to $\boldsymbol{\omega}$ is not investigated in the main body of the paper.

Proposition 7: If the model set is exact and only the prior model probability vector $\boldsymbol{\omega}$ is uncertain [i.e., the special ambiguity set (23) is investigated], the reformulated distributionally robust Bayesian estimation problem (24) can be further reformulated into a tractable quadratic fractional program

$$\begin{array}{ll}\max\limits_{\boldsymbol{\omega}} & -\dfrac{\boldsymbol{\omega}^\top(\boldsymbol{CAC}-\boldsymbol{pb}^\top\boldsymbol{C})\boldsymbol{\omega}}{\boldsymbol{\omega}^\top\boldsymbol{pp}^\top\boldsymbol{\omega}} \\ \textit{s.t.} & \begin{cases}\displaystyle\sum_{j=1}^N \omega_j = 1, \\ \omega_j \geq 0, \qquad \forall j \in [N], \\ \Delta_0(\boldsymbol{\omega}, \bar{\boldsymbol{\omega}}) \leq \theta_0,\end{cases}\end{array} \tag{52}$$

where $\boldsymbol{p} := [p_1(\boldsymbol{y}), p_2(\boldsymbol{y}), \ldots, p_N(\boldsymbol{y})]^\top$ denotes the likelihoods of the candidate models given the measurement $\boldsymbol{y}$ and $\boldsymbol{C} := \mathrm{diag}(\boldsymbol{p})$ is a diagonal matrix whose diagonal entries are elements of $\boldsymbol{p}$.

Proof: From (3), for every $j \in [N]$, we have $\mu_j = \frac{\omega_j p_j(\boldsymbol{y})}{\sum_{j=1}^N \omega_j p_j(\boldsymbol{y})} = \frac{\omega_j p_j(\boldsymbol{y})}{\boldsymbol{\omega}^\top\boldsymbol{p}}$, i.e., $\boldsymbol{\mu} = [\mu_1, \mu_2, \ldots, \mu_N]^\top = \frac{\boldsymbol{C\omega}}{\boldsymbol{\omega}^\top\boldsymbol{p}}$. Therefore, the problem (24) can be explicitly written as

$$\begin{array}{ll}\max\limits_{\boldsymbol{\omega}} & -\left(\dfrac{\boldsymbol{C\omega}}{\boldsymbol{\omega}^\top\boldsymbol{p}}\right)^\top \boldsymbol{A}\left(\dfrac{\boldsymbol{C\omega}}{\boldsymbol{\omega}^\top\boldsymbol{p}}\right) + \boldsymbol{b}^\top\left(\dfrac{\boldsymbol{C\omega}}{\boldsymbol{\omega}^\top\boldsymbol{p}}\right) \\ \textit{s.t.} & \begin{cases}\displaystyle\sum_{j=1}^N \omega_j = 1, \\ \omega_j \geq 0, \qquad \forall j \in [N], \\ \Delta_0(\boldsymbol{\omega}, \bar{\boldsymbol{\omega}}) \leq \theta_0,\end{cases}\end{array} \tag{53}$$

which can be rearranged into the quadratic fractional program (52). □

The problem (52) can be written in a compact form

$$\max_{\boldsymbol{\omega}\in\Omega} \frac{f_1(\boldsymbol{\omega})}{f_2(\boldsymbol{\omega})}, \tag{54}$$

where $f_1(\boldsymbol{\omega}) := -\boldsymbol{\omega}^\top(\boldsymbol{CAC}-\boldsymbol{pb}^\top\boldsymbol{C})\boldsymbol{\omega}$ denotes the numerator of the objective of (52), $f_2(\boldsymbol{\omega}) := \boldsymbol{\omega}^\top\boldsymbol{pp}^\top\boldsymbol{\omega}$ the denominator of the objective of (52), and Ω the feasible region of (52). One may verify that although $f_2(\boldsymbol{\omega})$ is convex, $f_1(\boldsymbol{\omega})$ is neither concave nor convex. However, $f_1(\boldsymbol{\omega}) \geq 0$ can be guaranteed because the objective of (19) is non-negative, as are those of (24) and (52). Complete (approximated) solutions to the problem (54) can be found in, e.g., [S1],[7] [S2],[8] where involved indefinite quadratic programs can be solved by the method in, e.g., [S3].[9] Numerically solving (54) is time-consuming due to the indefiniteness of $f_1(\boldsymbol{\omega})$. Therefore, in this paper, we do not proceed further for (54). Instead, we find a simplified alternative to the original problem (24) with respect to $\boldsymbol{\mu}$. Interested readers may implement solution methods in, e.g., [S3], to solve (54) themselves.

Appendix I
Solution to (26)

The Lagrangian of (26) is

$$\begin{array}{ll}\min\limits_{\lambda_0\geq 0,\lambda_1}\max\limits_{\boldsymbol{\mu}} & -\boldsymbol{\mu}^\top\boldsymbol{A\mu} + \boldsymbol{b}^\top\boldsymbol{\mu} + \lambda_1\cdot(1-\boldsymbol{1}^\top\boldsymbol{\mu}) + \\ & \lambda_0\cdot(\theta_0 - \boldsymbol{\mu}^\top\ln\boldsymbol{\mu} + \boldsymbol{\mu}^\top\ln\bar{\boldsymbol{\mu}}).\end{array} \tag{55}$$

For every $\lambda_0 \geq 0$ and λ_1, the maximum $\boldsymbol{\mu}$ satisfies the first-order optimality condition:

$$-2\boldsymbol{A\mu} + \boldsymbol{b} - \lambda_1\cdot\boldsymbol{1} + \lambda_0\cdot(-\ln\boldsymbol{\mu} - \boldsymbol{1} + \ln\bar{\boldsymbol{\mu}}) = \boldsymbol{0}, \tag{56}$$

[7] [S1] W. Dinkelbach, "On nonlinear fractional programming," Management Science, vol. 13, no. 7, pp. 492–498, 1967.

[8] [S2] A. T. Phillips, Quadratic Fractional Programming: Dinkelbach Method.Boston, MA: Springer US, 2001, pp. 2107–2110. [Online]. Available: https://doi.org/10.1007/0-306-48332-7_406.

[9] [S3] A. Phillips and J. Rosen, "Guaranteed ϵ-approximate solution for indefinite quadratic global minimization," Naval Research Logistics (NRL), vol. 37, no. 4, pp. 499–514, 1990.

which transforms (55) to

$$\min_{\lambda_0 \geq 0, \lambda_1} \quad \lambda_0\theta_0 + \lambda_1 + \boldsymbol{\mu}^\top \boldsymbol{A}\boldsymbol{\mu} + \lambda_0 \mathbf{1}^\top \boldsymbol{\mu}. \tag{57}$$

Since (26) is a convex program and $\bar{\boldsymbol{u}}$ is a relative interior point in the feasible set, there does not exist duality gap between (26) and (57). Since (57) is convex, any first-order gradient-based method, e.g., projected gradient descent, is applicable to solve it. Let the objective of (57) be denoted as $f(\boldsymbol{\lambda})$. From (56), we have $-2\boldsymbol{A}\frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\lambda_0} = \ln \boldsymbol{\mu} + \mathbf{1} - \ln \bar{\boldsymbol{\mu}} + \lambda_0 \frac{1}{\boldsymbol{\mu}} \odot \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\lambda_0}$, and $-2\boldsymbol{A}\frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\lambda_1} = \mathbf{1} + \lambda_0 \frac{1}{\boldsymbol{\mu}} \odot \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\lambda_1}$, where $\frac{1}{\boldsymbol{\mu}}$ means element-wise fraction, and $\odot$ denotes the Hadamard product (i.e., the element-wise product). The gradient of the objective of (57) with respect to λ_0 and λ_1 are given by

$$\begin{aligned} \frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_0} &= \theta_0 + 2\boldsymbol{\mu}^\top \boldsymbol{A} \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\lambda_0} + \mathbf{1}^\top \boldsymbol{\mu} + \lambda_0 \mathbf{1}^\top \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\lambda_0} \\ &= \theta_0 - \boldsymbol{\mu}^\top \ln \boldsymbol{\mu} + \boldsymbol{\mu}^\top \ln \bar{\boldsymbol{\mu}}, \end{aligned} \tag{58}$$

and

$$\frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_1} = 1 + 2\boldsymbol{\mu}^\top \boldsymbol{A} \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\lambda_0} + \lambda_0 \mathbf{1}^\top \frac{\mathrm{d}\boldsymbol{\mu}}{\mathrm{d}\lambda_1} = 1 - \mathbf{1}^\top \boldsymbol{\mu}. \tag{59}$$

respectively. Hence, when the optimality of (57) reaches, i.e., when the gradients with respect to λ_0 and λ_1 vanish, we have $1 = \sum_{j=1}^N \mu_j$ and $\theta_0 = \sum_{j=1}^N \mu_j \cdot \ln \frac{\mu_j}{\bar{\mu}_j}$. Specifically, it means $\boldsymbol{\mu}$ is indeed a distribution summed to unit and all the robustness budget θ_0 has been used. In summary, the solution to (26) is summarized in Algorithm 2. Since (26) is a convex program, every iteration improves the objective.

Algorithm 2 Solution to (26)

Definition: S as maximum allowed iteration steps and s the current iteration step; α as step size; ϵ as numerical precision threshold; $\mathrm{abs}(\cdot)$ returns absolute value.

Remark: Since (57) is convex, any initial values for $\lambda_0 \geq 0$ and λ_1 are acceptable. If early stopping is applied (i.e., S is not sufficiently large for time-saving purpose), a normalization procedure is necessary to guarantee $1 = \sum_j \mu_j$.

Input: $S, \alpha, \epsilon, \lambda_0, \lambda_1$

$s \leftarrow 0$;
while true **do**
  // Update $\boldsymbol{\mu}$
  Solve N-variable nonlinear root-finding sub-problem (56) to obtain $\boldsymbol{\mu}^{(s)}$ with current λ_0 and λ_1 (see Remark 8)
  // Gradient Descent to Update λ_0 *and* λ_1
  $\lambda_0 \leftarrow \lambda_0 - \alpha \cdot \frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_0}$ // See *(58)*
  $\lambda_1 \leftarrow \lambda_1 - \alpha \cdot \frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_1}$ // See *(59)*
  // Projection
  if $\lambda_0 < 0$ **then** $\lambda_0 \leftarrow 0$
  end if
  // Next Iteration
  $s \leftarrow s + 1$
  // Stopping Rule
  if $s > S$ **or** $\mathrm{abs}(\frac{\partial f(\boldsymbol{\lambda})}{\partial \lambda_1}) < \epsilon$ **then**
    if $1 \neq \sum_i \mu_i^{(s)}$ **then** *// Early Stopping Applied*
      $\mu_i^{(s)} \leftarrow \mu_i^{(s)} / \sum_j \mu_j^{(s)}, \quad \forall i \in [N]$,
    end if
    break while
  end if
end while

Output: $\boldsymbol{\mu}^{(s)}$

Remark 8: We discuss the N-variate root-finding problem $-2\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} - \lambda_1 \cdot \mathbf{1} + \lambda_0 \cdot (-\ln \boldsymbol{\mu} - \mathbf{1} + \ln \bar{\boldsymbol{\mu}}) = \mathbf{0}$ on $\boldsymbol{\mu} \geq \mathbf{0}$. Let $\boldsymbol{g}(\boldsymbol{\mu}) := -2\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b} - \lambda_1 \cdot \mathbf{1} + \lambda_0 \cdot (-\ln \boldsymbol{\mu} - \mathbf{1} + \ln \bar{\boldsymbol{\mu}})$. One may verify that $\mathrm{d}\boldsymbol{g}(\boldsymbol{\mu})/\mathrm{d}\boldsymbol{\mu} \prec \mathbf{0}$ (i.e., $\boldsymbol{g}$ is a monotonically decreasing function in $\boldsymbol{\mu}$), $\boldsymbol{g}(\mathbf{0}) \to \infty$, and $\boldsymbol{g}(\infty) \to -\infty$. Therefore, at least one root of $\boldsymbol{g}(\boldsymbol{\mu}) = \mathbf{0}$ exists and Newton's method can be used to find it. □

Remark 9: If the 2-norm constraint $\|\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}\|_2 \leq \theta_0$ is used to replace the KL divergence constraint, then the root-finding procedure would be significantly simplified. Therefore, in practice, to save computational time, one may choose the 2-norm constraint $(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}})^\top (\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}) \leq \theta_0^2$. Another choice to reduce the computational complexity is to use the Frank-Wolfe method (i.e., linearization of the objective function) as in Proposition 5. □

APPENDIX J
THE STANDARD IMM FILTER

The implementation details of the interactive multiple model (IMM) method is given in Algorithm 3. The results in Step 2 (see Line 25) are due to (2) and (4) where $\mu_{j,k|k-1}$ and $\mu_{j,k|k}$ are prior and posterior model probabilities of the j^{th} model, respectively. The prior model probability, model likelihood, and posterior model probability of the j^{th} model are calculated in Step 1.5 (see Line 18), Step 1.6 (see Line 20), and Step 1.7 (see Line 22), respectively. See [3], [5] (in the reference list of the main body of the paper) for more information.

Algorithm 3 Interactive Multiple Model Algorithm [3], [5]

Definition: Let $\hat{\boldsymbol{x}}_{j,k|k-1}$ denote the prior state estimate provided by the j^{th} model and $\boldsymbol{P}_{j,k|k-1}$ the corresponding state estimation error covariance. Let $\hat{\boldsymbol{x}}_{j,k|k}$ denote the posterior state estimate provided by the j^{th} model and $\boldsymbol{P}_{j,k|k}$ the corresponding state estimation error covariance; Let $\hat{\boldsymbol{x}}_{k|k}$ denote the combined posterior state estimate of the N models and $\boldsymbol{P}_{k|k}$ the corresponding state estimation error covariance; Let $\mu_{j,k|k-1}$ and $\mu_{j,k|k}$ be the prior and posterior model probability of the j^{th} model at the time k, respectively; Let $\{\pi_{ij}\}_{i,j=1,2,\ldots,N}$ be the model transition probability matrix.
Initialization: $\forall j \in [N]$, initialize $\mu_{j,0|0}$, $\hat{\boldsymbol{x}}_{j,0|0}$, and $\boldsymbol{P}_{j,0|0}$.
Remark: In literature, prior and posterior state estimate are also known as predicted and updated state estimate, respectively.

Input: $\boldsymbol{y}_k$, $k = 1, 2, 3, \ldots$

1: **while** true **do**
2:     *// (Step 1) At Time k*
3:     **for** $j = 1 : N$ **do**
4:         *// (Step 1.1) Transition Probability From i^{th} Model at Time $k-1$ To j^{th} Model at Time k*
5:             $\mu_{ij,k|k-1} = \frac{\pi_{ij} \cdot \mu_{i,k-1|k-1}}{\sum_{i=1}^{N} \pi_{ij} \cdot \mu_{i,k-1|k-1}}$
6:         *// (Step 1.2) Initialize the j^{th} Filter*
7:             $\hat{\boldsymbol{x}}^0_{j,k-1|k-1} = \sum_{i=1}^{N} \mu_{ij,k|k-1} \cdot \hat{\boldsymbol{x}}_{i,k-1|k-1}$
8:             $\boldsymbol{P}^0_{j,k-1|k-1} = \sum_{i=1}^{N} \mu_{ij,k|k-1} \cdot \Big\{ \boldsymbol{P}_{i,k-1|k-1} + (\hat{\boldsymbol{x}}_{i,k-1|k-1} - \hat{\boldsymbol{x}}^0_{j,k-1|k-1})(\hat{\boldsymbol{x}}_{i,k-1|k-1} - \hat{\boldsymbol{x}}^0_{j,k-1|k-1})^\top \Big\}$
9:         *// (Step 1.3) Prior Estimation of the j^{th} Filter (i.e., Time Update)*
10:             $\hat{\boldsymbol{x}}_{j,k|k-1} = \boldsymbol{F}_{j,k-1} \hat{\boldsymbol{x}}^0_{j,k-1|k-1}$
11:             $\boldsymbol{P}_{j,k|k-1} = \boldsymbol{F}_{j,k-1} \boldsymbol{P}^0_{j,k-1|k-1} \boldsymbol{F}^\top_{j,k-1} + \boldsymbol{G}_{j,k-1} \boldsymbol{Q}_{j,k-1} \boldsymbol{G}^\top_{j,k-1}$
12:         *// (Step 1.4) Posterior Estimation of the j^{th} Filter (i.e., Measurement Update)*
13:             $\boldsymbol{r}_{j,k} = \boldsymbol{y}_k - \boldsymbol{H}_{j,k} \hat{\boldsymbol{x}}_{j,k|k-1}$ *// Innovation*
14:             $\boldsymbol{S}_{j,k} = \boldsymbol{H}_{j,k} \boldsymbol{P}_{j,k|k-1} \boldsymbol{H}^\top_{j,k} + \boldsymbol{R}_{j,k}$ *// Innovation Covariance*
15:             $\boldsymbol{K}_{j,k} = \boldsymbol{P}_{j,k|k-1} \boldsymbol{H}^\top_{j,k} \boldsymbol{S}^{-1}_{j,k}$ *// Filter Gain*
16:             $\hat{\boldsymbol{x}}_{j,k|k} = \hat{\boldsymbol{x}}_{j,k|k-1} + \boldsymbol{K}_{j,k} \cdot \boldsymbol{r}_{j,k} = \hat{\boldsymbol{x}}_{j,k|k-1} + \boldsymbol{P}_{j,k|k-1} \boldsymbol{H}^\top_{j,k} \boldsymbol{S}^{-1}_{j,k} \cdot \big[\boldsymbol{y}(k) - \boldsymbol{H}_{j,k} \hat{\boldsymbol{x}}_{j,k|k-1} \big]$
17:             $\boldsymbol{P}_{j,k|k} = \boldsymbol{P}_{j,k|k-1} - \boldsymbol{P}_{j,k|k-1} \boldsymbol{H}^\top_{j,k} \boldsymbol{S}^{-1}_{j,k} \boldsymbol{H}^\top_{j,k} \boldsymbol{P}_{j,k|k-1}$
18:         *// (Step 1.5) Prior Probability of the j^{th} Model*
19:             $\mu_{j,k|k-1} = \sum_{i=1}^{N} \pi_{ij} \cdot \mu_{i,k-1|k-1}$
20:         *// (Step 1.6) Likelihood of the j^{th} Model*
21:             $\lambda_{j,k} = \mathcal{N}_n(\boldsymbol{r}_{j,k}; \boldsymbol{0}, \boldsymbol{S}_{j,k})$
22:         *// (Step 1.7) Posterior Probability of the j^{th} Model*
23:             $\mu_{j,k|k} = \frac{\mu_{j,k|k-1} \cdot \lambda_{j,k}}{\sum_{i=1}^{N} \mu_{j,k|k-1} \cdot \lambda_{j,k}}$
24:     **end for**
25:     *// (Step 2) Combined Posterior State Estimate*
26:         $\hat{\boldsymbol{x}}_{k|k} = \sum_{j=1}^{N} \mu_{j,k|k} \cdot \hat{\boldsymbol{x}}_{j,k|k}$
27:         $\boldsymbol{P}_{k|k} = \sum_{j=1}^{N} \mu_{j,k|k} \cdot \big\{ \boldsymbol{P}_{j,k|k} + (\hat{\boldsymbol{x}}_{j,k|k} - \hat{\boldsymbol{x}}_{k|k})(\hat{\boldsymbol{x}}_{j,k|k} - \hat{\boldsymbol{x}}_{k|k})^\top \big\}$
28:     *// (Step 3) Next Time Step*
29:         $k \leftarrow k + 1$
30: **end while**

Output: $\hat{\boldsymbol{x}}_{k|k}, \boldsymbol{P}_{k|k}, \mu_{j,k|k}$