OUTLIER-ROBUST VARIATIONAL STATE ESTIMATION OF MODEL-FREE PROCESS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop robust variational state estimation (RoVaS), a data-driven method for estimating the state of a model-free dynamical process from noisy measurements with outliers. We assume no knowledge of the dynamics of the process; hence, we do not have access to the state-transition model. The setting is challenging because outlier detection and dynamics learning are coupled: identifying corrupted measurements requires a model of the dynamics, while learning the dynamics requires knowing which measurements to trust. RoVaS introduces latent corruption indicators alongside the states and infers them jointly under a variational framework. A corruption gate, parameterized to saturate as the measurement magnitude grows, attenuates suspected outliers before they reach the state posterior; this yields a provable outlier-robustness guarantee at inference. On stochastic Lorenz-63 systems and real human motion-capture data, RoVaS degrades gracefully under corruption rates up to 25% and is competitive with a state-transition-model-informed robust filter, despite having no access to the state-transition model.

1 INTRODUCTION

The estimation of the state of a dynamical process from noisy measurements has many applications such as sequence modeling, tracking, navigation, and forecasting (Särkkä, 2013; Durbin & Koopman, 2012). Traditional state estimation methods such as the Kalman filter (KF) (Kalman, 1960), the extended Kalman filter (EKF) (Gelb & Corporation, 1974), the unscented Kalman filter (UKF) (Wan & Van Der Merwe, 2000; Julier & Uhlmann, 2004), the ensemble Kalman filter (EnKF) (Evensen, 2003), and the particle filter (PF) (Arulampalam et al., 2002) enjoy wide adoption across many applications. These traditional state estimation methods assume the underlying dynamical process to be Markov and rely on a state-transition model (STM) representing the Markov dynamics. The STM, denoted $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ where $\mathbf{x}_t$ is the state at time t, is typically given an analytical structure derived from the physics of the process or relevant domain knowledge. There exists a significant body of research on the design of STM-based state estimation methods; in this article, we address a setting in which no such STM is available and, additionally, some of the measurements are corrupted as outliers.

The STM-based state estimation methods have been further extended for robustness to outliers, that is, to settings in which some of the noisy measurements are corrupted. The corruption may take several forms; for example, some measurements are saturated, or replaced by spurious measurements, or the measurement noise has a heavy-tailed distribution. There exists a rich literature on outlier-robust STM-based state estimation methods (Karlgaard & Schaub, 2007; Chang, 2014; Gandhi & Mili, 2010; Piche et al., 2012; Roth et al., 2013; Huang et al., 2016). Recently, Duran-Martin et al. (2024) developed the weighted measurement likelihood filter (WoLF), in which suitable weighting functions are designed to use measurements with corruption. The WoLF requires explicit knowledge of the STM: at a given time t, the STM is used to predict the measurement that will be observed at time t; if the observed measurement deviates substantially from the predicted one, its influence is down-weighted. The WoLF has a provable robustness guarantee in the form of a bounded posterior influence function (PIF), and it has been experimentally demonstrated for designing outlier-robust KF, EKF, and EnKF.

A separate line of research addresses state estimation for model-free dynamical processes, in which no knowledge of the state dynamics is assumed and no analytically structured STM is available. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dynamical process can also be non-Markov. If the physics of the process is unknown or complex, then it is difficult to form an analytical STM, and traditional STM-based state estimation methods cannot be directly applied. Designing state estimation methods for model-free processes is therefore a challenging research problem.

Recently proposed data-driven state estimation methods for model-free processes use neural networks, whose parameters are optimized using a training dataset of noisy measurement sequences. Prominent examples in unsupervised and self-supervised setups are the recurrent filter (Ruhe & Forré, 2022), data-driven nonlinear state estimation (DANSE) (Ghosh et al., 2024), the gated inference network (GIN) (Hashempoor & Choi, 2024), the data-driven nonlinear smoother (DNS) (Cumlin et al., 2026), variational state estimation (VSE) (Norén et al., 2026), and differentiable particle filters (DPFs) (Corenflos et al., 2021; van Nierop et al., 2025). In supervised learning setups, examples include KalmanNet (Revach et al., 2022b), RTSNet (Ni et al., 2022), and recurrent Kalman networks (RKN) (Becker et al., 2019). There also exist state-space-model-based methods such as the deep Kalman filter (DKF) (Krishnan et al., 2015), the Kalman variational autoencoder (KVAE) (Fraccaro et al., 2017), the deep Markov model (DMM) (Krishnan et al., 2016), and deep variational Bayes filters (DVBF) (Karl et al., 2017), which are variants of dynamical variational autoencoders (DVAEs) (Girin et al., 2021) that use variational inference for sequential latent-variable models; DVAEs are mainly developed for measurement sequence modeling and latent representation learning, and not explicitly for state estimation. GIN and the DPF studied by van Nierop et al. (2025) explored gating of measurements, that is, preprocessing measurements before state estimation, but for denoising and dimensionality reduction, respectively. To the best of the authors' knowledge, none of the above-mentioned data-driven state estimation methods for model-free processes was developed for robustness to outlier measurements.

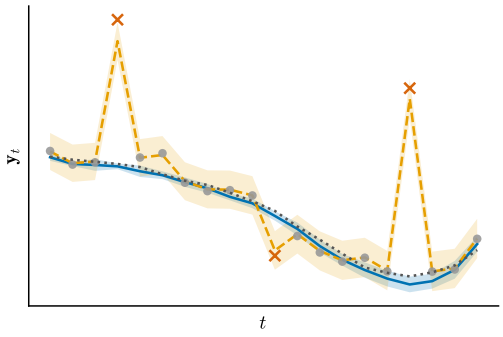


Figure 1: The black dotted line is a state trajectory, the grey dots are valid noisy linear measurements, and the red crosses are outliers. The orange dashed line is the non-robust DNS estimate (Cumlin et al., 2026); the blue solid line is the robust RoVaS estimate.

Designing an outlier-robust state estimation method for a model-free process is challenging because outlier detection and dynamics learning are coupled: identifying which measurements are corrupted requires a model of the underlying dynamics, while learning the dynamics from data requires deciding which measurements are trustworthy. STM-based outlier-robust filters such as the WoLF sidestep this coupling by assuming that the STM is known. In the model-free setting, no such assumption is available: WoLF-like weighting strategies cannot be used directly, and naive joint optimization of a variational objective for the state and an outlier indicator collapses to a degenerate solution in which all measurements are explained as outliers.

In this article, we develop robust variational state estimation (RoVaS), an outlier-robust state estimation method for model-free processes, designed for measurements that are randomly replaced by state-independent corruptions of arbitrary magnitude. In RoVaS, we introduce a latent corruption indicator variable $c_t \in \{0, 1\}$ that represents whether the measurement at time t is corrupted. The state $\mathbf{x}_t$ and the corruption indicator c_t are jointly treated as latent variables, and they are inferred under a variational inference framework for sequential latent-variable models. RoVaS uses a corruption gate that, by construction, saturates as the magnitude of the measurement grows, and it feeds the state posterior with correspondingly attenuated measurements; under mild regularity conditions on the encoder, this yields a provable outlier-robustness guarantee at inference, in the form of a bounded posterior influence function. To address the coupling between outlier detection and dynamics learning during training, RoVaS is optimized under a three-stage curriculum that learns an initial dynamics model before activating the corruption gate. Empirically, RoVaS outperforms non-robust data-driven baselines for model-free processes and is competitive with the STM-informed WoLF-based EKF, despite RoVaS having no access to the STM. A visual illustration of the robustness of RoVaS to outliers is shown in Figure 1.

2 OUTLIER-ROBUST SMOOTHING FOR MODEL-FREE PROCESS

2.1 PROBLEM SETUP

Let $\mathbf{x}_t \in \mathbb{R}^{d_x}$ be the state vector of a model-free dynamical process, where t denotes a discrete time index. A state trajectory of length T is denoted by $\mathbf{x}_{1:T} \triangleq \mathbf{x}_1, \ldots, \mathbf{x}_T$. The dynamical process $p(\mathbf{x}_{1:T})$ is assumed to be unknown, and can also be non-Markov. Following Ghosh et al. (2024); Cumlin et al. (2026), we call such a process model-free: no STM is available a priori. A model-free method may still learn a model of the dynamics from data; RoVaS does so through a learned dynamics prior (Section 2.2). At each time point t, we observe a noisy measurement $\mathbf{y}_t \in \mathbb{R}^{d_y}$. For the state trajectory $\mathbf{x}_{1:T}$, the corresponding measurement sequence of length T is denoted as $\mathbf{y}_{1:T} \triangleq \mathbf{y}_1, \ldots, \mathbf{y}_T$. Given the measurement sequence $\mathbf{y}_{1:T}$, the state estimation task is to compute the posterior $p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$.

In our case, some measurements over time are randomly corrupted as outliers. The valid (noisy) measurements follow a known measurement model, $\mathbf{y}_t|\mathbf{x}_t \sim p_{\text{noisy}}(\mathbf{y}_t|\mathbf{x}_t)$, whereas the corrupted measurements (outliers) follow an unknown corruption distribution, $\mathbf{y}_t|\mathbf{x}_t \sim p_{\text{corr}}(\mathbf{y}_t)$. The corrupted measurements are statistically independent of the underlying states; that is, we consider outliers that carry no information about the state.

We need a data-driven method for state estimation of the model-free process. We assume access to a training dataset $\mathcal{D} = \{\mathbf{y}_{1:T}^{(i)}\}_{i=1}^{N}$ of N measurement sequences of length T, where $\mathbf{y}_{1:T}^{(i)}$ corresponds to the (unobserved) state sequence $\mathbf{x}_{1:T}^{(i)}$. We develop the variational objective for a single sequence; extension to the full dataset $\mathcal{D}$ is immediate.

2.2 ROBUST VARIATIONAL STATE ESTIMATION (ROVAS)

To model corrupted measurements (outliers), we introduce a latent corruption indicator $c_t \in \{0, 1\}$, where $c_t = 0$ relates to a valid measurement sampled from $p_{\text{noisy}}(\mathbf{y}_t|\mathbf{x}_t)$ and $c_t = 1$ relates to a corrupted measurement (outlier). Conditioned on $(\mathbf{x}_t, c_t)$, the measurement model is

$$\mathbf{y}_t|\mathbf{x}_t, c_t \sim \begin{cases} p_{\text{noisy}}(\mathbf{y}_t|\mathbf{x}_t), & c_t = 0, \\ p_{\text{corr}}(\mathbf{y}_t), & c_t = 1. \end{cases} \tag{1}$$

The corruption indicator sequence is denoted by $c_{1:T} \triangleq c_1, c_2, \ldots, c_t, \ldots, c_T$; $c_{1:T}$ may have an arbitrary distribution $p(c_{1:T})$. The resulting joint distribution over states, corruption indicators, and measurements, factors as

$$p(\mathbf{x}_{1:T}, c_{1:T}, \mathbf{y}_{1:T}) = p(\mathbf{x}_{1:T})\, p(c_{1:T}) \prod_{t=1}^{T} p_{\text{noisy}}(\mathbf{y}_t|\mathbf{x}_t)^{1-c_t} p_{\text{corr}}(\mathbf{y}_t)^{c_t}. \tag{2}$$

2.2.1 VARIATIONAL OBJECTIVE

We seek a tractable lower bound to the evidence log-likelihood $\log p(\mathbf{y}_{1:T})$ that accounts for the latent corruption indicators $c_{1:T}$. The main challenge is to choose the modeling assumptions and the variational distribution such that the lower bound is tractable and the variational distributions can be learned stably.

We propose to use a variational posterior $q_{\boldsymbol{\theta}}$ to approximate the intractable joint posterior $p(\mathbf{x}_{1:T}, c_{1:T}|\mathbf{y}_{1:T})$. For notational simplicity, we define $\mathbf{y}_{1:T\setminus t} \triangleq \mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{y}_{t+1}, \ldots, \mathbf{y}_T$ and $\mathbf{x}_{1:0} = \emptyset$. The variational distribution is parameterized by $\boldsymbol{\theta}$ as follows:

$$q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}|\mathbf{y}_{1:T}) = \prod_{t=1}^{T} q_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{y}_{1:T\setminus t}, c_{1:T}) \prod_{t=1}^{T} q_{\boldsymbol{\theta}}(c_t|\mathbf{y}_{1:T}). \tag{3}$$

We highlight two aspects of the proposed variational distribution. First, the latent state variables are autoregressive, but the t'th factor omits $\mathbf{y}_t$: since $\mathbf{y}_t$ can be corrupted as an outlier, it is not used for the estimate of $\mathbf{x}_t$. Second, a mean-field factorization is used for the corruption indicators $c_{1:T}$; that is, they are treated as statistically independent over time. We model $q_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{y}_{1:T\setminus t}, c_{1:T})$ as a Gaussian distribution and $q_{\boldsymbol{\theta}}(c_t|\mathbf{y}_{1:T})$ as a Bernoulli distribution.

Since the state dynamics of the process is unknown, we model it with parameters $\boldsymbol{\phi}$; that is, the state dynamics is modeled by $p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T}) = \prod_{t=1}^{T} p_{\boldsymbol{\phi}}(\mathbf{x}_t|\mathbf{x}_{1:t-1})$. Here, $\boldsymbol{\phi}$ denotes the parameters of a neural network that maps the full state history $\mathbf{x}_{1:t-1}$ to the parameters of $p_{\boldsymbol{\phi}}(\mathbf{x}_t|\mathbf{x}_{1:t-1})$ (Appendix D.7); it is learned jointly with the variational posterior from the measurement data alone. The joint distribution factorizes as:

$$\begin{aligned} p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T}) &= p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, c_{1:T}) p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T}) p(c_{1:T}) \\ &= \prod_{t=1}^{T} p(\mathbf{y}_t|\mathbf{x}_t, c_t) \prod_{t=1}^{T} p_{\boldsymbol{\phi}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}) \prod_{t=1}^{T} p(c_t) \end{aligned} \tag{4}$$

Here, $c_t \sim \text{Bernoulli}(\varepsilon)$, where the prior corruption rate ε is a hyperparameter; it need not match the true corruption rate, which is unknown, and we use $\varepsilon = 0.01$ in all experiments. We assume that the latent states and latent corruption indicators are statistically independent, and the corruption indicator variables are independent and identically distributed (i.i.d) as $p(c)$ (that means $\forall t, p(c_t) = p(c)$). In our work, $p_{\boldsymbol{\phi}}(\mathbf{x}_t|\mathbf{x}_{1:t-1})$ is a Gaussian distribution. Its mean and diagonal covariance are nonlinear functions of the full history, so the learned prior is neither Markov nor jointly Gaussian; other conditional families could be used.

The standard evidence lower bound (ELBO) is given by:

$$\log p(\mathbf{y}_{1:T}) \geq \mathbb{E}_{q_{\boldsymbol{\theta}}}[\log p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T}) - \log q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})]\,. \tag{5}$$

Substituting (3) and (4) in (5) and grouping corresponding variational and prior terms, we have

$$\begin{aligned} &\mathcal{L}_{\text{ELBO}}^{\text{RoVaS}}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})}\big[\log p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, c_{1:T}) + \log p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T}, c_{1:T}) - \log q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})\big] \\ &= \mathbb{E}_{q_{\boldsymbol{\theta}}}[\log p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, c_{1:T})] - \mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\log \tfrac{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|c_{1:T})}{p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T})}\right] - \mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\log \tfrac{q_{\boldsymbol{\theta}}(c_{1:T})}{p(c_{1:T})}\right] \\ &= \underbrace{\mathbb{E}_{q_{\boldsymbol{\theta}}}[\log p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, c_{1:T})]}_{\mathcal{L}_{\text{fit}}(\boldsymbol{\theta})} - \underbrace{\mathbb{E}_{q_{\boldsymbol{\theta}}(c_{1:T})}[\text{KL}(q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|c_{1:T})\|p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T}))]}_{\mathcal{L}_{\text{dyn}}(\boldsymbol{\theta}, \boldsymbol{\phi})} - \underbrace{\text{KL}[q_{\boldsymbol{\theta}}(c_{1:T})\|p(c_{1:T})]}_{\mathcal{L}_{\text{corr}}(\boldsymbol{\theta})}. \end{aligned} \tag{6}$$

To avoid notational clutter, we write $q_{\boldsymbol{\theta}}(c_{1:T}) \triangleq q_{\boldsymbol{\theta}}(c_{1:T}|\mathbf{y}_{1:T})$, $q_{\boldsymbol{\theta}} \triangleq q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \triangleq q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}|\mathbf{y}_{1:T})$ and $q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|c_{1:T}) \triangleq q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|c_{1:T}, \mathbf{y}_{1:T})$. $\text{KL}(q\|p)$ denotes Kullback–Leibler divergence (KL-divergence) between q and p distributions. The bound decomposes into three interpretable terms. $\mathcal{L}_{\text{fit}}(\boldsymbol{\theta})$ is the expected data-likelihood fit that can be expressed as $\mathcal{L}_{\text{fit}}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \mathbb{E}_{q_{\boldsymbol{\theta}}}[\log p(\mathbf{y}_t|\mathbf{x}_t, c_t)]$. When $c_t = 0$ (valid measurement), the term $\log p_{\text{noisy}}(\mathbf{y}_t|\mathbf{x}_t)$ rewards state estimates that explain $\mathbf{y}_t$; when $c_t = 1$, the term $\log p_{\text{corr}}(\mathbf{y}_t)$ does not depend on $\mathbf{x}_t$, so a measurement flagged as corrupted does not constrain the state. $\mathcal{L}_{\text{dyn}}$ controls the dynamics, and encourages the variational distribution to estimate temporally coherent state distributions: since the dynamics prior $p_{\boldsymbol{\phi}}$ is shared across all training sequences, state trajectories that follow the measurement noise are poorly predicted from their own past and are therefore penalized. Lastly, $\mathcal{L}_{\text{corr}}$ regularizes the corruption variational distributions towards the prior $p(c)$, preventing the extreme situation that all the measurements are treated as corrupted. The derivation of the variational objective (6) is shown in Appendix A.

The variational objective of the RoVaS method is to maximize the ELBO

$$\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \mathcal{L}_{\text{ELBO}}^{\text{RoVaS}}(\boldsymbol{\theta}, \boldsymbol{\phi}) \implies \max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \{\mathcal{L}_{\text{fit}}(\boldsymbol{\theta}) - \mathcal{L}_{\text{dyn}}(\boldsymbol{\theta}, \boldsymbol{\phi}) - \mathcal{L}_{\text{corr}}(\boldsymbol{\theta})\}\,. \tag{7}$$

2.2.2 Training and Architectural Design

The variational objective (6) is optimized using Monte Carlo and the reparameterization trick (Kingma & Welling, 2014); the complete Monte Carlo estimator is given in Appendix A.

Corruption posterior: We model the corruption posterior as $q_{\boldsymbol{\theta}}(c_t|\mathbf{y}_{1:T}) = \text{Bernoulli}(\gamma_t)$ with

$$\gamma_t = \sigma\Big((g_{\boldsymbol{\theta}}(\mathbf{y}_{1:T}))_t + \lambda \, \|\mathbf{y}_t - \boldsymbol{\mu}_{\text{pool}}\|^2_{\boldsymbol{\Sigma}_{\text{pool}}^{-1}}\Big), \tag{8}$$

where $g_{\boldsymbol{\theta}} : \mathbb{R}^{d_y \times T} \to \mathbb{R}^T$ is a neural network, $\boldsymbol{\mu}_{\text{pool}}$ and $\boldsymbol{\Sigma}_{\text{pool}}$ are the empirical mean and covariance of the training measurements (the latter assumed to be positive definite), $\|\mathbf{v}\|^2_{\mathbf{M}} \triangleq \mathbf{v}^\top \mathbf{M}\, \mathbf{v}$ denotes the squared Mahalanobis norm, $\lambda > 0$ is a fixed scalar, and $\sigma : a \to 1/(1 + \exp(-a))$ is the sigmoid

map. Throughout, $\|\cdot\|$ denotes the ℓ_2-norm. The network $g_{\boldsymbol{\theta}}$ learns to detect corruptions from temporal context, while the Mahalanobis term acts as a scale-invariant base corruption score that grows quadratically with the magnitude of the measurement. This additive structure ensures that the gate logit satisfies the coercivity condition (Assumption 3) whenever $g_{\boldsymbol{\theta}}$ grows sub-quadratically in $\|\mathbf{y}_t\|$, for example when $g_{\boldsymbol{\theta}}$ is bounded, as for the GRU-based gate used in this work; together with a mild regularity condition on the state posterior, this guarantees outlier robustness at inference (Proposition 1). In practice, λ can be chosen small enough to have a negligible effect on γ_t for measurements of typical magnitude, while still providing the theoretical guarantee, which holds for every $\lambda > 0$.

State posterior: Since the Bernoulli distribution is not reparameterisable, we design the state posterior to condition on $\gamma_{1:T}$ directly rather than on samples of $c_{1:T}$. This avoids discrete sampling (e.g. Gumbel-Softmax (Jang et al., 2017)) while retaining the relevant information on corruption. The KL term $\mathcal{L}_{\text{corr}}$ admits a closed-form expression between two Bernoulli distributions. We model the state posterior as a Gaussian with diagonal covariance. Following the principle of robust filtering (Duran-Martin et al., 2024, Eq. (15)), the encoder receives gated measurements rather than raw measurements:

$$\mathbf{y}_t^{(\gamma)} = (1 - \gamma_t)\, \mathbf{y}_t, \tag{9}$$

so that measurements suspected to be corrupted ($\gamma_t \approx 1$) are attenuated before entering the proposal. The state posterior is then

$$q_{\boldsymbol{\theta}}\left(\mathbf{x}_t | \mathbf{x}_{1:t-1}, \mathbf{y}_{1:T\setminus t}, \gamma_{1:T}\right) = \mathcal{N}\Big(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(t)}\left(\mathbf{x}_{1:t-1}, \mathbf{y}_{1:T\setminus t}^{(\gamma)}, \gamma_{1:T}\right),\ \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t)}\left(\mathbf{x}_{1:t-1}, \mathbf{y}_{1:T\setminus t}^{(\gamma)}, \gamma_{1:T}\right)\Big). \tag{10}$$

Combined with the coercive gate, the gating mechanism keeps the encoder input in a bounded set, uniformly over all values of the measurement $\mathbf{y}_t$; in particular, $\mathbf{y}_t^{(\gamma)} \to \mathbf{0}$ as $\|\mathbf{y}_t\| \to \infty$ (Lemma 2). This ensures that the robustness guarantee of Proposition 1 holds for any encoder satisfying Assumption 4, without requiring the encoder to be bounded in its measurement input.

2.2.3 Amortized Inference

Amortized inference approximates the true posterior with the variational posterior as follows:

$$p(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}) \approx q(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}) = q(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}, \gamma_{1:T}) = \prod_{t=1}^{T} q_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{y}_{1:T\setminus t}, \gamma_{1:T}), \tag{11}$$

where $\gamma_{1:T}$ is computed from $\mathbf{y}_{1:T}$ with Eq. (8).

3 Theoretical Properties

To study the robustness of amortized inference, we use a variant of the classical influence function of Hampel (1974) (see also Huber, 1981; Hampel et al., 1986). The influence function was originally developed for statistical functionals, but has been generalized to Bayesian setups (Hooker & Vidyashankar, 2014; Ghosh & Basu, 2016; Nakagawa & Hashimoto, 2020; Matsubara et al., 2022), and, more recently, to Bayesian filtering (Altamirano et al., 2024; Duran-Martin et al., 2024). In Bayesian filtering, the posterior influence function (PIF) is operationalized as the KL-divergence between two posterior distributions, as follows.

Consider the amortized smoothing posterior produced by RoVaS (Eq. (11)): $q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}) = \prod_{t=1}^{T} q_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{y}_{1:T\setminus t}, \gamma_{1:T})$, where each factor is a Gaussian with mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(t)}$ and diagonal covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t)}$. Suppose the measurement $\mathbf{y}_t$ is replaced by an arbitrary corruption $\mathbf{y}_t^c \in \mathbb{R}^{d_y}$, with all other measurements held fixed. Applying RoVaS to the corrupted measurement sequence $\mathbf{y}_{1:T}^c \triangleq \mathbf{y}_1, \ldots, \mathbf{y}_{t-1}, \mathbf{y}_t^c, \mathbf{y}_{t+1}, \ldots, \mathbf{y}_T$ yields a perturbed posterior $q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}^c)$.

Definition 1 (Posterior Influence Function)**.** *The posterior influence function (PIF) is*

$$\text{PIF}_{\text{KL}}(\mathbf{y}_t^c) = \text{KL}\Big[q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}^c)\,\Big\|\,q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})\Big]. \tag{12}$$

The estimator is outlier-robust in the KL sense if, for every $t \in \{1, \ldots, T\}$ and every measurement sequence $\mathbf{y}_{1:T}$, $\sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \text{PIF}_{\text{KL}}(\mathbf{y}_t^c) < \infty$.

We establish the outlier robustness of the amortized inference of RoVaS under two structural conditions, which are sufficient but not necessary; both are stated formally in Appendix B. First, the corruption gate is coercive (Assumption 3); that is, the gate logit grows at least as fast as $\alpha\|\mathbf{y}_t\|^\beta$ for some $\alpha, \beta > 0$, which ensures $\gamma_t \to 1$ as the measurement becomes extreme. The Mahalanobis term in the gate (Eq. (8)) ensures this with $\beta = 2$ for any $\lambda > 0$, provided that $g_{\boldsymbol{\theta}}$ grows sub-quadratically in $\|\mathbf{y}_t\|$, which holds for GRUs, LSTMs, MLPs, and standard Transformers; hence, coercivity is imposed by construction rather than learned. Second, the encoder is regular (Assumption 4); that is, the measurements reach the state posterior only through the gated pair $(\mathbf{y}^{(\gamma)}_{1:T}, \gamma_{1:T})$, the mean of each state factor grows at most polynomially along the sampled state history, and the diagonal variances are bounded above and below. Polynomial growth holds with a linear rate for any network built from affine layers and Lipschitz activations, and trivially for a saturating state pathway such as the GRU decoder used here; the variance bounds hold whenever the variance head is a continuous function of a bounded representation, as for the GRU decoder with the exponential parameterisation used in this work, and can otherwise be imposed by clamping the log-variance head. Thus, encoder regularity is satisfied by most implementations, and both conditions hold for RoVaS as implemented (Corollary 2 in Appendix B).

Proposition 1. *For fixed trained parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$ of the RoVaS method, and under Assumptions 3 and 4, the amortized inference is outlier-robust in the KL sense; that is, for every $t \in \{1, \ldots, T\}$ and every measurement sequence $\mathbf{y}_{1:T}$,*

$$\sup_{\mathbf{y}^c_t \in \mathbb{R}^{d_y}} \mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}^c_t) < \infty. \tag{13}$$

The proof of Proposition 1 is given in Appendix B. The proof bounds each per-step KL divergence in expectation over the sampled state history, using moment bounds on the sampled trajectories that hold uniformly over corruptions (Lemma 1). We also show that, without the gated input, the guarantee fails for encoders whose mean is unbounded in the measurements (Proposition 3). Since the amortized inference involves only $q_{\boldsymbol{\theta}}$, the guarantee does not depend on the form of the learned dynamics prior $p_{\boldsymbol{\phi}}$. An empirical study is provided in Appendix C.

4 Experiments

In this section, we evaluate the robustness of RoVaS using synthetically generated data from a stochastic Lorenz-63 system, with a linear and a nonlinear camera measurement setup, and real human motion data with a linear measurement setup. The stochastic Lorenz-63 system is a well-studied benchmark for state estimation (van Nierop et al., 2025; Ghosh et al., 2024; 2025; Cumlin et al., 2026; Norén et al., 2026; Ruhe & Forré, 2022; Ni et al., 2022; Revach et al., 2022b;a); its state $\mathbf{x}_t = [x_{t,1}\, x_{t,2}\, x_{t,3}]^\top \in \mathbb{R}^3$ is three-dimensional, and its STM is given in Appendix D.1. Our objectives are threefold: (i) assess state estimation performance across measurement noise levels and corruption rates, (ii) evaluate the gating mechanism as an outlier detector, and (iii) compare against both robust and non-robust baseline state estimation methods. Details on experiments, training, architectures, and dataset preparation are provided in Appendix D[1]. Direct optimization of $\mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ (Eq. (6)) results in a degenerate corruption gate that predicts all measurements as corrupted. Therefore, we employ a three-stage training curriculum, similar to typical KL annealing in DVAE training (see, e.g., Bowman et al., 2015; Sønderby et al., 2016), which is explained and discussed in Appendix E.

4.1 Experimental Setup

We measure the state estimation (smoothing) performance by the normalized mean-squared error (NMSE in dB) at varying signal-to-measurement-noise ratios (SMNR in dB) and corruption levels, following Ghosh et al. (2024); definitions are given in Appendix D.2. A low NMSE is desirable, and a low SMNR means that the measurement noise is strong, which makes a low NMSE hard to achieve. We evaluate the corruption gate by the area under the receiver operating characteristic curve (AUROC), treating the ground-truth corruption indicator as the binary label and the gate output γ_t as the score.

[1] Code will be made available upon acceptance.

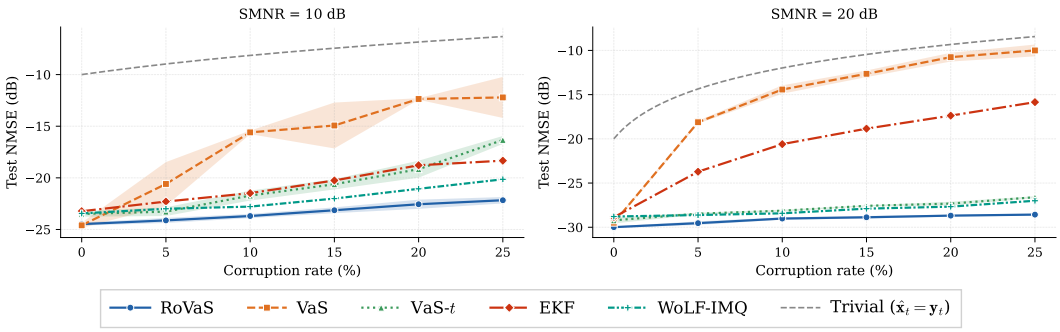


Figure 2: NMSE (dB) versus corruption rate p_ϵ for the stochastic Lorenz-63 system with linear measurements at SMNR=10 and 20 dB. Shaded regions indicate ± 1 standard deviation over 5 experimental trials. Here, the trivial estimator uses the estimate as $\hat{\mathbf{x}}_t = \mathbf{y}_t$.

RoVaS is compared against four methods: VaS (the architecture of RoVaS without the gating mechanism), VaS-t (VaS with a heavy-tailed Student-t measurement likelihood instead of the Gaussian one), the extended Kalman filter (EKF), and WoLF-IMQ (Duran-Martin et al., 2024), a provably robust WoLF-based EKF that uses generalized Bayes with inverse multi-quadratic (IMQ) weighting. EKF and WoLF-IMQ are STM-based methods, and VaS and EKF serve as non-robust baselines for the robust methods WoLF-IMQ, VaS-t, and RoVaS. Further, we also evaluated three data-driven model-free methods, DANSE (Ghosh et al., 2024), DNS (Cumlin et al., 2026), and VSE (Norén et al., 2026), and found them to be non-robust to measurement corruption (Appendix D.3). RoVaS is given neither the true corruption rate p_ϵ nor the corruption-generating distribution. In all experiments, RoVaS uses the fixed prior corruption rate $\varepsilon = 0.01$ and the surrogate p_{corr} of Appendix D.10; Appendix F.2 shows that the results are relatively insensitive to both choices.

4.2 Experiments for Stochastic Lorenz-63 System

4.2.1 Linear Measurement Setup

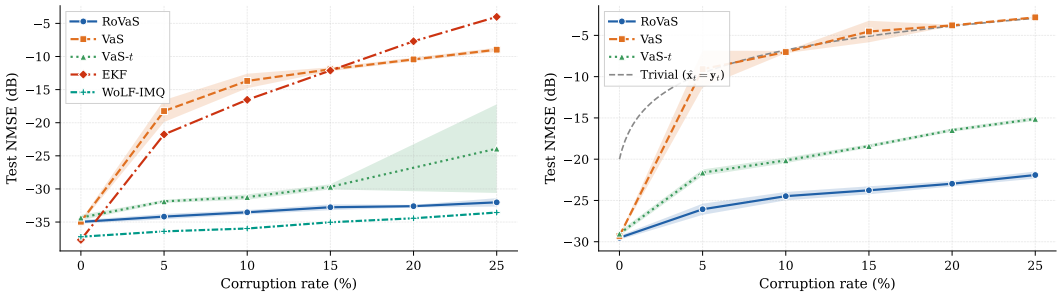
We use a linear measurement model $\mathbf{y}_t = \mathbf{x}_t + \mathbf{w}_t$ where $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I})$ and the choice of σ_w^2 determines an SMNR level. We use $N = 1000$ training sequences of length $T = 100$, and compute the NMSE on 100 test sequences of the same length, averaged over 5 runs(/seeds). At each time step, measurements are independently corrupted with probability p_ϵ, which varies from 0% to 25%. The corrupted measurements (outliers) are replaced by samples drawn, independently of the state, from a fixed corruption distribution: a Gaussian fitted to the pooled measurements (Appendix D.12).

Figure 5 in Appendix D.4 shows a state trajectory, corrupted measurements, and the estimated trajectories of all methods: WoLF-IMQ provides robustness to EKF, both of which use the STM of the stochastic Lorenz-63 system, and among the data-driven methods unaware of the STM, VaS is non-robust, whereas VaS-t and RoVaS are robust. Figure 2 shows the NMSE versus corruption rate for SMNR = 10 and 20 dB. At 0% corruption (no outliers), all competing methods perform comparably. RoVaS shows the best graceful degradation in NMSE performance as the corruption rate increases. WoLF-IMQ is comparable to VaS-t at 20 dB SMNR and better at 10 dB SMNR. EKF and VaS are non-robust. For example, at 20 dB SMNR and 20–25% corruption rate, RoVaS improves the NMSE by around 13–14 dB, 17–18 dB and 3 dB over EKF, VaS and VaS-t, respectively. Figure 2 also includes the trivial estimator $\hat{\mathbf{x}}_t = \mathbf{y}_t$; derivation of the NMSE of the trivial estimator with respect to SMNR and corruption rate is given in Appendix G. Overall, RoVaS combines the most graceful degradation in NMSE with a provable bound on the PIF (Proposition 1), whereas the robustness of VaS-t is empirical only. An empirical study of the PIF is given in Appendix C.

Table 1 reports the AUROC of the gate as a binary outlier detector, which measures how well γ_t ranks corrupted measurements above valid ones. At 20 dB SMNR, the gate is a near-perfect detector across all corruption rates, with AUROC at or above 0.97; at 10 dB SMNR, where the measurement noise is strong, it retains AUROC above 0.91. The detection performance is largely insensitive to the corruption rate p_ϵ, which is consistent with the gate's training signal: the corruption-KL term $\mathcal{L}_{\text{corr}}$ regularises γ_t towards the prior ε at clean indices and is overridden at corrupted indices by the

Table 1: AUROC of the corruption parameter γ_t with respect to the ground truth corruption indicators on the Lorenz-63 system (mean $\pm$ std over 5 seeds).

	Corruption rate p_ϵ				
SMNR	5%	10%	15%	20%	25%
10 dB	$0.92_{\pm 0.01}$	$0.93_{\pm 0.00}$	$0.92_{\pm 0.02}$	$0.92_{\pm 0.00}$	$0.91_{\pm 0.02}$
20 dB	$0.97_{\pm 0.01}$	$0.98_{\pm 0.02}$	$0.99_{\pm 0.01}$	$0.99_{\pm 0.00}$	$0.99_{\pm 0.01}$



(a) Lorenz-63, 2D camera model.

(b) BMLmovi dataset.

Figure 3: NMSE versus corruption rate p_ϵ at 20 dB SMNR for (a) the stochastic Lorenz-63 system with the 2D camera model and (b) the BMLmovi dataset. Shaded regions indicate ± 1 standard deviations over 5 experimental trials(/seeds).

data-fit term $\mathcal{L}_{\text{fit}}$, so the gate's discrimination depends on the SMNR rather than on the prevalence of corruptions in the sequence.

4.2.2 Nonlinear Measurement Setup

For the nonlinear measurement setup, we use a 2D camera model with Gaussian point spread functions over an 8×8 pixel grid, following Buchnik et al. (2024); van Nierop et al. (2025); Norén et al. (2026). The grid points $\{g_i\}_{i=1}^{64}$ are uniformly distributed in the rectangle $[-30, 30] \times [-40, 40]$. The state is projected into the camera plane according to its first two coordinates, with the third coordinate representing depth. The measurement function $\mathbf{h} : \mathbb{R}^3 \to \mathbb{R}^{64}$ is given by

$$\mathbf{h}(\mathbf{x}_t)_i = 10 \exp\left(-\frac{2}{7 + x_{t,3}} \left\| g_i - \begin{bmatrix} x_{t,1} \\ x_{t,2} \end{bmatrix} \right\|_2^2 \right), \tag{14}$$

where $\mathbf{h}(\mathbf{x}_t)_i$ represents the intensity at pixel i. The measurements are given by $\mathbf{y}_t = \mathbf{h}(\mathbf{x}_t) + \mathbf{w}_t$ with $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I})$, where the choice of σ_w^2 determines the SMNR. The dataset sizes are as in the linear setup. Figure 3a reports the NMSE versus corruption rate at 20 dB SMNR. The model-based WoLF-IMQ performs the best, and RoVaS performs the best among the model-free methods, whereas EKF and VaS quickly degrade as the corruption rate increases. A visual illustration is given in Figure 6 (Appendix D.5).

4.3 Experiments for Real Data

We further evaluate RoVaS on real human motion data using the BMLmovi dataset (Ghorbani et al., 2021), distributed as part of the AMASS motion capture corpus (Mahmood et al., 2019). BMLmovi dataset contains recordings of 89 actors performing a fixed set of 20 everyday actions and sports movements (walking, jogging, jumping, kicking, sitting, clapping, throwing, etc.), captured by an optical motion capture system. From the raw recordings, the dataset provides the joint angles of 21 body joints (shoulders, elbows, knees, hips, spine, etc.) over time. Each joint has 3 rotational degrees of freedom, so the state vector at time t is $\mathbf{x}_t \in \mathbb{R}^{63}$. The recordings are sampled at 120 Hz, and we extract sequences of length $T = 100$, corresponding to approximately 0.83 s of motion per sequence. We use subjects 1–79 for training and validation, and subjects 80–89 for testing. We use $N = 1000$ training sequences and 100 testing sequences. The STM of human motion is unknown and highly

nonlinear, which makes the BMLmovi dataset a realistic test case for model-free state estimation, where STM-based methods cannot be readily used.

For our experiments, we use the linear measurement model $\mathbf{y}_t = \mathbf{x}_t + \mathbf{w}_t$, where $\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I})$ and SMNR is $20\,\text{dB}$. The corruption rate, represented by p_ϵ, varies between the levels $0 - 25\%$. As for Lorenz-63, corruptions are drawn from a Gaussian fitted to the pooled measurements. In RoVaS, we use the surrogate $p_{\text{corr}} = \mathcal{N}(\boldsymbol{\mu}_{\text{pool}}, 5\boldsymbol{\sigma}_w^2 \mathbf{I})$, where $\boldsymbol{\mu}_{\text{pool}}$ is the mean of the measurements, instead of the pooled Gaussian, since the latter led to the degenerate all-outlier solution for some seeds (Appendix D.10). Figure 3b reports the NMSE as a function of the corruption rate, averaged over 5 experimental trials. RoVaS provides the best performance, whereas VaS quickly deteriorates and behaves like a trivial estimator at high corruption rates. At the 25% corruption rate, the NMSE gap between RoVaS and VaS-t is greater than 6 dB, which shows the importance of the corruption gate of RoVaS compared to heavy-tailed likelihood modeling with a Student-t distribution. A visual illustration of the BMLmovi dataset is given in Figure 7 (Appendix D.6).

4.4 Ablation Study

We ablate the three core design choices of RoVaS on the linear Lorenz-63 setup at SMNR $= 20$ dB and $p_\varepsilon = 10\%$ using 5 experimental trials. Removing the $\mathbf{y}_t$-exclusion from the t-th state factor (Eq. (3)) degrades NMSE by 15.6 dB ($-29.0 \to -13.4$ dB) and collapses the gate to $\mathbb{E}[\gamma_t] = 0$: with direct access to $\mathbf{y}_t$, the encoder fits each measurement independently and the gate has no incentive to fire. Setting $\lambda = 0$ or removing the gated input $\mathbf{y}_t^{(\gamma)}$ leaves NMSE unchanged, since the corruptions in this benchmark are drawn from a Gaussian fitted to the measurements and are thus of typical magnitude, whereas both mechanisms target the regime $\|\mathbf{y}_t\| \to \infty$. Both are nevertheless part of the sufficient conditions of Proposition 1: $\lambda > 0$ makes the gate coercive, and the gated input makes the guarantee independent of the encoder architecture. Full results, together with sensitivity analyses on the prior corruption rate ε, the corruption loss weight β_0, and the corruption-likelihood scale k, are reported in Appendix F; NMSE varies by at most 1.3 dB across order-of-magnitude sweeps of any of them.

5 Conclusion

We conclude that it is possible to design outlier-robust data-driven methods for the state estimation of model-free processes. Of the two robust methods developed in this article, VaS-t and RoVaS, only RoVaS has a provable robustness guarantee: its posterior influence function is bounded by construction of the coercive gate and the gated input, for any encoder satisfying a mild regularity condition. The success of RoVaS is due to the indicator variables that model the occurrence of measurement outliers, the variational objective, the gating mechanism, and the three-stage training curriculum. The main limitation of the developed methods is the necessity of knowing the measurement model for valid measurements; if it is not known, the methodology could be extended to outlier-robust DVAEs for sequence modeling and representation learning by introducing neural network-based decoders. RoVaS also assumes that corruptions are independent of the state, and our experiments use randomly replaced measurements; heavy-tailed noise and partial corruptions, which remain informative about the state, are not evaluated, and a heavy-tailed likelihood as in VaS-t may be more adequate for them. Finally, our guarantee establishes that the posterior influence function is bounded, but does not quantify the bound, whose size is only assessed empirically.

AI use statement

In this work, we used generative AI tools for editing (grammar, spelling, and word choice); for improving the clarity and consistency of the text, including the presentation of the proofs in Appendix B; for facilitating and running experiments; feedback on research methodology and experiments; and for analysis of the results of PIF under bounded corruption . We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, implement methods, assist with translation, or clean and reformat data sets. We have reviewed all AI-assisted work: the authors verified all derivations and proofs, as well as the experimental code and results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

The generative model, the variational distribution, and the training objective are specified in Section 2.2, with the full derivation and the Monte Carlo estimator in Appendix A. All assumptions and complete proofs of the theoretical results are given in Appendix B. The architectures, hyperparameters, optimisation settings, the specification of p_{corr}, and the data generation and corruption procedures are described in Appendix D, and the training curriculum in Appendix E. Code and configuration files will be released upon acceptance.

References

Matias Altamirano, François-Xavier Briol, and Jeremias Knoblauch. Robust and conjugate gaussian process regression. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, 2002. doi: 10.1109/78.978374.

Philipp Becker, Harit Pandya, Gregor H. W. Gebhardt, Cheng Zhao, C. James Taylor, and Gerhard Neumann. Recurrent kalman networks: Factorized inference in high-dimensional deep feature spaces. In *ICML*, pp. 544–552, 2019. URL http://proceedings.mlr.press/v97/becker19a.html.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Conference on Computational Natural Language Learning*, 2015. URL https://api.semanticscholar.org/CorpusID:748227.

Itay Buchnik, Guy Revach, Damiano Steger, Ruud J. G. van Sloun, Tirza Routtenberg, and Nir Shlezinger. Latent-kalmannet: Learned kalman filtering for tracking from high-dimensional signals. *Trans. Sig. Proc.*, 72:352–367, January 2024. ISSN 1053-587X. doi: 10.1109/TSP.2023.3344360. URL https://doi.org/10.1109/TSP.2023.3344360.

Guobin Chang. Robust Kalman filtering based on Mahalanobis distance as outlier judging criterion. *Journal of Geodesy*, 88:391–401, 2014.

Adrien Corenflos, James Thornton, George Deligiannidis, and Arnaud Doucet. Differentiable particle filtering via entropy-regularized optimal transport. In *International Conference on Machine Learning*, pp. 2100–2111. PMLR, 2021.

Fredrik Cumlin, Anubhab Ghosh, and Saikat Chatterjee. Dns: Data-driven nonlinear smoother for complex model-free process. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 356–360, 2026. doi: 10.1109/ICASSP55912.2026.11462991.

Gerardo Duran-Martin, Matias Altamirano, Alex Shestopaloff, Leandro Sánchez-Betancourt, Jeremias Knoblauch, Matt Jones, Francois-Xavier Briol, and Kevin Patrick Murphy. Outlier-robust kalman filtering through generalised Bayes. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12138–12171. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/duran-martin24a.html.

James Durbin and Siem Jan Koopman. *Time Series Analysis by State Space Methods*. Number 9780199641178 in OUP Catalogue. Oxford University Press, 2 edition, None 2012. doi: None. URL https://ideas.repec.org/b/oxp/obooks/9780199641178.html.

Geir Evensen. The ensemble kalman filter: Theoretical formulation and practical implementation. *Ocean dynamics*, 53(4):343–367, 2003.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3604–3613, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Mital A. Gandhi and Lamine Mili. Robust kalman filter based on a generalized maximum-likelihood-type estimator. *IEEE Transactions on Signal Processing*, 58(5):2509–2520, 2010. doi: 10.1109/TSP.2009.2039731.

Arthur Gelb and The Analytic Sciences Corporation. *Applied Optimal Estimation*. The MIT Press, 1974. ISBN 0262570483.

Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F. Troje. Movi: A large multi-purpose human motion and video dataset. *PLOS ONE*, 16(6):1–15, 06 2021. doi: 10.1371/journal.pone.0253157. URL https://doi.org/10.1371/journal.pone.0253157.

Abhik Ghosh and Ayanendranath Basu. Robust bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68(2):413–437, 2016.

Anubhab Ghosh, Antoine Honoré, and Saikat Chatterjee. DANSE: Data-Driven Non-Linear State Estimation of Model-Free Process in Unsupervised Learning Setup. *IEEE Transactions on Signal Processing*, 72:1824–1838, 2024. doi: 10.1109/TSP.2024.3383277.

Anubhab Ghosh, Yonina C. Eldar, and Saikat Chatterjee. Particle-based data-driven nonlinear state estimation of model-free process from nonlinear measurements. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888810.

Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *Found. Trends Mach. Learn.*, 15(1–2):1–175, December 2021. ISSN 1935-8237. doi: 10.1561/2200000089. URL https://doi.org/10.1561/2200000089.

Frank Hampel, Elvezio Ronchetti, Peter Rousseeuw, and Werner Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 03 1986. ISBN 9780471735779. doi: 10.1002/9781118186435.

Frank R. Hampel. The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393, 1974.

Hamidreza Hashempoor and Wan Choi. Gated inference network: Inference and learning state-space models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=z4duW3KzlD.

Giles Hooker and Anand Vidyashankar. Bayesian model robustness via disparities. *TEST*, 23, 3 2014. doi: 10.1007/s11749-014-0360-z.

Yulong Huang, Yonggang Zhang, Ning Li, and Jonathon A. Chambers. Robust student's t based nonlinear filter and smoother. *IEEE Transactions on Aerospace and Electronic Systems*, 52:2586–2596, 2016. URL https://api.semanticscholar.org/CorpusID:32591422.

Peter J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley, 1981.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

S.J. Julier and J.K. Uhlmann. Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92(3):401–422, 2004. doi: 10.1109/JPROC.2003.823141.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

Maximilian Karl, Maximilian Soelch, Justin Bayer, and Patrick van der Smagt. Deep variational bayes filters: Unsupervised learning of state space models from raw data. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyTqHL5xg.

Christopher D. Karlgaard and Hanspeter Schaub. Huber-based divided difference filtering. *Journal of Guidance, Control, and Dynamics*, 30(3):885–891, 2007. doi: 10.2514/1.27968. URL https://doi.org/10.2514/1.27968.

Hyunjik Kim, George Papamakarios, and Andriy Mnih. The lipschitz constant of self-attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5562–5571. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/kim21i.html.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6114.

Rahul G. Krishnan, Uri Shalit, and David Sontag. Deep kalman filters, 2015. URL https://arxiv.org/abs/1511.05121.

Rahul G. Krishnan, Uri Shalit, and David A. Sontag. Structured inference networks for nonlinear state space models. In *AAAI Conference on Artificial Intelligence*, 2016. URL https://api.semanticscholar.org/CorpusID:2901305.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Takuo Matsubara, Jeremias Knoblauch, François-Xavier Briol, and Chris J. Oates. Robust generalised bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022, 04 2022. ISSN 1369-7412. doi: 10.1111/rssb.12500. URL https://doi.org/10.1111/rssb.12500.

Tomoyuki Nakagawa and Shintaro Hashimoto. Robust bayesian inference via gamma-divergence. *Communications in Statistics - Theory and Methods*, 49(2):343–360, 2020. doi: 10.1080/03610926.2018.1543765. URL https://doi.org/10.1080/03610926.2018.1543765.

Xiaoyong Ni, Guy Revach, Nir Shlezinger, Ruud J. G. van Sloun, and Yonina C. Eldar. Rtsnet: Deep learning aided kalman smoothing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5902–5906, 2022. doi: 10.1109/ICASSP43922.2022.9746487.

Gustav Norén, Anubhab Ghosh, Fredrik Cumlin, and Saikat Chatterjee. Vse: Variational state estimation of complex model-free process. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4921–4925, 2026. doi: 10.1109/ICASSP55912.2026.11463548.

Robert Piche, Simo Särkkä, and Jouni Hartikainen. Recursive outlier-robust filtering and smoothing for nonlinear systems using the multivariate student-t distribution. In *IEEE International Workshop on Machine Learning for Signal Processing, Santander, Spain, September 23-26, 2012*, 2012.

Guy Revach, Nir Shlezinger, Timur Locher, Xiaoyong Ni, Ruud J. G. van Sloun, and Yonina C. Eldar. Unsupervised learned kalman filtering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1571–1575, 2022a. doi: 10.23919/EUSIPCO55093.2022.9909801.

Guy Revach, Nir Shlezinger, Xiaoyong Ni, Adrià López Escoriza, Ruud J. G. van Sloun, and Yonina C. Eldar. KalmanNet: Neural Network Aided Kalman Filtering for Partially Known Dynamics. *IEEE Transactions on Signal Processing*, 70:1532–1547, 2022b. doi: 10.1109/TSP.2022.3158588.

Michael Roth, Emre Özkan, and Fredrik Gustafsson. A student's t filter for heavy tailed process and measurement noise. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5770–5774, 2013. doi: 10.1109/ICASSP.2013.6638770.

David Ruhe and Patrick Forré. Self-supervised inference in state-space models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=VPjw9KPWRSK.

Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3745–3753, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Simo Särkkä. *Bayesian Filtering and Smoothing*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, 2013.

Wessel L. van Nierop, Nir Shlezinger, and Ruud J.G. van Sloun. Deep variational sequential monte carlo for high-dimensional observations. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889044.

E.A. Wan and R. Van Der Merwe. The unscented kalman filter for nonlinear estimation. In *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, pp. 153–158, 2000. doi: 10.1109/ASSPCC.2000.882463.

A Derivation of the RoVaS Objective

In this section we derive the RoVaS objective $\mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi})$ and provide a bound between the evidence log-likelihood and $\mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi})$; the results follow standard variational arguments, with extension of marginalizing over the latent corruption variables $c_{1:T}$. We finish by showing the Monte Carlo approximation of $\mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi})$.

$$
\begin{aligned}
\log p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}) &= \log \sum_{c_{1:T}} \int p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T})\, d\mathbf{x}_{1:T} \\
&= \log \sum_{c_{1:T}} \int q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \frac{p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T})}{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})}\, d\mathbf{x}_{1:T} \\
&\geq \sum_{c_{1:T}} \int q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \log \frac{p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T})}{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})}\, d\mathbf{x}_{1:T} \\
&= \mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\log \frac{p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, c_{1:T})\, p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T})\, p(c_{1:T})}{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|c_{1:T})\, q_{\boldsymbol{\theta}}(c_{1:T})}\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\theta}}}[\log p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, c_{1:T})] - \mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\log \frac{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|c_{1:T})}{p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T})}\right] - \mathbb{E}_{q_{\boldsymbol{\theta}}}\left[\log \frac{q_{\boldsymbol{\theta}}(c_{1:T})}{p(c_{1:T})}\right] \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{\theta}}}[\log p(\mathbf{y}_{1:T}|\mathbf{x}_{1:T}, c_{1:T})]}_{\mathcal{L}_{\mathrm{fit}}(\boldsymbol{\theta})} - \underbrace{\mathbb{E}_{q_{\boldsymbol{\theta}}(c_{1:T})}[\mathrm{KL}[q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|c_{1:T}) \parallel p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T})]]}_{\mathcal{L}_{\mathrm{dyn}}(\boldsymbol{\theta}, \boldsymbol{\phi})} \\
&\quad - \underbrace{\mathrm{KL}[q_{\boldsymbol{\theta}}(c_{1:T}) \parallel p(c_{1:T})]}_{\mathcal{L}_{\mathrm{corr}}(\boldsymbol{\theta})} \\
&= \mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta}, \boldsymbol{\phi}),
\end{aligned}
\tag{15}
$$

where the inequality follows from Jensen's inequality, noting that $\log$ is a concave function. Further, the gap is

$$\begin{aligned}
\log p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}) - \mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta},\boldsymbol{\phi}) &= \log p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}) - \sum_{c_{1:T}} \int q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \log \frac{p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T})}{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})}\, d\mathbf{x}_{1:T} \\
&= \sum_{c_{1:T}} \int q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \log p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}) d\mathbf{x}_{1:T} \\
&\quad - \sum_{c_{1:T}} \int q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \log \frac{p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T})}{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})}\, d\mathbf{x}_{1:T} \\
&= \sum_{c_{1:T}} \int q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \left[\log p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}) - \log \frac{p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T})}{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})} \right] d\mathbf{x}_{1:T} \\
&= \sum_{c_{1:T}} \int q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \log \frac{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})\, p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T})}{p_{\boldsymbol{\phi}}(\mathbf{y}_{1:T}, \mathbf{x}_{1:T}, c_{1:T})}\, d\mathbf{x}_{1:T} \\
&= \sum_{c_{1:T}} \int q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \log \frac{q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T})}{p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T}, c_{1:T} | \mathbf{y}_{1:T})}\, d\mathbf{x}_{1:T} \\
&= \mathrm{KL}[q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) \parallel p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T}, c_{1:T} | \mathbf{y}_{1:T})] \geq 0,
\end{aligned} \tag{16}$$

with equality if and only if $q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}, c_{1:T}) = p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T}, c_{1:T} | \mathbf{y}_{1:T})$ almost everywhere.

We continue with deriving the Monte Carlo estimate of $\mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta},\boldsymbol{\phi})$. The estimate is computed by using the variational posterior to estimate the parameters of the underlying distributions, as well as simulating state trajectories. Let $\mathbf{y}_{1:T}$ be a measured trajectory. The Monte Carlo estimate of $\mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta},\boldsymbol{\phi})(\mathbf{y}_{1:T})$ is as follows.

Corruption posterior parameter computation: Given a measurement sequence $\mathbf{y}_{1:T}$, we compute the corruption parameters $\gamma_t = q_{\boldsymbol{\theta}}(c_t{=}1 | \mathbf{y}_{1:T})$ for all $t = 1, \ldots, T$ using (8).

State posterior parameter computation and state trajectory estimation: As discussed in Section 2.2.2, the state posterior conditions on $\gamma_{1:T}$ instead of the discrete samples $c_{1:T}$, to avoid the need for discrete reparameterisation. Let L be the number of state trajectories to sample. Then, for each $i = 1, \ldots, L$, we get a state trajectory $\mathbf{x}^{(i)}_{1:T}$ and associated parameters $\{(\boldsymbol{\mu}^{(t,i)}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}^{(t,i)}_{\boldsymbol{\theta}})\}_{t=1}^{T}$ by the following recursive steps. At step $t = 1$ we use (10) to get

$$q_{\boldsymbol{\theta}}\left(\mathbf{x}_1 | \mathbf{y}_{1:T\setminus 1}, \gamma_{1:T}\right) = \mathcal{N}\Big(\boldsymbol{\mu}^{(1,i)}_{\boldsymbol{\theta}}\left(\mathbf{y}^{(\gamma)}_{1:T\setminus 1}, \gamma_{1:T}\right),\ \boldsymbol{\Sigma}^{(1,i)}_{\boldsymbol{\theta}}\left(\mathbf{y}^{(\gamma)}_{1:T\setminus 1}, \gamma_{1:T}\right)\Big).$$

We then reparameterize the Gaussian to sample $\mathbf{x}^{(i)}_1$: we draw $\boldsymbol{\varepsilon}^{(i)}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and compute $\mathbf{x}^{(i)}_1 = \boldsymbol{\mu}^{(1,i)}_{\boldsymbol{\theta}} + (\boldsymbol{\Sigma}^{(1,i)}_{\boldsymbol{\theta}})^{1/2} \boldsymbol{\varepsilon}^{(i)}_1$, where $(\boldsymbol{\Sigma}^{(1,i)}_{\boldsymbol{\theta}})^{1/2}$ is the diagonal matrix of standard deviations.

Assume we have sampled the states $\mathbf{x}^{(i)}_{1:t-1}$. To sample $\mathbf{x}^{(i)}_t$, we use (10) to get

$$q_{\boldsymbol{\theta}}\left(\mathbf{x}_t | \mathbf{x}^{(i)}_{1:t-1}, \mathbf{y}_{1:T\setminus t}, \gamma_{1:T}\right) = \mathcal{N}\Big(\boldsymbol{\mu}^{(t,i)}_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}_{1:t-1}, \mathbf{y}^{(\gamma)}_{1:T\setminus t}, \gamma_{1:T}\right),\ \boldsymbol{\Sigma}^{(t,i)}_{\boldsymbol{\theta}}\left(\mathbf{x}^{(i)}_{1:t-1}, \mathbf{y}^{(\gamma)}_{1:T\setminus t}, \gamma_{1:T}\right)\Big).$$

We reparameterize the Gaussian: draw $\boldsymbol{\varepsilon}^{(i)}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and compute $\mathbf{x}^{(i)}_t = \boldsymbol{\mu}^{(t,i)}_{\boldsymbol{\theta}} + (\boldsymbol{\Sigma}^{(t,i)}_{\boldsymbol{\theta}})^{1/2} \boldsymbol{\varepsilon}^{(i)}_t$.

Monte Carlo estimate of $\mathcal{L}^{\mathrm{RoVaS}}_{\mathrm{ELBO}}(\boldsymbol{\theta},\boldsymbol{\phi})$: Given the corruption parameters $\gamma_{1:T}$, the state trajectory parameters $\{(\boldsymbol{\mu}^{(t,i)}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}^{(t,i)}_{\boldsymbol{\theta}})\}_{t=1,i=1}^{T,L}$, and the sampled trajectories $\{\mathbf{x}^{(i)}_{1:T}\}_{i=1}^{L}$, we estimate each ELBO term.

The data-fit term is estimated as:

$$\hat{\mathcal{L}}_{\mathrm{fit}} = \frac{1}{L} \sum_{i=1}^{L} \sum_{t=1}^{T} \Big[(1 - \gamma_t) \log p_{\mathrm{noisy}}(\mathbf{y}_t | \mathbf{x}^{(i)}_t) + \gamma_t \log p_{\mathrm{corr}}(\mathbf{y}_t) \Big], \tag{17}$$

where we have used $\mathbb{E}_{q_{\boldsymbol{\theta}}(c_t)}[\log p(\mathbf{y}_t | \mathbf{x}_t, c_t)] = (1 - \gamma_t) \log p_{\mathrm{noisy}}(\mathbf{y}_t | \mathbf{x}_t) + \gamma_t \log p_{\mathrm{corr}}(\mathbf{y}_t)$.

For the dynamics term, since both the variational posterior and the dynamics are Gaussian, the KL divergence at each time step has the closed form:

$$\begin{aligned}\hat{\mathcal{L}}_{\text{dyn}} &= \frac{1}{L}\sum_{i=1}^{L}\sum_{t=1}^{T}\mathrm{KL}\Big[\mathcal{N}\Big(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(t,i)},\,\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t,i)}\Big)\;\|\;\mathcal{N}\Big(\boldsymbol{\mu}_{\boldsymbol{\phi}}^{(t,i)},\,\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{(t,i)}\Big)\Big] \\ &= \frac{1}{2L}\sum_{i=1}^{L}\sum_{t=1}^{T}\Bigg[\log\frac{|\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{(t,i)}|}{|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t,i)}|}+\mathrm{tr}\Big((\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{(t,i)})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(t,i)}\Big) \\ &\qquad +(\boldsymbol{\mu}_{\boldsymbol{\phi}}^{(t,i)}-\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(t,i)})^{\top}(\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{(t,i)})^{-1}(\boldsymbol{\mu}_{\boldsymbol{\phi}}^{(t,i)}-\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(t,i)})-d_x\Bigg]\end{aligned} \tag{18}$$

where we have used the prior:

$$p_{\boldsymbol{\phi}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}^{(i)})=\mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}}^{(t,i)}(\mathbf{x}_{1:t-1}^{(i)}),\boldsymbol{\Sigma}_{\boldsymbol{\phi}}^{(t,i)}(\mathbf{x}_{1:t-1}^{(i)})).$$

For the corruption term, since $q_{\boldsymbol{\theta}}(c_t|\mathbf{y}_{1:T})=\mathrm{Bernoulli}(\gamma_t)$, $\mathcal{L}_{\text{corr}}$ admits a closed-form expression:

$$\mathcal{L}_{\text{corr}}(\boldsymbol{\theta})=\sum_{t=1}^{T}\mathrm{KL}[\mathrm{Bernoulli}(\gamma_t)\;\|\;\mathrm{Bernoulli}(\varepsilon)]=\sum_{t=1}^{T}\left[\gamma_t\log\frac{\gamma_t}{\varepsilon}+(1-\gamma_t)\log\frac{1-\gamma_t}{1-\varepsilon}\right], \tag{19}$$

where ε is the prior corruption parameter.

The full Monte Carlo objective to be maximised is:

$$\hat{\mathcal{L}}_{\text{ELBO}}^{\text{RoVaS}}(\boldsymbol{\theta},\boldsymbol{\phi})=\hat{\mathcal{L}}_{\text{fit}}(\boldsymbol{\theta})-\hat{\mathcal{L}}_{\text{dyn}}(\boldsymbol{\theta},\boldsymbol{\phi})-\mathcal{L}_{\text{corr}}(\boldsymbol{\theta}). \tag{20}$$

Note that $\mathcal{L}_{\text{corr}}$ requires no Monte Carlo estimation but admits an exact expression. Gradients with respect to $(\boldsymbol{\theta},\boldsymbol{\phi})$ are computed via backpropagation through the reparameterised samples.

B Proof of Proposition 1

In this section, we prove Proposition 1. The proof has four parts. First, we give general conditions on the variational state posterior under which the amortized inference is outlier-robust in the KL sense (Appendix B.1). Second, we show that these conditions follow from two assumptions that are easy to check, one on the corruption gate and one on the encoder (Appendix B.2). Third, we show that RoVaS satisfies both assumptions (Appendix B.3). Lastly, we show that without the gated input, the guarantee fails for standard non-saturating encoders (Appendix B.4). An ungated smoother can therefore only be robust if its encoder happens to saturate, while for RoVaS, the robustness follows from the gating mechanism and holds for any encoder that satisfies a mild regularity condition (Assumption 4).

Notation. Throughout this appendix, we fix the corrupted time index t and the other measurements $\mathbf{y}_{s\neq t}$. We write $q_{\boldsymbol{\theta}}\triangleq q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{y}_{1:T})$ for the reference posterior, computed from the clean sequence, and $q_{\boldsymbol{\theta}}^{c}\triangleq q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T}|\mathbf{y}_{1:T}^{c})$ for the perturbed posterior, computed from the corrupted sequence. For a state history $\mathbf{x}_{1:s-1}$, the s-th factor $q_{\boldsymbol{\theta}}(\mathbf{x}_s|\mathbf{x}_{1:s-1})$ of the reference posterior has mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s)}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s)}$, and the s-th factor $q_{\boldsymbol{\theta}}^{c}(\mathbf{x}_s|\mathbf{x}_{1:s-1})$ of the perturbed posterior has mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s,c)}$; both factors are evaluated at the same history. We denote the d-th diagonal entries of the covariances by $\sigma_{\boldsymbol{\theta},d}^{(s)\,2}$ and $\sigma_{\boldsymbol{\theta},d}^{(s,c)\,2}$. For $s=1$, the history is empty, and we use the convention $\max_{j<1}\|\mathbf{x}_j\|=0$.

B.1 A general sufficient condition

We start with two assumptions on the variational state posterior, which together are sufficient for outlier robustness. Since each variational factor is a Gaussian with positive variances, the sampled states can take any value, and we cannot assume that the sampled trajectories stay in a compact set. Instead, we assume that the encoder outputs grow in a controlled way along the sampled trajectories, and we bound the per-step KL divergences in expectation over the sampled history.

Assumption 1 (Polynomial growth of the encoder mean)**.** *There exist constants $a, b \geq 0$ and $p \geq 1$, independent of $\mathbf{y}_t^c \in \mathbb{R}^{d_y}$, such that for every $s = 1, \ldots, T$ and every state history $\mathbf{x}_{1:s-1} \in \mathbb{R}^{(s-1)d_x}$,*

$$\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s)}\| \;\leq\; a + b\left(1 + \max_{j<s} \|\mathbf{x}_j\|\right)^p, \qquad \|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}\| \;\leq\; a + b\left(1 + \max_{j<s} \|\mathbf{x}_j\|\right)^p, \tag{21}$$

where the bound holds uniformly over $\mathbf{y}_t^c \in \mathbb{R}^{d_y}$. For $s = 1$ the history is empty and the bound reads $\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1)}\|, \|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1,c)}\| \leq a + b$.

Assumption 2 (Uniformly bounded variance)**.** *The covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s,c)}$ are diagonal for all $s = 1, \ldots, T$, and their diagonal entries admit uniform positive lower and upper bounds: there exist $0 < \underline{\sigma}^2 \leq \bar{\sigma}^2 < \infty$ such that*

$$\underline{\sigma}^2 \;\leq\; \sigma_{\boldsymbol{\theta},d}^{(s)\,2},\; \sigma_{\boldsymbol{\theta},d}^{(s,c)\,2} \;\leq\; \bar{\sigma}^2 \tag{22}$$

for all s, all d, all state histories, and uniformly over $\mathbf{y}_t^c \in \mathbb{R}^{d_y}$.

Note that the two assumptions are sufficient, but not necessary, for outlier robustness, and that both allow for Gaussian variational factors with positive variances. Assumption 1 allows the encoder mean to grow polynomially with the sampled history. Assumption 2 holds if the variance head is a continuous function of a bounded representation, and it can always be enforced by clamping the log-variance head to $[\log \underline{\sigma}^2, \log \bar{\sigma}^2]$.

Under these assumptions, all moments of the sampled states are finite, with bounds that do not depend on the corruption. We show this in the following lemma.

Lemma 1 (Moment propagation)**.** *Under Assumptions 1 and 2, for every $s = 1, \ldots, T$ and every $m \geq 1$,*

$$\sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \mathbb{E}_{q_{\boldsymbol{\theta}}^c}[\|\mathbf{x}_s\|^m] < \infty. \tag{23}$$

Proof. We use induction over the time index s. The bound for $\mathbf{x}_s$ will depend on higher moments of the earlier states $\mathbf{x}_1, \ldots, \mathbf{x}_{s-1}$, so we prove the statement for all orders m at the same time. That is, for $s = 1, \ldots, T$, let $\mathcal{S}(s)$ be the statement

$$\mathcal{S}(s): \qquad \sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \mathbb{E}_{q_{\boldsymbol{\theta}}^c}[\|\mathbf{x}_s\|^m] < \infty \quad \text{for every } m \geq 1. \tag{24}$$

We first show $\mathcal{S}(1)$, and then that $\mathcal{S}(1), \ldots, \mathcal{S}(s-1)$ together imply $\mathcal{S}(s)$. Since there are only T time steps, this proves the lemma.

We begin by simplifying $\mathcal{S}(s)$. Let $M_s \triangleq \max_{j<s} \|\mathbf{x}_j\|$ be the largest state norm in the history, with $M_1 = 0$. By the reparameterisation of the Gaussian factor (Appendix A), we can write $\mathbf{x}_s = \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)} + (\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s,c)})^{1/2}\boldsymbol{\varepsilon}_s$, where the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}$ is a function of the history $\mathbf{x}_{1:s-1}$, and $\boldsymbol{\varepsilon}_s \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is independent of the history. The covariance is diagonal with entries at most $\bar{\sigma}^2$ (Assumption 2), so $\|(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s,c)})^{1/2}\boldsymbol{\varepsilon}_s\| \leq \bar{\sigma}\|\boldsymbol{\varepsilon}_s\|$. Using the triangle inequality and $(u+v)^m \leq 2^{m-1}(u^m + v^m)$ for $u, v \geq 0$ and $m \geq 1$, we get

$$\mathbb{E}_{q_{\boldsymbol{\theta}}^c}\|\mathbf{x}_s\|^m \;\leq\; 2^{m-1}\left(\mathbb{E}_{q_{\boldsymbol{\theta}}^c}\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}\|^m + \bar{\sigma}^m\,\mathbb{E}\|\boldsymbol{\varepsilon}_s\|^m\right). \tag{25}$$

The Gaussian moment $\mathbb{E}\|\boldsymbol{\varepsilon}_s\|^m$ is finite and does not depend on $\mathbf{y}_t^c$. Hence, to show $\mathcal{S}(s)$, it suffices to bound the moments of the mean.

Base case ($s = 1$). The history is empty, and Assumption 1 gives $\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1,c)}\| \leq a + b$ for every $\mathbf{y}_t^c$. This implies $\mathbb{E}_{q_{\boldsymbol{\theta}}^c}\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}\|^m \leq (a+b)^m$ for every $m \geq 1$, as a, b does not depend on $\mathbf{y}_t^c$. $\mathcal{S}(1)$ holds.

Inductive step. Let $s \geq 2$, and assume that $\mathcal{S}(j)$ holds for all $j < s$. Fix $m \geq 1$. By Assumption 1, $\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}\| \leq a + b\,(1 + M_s)^p$, and using $(u+v)^m \leq 2^{m-1}(u^m + v^m)$ once more, we get

$$\mathbb{E}_{q_{\boldsymbol{\theta}}^c}\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}\|^m \;\leq\; 2^{m-1}\Big(a^m + b^m\,\mathbb{E}_{q_{\boldsymbol{\theta}}^c}\big[(1 + M_s)^{pm}\big]\Big). \tag{26}$$

Further, since $M_s^{pm} = \max_{j<s} \|\mathbf{x}_j\|^{pm} \leq \sum_{j<s} \|\mathbf{x}_j\|^{pm}$, we have

$$\mathbb{E}_{q_{\boldsymbol{\theta}}^c}\big[(1 + M_s)^{pm}\big] \;\leq\; 2^{pm-1}\Big(1 + \mathbb{E}_{q_{\boldsymbol{\theta}}^c} M_s^{pm}\Big) \;\leq\; 2^{pm-1}\Big(1 + \sum_{j<s} \mathbb{E}_{q_{\boldsymbol{\theta}}^c}\|\mathbf{x}_j\|^{pm}\Big). \tag{27}$$

The right-hand side of (27) only contains moments of order pm of the earlier states $\mathbf{x}_1, \dots, \mathbf{x}_{s-1}$. By the induction hypothesis, applied with the order pm instead of m, these moments are bounded uniformly over $\mathbf{y}_t^c$. Inserting (27) into (26) bounds the moments of the mean at time s, and thus bounds $\mathbb{E}_{q_{\boldsymbol{\theta}}^c}\|\mathbf{x}_s\|^m$ by a constant that depends on a, b, p, $\bar{\sigma}$, m, and s, but not on $\mathbf{y}_t^c$. Since m was arbitrary, $\mathcal{S}(s)$ holds, which completes the induction. □

Proposition 2. *Assume that the variational state posterior satisfies Assumptions 1 and 2. Then, the amortized inference is outlier-robust in the KL sense; that is,*

$$\sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) < \infty. \tag{28}$$

Proof. By the chain rule for KL divergence, we have

$$\mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) = \sum_{s=1}^{T} \mathbb{E}_{q_{\boldsymbol{\theta}}^c(\mathbf{x}_{1:s-1})}\Big[\mathrm{KL}\Big(q_{\boldsymbol{\theta}}^c(\mathbf{x}_s|\mathbf{x}_{1:s-1}) \;\Big\|\; q_{\boldsymbol{\theta}}(\mathbf{x}_s|\mathbf{x}_{1:s-1})\Big)\Big]. \tag{29}$$

Since the sum has finitely many terms, it is enough to bound each expected per-step KL uniformly over $\mathbf{y}_t^c \in \mathbb{R}^{d_y}$. Note that in each term, the corrupted factor $q_{\boldsymbol{\theta}}^c(\mathbf{x}_s|\mathbf{x}_{1:s-1})$ and the reference factor $q_{\boldsymbol{\theta}}(\mathbf{x}_s|\mathbf{x}_{1:s-1})$ are evaluated at the *same* history $\mathbf{x}_{1:s-1} \sim q_{\boldsymbol{\theta}}^c$, so they only differ through the measurement inputs.

Each per-step KL is between two Gaussians with diagonal covariances, and has the closed form

$$\begin{aligned}&\mathrm{KL}\Big(q_{\boldsymbol{\theta}}^c(\mathbf{x}_s|\mathbf{x}_{1:s-1}) \;\Big\|\; q_{\boldsymbol{\theta}}(\mathbf{x}_s|\mathbf{x}_{1:s-1})\Big) = \\ &\tfrac{1}{2}\left[\mathrm{tr}\Big((\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s)})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s,c)}\Big) + (\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)})^\top(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s)})^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}) - d_x + \log\frac{|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s)}|}{|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s,c)}|}\right].\end{aligned} \tag{30}$$

We bound the terms one by one. By Assumption 2, the trace term satisfies $\mathrm{tr}((\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s)})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s,c)}) \le d_x\,\bar{\sigma}^2/\underline{\sigma}^2$, and the log-determinant ratio satisfies $|\log(|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s)}|/|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s,c)}|)| \le d_x \log(\bar{\sigma}^2/\underline{\sigma}^2)$; both of these bounds are deterministic and hold for all $\mathbf{y}_t^c$. For the quadratic form, we have

$$\begin{aligned}(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)})^\top(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(s)})^{-1}(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}) &\le \frac{\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}\|^2}{\underline{\sigma}^2} \\ &\le \frac{2\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s)}\|^2 + 2\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(s,c)}\|^2}{\underline{\sigma}^2} \\ &\le \frac{4\big(a + b\,(1 + \max_{j<s}\|\mathbf{x}_j\|)^p\big)^2}{\underline{\sigma}^2},\end{aligned} \tag{31}$$

where the last inequality uses Assumption 1 for both means at the common history $\mathbf{x}_{1:s-1}$. We now take the expectation over the history $\mathbf{x}_{1:s-1} \sim q_{\boldsymbol{\theta}}^c$. Using $\big(a + b\,(1 + \max_{j<s}\|\mathbf{x}_j\|)^p\big)^2 \le 2a^2 + 2b^2\,(1+\max_{j<s}\|\mathbf{x}_j\|)^{2p}$, together with $\mathbb{E}(1+\max_{j<s}\|\mathbf{x}_j\|)^{2p} \le 2^{2p-1}\big(1+\sum_{j<s}\mathbb{E}\|\mathbf{x}_j\|^{2p}\big)$ and Lemma 1 for the order $2p$, we obtain

$$\sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \mathbb{E}_{q_{\boldsymbol{\theta}}^c(\mathbf{x}_{1:s-1})}\Big[\mathrm{KL}\Big(q_{\boldsymbol{\theta}}^c(\mathbf{x}_s|\mathbf{x}_{1:s-1}) \;\Big\|\; q_{\boldsymbol{\theta}}(\mathbf{x}_s|\mathbf{x}_{1:s-1})\Big)\Big] \triangleq B_s < \infty. \tag{32}$$

Inserting this into (29) gives

$$\sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) \le \sum_{s=1}^{T} B_s < \infty. \tag{33}$$

□

Remark 1. *The main restriction in Assumption 1 is that the constants (a, b) must be the same for all corruptions $\mathbf{y}_t^c \in \mathbb{R}^{d_y}$. If the encoder receives the raw measurement $\mathbf{y}_t$, this is in general not the case. For example, for a dense neural network with ReLU activations, the mean grows without bound as $\|\mathbf{y}_t^c\| \to \infty$, and no finite (a, b) exist (see Proposition 3 below). In the next subsection, we show that the corruption gate solves this problem: the gated input stays bounded for all corruptions, so the corruption can only affect the encoder through a compact set of inputs.*

B.2 SUFFICIENT CONDITIONS VIA THE CORRUPTION GATE

We now give two assumptions, one on the corruption gate and one on the encoder, under which Assumptions 1 and 2 hold.

Assumption 3 (Coercive gate)**.** *Let the variational corruption posterior be given by $q_{\boldsymbol{\theta}}(c_t = 1|\mathbf{y}_{1:T}) = \gamma_t = \sigma(h_{\boldsymbol{\theta}}(\mathbf{y}_{1:T})_t)$, where $h_{\boldsymbol{\theta}}(\mathbf{y}_{1:T})_t$ is the gate logit at time t. The gate logit satisfies*

$$h_{\boldsymbol{\theta}}(\mathbf{y}_{1:T})_t \geq \alpha \left\|\mathbf{y}_t\right\|^{\beta} - C \qquad \text{for all } \mathbf{y}_t \in \mathbb{R}^{d_y}, \tag{34}$$

where $\alpha > 0$, $\beta > 0$, and $C \geq 0$ are constants that may depend on $\boldsymbol{\theta}$ and $\mathbf{y}_{s\neq t}$ but not on $\mathbf{y}_t$.

In words, the coercivity assumption says that for extreme measurements $\mathbf{y}_t$, the corruption posterior eventually predicts $\gamma_t \approx 1$. The assumption can be seen as the analogue of the condition $\sup_{\mathbf{y}_t} W^2(\mathbf{y}_t, \hat{\mathbf{y}}_t) \left\|\mathbf{y}_t\right\|^2 < \infty$ on the weighting function W in WoLF (Duran-Martin et al., 2024, Theorem 3.2). Indeed, Lemma 2 below shows that coercivity gives $\sup_{\mathbf{y}_t}(1-\gamma_t)^2 \left\|\mathbf{y}_t\right\|^2 < \infty$, which is the same condition with W replaced by the weight $1-\gamma_t$ of the gated input (Eq. (9)).

For RoVaS (Eq. (8)), the gate logit is

$$h_{\boldsymbol{\theta}}(\mathbf{y}_{1:T})_t = g_{\boldsymbol{\theta}}(\mathbf{y}_{1:T})_t + \lambda \left\|\mathbf{y}_t - \boldsymbol{\mu}_{\text{pool}}\right\|^2_{\boldsymbol{\Sigma}^{-1}_{\text{pool}}}. \tag{35}$$

Let $\lambda_{\min} \triangleq \lambda_{\min}(\boldsymbol{\Sigma}^{-1}_{\text{pool}}) > 0$. Since $\left\|\mathbf{y}_t - \boldsymbol{\mu}_{\text{pool}}\right\|^2 \geq \frac{1}{2}\left\|\mathbf{y}_t\right\|^2 - \left\|\boldsymbol{\mu}_{\text{pool}}\right\|^2$, the Mahalanobis term satisfies

$$\lambda \left\|\mathbf{y}_t - \boldsymbol{\mu}_{\text{pool}}\right\|^2_{\boldsymbol{\Sigma}^{-1}_{\text{pool}}} \;\geq\; \tfrac{1}{2}\lambda\lambda_{\min}\left\|\mathbf{y}_t\right\|^2 - \lambda\lambda_{\min}\left\|\boldsymbol{\mu}_{\text{pool}}\right\|^2. \tag{36}$$

This quadratic term dominates any negative part of $g_{\boldsymbol{\theta}}$ that grows slower than quadratically in $\|\mathbf{y}_t\|$. Therefore, coercivity holds with $\beta = 2$ whenever $g_{\boldsymbol{\theta}}(\mathbf{y}_{1:T})_t \geq -A - B\left\|\mathbf{y}_t\right\|^r$ for some constants $A, B \geq 0$ and $0 \leq r < 2$ that do not depend on $\mathbf{y}_t$. In particular, if $g_{\boldsymbol{\theta}}$ is bounded, $|g_{\boldsymbol{\theta}}(\mathbf{y}_{1:T})_t| \leq G$, then Assumption 3 holds with $\alpha = \frac{1}{2}\lambda\lambda_{\min}$, $\beta = 2$, and $C = G + \lambda\lambda_{\min}\left\|\boldsymbol{\mu}_{\text{pool}}\right\|^2$. This is the case for a GRU or an LSTM followed by a continuous read-out, since their hidden states lie in $[-1,1]^H$. Linear growth ($r = 1$) is also enough. This holds for MLPs with Lipschitz activations, and for Transformers built from affine layers, such MLPs, dot-product attention, residual connections, and layer normalisation; note that the output of dot-product attention is a convex combination of value vectors, which are linear in the inputs.

Assumption 4 (Encoder regularity)**.** *The variational state posterior receives the measurements and gate values only through the pair $u \triangleq (\mathbf{y}^{(\gamma)}_{1:T}, \gamma_{1:T})$, where $\mathbf{y}^{(\gamma)}_t = (1-\gamma_t)\mathbf{y}_t$, so that the factors $q_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{y}^{(\gamma)}_{1:T\setminus t}, \gamma_{1:T})$ in the encoder are given by a map $(\mathbf{x}_{1:t-1}, u) \mapsto (\boldsymbol{\mu}^{(t)}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\theta}})$ with $\boldsymbol{\Sigma}^{(t)}_{\boldsymbol{\theta}}$ diagonal. There is an exponent $p \geq 1$ such that, for every compact set $K \subset \mathbb{R}^{Td_y} \times [0,1]^T$, the following two conditions hold for all $t = 1, \ldots, T$, all $d \in \{1, \ldots, d_x\}$, all state histories $\mathbf{x}_{1:t-1} \in \mathbb{R}^{(t-1)d_x}$ and all $u \in K$.*

(i) (Polynomial growth of the mean.) *There exist finite constants $a_K, b_K \geq 0$ such that*

$$\left\|\boldsymbol{\mu}^{(t)}_{\boldsymbol{\theta}}(\mathbf{x}_{1:t-1}, u)\right\| \;\leq\; a_K + b_K\left(1 + \max_{j<t}\left\|\mathbf{x}_j\right\|\right)^p. \tag{37}$$

(ii) (Two-sided variance bounds.) *There exist constants $0 < \underline{\sigma}^2_K \leq \bar{\sigma}^2_K < \infty$ such that $\underline{\sigma}^2_K \leq \sigma^{(t)\,2}_{\boldsymbol{\theta},d}(\mathbf{x}_{1:t-1}, u) \leq \bar{\sigma}^2_K$.*

Condition (i) holds for most standard neural-network architectures. If the network is globally Lipschitz in the state history and continuous in u, then (i) holds with $p = 1$; this is the case for MLPs and RNNs built from affine layers and Lipschitz activations. If the state pathway saturates, as for the GRU decoder used in RoVaS, whose hidden state lies in $[-1,1]^H$ (Appendix D.7), then (i) holds with $b_K \equiv 0$. Note that global Lipschitz continuity is not needed. For example, dot-product self-attention is not Lipschitz on an unbounded domain (Kim et al., 2021), but a Transformer state pathway built from attention, residual connections, layer normalisation, and MLPs with Lipschitz activations still satisfies (i) with $p = 1$, since the output of an attention layer is a convex combination of value vectors

that are linear in the inputs. Hence, condition (i) depends on the architecture, and not on the trained weights.

By the Extreme Value Theorem, condition (ii) holds whenever the variance head is a continuous function of a bounded state representation and of the gated input. This is the case for RoVaS (see Appendix B.3). If the state pathway is unbounded, condition (ii) can be enforced by clamping the log-variance head to $[\log \underline{\sigma}^2, \log \bar{\sigma}^2]$, which is a simple change in the implementation and common practice for numerical stability.

The following two lemmas show that Assumptions 3 and 4 together imply Assumptions 1 and 2.

Lemma 2. *Under Assumption 3 and with the measurements $\mathbf{y}_{s\neq t}$ fixed, the gated measurement sequence and the gates, computed from $\mathbf{y}^c_{1:T}$, are uniformly bounded over corruptions; that is, the set*

$$\left\{(\mathbf{y}^{(\gamma)}_{1:T}, \gamma_{1:T}) : \mathbf{y}^c_t \in \mathbb{R}^{d_y}\right\} \tag{38}$$

is contained in a compact subset of $\mathbb{R}^{Td_y} \times [0,1]^T$. Moreover, $\mathbf{y}^{(\gamma)}_t \to \mathbf{0}$ as $\|\mathbf{y}^c_t\| \to \infty$.

Proof. Throughout the proof, $\gamma^c_{1:T}$ denotes the gates computed from $\mathbf{y}^c_{1:T}$, so that $\mathbf{y}^{(\gamma)}_s = (1-\gamma^c_s)\mathbf{y}_s$ for $s \neq t$ and $\mathbf{y}^{(\gamma)}_t = (1-\gamma^c_t)\mathbf{y}^c_t$. Since $\gamma^c_{1:T} \in [0,1]^T$ and $[0,1]^T$ is compact, it suffices to show that $\mathbf{y}^{(\gamma)}_{1:T}$ is contained in a compact set uniformly over $\mathbf{y}^c_t$ (since a product of compact sets is compact). For $s \neq t$, the measurement $\mathbf{y}_s$ is fixed, and since $\gamma^c_s \in [0,1]$,

$$\|\mathbf{y}^{(\gamma)}_s\| = \|(1-\gamma^c_s)\mathbf{y}_s\| \leq \|\mathbf{y}_s\|, \tag{39}$$

which is a constant depending only on the fixed measurements $\mathbf{y}_{s\neq t}$.

It remains to bound $\mathbf{y}^{(\gamma)}_t = (1-\gamma^c_t)\mathbf{y}^c_t$ uniformly in $\mathbf{y}^c_t \in \mathbb{R}^{d_y}$. By Assumption 3, there exist constants $\alpha, \beta > 0$ and $C \geq 0$ such that $h_{\boldsymbol{\theta}}(\mathbf{y}^c_{1:T})_t \geq \alpha\|\mathbf{y}^c_t\|^\beta - C$. Using the inequality $1-\sigma(x) = (1+e^x)^{-1} \leq e^{-x}$, which holds for every $x \in \mathbb{R}$, with $x = h_{\boldsymbol{\theta}}(\mathbf{y}^c_{1:T})_t$, gives

$$\begin{aligned}
\|(1-\gamma^c_t)\mathbf{y}^c_t\| &= (1-\sigma(h_{\boldsymbol{\theta}}(\mathbf{y}^c_{1:T})_t))\|\mathbf{y}^c_t\| \\
&\leq \exp(-h_{\boldsymbol{\theta}}(\mathbf{y}^c_{1:T})_t)\,\|\mathbf{y}^c_t\| \\
&\leq \exp\big(-\alpha\|\mathbf{y}^c_t\|^\beta + C\big)\,\|\mathbf{y}^c_t\| \\
&= \exp\big(-\alpha r^\beta + C\big)\, r,
\end{aligned} \tag{40}$$

where we have set $r = \|\mathbf{y}^c_t\|$. Now, consider the function $\varphi(r) = r\, e^{-\alpha r^\beta + C}$ defined for $r \in [0,\infty)$. The function φ is continuous on $[0,\infty)$ and satisfies $\varphi(r) \to 0$ as $r \to \infty$ (since the exponential dominates the polynomial). Hence φ attains a finite maximum on $[0,\infty)$, and so $\|\mathbf{y}^{(\gamma)}_t\| \leq \sup_{r\geq 0}\varphi(r) < \infty$ for every $\mathbf{y}^c_t \in \mathbb{R}^{d_y}$; moreover, $\|\mathbf{y}^{(\gamma)}_t\| \leq \varphi(\|\mathbf{y}^c_t\|) \to 0$ as $\|\mathbf{y}^c_t\| \to \infty$. The set $\{(\mathbf{y}^{(\gamma)}_{1:T}, \gamma_{1:T}) : \mathbf{y}^c_t \in \mathbb{R}^{d_y}\}$ is therefore bounded, and by the Heine–Borel Theorem, contained in a compact set. □

Lemma 3. *Under Assumptions 3 and 4 and with the measurements $\mathbf{y}_{s\neq t}$ fixed, Assumptions 1 and 2 hold; that is, the encoder means satisfy the polynomial growth bound* (21) *with constants (a, b, p) uniform over $\mathbf{y}^c_t \in \mathbb{R}^{d_y}$, and the variances are uniformly bounded above and below.*

Proof. By Lemma 2, the gated input $u = (\mathbf{y}^{(\gamma)}_{1:T}, \gamma_{1:T})$ lies in a compact set K_y for all corruptions $\mathbf{y}^c_t$. This is the main point of the proof: the corruption can only affect the encoder through the gated input, and the gated input stays in a compact set, however large the corruption is. The clean sequence corresponds to the choice $\mathbf{y}^c_t = \mathbf{y}_t$, so the clean gated input also lies in K_y.

We now apply Assumption 4 with $K = K_y$. Condition (i) gives finite constants a_{K_y}, b_{K_y} such that, for every s, every state history, and every $\mathbf{y}^c_t \in \mathbb{R}^{d_y}$,

$$\|\boldsymbol{\mu}^{(s,c)}_{\boldsymbol{\theta}}\| \;\leq\; a_{K_y} + b_{K_y}\left(1 + \max_{j<s}\|\mathbf{x}_j\|\right)^p. \tag{41}$$

The same bound holds for the reference mean $\boldsymbol{\mu}^{(s)}_{\boldsymbol{\theta}}$, since the clean gated input also lies in K_y. This is Assumption 1 with the constants (a_{K_y}, b_{K_y}, p). Similarly, condition (ii) gives constants $0 < \underline{\sigma}^2_{K_y} \leq \bar{\sigma}^2_{K_y} < \infty$ that bound the diagonal variances of both the corrupted and the reference factors, for all state histories and all $\mathbf{y}^c_t$, which is Assumption 2. □

We are now ready to combine the results. The following corollary states that Assumptions 3 and 4 together imply outlier robustness in the KL sense.

Corollary 1. *Under Assumptions 3 and 4, the amortized inference is outlier-robust in the KL sense; that is,*

$$\sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) < \infty. \tag{42}$$

Proof. By Lemma 3, Assumptions 1 and 2 hold. The result follows from Proposition 2. □

Since t and the clean sequence $\mathbf{y}_{1:T}$ were arbitrary, Corollary 1 is Proposition 1 of the main text, which completes its proof.

B.3 RoVaS is outlier-robust

Finally, we specialise to RoVaS. The corruption gate of RoVaS is given by Eq. (8), and the gated input by Eq. (9). The gate network $g_{\boldsymbol{\theta}}$ is a bidirectional GRU over the measurement sequence, followed by a two-layer projection (Appendix D.7). The GRU hidden states lie in $[-1,1]^H$ and the projection is continuous, so $g_{\boldsymbol{\theta}}$ is bounded. As discussed after Assumption 3, the gate of RoVaS is therefore coercive with $\beta = 2$ and $\alpha = \frac{1}{2}\lambda\lambda_{\min}(\boldsymbol{\Sigma}_{\mathrm{pool}}^{-1})$, since the additive Mahalanobis term $\lambda \left\|\mathbf{y}_t - \boldsymbol{\mu}_{\mathrm{pool}}\right\|^2_{\boldsymbol{\Sigma}_{\mathrm{pool}}^{-1}}$ makes the gate logit grow at least quadratically in $\|\mathbf{y}_t\|$.

For the encoder, we check both conditions of Assumption 4 directly on the architecture in Appendix D.7, for an arbitrary compact set K of gated inputs. The sampled states only enter the autoregressive decoder through a GRU hidden state $\mathbf{h}$, which lies in $[-1,1]^H$ for any input. The context $\mathbf{z}_t$ is a continuous function of the gated input, so for $u \in K$, the pairs $(\mathbf{h}, \mathbf{z}_t)$ lie in a compact subset of $[-1,1]^H \times \mathbb{R}^H$. The mean and log-variance heads are a two-layer MLP applied to this pair, and are therefore continuous. By the Extreme Value Theorem, their outputs are bounded. The same argument covers the initialisation head for the first state, which is a continuous function of $\mathbf{z}_1$ only. Therefore, condition (i) holds with $b_K \equiv 0$ and any $p \geq 1$, and condition (ii) holds with $\underline{\sigma}_K^2 = \exp(\min_d \ell_{\boldsymbol{\theta},d})$ and $\bar{\sigma}_K^2 = \exp(\max_d \ell_{\boldsymbol{\theta},d})$ for the exponential parameterisation $\sigma^2_{\boldsymbol{\theta},d} = \exp(\ell_{\boldsymbol{\theta},d})$ used in this work, where the extrema are taken over the same compact set. Note that RoVaS does not need a clamped log-variance head; clamping is only needed for architectures with an unbounded state pathway. Hence, Assumption 4 holds for RoVaS.

Corollary 2 (RoVaS is outlier-robust)**.** *Let $(\boldsymbol{\theta}, \boldsymbol{\phi})$ be fixed trained parameters of RoVaS. The amortized inference of RoVaS is outlier-robust in the KL sense; that is,*

$$\sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) < \infty. \tag{43}$$

Proof. Assumptions 3 and 4 hold for RoVaS as discussed above. The result follows from Corollary 1. □

Remark 2. *For RoVaS, we have $b_K \equiv 0$ above; that is, the encoder mean is bounded along the sampled history. The polynomial growth in Assumption 1 is therefore not needed for the architecture used in this work. We keep it since it allows the guarantee to also cover encoders with an unbounded state pathway, where the state history has no compact representation and the supremum of the mean over all histories can be infinite. Further, since the GRUs of the encoder also saturate, $\mathbf{z}_t$ is in fact bounded even for unbounded inputs. We do not use this property above, and the guarantee therefore does not depend on it (see Appendix B.4).*

B.4 Without the gated input, the guarantee fails

Lastly, we show that the gated input is not only a tool for the proof. If the encoder receives the raw measurement $\mathbf{y}_t$ directly, outlier robustness in the KL sense does not hold for standard non-saturating architectures. Together with Corollary 1, this shows that the gating mechanism (Eq. (9)) and the coercive gate (Assumption 3) give the guarantee by design, while an ungated smoother is only robust if its encoder happens to saturate.

Proposition 3 (Non-robustness of ungated encoders)**.** *Suppose the encoder receives the raw measurement sequence, so that the first variational factor is* $q_{\boldsymbol{\theta}}(\mathbf{x}_1|\mathbf{y}_{1:T\setminus 1}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1)}(\mathbf{y}_{1:T\setminus 1}), \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1)}(\mathbf{y}_{1:T\setminus 1}))$ *with* $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1)}$ *diagonal and positive definite. Fix* $t \neq 1$ *and suppose the mean map is unbounded in* $\mathbf{y}_t$*; that is,*

$$\limsup_{\|\mathbf{y}_t^c\|\to\infty} \|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1)}(\mathbf{y}_1, \ldots, \mathbf{y}_t^c, \ldots, \mathbf{y}_T)\| = \infty. \tag{44}$$

Then the amortized inference is not outlier-robust in the KL sense:

$$\sup_{\mathbf{y}_t^c \in \mathbb{R}^{d_y}} \mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) = \infty. \tag{45}$$

Proof. All terms in the chain-rule decomposition (29) are non-negative. Keeping only the term for $s = 1$, where the history is empty and there is no expectation over histories, gives

$$\begin{aligned} \mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) &\geq \mathrm{KL}\Big(q_{\boldsymbol{\theta}}^c(\mathbf{x}_1) \,\Big\|\, q_{\boldsymbol{\theta}}(\mathbf{x}_1)\Big) \\ &\geq \tfrac{1}{2}\,(\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1,c)})^\top (\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1)})^{-1} (\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1,c)}) \\ &\geq \frac{\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1,c)}\|^2}{2\,\bar{\sigma}_{\mathrm{ref}}^2}, \end{aligned} \tag{46}$$

In the second inequality, we dropped the remaining terms of the Gaussian KL, namely $\frac{1}{2}[\mathrm{tr}((\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1)})^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1,c)}) - d_x + \log(|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1)}|/|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1,c)}|)]$. Together, these terms are the KL between two zero-mean Gaussians with covariances $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1,c)}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1)}$, and are therefore non-negative. In the third inequality, $\bar{\sigma}_{\mathrm{ref}}^2$ is the largest diagonal entry of $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1)}$. This is a finite positive constant, and we need no assumption for it, since $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{(1)}$ is the *reference* covariance, computed from the clean sequence, and hence a fixed matrix that does not depend on $\mathbf{y}_t^c$. The reference mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1)}$ is also fixed. By (44), we can choose a sequence of corruptions $\mathbf{y}_t^{c,(n)}$ such that $\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1,c)}\| \to \infty$. Then $\|\boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1)} - \boldsymbol{\mu}_{\boldsymbol{\theta}}^{(1,c)}\| \to \infty$, and the lower bound goes to infinity. □

Remark 3. *Condition* (44) *holds for typical non-saturating architectures. For example, the output of a linear encoder with non-zero gain on* $\mathbf{y}_t$ *is unbounded in the input, and this is typically also the case for MLPs with ReLU activations. The condition does not hold for architectures where the measurement pathway saturates, such as a GRU with* $\tanh$ *activations. This explains why the ungated ablations VaS and VaS-*t *also reach a PIF plateau in Appendix C, although at a level roughly three orders of magnitude higher than RoVaS. For VaS and VaS-*t*, the boundedness comes from the GRU encoder, and it is lost if the encoder is replaced by an architecture that does not saturate in the measurements. Note that this also applies to VaS-*t*, since its Student-*t *likelihood is only used in the training objective, and at inference, its amortized posterior is an ungated encoder. For RoVaS, on the other hand, the guarantee of Corollary 1 comes from the gate, and holds for any encoder that satisfies Assumption 4.*

C Empirical Study of Outlier Robustness

In this section, we empirically illustrate the robustness guarantee of Proposition 1 and examine the role of the encoder architecture. We perform two experiments. First, we evaluate $\mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c)$ for VaS, VaS-t, and RoVaS as the perturbation magnitude varies. Second, we examine the per-step factor of $\mathrm{PIF}_{\mathrm{KL}}$ from Eq. (29), to study how the corruption's influence localises in time.

PIF as a function of perturbation magnitude. We report the average and standard deviation of $\mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c)$ over 50 trajectories of length $T = 100$, with t uniformly drawn from 1 to 100. The perturbed (or corrupted) $\mathbf{y}_t^c$ is defined to be

$$\mathbf{y}_t^c = \mathbf{y}_t + C\frac{\boldsymbol{\varepsilon}_t}{\|\boldsymbol{\varepsilon}_t\|}, \tag{47}$$

where $\boldsymbol{\varepsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ gives a random direction and $C > 0$ controls the magnitude of the perturbation. We test with $C \in \{1, 10, 10^2, 10^3, 10^4, 10^5\}$. Note that at $C = 0$ we recover $\mathbf{y}_t$; $\mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) = 0$ in this case.

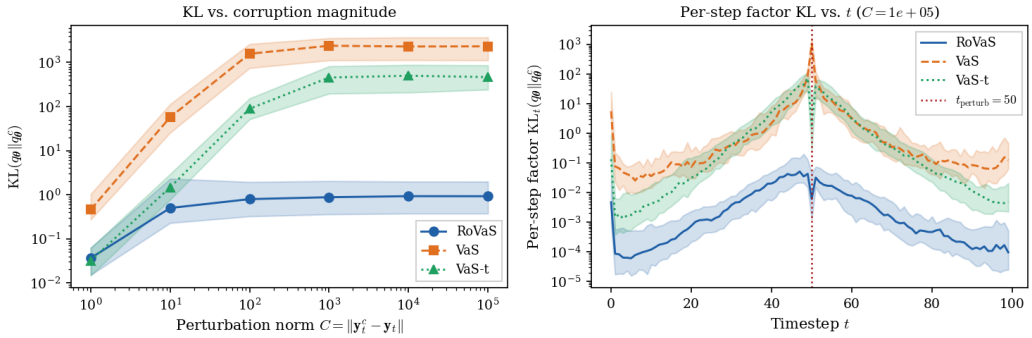


Figure 4: *Left*: Mean and standard deviation of $\mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c) = \mathrm{KL}(q_{\boldsymbol{\theta}}^c \| q_{\boldsymbol{\theta}})$ as C varies for 50 trajectories of length 100. *Right*: Per-step factor KL when introducing corruption at $t = 50$ with $C = 10^5$. Shaded regions show ± 1 standard deviation over 50 runs.

Figure 4 (left) shows $\mathrm{PIF}_{\mathrm{KL}}(\mathbf{y}_t^c)$ averaged over 50 trajectories as C varies. All three methods plateau within the chosen range of C: RoVaS at ≈ 1, VaS-t at $\approx 5 \times 10^2$, and VaS at $\approx 2 \times 10^3$. That is, the PIF of RoVaS is roughly three orders of magnitude lower than for the ablations. The plateau of RoVaS agrees with Proposition 1. For VaS and VaS-t, the plateaus come from the saturation of the GRU encoder (Appendix B.4).

Note that Proposition 1 only says that the PIF of RoVaS is bounded; it does not say how large the bound is. In practice, the level of the plateau also matters, since it determines how much a corruption can move the posterior. For RoVaS, the PIF is bounded because of the corruption gate (Eq. (8)) and the gated input (Eq. (9)), and this holds for any encoder that satisfies Assumption 4. VaS and VaS-t also plateau, but for a different reason: their GRU encoder uses saturating $\tanh$ activations, which bound the encoder output. If the GRU is replaced by an encoder whose mean is unbounded in the measurements, such as an MLP with ReLU activations, then Proposition 3 shows that the PIF of VaS and VaS-t becomes unbounded. The guarantee for RoVaS still holds for such an encoder, as long as it satisfies Assumption 4 (Corollary 1), since the gated input stays bounded for any encoder.

Per-step factor of the PIF. We study the per-step factor

$$\mathbb{E}_{q_{\boldsymbol{\theta}}^c(\mathbf{x}_{1:s-1})}\left[\mathrm{KL}\left(q_{\boldsymbol{\theta}}^c(\mathbf{x}_s|\mathbf{x}_{1:s-1}) \;\middle\|\; q_{\boldsymbol{\theta}}(\mathbf{x}_s|\mathbf{x}_{1:s-1})\right)\right]$$

for $s = 1, \ldots, 100$, with the corrupted measurement applied at $t = 50$ using $C = 10^5$. The result is shown in Figure 4 (right). There are three qualitative observations.

First, all methods exhibit elevated per-step KL at $t = 1$. This is an artifact of the amortized inference: $q_{\boldsymbol{\theta}}(\mathbf{x}_1)$ is conditioned on $\mathbf{y}_{1:T\setminus 1}$ but not on any sampled state, so the prediction is more uncertain than at later timesteps where $\mathbf{x}_{1:s-1}$ is available.

Second, at the corrupted timestep $t = 50$, all methods exhibit a change in per-step KL, but the direction differs. RoVaS shows a slight decrease, which might be expected; if a similar state history has been sampled, the input to the encoder is almost the same, since the corrupted measurement is not used by the 50th factor for either sequence. VaS-t exhibits the same behaviour but at a larger magnitude. VaS, in contrast, shows an increase, indicating its state factors at $t = 50$ differ significantly between the valid and corrupted measurement sequences.

Third, the magnitude of the per-step KL is significantly lower for RoVaS compared to VaS-t and VaS; almost four orders of magnitude. Furthermore, the KL values for VaS-t at timesteps 30 to 70 are almost identical to VaS (except at $t = 50$). This empirically suggests that VaS-t is as sensitive to large corrupted measurements as VaS at neighbouring timesteps.

In conclusion, the empirical results agree with Proposition 1: the PIF of RoVaS plateaus as the perturbation magnitude grows. VaS and VaS-t also plateau with the GRU encoder used here, but this is a property of the architecture and not a guarantee (Proposition 3). The gap of roughly three orders of magnitude between the plateaus, together with the difference in how the influence is localised in

time (Fig. 4, right), suggests that attenuating the measurements with a gate gives a different and more reliable form of robustness than using a heavy-tailed distribution such as the Student-t.

D On experimental setup and additional experimental results

D.1 On stochastic Lorenz-63 system

The stochastic Lorenz-63 system is a 3-dimensional stochastic dynamical system given by the discretised dynamics

$$\mathbf{x}_{t+1} = F(\mathbf{x}_t)\,\mathbf{x}_t + \mathbf{e}_t, \quad \mathbf{e}_t \sim \mathcal{N}(\mathbf{0},\, \sigma_e^2\,\mathbf{I}), \tag{48}$$

where $F(\mathbf{x}_t) = \exp\left(\begin{bmatrix} -10 & 10 & 0 \\ 28 & -1 & -x_{t,1} \\ 0 & x_{t,1} & -8/3 \end{bmatrix} \Delta\tau\right)$, with $\Delta\tau = 0.02$ and $\sigma_e^2 = 0.1$ corresponding to -10 dB process noise. The stochastic Lorenz-63 system is used in many works on state estimation, such as (van Nierop et al., 2025; Ghosh et al., 2024; 2025; Cumlin et al., 2026; Norén et al., 2026; Ruhe & Forré, 2022; Ni et al., 2022; Revach et al., 2022b;a).

D.2 Definitions of NMSE and SMNR

The signal-to-measurement noise ratio (SMNR) is defined as

$$\text{SMNR} = 10\log_{10}\left(\frac{\sum_{t=1}^{T}\mathbb{E}\left[\left\|\mathbf{h}(\mathbf{x}_t) - \mathbb{E}[\mathbf{h}(\mathbf{x}_t)]\right\|_2^2\right]}{T d_y \sigma_w^2}\right), \tag{49}$$

where d_y denotes the measurement dimension, σ_w^2 the per-dimension measurement noise variance, and $\mathbf{h}(\mathbf{x}_t^{(j)})$ the noiseless measurement. Higher measurement noise yields a lower SMNR.

State estimation accuracy is measured by the normalised mean squared error (NMSE) in dB,

$$\text{NMSE} = 10\log_{10}\left(\frac{\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t - \hat{\mathbf{x}}_t\|_2^2}{\frac{1}{T}\sum_{t=1}^{T}\|\mathbf{x}_t\|_2^2}\right), \tag{50}$$

where $\mathbf{x}_t$ is the true state, $\hat{\mathbf{x}}_t$ the estimated state, and T the trajectory length. Smaller values indicate better estimation. Here, we use $\hat{\mathbf{x}}_t$ as the mean of the state's posterior or the approximate posterior.

D.3 Results of DANSE, DNS, and VSE

We evaluate three recent data-driven model-free state estimators: DANSE (Ghosh et al., 2024), DNS (Cumlin et al., 2026), and VSE (Norén et al., 2026). We use the Lorenz-63 identity setup as done in Section 4.2.1, at SMNR $= 20$ dB and corruption rate $p_\varepsilon = 10\%$. Note that these methods are not designed for the corruption scenario, and we expect them to degrade significantly under corruption.

Table 2 reports NMSE averaged over five seeds. All three methods achieve an NMSE between -13.6 and -12.6 dB, which is close to the trivial upper bound of -12.0 dB (Appendix G); thus, the methods provide essentially no improvement over the raw measurement. By contrast, RoVaS achieves -29.0 dB, an improvement of more than 15 dB. This confirms that adding outlier-robustness to model-free methods is non-trivial; it is not enough to use a model-free state estimator on corrupted data, since these methods learn from the corrupted measurements and thus inherit the corruption.

D.4 Visual illustration for Lorenz-63 linear

A visual illustration of a Lorenz-63 trajectory with linear measurements, together with the estimated trajectories of the compared methods, is shown in Figure 5.

D.5 Visual illustration for Lorenz-63 nonlinear

A visual illustration of a Lorenz-63 trajectory with nonlinear measurements, together with the estimated trajectories of the compared methods, is shown in Figure 6.

Table 2: NMSE of recent model-free state estimators on Lorenz-63 (identity, SMNR $=$ 20 dB, $p_\varepsilon = 10\%$). Mean $\pm$ std over 5 seeds.

Method	NMSE (dB)
DANSE (Ghosh et al., 2024)	-13.57 ± 0.11
DNS (Cumlin et al., 2026)	-13.63 ± 0.19
VSE (Norén et al., 2026)	-12.57 ± 0.10
Trivial Estimator (App. G)	-12.00
RoVaS (ours)	-29.02 ± 0.13

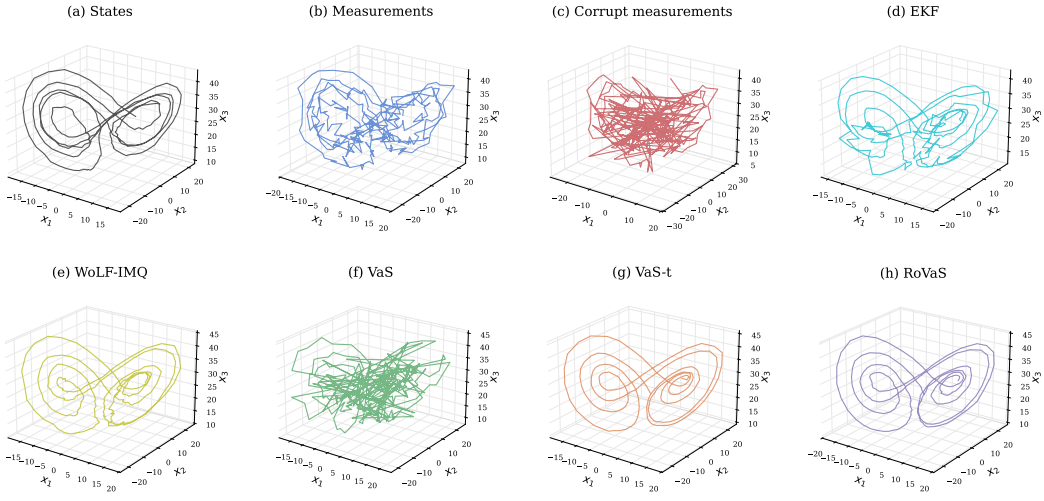


Figure 5: Visual illustration for stochastic Lorenz-63 system in the linear measurement setup. First row of the figure: (a) a state trajectory; (b) noisy measurements at 20 dB SMNR; (c) corrupted measurements with 20% corruption rate; (d) estimated trajectory using EKF. Second row of the figure: (e) estimated state trajectory using WoLF-IMQ; (f) estimated state trajectory using VaS. (g) estimated state trajectory using VaS-t; (h) estimated state trajectory using RoVaS.

D.6 Visual illustration for BMLMovi dataset

A visual illustration of the joints, measurements and estimated joints is shown in Figure 7. Note that some of the corrupted frames partially resemble the true pose. This is because the corruptions are drawn, independently of the state, from a Gaussian fitted to the pooled measurements. Joint angles with a small variance across the dataset are therefore replaced by values close to their mean, which are often close to the true values.

D.7 Architecture

RoVaS comprises three jointly-trained modules: the state variational distribution $q_{\boldsymbol{\theta}}(\mathbf{x}_{1:T})$, the dynamics prior $p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T})$, and the corruption gate $g_{\boldsymbol{\theta}}$. We describe each in turn.

State variational distribution. The state variational distribution is structured as a shifted bidirectional GRU encoder following Cumlin et al. (2026). Following the encoder is an autoregressive decoder that maps the history of drawn states as well as the current context to the Gaussian parameters over the current state. The encoder receives the gated measurement $\mathbf{y}_t^{(\gamma)} = (1 - \gamma_t)\mathbf{y}_t$ concatenated with the gate output γ_t, giving an input dimension of $d_y + 1$. A separate forward GRU and backward GRU are run over the input sequence; the encoded context $\mathbf{z}_t$ is formed by concatenating the forward hidden state at time $t-1$ with the backward hidden state at time $t+1$, ensuring that $\mathbf{z}_t$ does not depend directly on $\mathbf{y}_t$. At the boundaries ($t = 1$ and $t = T$), the available half-context is mapped

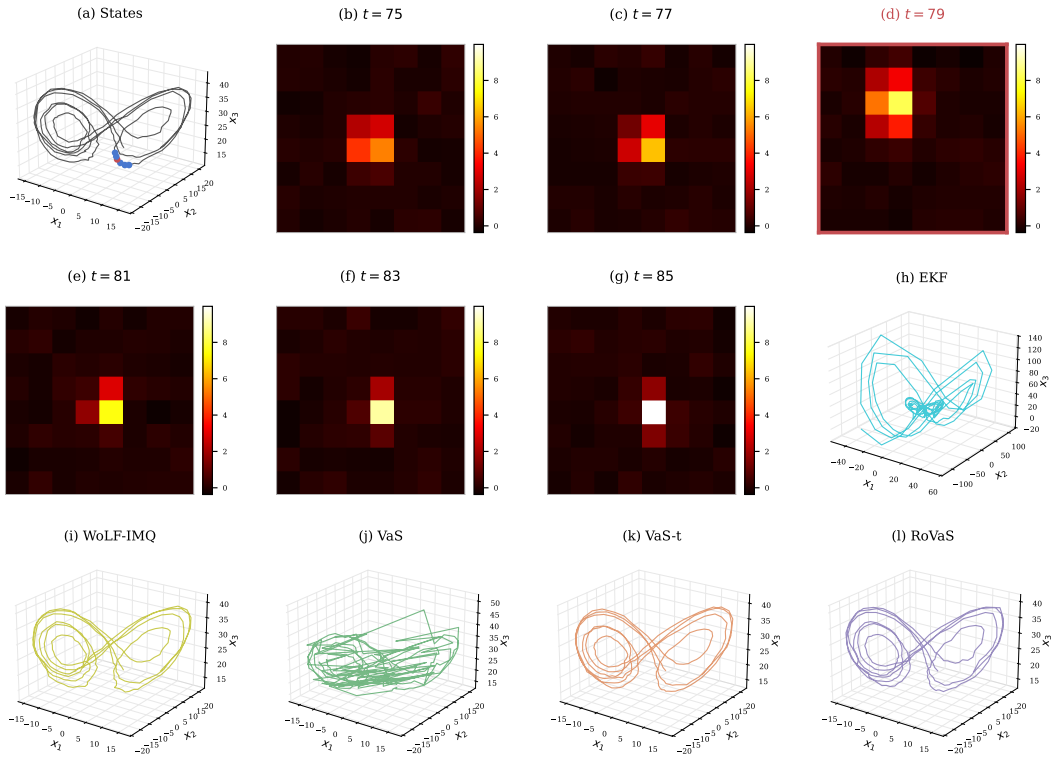


Figure 6: Visual illustration for the stochastic Lorenz-63 system in the nonlinear camera measurement setup at 20 dB SMNR with 20% corruption rate. First row of the figure: (a) a state trajectory; (b)–(d) camera measurements at $t = 75$, $t = 77$, and $t = 79$, where the frame at $t = 79$ (highlighted in red) is a corrupted measurement. Second row of the figure: (e)–(g) camera measurements at $t = 81$, $t = 83$, and $t = 85$; (h) estimated state trajectory using EKF. Third row of the figure: (i) estimated state trajectory using WoLF-IMQ; (j) estimated state trajectory using VaS; (k) estimated state trajectory using VaS-t; (l) estimated state trajectory using RoVaS.

through a learned linear projection to match the inner dimension. A linear context projection followed by a ReLU produces the final $\mathbf{z}_{1:T}$ at hidden dimension H.

The autoregressive decoder maintains a GRU hidden state over the sampled state trajectory $\mathbf{x}_{1:t-1}$. At time $t \geq 2$, the previous sample $\mathbf{x}_{t-1}$ is fed through the GRU, and the output is concatenated with $\mathbf{z}_t$ and passed through a two-layer MLP that outputs the mean and diagonal log-variance of $q_{\boldsymbol{\theta}}(\mathbf{x}_t|\mathbf{x}_{1:t-1}, \mathbf{y}_{1:T\setminus t}, \gamma_{1:T})$. The state at $t = 1$ is generated by a separate initialisation head that maps $\mathbf{z}_1$ alone (without state context) to a mean and log-variance. Reparameterised samples $\mathbf{x}_t$ are drawn at every time step.

Dynamics prior. The dynamics prior $p_{\boldsymbol{\phi}}(\mathbf{x}_t|\mathbf{x}_{1:t-1})$ is non-Markovian. A causal one-dimensional convolutional stack with two layers and kernel size 3 (receptive field 5) produces a feature representation of the recent state history, which is fed into a single-layer GRU to accumulate the full history (as proposed by Cumlin et al., 2026). The GRU output is mapped through linear heads to the mean and diagonal log-variance of a Gaussian transition. We predict a diagonal covariance.

Corruption gate. The corruption gate is a bidirectional GRU over the raw measurement sequence, followed by a two-layer projection to a per-time scalar logit. The output bias is initialised to -5 to ensure the gate fires sparsely at initialisation (so $\gamma_t \approx \sigma(-5) \approx 0.007$). The final gate output is

$$\gamma_t = \sigma\Big(g_{\boldsymbol{\theta}}(\mathbf{y}_{1:T})_t + \lambda \, \|\mathbf{y}_t - \boldsymbol{\mu}_{\text{pool}}\|^2_{\boldsymbol{\Sigma}^{-1}_{\text{pool}}}\Big), \tag{51}$$

where $\boldsymbol{\mu}_{\text{pool}}$ and $\boldsymbol{\Sigma}_{\text{pool}}$ are the empirical mean and covariance of the training measurements and λ is a fixed coefficient that makes the gate coercive (Assumption 3; see Appendix B.3).

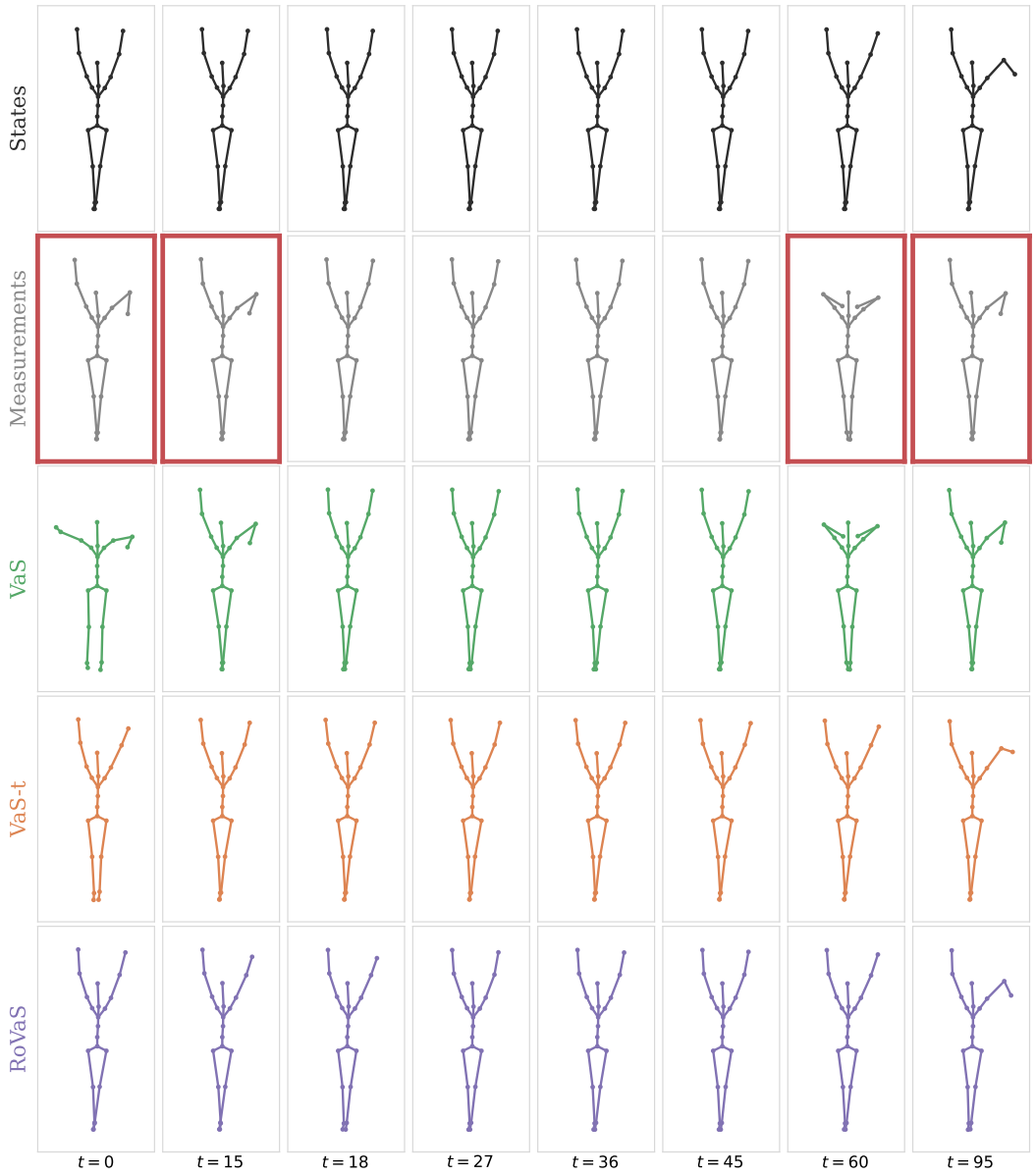


Figure 7: Qualitative results on a BMLMoVi test sequence at 20% corruption rate (at SMNR = 20 dB). Top to bottom: states (ground-truth), noisy measurements (corrupted frames marked in red), and state estimates from VaS, VaS-t, and RoVaS. The skeleton-like figures are projected onto the xy-plane for easier visualization.

Architectural hyperparameters by experiment. Table 3 summarises architecture sizes across experiments. The smoother and gate widths scale with the measurement dimension. The Lorenz-63 camera setup uses a higher-dimensional measurement ($d_y = 64$) requiring wider gate and encoder; the BML pose data uses a $d_y = 63$ measurement with wider hidden dimensions throughout to match the higher-dimensional state space.

D.8 Training Curriculum

Direct optimisation of the RoVaS ELBO from random initialisation collapses to the degenerate solution $\gamma_t \to 1$ everywhere; the model explains every measurement as corrupt. We resolve this with a three-stage curriculum that sequences the optimisation order so that the dynamics and the

Table 3: Architectural hyperparameters by experiment.

	Lorenz-63 (identity)	Lorenz-63 (camera)	BML pose
d_x (state)	3	3	63
d_y (measurement)	3	64	63
Smoother hidden dim	64	64	256
Smoother dense dim	64	64	256
STM hidden dim	64	32	256
STM CNN channels	$[16, 16]$	$[16, 16]$	$[32, 32]$
Gate hidden dim	64	64	256
Gate layers	2	2	2

encoder reach a usable state before the gate is allowed to fire. This training curriculum is described and analyzed in detail in Appendix E.

D.9 Loss and Optimisation

The training loss is the negative ELBO, $-\mathcal{L}_{\mathrm{ELBO}}^{\mathrm{RoVaS}}(\boldsymbol{\theta}, \boldsymbol{\phi})$. This is optimized using Monte-Carlo and the reparameterization trick. This is described in detail in Appendix A.

We optimise with Adam at learning rate 10^{-3}, batch size 64, and zero weight decay. A step learning-rate schedule decays the rate by a factor of 0.9 every $\lceil \text{num_epochs}/6 \rceil$ epochs, following previous works such as Ghosh et al. (2024); Cumlin et al. (2026). Gradients are clipped to a maximum norm of 5. Training runs for 3000 epochs in all experiments with early stopping of the loss on the validation data after 500 epochs. We report the results from the final checkpoint.

D.10 Specification of p_{corr}

The corruption likelihood p_{corr} is a user-specified modelling choice as it is assumed to be unknown. We use two variants for our experiments. One parameterization is to use $p_{\mathrm{corr}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{pool}}, \boldsymbol{\Sigma}_{\mathrm{pool}})$, where the moments are estimated from the measurements; including corrupted and non-corrupted measurements. Another parameterization is $p_{\mathrm{corr}} = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{pool}}, k\boldsymbol{\Sigma}_w)$ where k is a constant and $\boldsymbol{\Sigma}_w$ is the (known) measurement noise covariance. On Lorenz-63 with both identity map and Gaussian point spread map, the empirical pool covariance worked well. Note that in this case, p_{corr} has the same Gaussian form as the distribution that generates the corruptions (Appendix D.12), and changing its scale over two orders of magnitude changes the NMSE by at most 0.4 dB (Table 9). However, on BMLMoVi, training resulted for some seeds in the degenerate solution of explaining all measurements as outliers. Therefore, on BMLMoVi, we reduced the covariance of p_{corr}; we found that $k = 5$ worked well. A sensitivity analysis on k is reported in Appendix F.2.

D.11 Training Hyperparameters by Experiment

Table 4 summarises training hyperparameters across our three experimental settings. The dynamics-KL annealing schedule ($T_1 = 800$ steps from $\beta_{\mathrm{dyn}} = 0$ to 1), the corruption-KL anneal length ($T_2 = 800$ steps from β_0 to 1), the prior corruption rate $\varepsilon = 0.01$, the gate's coercivity coefficient λ, and the optimisation hyperparameters are held fixed across SMNR levels and corruption rates. The single hyperparameter that varies across experiments is β_0, which scales with the measurement dimension and corruption likelihood specification; see Table 5.

D.12 Datasets

Stochastic Lorenz-63 system and linear measurement system. We integrate the stochastic Lorenz-63 system with parameters $\sigma = 10$, $\rho = 28$, $\beta = 8/3$, an integration step $\Delta t = 0.02$, and additive process noise $\eta_t \sim \mathcal{N}(0, 0.01I)$. After a burn-in of 500 steps to enter the attractor, we record sequences of length $T = 100$. Measurements are $\mathbf{y}_t = \mathbf{x}_t + \mathbf{w}_t$ with $\mathbf{w}_t \sim \mathcal{N}(0, \sigma_w^2 I)$, where σ_w^2 is set to match the target SMNR. Corrupted measurements are drawn i.i.d. from $\boldsymbol{\mu}_{\mathrm{pool}} + \mathcal{N}(0, \boldsymbol{\Sigma}_{\mathrm{pool}})$, the empirical mean and covariance of clean measurements. The training, validation, and test sets contain 1000, 100, and 100 sequences respectively.

Table 4: Training hyperparameters by experiment. T_1 and T_2 denote the number of steps of curriculum stages 1 and 2 in optimisation steps; β_0 is the initial corruption-KL weight; ε is the prior corruption rate; λ is the gate coercivity coefficient. If k is none (as for the Lorenz-63 datasets), the Gaussian pool is used.

Setting	Lorenz-63 (identity)	Lorenz-63 (camera)	BMLMoVi
SMNR (dB)	10, 20	20	20
Corruption rate p_ε	0–25%	0–25%	0–25%
T_1 (steps)	800	800	800
T_2 (steps)	800	800	800
β_0	5	200	1000
ε	0.01	0.01	0.01
λ	0.01	0.01	0.01
p_{corr} scale k	None	None	5
Learning rate	10^{-3}	10^{-3}	10^{-3}
Batch size	64	64	64
Epochs	3000	3000	3000
Sequences	1000	1000	1000
Sequence length T	100	100	100

Stochastic Lorenz-63 system and nonlinear measurement system. The state is the same Lorenz-63 trajectory; the measurement is a $d_y = 64$ Gaussian point-spread-function rendering of the projected state, simulating a camera measurement. Other dataset parameters match the linear setting.

BMLMoVi and linear measurement system. We use motion capture sequences from the BML-movi dataset, projected to a $d_x = d_y = 63$ pose representation. Sequences of length $T = 100$ are sampled from continuous motion segments. The measurement model is $\mathbf{y}_t = \mathbf{x}_t + \mathbf{w}_t$ with Gaussian noise. Unlike Lorenz-63, the corruption-generating distribution is not directly known; we apply the same i.i.d. corruption procedure with corruption draws from a Gaussian fitted to the empirical pool.

D.13 Reproducibility

All experiments are run on a single NVIDIA A40 GPU. Each reported number is the mean over five seeds; standard deviations are reported as appropriate. We provide the Gin configuration file for each experiment on GitHub, shared upon acceptance.

E Staged Training Procedure

The RoVaS objective (Eq. (6)) is difficult to optimise, as mentioned in Section 1: detecting a corruption requires a reasonable dynamics model, and learning the dynamics requires knowing which measurements to trust. Direct optimisation of $\mathcal{L}_{\text{ELBO}}^{\text{RoVaS}}$ from random initialisation gives the optimiser no way to break this circularity; in practice, everything is predicted as a corruption, and the variational posterior collapses to an uninformative prior $p_{\boldsymbol{\phi}}(\mathbf{x}_{1:T})$.

We resolve this with a three-stage training approach. Each stage reweights two of the three ELBO terms,

$$\mathcal{L}_{\text{ELBO}}^{\text{RoVaS}}\big(\boldsymbol{\theta}, \boldsymbol{\phi}; \beta_{\text{dyn}}, \beta_{\text{corr}}\big) = \mathcal{L}_{\text{fit}}(\boldsymbol{\theta}) - \beta_{\text{dyn}}\, \mathcal{L}_{\text{dyn}}(\boldsymbol{\theta}, \boldsymbol{\phi}) - \beta_{\text{corr}}\, \mathcal{L}_{\text{corr}}(\boldsymbol{\theta}), \tag{52}$$

where $\beta_{\text{dyn}}, \beta_{\text{corr}} \geq 0$ are scheduled scalars. The unmodified ELBO corresponds to $\beta_{\text{dyn}} = \beta_{\text{corr}} = 1$. The three stages are as follows.

Stage 1: Dynamics warm-up (epochs 0 to 50; gradient steps 0 to 800). We set $\beta_{\text{corr}} = 0$ and anneal β_{dyn} linearly from 0 to 1. We also clamp the gate output to a small constant, $\gamma_t \equiv \gamma_0$ for all t, so the encoder receives a consistent near-zero gating signal throughout Stage 1. This prevents the gate from drifting during a stage in which the loss does not constrain it. The model thus learns a state posterior and dynamics prior that explain all measurements, including corruptions. This is deliberately non-robust. The goal is a usable posterior before the gate is asked to do anything. Annealing β_{dyn} from zero rather than starting at 1 is the standard approach to avoid posterior collapse in variational sequence models (see, e.g., Bowman et al., 2015; Sønderby et al., 2016).

Stage 2: Gate activation (epochs 50 to 100; gradient steps 800 to 1600). We release the gate so that γ_t is computed from Eq. (8), and keep $\beta_{\mathrm{dyn}} = 1$, and start to ramp β_{corr} linearly from an initial value β_0 down to 1. The initial β_{corr} results in a gate near the prior ($\gamma_t \approx \varepsilon$), which blocks the pathological solution in which every hard-to-fit measurement is marked corrupt. As β_{corr} relaxes, the gate gains freedom to deviate from the corruption prior. The initial value β_0 is the main sensitive hyperparameter of the approach and depends on both the dataset and the SMNR level. Higher SMNR requires a larger β_0: small prediction errors in a very clean environment can be heavily penalised by the clean measurement likelihood, resulting in all being marked as corrupt. Therefore, we need to pin to the corruption prior with a higher β_0 at high SMNR. We report the values used in Table 5.

Stage 3: Full ELBO (epochs 100 to 3000; gradient steps 1600 onwards). We fix $\beta_{\mathrm{dyn}} = \beta_{\mathrm{corr}} = 1$ and optimise the unmodified ELBO for the remainder of training. This is the only stage whose objective matches the variational bound. Refinement in this stage is small; see Figure 9.

Table 5: Initial corruption-KL weight β_0 at the start of Stage 2.

Dataset	SMNR	β_0
Lorenz-63 (identity obs.)	$10 - 20$ dB	5
Lorenz-63 (Gaussian PSF)	20 dB	200
BML	20 dB	1000

To visualise the different stages of training, we show both example trajectories as well as training loss and validation NMSE. In Fig. 8 we show the smoothing estimate on a Lorenz-63 sequence ($p_\varepsilon = 10\%$, SMNR $= 20$ dB) at two checkpoints per stage. Within Stage 1 (top row, epochs 10, 50) the estimate develops from near-interpolation of the measurements into a trajectory that maps the corruptions; the variational posterior is not robust and overfits to the data. The gate is clamped to γ_0 throughout and carries no information.

Within Stage 2 (middle row, epochs 60, 100) the gate learning initiates and β_{corr} relaxes. At epoch 60 the estimate looks similar to the end of Stage 1 and the corruptions are starting to be classified as such. By epoch 100 the gate identifies the corruptions cleanly and the estimate tracks the true state. Within Stage 3 (bottom row, epochs 110, 3000) the change is minor: estimates tighten and γ_t becomes crisper, but no qualitative transition occurs. This matches the quantitative behaviour in Figure 9.

Figure 9 shows the training loss (negative ELBO) and the validation NMSE across all 3000 epochs, with stage boundaries marked. Both curves exhibit the expected three-phase structure. During Stage 1 the ELBO improves rapidly as the dynamics model fits, but the validation NMSE first improves and then degrades: the encoder is committing to a non-robust solution that maps the corruptions. At the Stage 1/Stage 2 boundary, releasing the gate and activating $\mathcal{L}_{\mathrm{corr}}$ with a large β_0 causes a transient increase in the training loss (the corruption KL now contributes a large penalty), but the validation NMSE drops sharply as the gate identifies corruptions. During Stage 3 both curves flatten. The validation NMSE at epoch 3000 is within 2 dB of the value at epoch 100, indicating that most of the useful learning is complete by the end of Stage 2 and the long Stage 3 tail contributes only mild refinement.

F Ablations

In this section, we ablate the design choices of the variational distribution (F.1) and the hyperparameters governing the training of the corruption posterior (F.2). All ablations are done on Lorenz-63 dataset with identity measurement, at SMNR $= 20$ dB, and corruption rate $p_\varepsilon = 10\%$, with five seeds per configuration; all other settings match those outlined in Appendix D.

F.1 Design Choices of the Variational Distribution

We will now ablate three design choices of the variational distribution: the gated input $\mathbf{y}_t^{(\gamma)} = (1 - \gamma_t)\mathbf{y}_t$ (see Eq. (9)), the exclusion of $\mathbf{y}_t$ from the t-th factor of the state posterior (see Eq. (3)), and the coercivity coefficient λ in the gate (see Eq. (8)). Results are reported on Lorenz-63 (identity

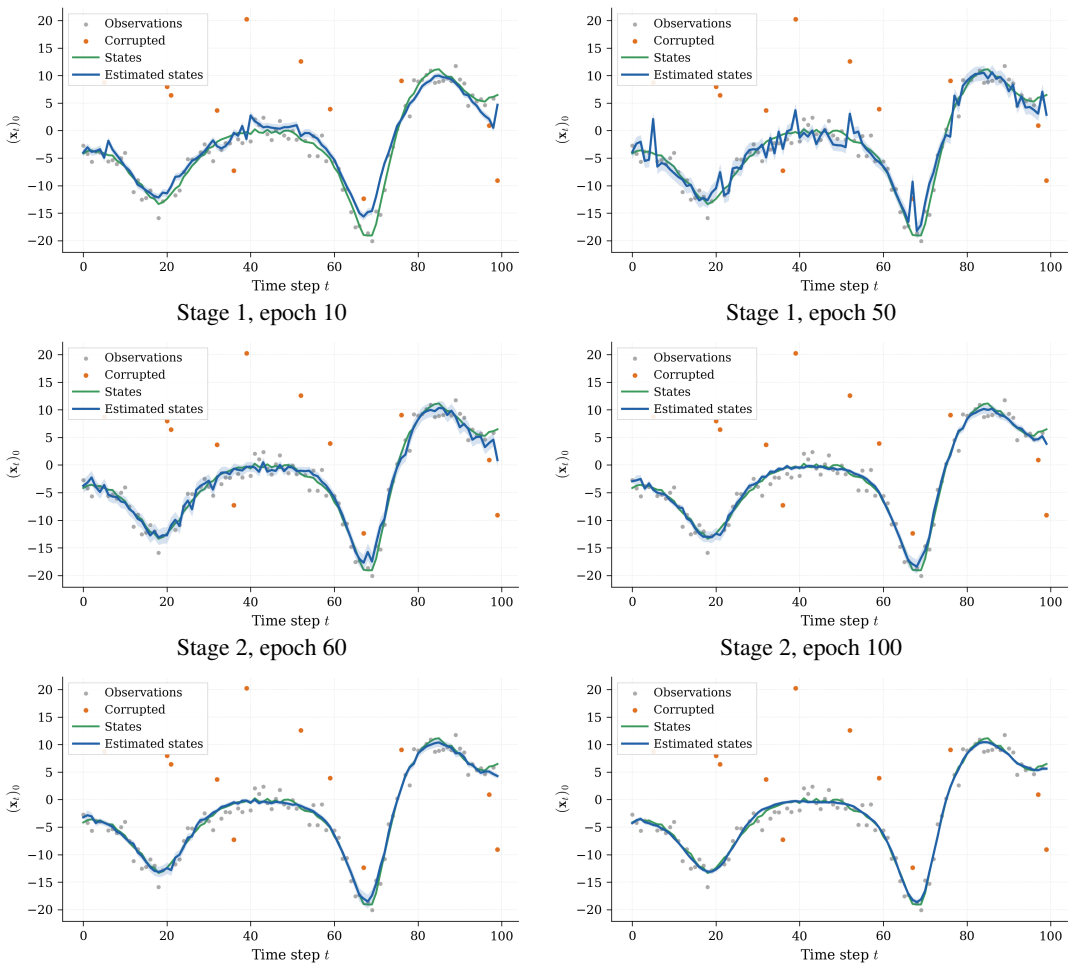


Figure 8: Within-stage training progression on a held-out Lorenz-63 sequence ($p_\varepsilon = 10\%$, SMNR $= 20$ dB). Rows correspond to the three stages. Columns show the early and late checkpoint within each stage. Each panel plots the smoothing estimate (blue) against the ground-truth state (green) and the measurements (grey, with corruptions marked). The gate is clamped during Stage 1 and carries no information; it is initiated during Stage 2 and refines during Stage 3.

measurement) at SMNR $= 20$ dB and $p_\varepsilon = 10\%$, averaged over five seeds. We also report $\mathbb{E}[\gamma_t]$ in addition to AUROC; this helps in understanding the average number of predicted corruptions.

Table 6: Ablation of the variational distribution's design choices on Lorenz-63 (identity, SMNR $= 20$ dB, $p_\varepsilon = 10\%$). Mean $\pm$ std over 5 seeds. The true corruption rate is 0.10.

	NMSE (dB)	AUROC	$\mathbb{E}[\gamma_t]$
RoVaS (full)	-29.02 ± 0.13	0.99 ± 0.00	0.10 ± 0.00
$\lambda = 0$	-29.18 ± 0.23	0.98 ± 0.02	0.10 ± 0.00
No gated input	-29.03 ± 0.26	0.97 ± 0.00	0.10 ± 0.00
No $\mathbf{y}_t$ exclusion	-13.40 ± 0.17	0.72 ± 0.03	0.00 ± 0.00

Setting $\lambda = 0$ and removing the gated input both leave NMSE, AUROC, and gate calibration unchanged. This might be expected, since the corruptions in this benchmark are drawn from a Gaussian fitted to the measurements, and are therefore of typical magnitude. Both mechanisms are instead designed for very large corruptions, $\|\mathbf{y}_t\| \to \infty$, which is the case covered by the theory: $\lambda > 0$ makes the gate coercive (Assumption 3), and, given coercivity, the gated input keeps the encoder

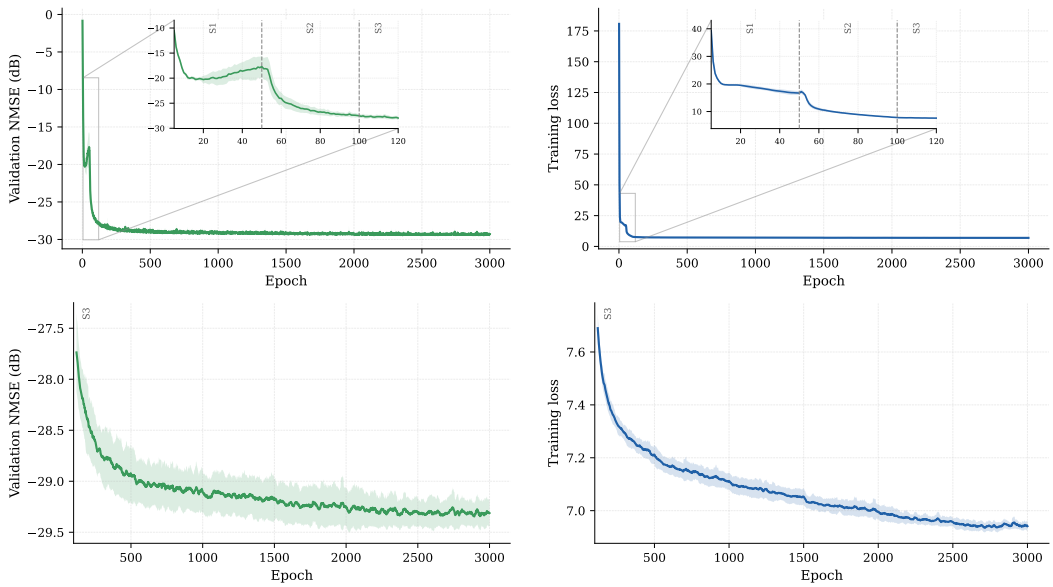


Figure 9: Validation NMSE (left) and training loss (right) over 3000 epochs on Lorenz-63 ($p_\varepsilon = 10\%$, SMNR $= 20$ dB), mean ± 1 standard deviation over $n = 5$ seeds. The training loss is the negative ELBO with the current stage's β values. **Top row:** full training curves with dashed lines at the stage boundaries (epochs 50 and 100). Both curves transition smoothly through the stage boundaries. **Bottom row:** Stage 3 only (epochs 100–3000), where the objective is the unmodified negative ELBO.

input bounded for all corruptions (Lemma 2). Together, they make the guarantee of Proposition 1 independent of the encoder architecture. Without them, the PIF of our GRU-based implementation would only be bounded because the GRUs saturate, as for VaS and VaS-t (Appendix B.4). Removing the $\mathbf{y}_t$-exclusion has a much larger effect: NMSE degrades by 15.6 dB, and the gate collapses to $\mathbb{E}[\gamma_t] = 0$. With direct access to $\mathbf{y}_t$, the encoder fits each measurement independently, and the gate has no incentive to fire. Hence, the $\mathbf{y}_t$-exclusion is needed for the gate to train.

F.2 Training-robustness of the Corruption Variational Distribution

The corruption variational distribution $q_{\boldsymbol{\theta}}(c_t|\mathbf{y}_{1:T})$ is trained jointly with the state variational distribution under the curriculum of Appendix E. Three hyperparameters mainly control the training of the corruption variational distribution: the prior corruption rate ε, the initial corruption-KL weight β_0 in stage 2 of the curriculum, and the scale k of the corruption likelihood $p_{\text{corr}} = \mathcal{N}(\boldsymbol{\mu}_{\text{pool}}, k \cdot \boldsymbol{\Sigma}_w)$. Note that none of them is given by the data-generating process; in particular, RoVaS does not know the true corruption rate. We sweep each of them independently on Lorenz-63 (identity, SMNR $= 20$ dB, $p_\varepsilon = 10\%$) over five seeds, with all other settings as in Appendix D.

Table 7: Sensitivity to the prior corruption rate ε of the corruption prior; the true corruption rate is 10%. Mean $\pm$ std over 5 seeds.

ε	NMSE (dB)	AUROC
0.1%	-29.03 ± 0.07	0.995 ± 0.001
1% (default)	-29.08 ± 0.18	0.978 ± 0.019
5%	-29.23 ± 0.15	0.966 ± 0.003
10% (true rate)	-29.29 ± 0.10	0.968 ± 0.007
25%	-29.34 ± 0.10	0.975 ± 0.006

In Table 7, NMSE varies by at most 0.31 dB across two and a half orders of magnitude in ε, and the AUROC stays at or above 0.966. Using the actual (true) prior parameter, $\varepsilon = 10\%$, we get little to no improvement over the value used in our experiments, $\varepsilon = 1\%$. These results suggest that RoVaS is

robust to the assumed prior corruption rate ε over the tested range. Note, however, that removing the prior corruption model, i.e., removing $\mathcal{L}_{\text{corr}}$ from the training objective (see Eqs. (6) and (20)), results in a degenerate solution: all measurements are predicted to be corrupt, which leads to poor performance.

Table 8: Sensitivity to the initial corruption-KL weight β_0 at $k = 80$. Mean $\pm$ std over 5 seeds.

β_0	NMSE (dB)	AUROC	$\mathbb{E}[\gamma_t]$
0.1	-28.04 ± 0.36	0.99 ± 0.00	0.07 ± 0.01
5 (default)	-29.02 ± 0.13	0.99 ± 0.00	0.10 ± 0.00
10	-29.14 ± 0.18	0.98 ± 0.02	0.10 ± 0.00
100	-29.27 ± 0.31	0.97 ± 0.01	0.10 ± 0.00
1000	-29.00 ± 0.27	0.98 ± 0.02	0.10 ± 0.00

Table 9: Sensitivity to the corruption-likelihood scale k in $p_{\text{corr}} = \mathcal{N}(\boldsymbol{\mu}_{\text{pool}}, k\boldsymbol{\Sigma}_w)$ at $\beta_0 = 5$. The empirical pool covariance corresponds to $k \approx 80$. Mean $\pm$ std over 5 seeds.

k	NMSE (dB)	AUROC	$\mathbb{E}[\gamma_t]$
5	-29.22 ± 0.19	0.98 ± 0.01	0.10 ± 0.00
80 (default)	-29.02 ± 0.13	0.99 ± 0.00	0.10 ± 0.00
500	-28.86 ± 0.11	0.99 ± 0.00	0.09 ± 0.00

NMSE varies by at most 1.3 dB across the β_0 sweep spanning four orders of magnitude, and by at most 0.4 dB across the k sweep spanning two orders of magnitude. Further, it seems that the gate can accurately predict corrupted measurements for all values of β_0 and k, except for very small β_0; there, the gate slightly under-fires ($\mathbb{E}[\gamma_t] = 0.07$), and NMSE degrades by ~ 1 dB. Thus, we conclude that the training of the corruption posterior is robust to order-of-magnitude misspecification of p_{corr} and of the prior corruption rate ε, as well as to different choices of β_0. Hence, if only rough prior knowledge is given on the rate of corruption and the corruption distribution, RoVaS seems robust to potential misalignment.

G Trivial Upper Bound on NMSE

For state space models under the identity measurement map and Gaussian corruption modeling, we use the trivial estimator $\hat{\mathbf{x}}_t = \mathbf{y}_t$ as a reference; that is, we use the raw measurement as the state estimate without any filtering or smoothing. The NMSE achieved by this estimator is a trivial upper bound for any reasonable smoother.

Let $\mathbf{x}_t \in \mathbb{R}^d$ denote the latent state, where $d = d_x = d_y$. Define $\boldsymbol{\mu}_x \triangleq \mathbb{E}[\mathbf{x}_t]$, $P_x \triangleq \mathbb{E}[\|\mathbf{x}_t\|^2]$, and let a denote the SMNR in dB, and $\epsilon \in [0, 1)$ the corruption probability.

At each time step, with probability $(1 - \epsilon)$ the measurement is valid; that is,

$$\mathbf{y}_t = \mathbf{x}_t + \mathbf{w}_t, \qquad \mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, \sigma_w^2 \mathbf{I}_d), \tag{53}$$

and with probability ϵ, the measurement is drawn independently from the corruption distribution with mean $\boldsymbol{\mu}_{\text{c}}$ and covariance $\boldsymbol{\Sigma}_{\text{c}}$. The SMNR is

$$a = 10 \log_{10} \frac{P_x}{d\,\sigma_w^2} \quad \text{(dB)}, \tag{54}$$

or, equivalently, $d\,\sigma_w^2 = P_x \cdot 10^{-a/10}$.

The MSE of the trivial estimator decomposes over valid and corrupted measurements as

$$\mathbb{E}\big[\|\mathbf{y}_t - \mathbf{x}_t\|^2\big] = (1 - \epsilon)\,\mathbb{E}\big[\|\mathbf{w}_t\|^2\big] + \epsilon\,\mathbb{E}\big[\|\mathbf{y}_{\text{c}} - \mathbf{x}_t\|^2\big]. \tag{55}$$

For the first term, $\mathbb{E}[\|\mathbf{w}_t\|^2] = d\,\sigma_w^2$. For the second term, $\mathbf{y}_{\text{c}}$ and $\mathbf{x}_t$ are independent, thus,

$$\begin{aligned} \mathbb{E}\big[\|\mathbf{y}_{\text{c}} - \mathbf{x}_t\|^2\big] &= \mathbb{E}\big[\|\mathbf{y}_{\text{c}}\|^2\big] - 2\,\boldsymbol{\mu}_{\text{c}}^\top \boldsymbol{\mu}_x + P_x \\ &= \operatorname{tr}(\boldsymbol{\Sigma}_{\text{c}}) + \|\boldsymbol{\mu}_{\text{c}}\|^2 - 2\,\boldsymbol{\mu}_{\text{c}}^\top \boldsymbol{\mu}_x + P_x. \end{aligned} \tag{56}$$

Since the corruption distribution is the empirical pool of all measurements, its mean satisfies $\boldsymbol{\mu}_{\mathrm{c}} = (1-\epsilon)\,\boldsymbol{\mu}_x + \epsilon\,\boldsymbol{\mu}_{\mathrm{c}}$; hence, $\boldsymbol{\mu}_{\mathrm{c}} = \boldsymbol{\mu}_x$. Substituting, we have

$$\mathbb{E}\big[\|\mathbf{y}_{\mathrm{c}} - \mathbf{x}_t\|^2\big] = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{c}}) + P_x - \|\boldsymbol{\mu}_x\|^2 = \mathrm{tr}(\boldsymbol{\Sigma}_{\mathrm{c}}) + \mathrm{tr}(\boldsymbol{\Sigma}_x), \tag{57}$$

where $\boldsymbol{\Sigma}_x \triangleq \mathrm{Cov}(\mathbf{x}_t) = \mathbb{E}[\mathbf{x}_t\mathbf{x}_t^\top] - \boldsymbol{\mu}_x\boldsymbol{\mu}_x^\top$ is the state covariance. Further, as $\mathbf{y}_t = \mathbf{x}_t + \mathbf{w}_t$, and $\mathbf{x}_t$ and $\mathbf{w}_t$ are independent, we have $\mathrm{Cov}(\mathbf{y}_t) = \mathrm{Cov}(\mathbf{x}_t) + \sigma_w^2\mathbf{I} = \boldsymbol{\Sigma}_x + \sigma_w^2\mathbf{I}$. Since we define $\boldsymbol{\Sigma}_c = \mathrm{Cov}(\mathbf{y}_t)$, it follows that $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}_x + \sigma_w^2\mathbf{I}$, and

$$\mathbb{E}\big[\|\mathbf{y}_{\mathrm{c}} - \mathbf{x}_t\|^2\big] = 2\mathrm{tr}(\boldsymbol{\Sigma}_x) + d\sigma_w^2 = 2\mathrm{tr}(\boldsymbol{\Sigma}_x) + P_x 10^{-a/10} \tag{58}$$

The NMSE of the trivial estimator is thus

$$\begin{aligned}\mathrm{NMSE}_{\mathrm{trivial}} &= (1-\epsilon)\cdot 10^{-a/10} + \epsilon\frac{2\mathrm{tr}(\boldsymbol{\Sigma}_x) + P_x 10^{-a/10}}{P_x} \\ &= 10^{-a/10} + \epsilon\frac{2\mathrm{tr}(\boldsymbol{\Sigma}_x)}{P_x}\end{aligned} \tag{59}$$

or, in dB,

$$\mathrm{NMSE}^{\mathrm{dB}}_{\mathrm{trivial}} = 10\log_{10}\left(10^{-a/10} + \epsilon\frac{2\mathrm{tr}(\boldsymbol{\Sigma}_x)}{P_x}\right). \tag{60}$$

The quantities (P_x, $\boldsymbol{\Sigma}_x$) can be computed from the stationary distribution of the Lorenz-63 system or estimated empirically; we do the latter.

Note that under no corruption, the trivial estimator gives $\mathrm{NMSE}^{\mathrm{dB}}_{\mathrm{trivial}} = -a = -\mathrm{SMNR}$.