Robotics: Science and Systems 2025
Los Angeles, CA, USA, June 21–25, 2025

Joint State and Noise Covariance Estimation

Kasra Khosoussi*
School of Electrical Engineering and Computer Science
The University of Queensland
St Lucia, QLD, Australia
k.khosoussi@uq.edu.au

Iman Shames
School of Engineering
The Australian National University
Canberra, ACT, Australia
iman.shames@anu.edu.au

***Abstract*—This paper tackles the problem of jointly estimating the noise covariance matrix alongside states (parameters such as poses and points) from measurements corrupted by Gaussian noise and, if available, prior information. In such settings, the noise covariance matrix determines the weights assigned to individual measurements in the least squares problem. We show that the joint problem exhibits a convex structure and provide a full characterization of the optimal noise covariance estimate (with analytical solutions) within joint maximum *a posteriori* and likelihood frameworks and several variants. Leveraging this theoretical result, we propose two novel algorithms that jointly estimate the primary parameters and the noise covariance matrix. Our BCD algorithm can be easily integrated into existing nonlinear least squares solvers, with negligible per-iteration computational overhead. To validate our approach, we conduct extensive experiments across diverse scenarios and offer practical insights into their application in robotics and computer vision estimation problems with a particular focus on SLAM.**

I. Introduction

Maximum likelihood (ML) and maximum *a posteriori* (MAP) are the two most common point-estimation criteria in robotics and computer vision applications such as all variants of simultaneous localization and mapping (SLAM) and bundle adjustment. Under the standard assumption of zero-mean Gaussian measurement noise—and, for MAP, Gaussian priors—these estimation problems reduce to least squares; i.e., finding estimates that minimize the weighted sum of squared errors between observed and expected measurements (and, when priors are available, the weighted squared errors between parameter values and their expected priors).

Each measurement error in the least squares objective is weighted by its corresponding noise information matrix (i.e., the inverse of the covariance matrix). Intuitively, more precise sensors receive "larger" weights, thus exerting greater influence on the final estimate. Moreover, the weight matrices enable the estimator to account for correlations between different components of measurements, preventing "double counting" of information. Therefore, obtaining an accurate estimate of the noise covariance matrix is critical for achieving high estimation accuracy. In addition, the (estimated) noise covariance matrix also determines the (estimated) covariance of the estimated parameter values often used for control and decision making in robotics. Consequently, an inaccurate noise covariance estimate can cause overconfidence or underconfidence in state estimates, potentially leading to poor or even catastrophic decisions.

In principle, the noise covariance matrix can be estimated *a priori* (offline) using a calibration dataset where the true values of the primary parameters (e.g., robot poses) are known (see Remark 5). This can be done either by the sensor manufacturer or the end user. However, in practice, several challenges arise:

1) Calibration is a labor-intensive process and may not always be feasible, particularly when obtaining ground truth for primary parameters requires additional instrumentation.
2) In many cases, raw measurements are preprocessed by intermediate algorithms before being used in the estimation problem (e.g., in a SLAM front-end), making it difficult to model their noise characteristics.
3) The noise characteristics may evolve over time (e.g., due to dynamic environmental factors such as temperature), making the pre-calibrated noise model obsolete.

Due to these challenges, many applications rely on ad hoc noise covariances, such as arbitrary isotropic (or diagonal) covariances, which are either manually set by experts or determined through trial-and-error tuning. Despite being recognized as one of the most critical and widely acknowledged challenges in SLAM [14, Sections III.B, III.G, and V], the problem of noise covariance estimation remains unsolved and understudied in robotics and computer vision literature.

We present, to the best of our knowledge, the first algorithms for online (i.e., during deployment) *joint* ML/MAP estimation of states and noise covariance matrices from noisy measurements and, when available, prior information. Our approach is general and eliminates the need for a separate calibration stage across a broad class of estimation problems beyond SLAM. We analyze the convergence properties of the proposed algorithm and demonstrate that it can be seamlessly integrated into existing sparse nonlinear least squares solvers [3, 12, 19], with negligible computational overhead.

Notation

We use $[n]$ to denote the set of integers from 1 to n. The abbreviated notation $x_{1:n}$ is used to denote $x_1, \ldots, x_n$. The zero matrix (and vector) is denote by 0 where the size should be clear from the context. $\mathbb{S}^d_{\succeq 0}$ and $\mathbb{S}^d_{\succ 0}$ denote the sets of $d \times d$ symmetric positive semidefinite and positive definite real matrices, respectively. For two symmetric real matrices A and B, $A \succeq B$ (resp., $A \succ B$) means $A - B$ is positive

*Also affiliated with CSIRO Robotics, Data61.

semidefinite (resp., positive definite). A_{ij} denotes the (i,j) element of matrix A, and $\mathrm{Diag}(A)$ denotes the diagonal matrix obtained by zeroing out the off-diagonal elements of A. The standard (Frobenius) inner product between $n \times n$ real matrices A and B is denoted by $\langle A, B \rangle \triangleq \mathrm{trace}(A^\top B)$. The Frobenius norm of A is denoted by $\|A\| = \sqrt{\langle A, A \rangle}$. The weighted Euclidean norm of x given a weight matrix $W \succ 0$ is denoted by $\|x\|_W \triangleq \sqrt{x^\top W x}$. The probability density function of the multivariate normal distribution of random variable x with mean vector μ and covariance matrix Σ is denoted by $\mathcal{N}(x; \mu, \Sigma)$.

II. Related Works

We refer the reader to [8, 13, 14, 29] for comprehensive reviews of state-of-the-art estimation frameworks in robotics and computer vision. Sparse nonlinear least squares solvers for solving these estimation problems can be found in [3, 12, 19]. However, all of these works assume that the noise covariance is known beforehand. In contrast, our approach simultaneously estimates both the primary parameters (e.g., robot's trajectory) and the noise covariance matrix directly from noisy measurements. The importance of automatic hyperparameter tuning has recently gained recognition in the SLAM literature; see, e.g., [14, Section V] and [15, 16]. We share this perspective and present, to the best of our knowledge, the first principled approach for ML/MAP measurement covariance estimation in SLAM and related problems.

A. Optimal Covariance Estimation via Convex Optimization

ML estimation of the mean and covariance from independent and identically distributed (i.i.d.) Gaussian samples using the sample mean and sample covariance is a classic example found in textbooks. Boyd and Vandenberghe [7, Chapter 7.1.1] show that covariance estimation in this standard setting and several of its variants can be formulated as convex optimization problems. However, many estimation problems that arise in robotics and other engineering disciplines extend beyond the standard setting. While *noise samples* are assumed to be i.i.d. for each measurement *type* (Section IV-C1), the measurements themselves are *not* identically distributed. Each measurement follows a Gaussian distribution with the corresponding noise covariance matrix and a *unique* mean that depends on an unknown parameter belonging to a manifold. Furthermore, the measurement function varies across different measurements and is often nonlinear. We demonstrate that the noise covariance estimation problem in this more general setting can also be formulated as a convex optimization problem. Similar to [7], we explore several problem variants that incorporate prior information and additional structural constraints on the noise covariance matrix. These variants differ from those studied in [7, Chapter 7.1.1] and admit analytical (closed-form) solutions.

B. Covariance Estimation in Robotics and Computer Vision

Zhan et al. [32] propose a joint pose and noise covariance estimation method for the perspective-n-point (PnP) problem in computer vision. Their approach is based on the iterated (or iterative) generalized least squares (IGLS) method (see [27, Chapter 12.5] and references therein), alternating between pose and noise covariance estimation. They report improvements in estimation accuracy ranging from 2% to 34% compared to baseline methods that assume a fixed isotropic noise covariance. To the best of our knowledge, before [32], IGLS had not been applied in robotics or computer vision.[1] Our work (developed concurrently with [32]) generalizes and extends both IGLS and [32] in several keys ways. First, we prove that the joint ML estimation problem is ill-posed when the sample covariance matrix is singular. This critical problem arises frequently in real-world applications, where the sample covariance matrix can be singular or poorly conditioned (Remark 4). We address this critical issue by (i) constraining the minimum eigenvalue of the noise covariance matrix, and, in our MAP formulation, (ii) imposing a Wishart prior on the noise information matrix. We derive analytical optimal solutions for the noise covariance matrix, conditioned on fixed values of the primary parameters, for MAP and ML joint estimation problems and several of their constrained variants (Theorem 1). These enable the end user to leverage prior information about the noise covariance matrix (from, e.g., manufacturer's calibration) in addition to noisy measurements to solve the joint estimation problem. We propose several algorithms for solving these joint estimation problems and present a rigorous theoretical analysis of their convergence properties. Our formulation is more general than IGLS and we show how our framework can be extended to heteroscedastic measurements, nonlinear (with respect to) noise models, and manifold-valued parameters. Finally, we provide insights into the application of our approach to "graph-structured" estimation problems [18], such as PGO and other SLAM variants. These problems present additional challenges compared to the PnP problem due to the increasing number of primary parameters (thousands of poses in PGO vs. a single pose in PnP) and the sparsity of measurements, which reduce the effective signal-to-noise ratio.

Barfoot et al. [4] and Wong et al. [31] propose an EM-type method for learning the noise covariance matrix as part of a variational inference framework. Similar to our work, they used an Inverse-Wishart prior on the covariance matrix. However, we estimate the covariance matrix by solving the joint MAP/ML estimation problems and their constrained variants. While the covariance estimation formulation and proposed techniques share similarities, our method focuses on widely used MAP/ML *point estimation* rather than obtaining an analytical approximation of the entire posterior, yielding significantly faster solutions (by up to several orders of magnitude based on the statistics reported in [4]). As a result, unlike [4, 31], our method can be readily integrated into existing nonlinear least squares solvers such as [3, 12, 19] in both online and offline settings with a negligible computational overhead.

Lu et al. [20] introduce a covariance autotuning method for object measurements in SLAM, employing an alternating

[1] We were unable to find the original reference for IGLS. However, the concept was already known and analyzed in the econometrics literature in 1970s [21]. IGLS is closely related to the feasible generalized least squares (FGLS) method which also has a long history in econometrics and regression.

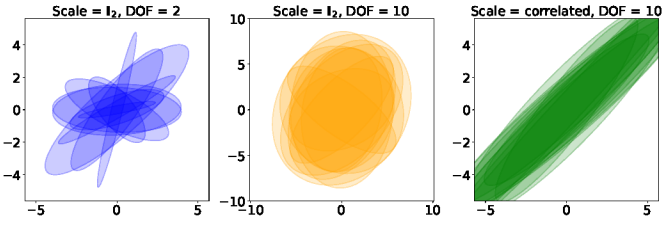


Fig. 1: Confidence ellipses for 10 samples drawn from the Wishart distribution $\mathcal{W}(P; V, \nu)$ with different parameters. The scale matrix V is set to identity in the left and middle plots, and to a correlated matrix in the right plot. The degrees of freedom ν are set to 2 (left) and 10 (middle and right). As ν increases, the samples become more concentrated. Increasing ν changes the scale of samples as well (cf. left and middle).

optimization scheme over states and the variances of a *diagonal* covariance matrix. Despite similarities between [20] and our BCD algorithm, the cost function in [20] differs from standard MAP/ML formulation. Additionally, unlike [20], our approach does not assume that the covariance matrix is diagonal.

Qadri et al. [26] propose a bilevel optimization framework to learn measurement covariances from a calibration dataset with known ground truth. Specifically, they seek the covariance estimate (outer problem) that minimizes the state estimation (inner problem) error. Unlike [26], our method does not directly minimize the estimation error and thus does not require access to the ground truth. Instead, we jointly estimate both states and covariances directly from observed data (and, optionally, prior information on the covariance) in a joint MAP/ML framework. As a result, our method can be used during deployment to learn measurement covariances based on the collected measurements.

C. Noise Covariance Estimation in Kalman Filtering

The problem of identifying process and measurement noise models in Kalman filtering (often referred to as *adaptive* Kalman filtering) has been extensively studied since the late 1960s; see, e.g., [1, 22, 33, 11, 17] and references therein. While our work shares certain similarities with these approaches and their underlying principles, such methods are specifically designed for (approximate, when models are nonlinear) recursive MAP estimation in linear(ized) models within the filtering setting. As a result, they are not readily applicable to batch and smoothing formulations, nonlinear measurement models, sparse large-scale problems, or (nonlinear) manifold-valued states. These features are essential for addressing many estimation problems in robotics and computer vision (see, e.g., state-of-the-art estimation frameworks for SLAM [13]).

III. Problem Statement

Consider the standard problem of estimating an unknown vector $x_{\text{true}} \in \mathcal{M}$ given k noisy m-dimensional measurements $\mathcal{Z} \triangleq \{z_i\}_{i=1}^k$ corrupted by i.i.d. zero-mean Gaussian noise:

$$z_i = h_i(x_{\text{true}}) \boxplus \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0_m, \Sigma_{\text{true}}), \tag{1}$$

where Σ_{true} is the unknown noise covariance. For Euclidean-valued measurements (such as relative position of a landmark with respect to a robot pose), $\boxplus$ reduces to addition in $\mathbb{R}^m$. For matrix Lie group-valued measurements (such as relative pose or orientation between two poses), $\boxplus$ is equivalent to multiplication by $\operatorname{Exp}(\epsilon_i)$ where Exp denotes the matrix exponential composed with the so-called hat operator. We denote the *residual* of measurement z_i evaluated at $x \in \mathcal{M}$ with $r_i(x) \triangleq z_i \boxminus h_i(x)$. As above, $\boxminus$ is subtraction in $\mathbb{R}^m$ for Euclidean measurements, and $\operatorname{Log}(h_i(x)^{-1} z_i)$ in the case of matrix Lie-group-valued measurements where Log is the matrix logarithm composed with the so-called vee operator (in this case, m refers to the dimension of Lie algebra).

In this paper, we refer to x_{true} as the *primary* "parameters" to distinguish them from Σ_{true}. To simplify the discussion, we first consider the case where all measurements share the same noise distribution, meaning there is a single noise covariance matrix Σ_{true} in (1). See Section IV-C1 for extensions to more general cases. In robotics and computer vision applications, $\mathcal{M}$ is typically a (smooth) product manifold comprised of $\mathbb{R}^d$ and $\mathrm{SO}(d)$ components ($d \in \{2, 3\}$) and other real components (e.g., time offsets, IMU biases). We assume the measurement functions $h_i : \mathcal{M} \to \mathbb{R}^m$ are smooth. This standard model (along with extensions in Section IV) is quite general, capturing many estimation problems such as SLAM (with various sensing modalities and variants), PGO, point cloud registration (with known correspondences), perspective-n-point, and bundle adjustment.

In this paper, we are interested in the setting where the noise covariance matrix $\Sigma_{\text{true}} \succ 0$ is unknown and must be estimated *jointly* with x_{true} based on the collected measurements $\{z_i\}_{i=1}^k$. For convenience, we formulate the problem in the *information form* and estimate the noise information (or precision) matrix $P_{\text{true}} \triangleq \Sigma_{\text{true}}^{-1}$. Without loss of generality, we assume a non-informative prior on x_{true} which is almost always the case in real-world applications (effectively treating it as an unknown "parameter"). We assume a Wishart prior on the noise information matrix P_{true} and denote its probability density function with $\mathcal{W}(P; V, \nu)$ where $V \in \mathbb{S}^m_{\succ 0}$ is the scale matrix and the integer $\nu \geq m + 1$ is the number of degrees of freedom; Figure 1 illustrates random samples drawn from $\mathcal{W}(P; V, \nu)$. The Wishart distribution is the standard choice for prior in Bayesian statistics for estimating the information matrix from multivariate Gaussian data (in part due to conjugacy); see, e.g., [5, Eq. (2.155)]. In Algorithm 1, we propose a procedure for setting the parameters of the prior, V and ν, when the user has access to a prior estimate Σ_0 for the noise covariance matrix Σ_{true} (e.g., from prior calibration by the manufacturer of the sensor); see Appendix A for a justification.

The joint MAP estimator for x_{true} and P_{true} are the maximizers of the posterior. Invoking the Bayes' Theorem (and omitting

Algorithm 1 Setting Wishart Parameters via Mode Matching

procedure PRIORMODEMATCHING(Σ_0, w_{prior}, k)
  // $\Sigma_0 \succ 0$ is a prior estimate for Σ_{true}
  // $w_{\text{prior}} > 0$ is the weight assigned to prior relative to measurement likelihood
  // m is the dimension of Σ_0
  // k is the number of measurements
  $V \leftarrow (w_{\text{prior}}\, k\, \Sigma_0)^{-1}$
  $\nu \leftarrow w_{\text{prior}}\, k + m + 1$
  return (V, ν)
end procedure

the normalizing constant) results in the following problem:

$$\underset{x\in\mathcal{M},P\succeq 0}{\text{maximize}} \quad \underbrace{\mathcal{W}(P;V,\nu)}_{\text{prior on } P_{\text{true}}} \overbrace{\prod_{i=1}^{k}}^{\text{measurement likelihood}} \mathcal{N}(z_i; h_i(x), P^{-1})\,. \tag{2}$$

For any $x \in \mathcal{M}$, define the *sample covariance at* x as follows:

$$S(x) \triangleq \frac{1}{k}\sum_{i=1}^{k} r_i(x) r_i(x)^\top \succeq 0. \tag{3}$$

Then, computing the negative log-posterior, omitting normalizing constants, dividing the objective by $k+\nu-m-1$, and using the cyclic property of trace yields the following equivalent problem.

Problem 1 (Joint MAP)**.**

$$\underset{x\in\mathcal{M},P\succeq 0}{\text{minimize}}\; F(x,P) \triangleq -\log\det P + \langle M(x), P\rangle\,, \tag{4}$$

where

$$M(x) \triangleq \frac{kS(x) + V^{-1}}{k+\nu-m-1} \succ 0. \tag{5}$$

We also introduce and study three new variants of Problem 1 by imposing additional (hard) constraints on the noise covariance matrix Σ. These constraints enable the user to enforce prior structural information about the covariance.

In the first variant, Σ (and thus P) is forced to be diagonal. This allows the user to enforce independence between noise components.

Problem 2 (Diagonal Joint MAP)**.**

$$\begin{aligned} \underset{x\in\mathcal{M},P\succeq 0}{\text{minimize}} \quad & F(x,P) \\ \text{subject to} \quad & P \text{ is diagonal.} \end{aligned} \tag{6}$$

In the second variant, we constrain the eigenvalues of $\Sigma = P^{-1}$ to $[\lambda_{\min}, \lambda_{\max}]$ where $\lambda_{\max} \geq \lambda_{\min} > 0$. These eigenvalues specify the minimum and maximum variance along all directions in $\mathbb{R}^m$ (i.e., variance of all normalized linear combinations of noise components), Therefore, this constraint allows the user to incorporate prior knowledge about the sensor's noise limits. We will also show that in many real-world instances (especially with a weak or no prior), constraining the smallest eigenvalue of Σ (or largest eigenvalue of P) is essential for preventing Σ from collapsing to zero.

Problem 3 (Eigenvalue-constrained Joint MAP)**.**

$$\begin{aligned} \underset{x\in\mathcal{M},P\succeq 0}{\text{minimize}} \quad & F(x,P) \\ \text{subject to} \quad & \lambda_{\max}^{-1} I \preceq P \preceq \lambda_{\min}^{-1} I. \end{aligned} \tag{7}$$

The last variant imposes both constraints simultaneously.

Problem 4 (Diagonal Eigenvalue-constrained Joint MAP)**.**

$$\begin{aligned} \underset{x\in\mathcal{M},P\succeq 0}{\text{minimize}} \quad & F(x,P) \\ \text{subject to} \quad & \lambda_{\max}^{-1} I \preceq P \preceq \lambda_{\min}^{-1} I, \\ & P \text{ is diagonal.} \end{aligned} \tag{8}$$

Remark 1 (Joint ML Estimation)**.** One can resort to joint ML estimation when a prior distribution is not available. The negative log-likelihood cost function arising in the joint ML estimation problem and its constrained variants (i.e., with eigenvalue and/or diagonal constraints) takes a form similar to $F(x,P)$ defined in (4):

$$F_{\text{ML}}(x,P) \triangleq -\log\det P + \langle S(x), P\rangle, \tag{9}$$

where $S(x)$ denotes the sample covariance matrix (3). The *unconstrained* joint ML problem was also derived in [32].

Remark 2 (Fixing Noise Covariance)**.** It is worth noting that by fixing $P = P_0$, Problem 1 reduces to nonlinear least squares over x and can be (locally) solved using existing solvers [3, 12, 19]. To see this, note that the only term in the objective function (4) that is a function of x can be written as:

$$\begin{aligned} \langle M(x), P_0\rangle &= \gamma^{-1}\langle kS(x), P_0\rangle + \text{const} && (10)\\ &= \gamma^{-1}\left\langle \sum_{i=1}^{k} r_i(x) r_i(x)^\top, P_0 \right\rangle + \text{const} && (11)\\ &= \gamma^{-1}\sum_{i=1}^{k} \|r_i(x)\|_{P_0}^2 + \text{const}, && (12) \end{aligned}$$

where $\gamma \triangleq k+\nu-m-1$ is the constant that appears in the denominator of (5).

IV. Optimal Information Matrix Estimation

Problems 1-4 are in general *non*-convex in x because of the residuals r_i's and, when estimating rotations (e.g., in SLAM), the $\mathrm{SO}(d)$ constraints imposed on rotational components of x. In this section, we reveal a convexity structure in these problems and provide analytical (globally) optimal solutions for estimating the covariance matrix for a given $x \in \mathcal{M}$.

A. Inner Subproblem: Covariance Estimation

Problems 1-4 can be separated into two nested subproblems: an inner subproblem and an outer subproblem, i.e.,

$$\begin{aligned} \underset{x\in\mathcal{M}}{\text{minimize}} \quad & \min_{P\succeq 0} \quad F(x,P) \\ & \text{subject to} \quad \text{appropriate constraints on } P, \end{aligned} \tag{13}$$

where the constraints for each problem are given in Problems 1-4. The inner subproblem focuses on minimizing the objective function over the information matrix P and as a function of $x \in \mathcal{M}$. The outer subproblem minimizes the overall objective function by optimizing over $x \in \mathcal{M}$ when the objective is evaluated at the optimal information matrix obtained from the inner subproblem.

Remark 3. For (13) to be well defined, we require the constraint set to be closed and the cost function to be bounded from below. As $M(x)$ is positive definite by construction, the cost is bounded from below. The positive semidefinite constraint guarantees the closedness of the constraint set. However, due to the presence of the $\log\det P$ term in the cost function, if a solution exists, the cone constraints are not active at this solution. Thus, a singular $P^\star$ can never be a solution to this problem.

B. Analytical Solution to the Inner Subproblem

The inner subproblem for a fixed x can be written as

$$\begin{aligned} \underset{P\succeq 0}{\text{minimize}} \quad & -\log\det P + \langle M(x), P\rangle \\ \text{subject to} \quad & \text{appropriate constraints on } P. \end{aligned} \tag{14}$$

Proposition 1. *For any $x \in \mathcal{M}$, the inner subproblem (14) is a convex optimization problem and has at most one optimal solution.*

Proof: See Appendix B. ∎

The following theorem provides analytical expressions for the unique optimal solutions to the inner subproblems (14) in Problems 1-4.

Theorem 1 (Analytical Solution to the Inner Problem)**.** *Consider the inner problem (14) for a given $x \in \mathcal{M}$. The following statements hold:*

1) *Inner Subproblem in Problem 1: The optimal solution is given by*

$$P^\star(x) = M(x)^{-1}. \tag{15}$$

2) *Inner Subproblem in Problem 2: The optimal solution is given by*

$$P^\star(x) = \mathrm{Diag}(M(x))^{-1} \tag{16}$$

3) *Inner Subproblem in Problem 3: Let*

$$M(x) = U(x)D(x)U(x)^\top \tag{17}$$

be an eigendecomposition of $M(x)$ where $U(x)$ and $D(x)$ are orthogonal and diagonal, respectively. The optimal solution is given by

$$P^\star(x) = U(x)\Lambda(x)U(x)^\top \tag{18}$$

where $\Lambda(x)$ is a diagonal matrix with the following elements:

$$\Lambda_{ii}(x) = \begin{cases} \lambda_{\min}^{-1} & D_{ii}(x) \in [0, \lambda_{\min}], \\ D_{ii}(x)^{-1} & D_{ii}(x) \in (\lambda_{\min}, \lambda_{\max}), \\ \lambda_{\max}^{-1} & D_{ii}(x) \in [\lambda_{\max}, \infty). \end{cases} \tag{19}$$

4) *Inner Subproblem in Problem 4: The optimal solution is a diagonal matrix with the following elements:*

$$P^\star_{ii}(x) = \begin{cases} \lambda_{\min}^{-1} & M_{ii}(x) \in [0, \lambda_{\min}], \\ M_{ii}(x)^{-1} & M_{ii}(x) \in (\lambda_{\min}, \lambda_{\max}), \\ \lambda_{\max}^{-1} & M_{ii}(x) \in [\lambda_{\max}, \infty). \end{cases} \tag{20}$$

Proof: See Appendix C. ∎

As we saw in Remark 1, the objective function in the ML formulation for estimating the noise covariance can be written as $-\log\det P + \langle S(x), P\rangle$ which has a similar form to the cost function in (14). However, unlike $M(x)$, the sample covariance $S(x)$ can become singular. Therefore, Theorem 1 readily applies to the ML case (and its constrained variants) with an important exception: without the constraint $\Sigma \succeq \lambda_{\min} I$, if $S(x)$ becomes singular, the problem becomes unbounded from below and thus ML estimation becomes ill-posed. This is formally proved in the following theorem.

Theorem 2. *Consider the (unconstrained) joint ML estimation problem (Remark 1),*

$$\underset{x\in\mathcal{M}, P\succeq 0}{\text{minimize}} \quad -\log\det P + \langle S(x), P\rangle\,. \tag{21}$$

If $S(x)$ is singular, this problem (and the corresponding "inner subproblem") is unbounded below and thus does not have a solution.

Proof: See Appendix D. ∎

Remark 4. Theorem 2 shows that, without the Wishart prior, if $S(x)$ is singular, the unconstrained joint estimation problem (specifically, without $\Sigma \succeq \lambda_{\min} I$) becomes ill-posed. Similarly, singularity of $S(x)$ can make Problems 1 and 2 ill-conditioned when the prior is weak (i.e., w_{prior} is small) as the corresponding objective function becomes very sensitive in certain directions that correspond to the smallest eigenvalues of $M(x)$. The sample covariance matrix $S(x)$ could be singular in many different situations, but two particular cases that can lead to singularity are as follows:

1) When $k < m$ (i.e., there are not enough measurements to estimate Σ_{true}), $S(x)$ will be singular at any $x \in \mathcal{M}$.
2) Specific values of $x \in \mathcal{M}$ can lead to singularity in some problems. For example, consider the problem of estimating odometry noise covariance matrix in PGO. Let x_{odo} be the odometry estimate obtained from composing odometry

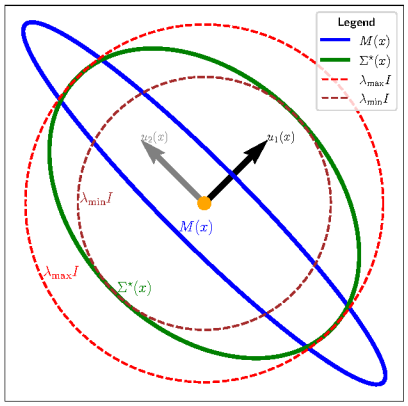


Fig. 2: The confidence ellipses corresponding to $M(x)$ (in blue) and the optimal covariance matrix $\Sigma^\star(x) = P^\star(x)^{-1}$ as given in (18) for Problem 3 (in green). The circles show the confidence ellipses associated to $\lambda_{\min} I$ and $\lambda_{\max} I$. Note that the principal axes $u_1(x)$ and $u_2(x)$ of $M(x)$ remain unchanged, while its radii (along the principal axes) are adjusted to fit within the bounds of the circles.

measurements. At $x = x_{\text{odo}}$, the residuals $r_i(x_{\text{odo}})$ (and thus $S(x_{\text{odo}})$) will be zero.

Theorem 1 has a clear geometric interpretation. For instance, the optimal covariance matrix $\Sigma^\star(x) = P^\star(x)^{-1}$ for Problem 3 (18) has the following properties: (i) it preserves the eigenvectors of $M(x)$, which define the principal axes of the associated confidence ellipsoid; (ii) it matches $M(x)$ along directions corresponding to eigenvalues within the range $[\lambda_{\min}, \lambda_{\max}]$; and (iii) it only adjusts the radii of the ellipsoid along the remaining principal axes to satisfy the constraint. This is visualized in Figure 2. Similarly, the confidence ellipsoid associated to (16) is obtained by projecting $M(x)$ onto the set of diagonal matrices (i.e., axis-aligned ellipsoids). Finally, in the case of (20), the radii of the axis-aligned ellipsoid are adjusted to satisfy the eigenvalue constraint.

Remark 5 (Calibration)**.** The inner subproblem (14) also arises in offline calibration. In this context, a calibration dataset $\mathcal{Z}^{\text{cal}}$ is provided with ground truth $x^{\text{cal}}_{\text{true}}$ (or a close approximation). The objective is to estimate the noise covariance matrix Σ_{true} based on the calibration dataset, which can then be used for future datasets collected with the same sensors. Note that this approach differs from the joint problems (Problems 1-4), where x_{true} and Σ_{true} must be estimated *simultaneously* without the aid of a calibration dataset containing ground truth information. Applying Theorem 1 at $x = x^{\text{cal}}_{\text{true}}$ directly provides the optimal noise covariance matrix for this scenario. If calibration is performed using an approximate ground truth $x^{\text{cal}} \approx x^{\text{cal}}_{\text{true}}$, the estimated noise covariance matrix will be biased.

C. Two Important Extensions

1) **Heteroscedastic Measurements**: For simplicity, we have so far assumed that measurement noises are identically distributed (i.e., homoscedastic). However, in general, there may be T distinct *types* of measurements (e.g., obtained using different sensors in sensor fusion problems, odometry vs. loop closure in PGO, etc), where the noises corrupting each type are identically distributed. In such cases, the inner problem (14) decomposes into T independent problems (with different $M(x)$), solving each yields one of the T noise information matrices. Theorem 1 can then be applied independently to each of these problems to find the analytical optimal solution for each noise information matrix as a function of x.

2) **Preprocessed Measurements and Non-Additive Noise**: In SLAM, raw measurements are often preprocessed and transformed *nonlinearly* into standard models supported by popular solvers. For instance, raw range-bearing measurements (corrupted by additive noise with covariance Σ_{true}) are often expressed in Cartesian coordinates. As a result, although the raw measurements generated by the sensor have identically distributed noise, the *transformed measurements* that appear in the least squares problem may have different covariances because of the nonlinear transformation. Let Σ_{true} be the covariance of raw measurements, and Σ_i be the covariance of the ith transformed measurement. In practice, Σ_i is approximated by linearization, i.e., $\Sigma_i \approx J_i \Sigma_{\text{true}} J_i^\top$ in which J_i is the (known) Jacobian of the transformation. It is easy to verify that Theorem 1 readily extends to this case when J_i's are full-rank square matrices by replacing the sample covariance $S(x)$ as defined in (3) with

$$\widetilde{S}(x) \triangleq \frac{1}{k} \sum_{i=1}^{k} J_i^{-1} r_i(x) r_i(x)^\top J_i^{-\top}. \tag{22}$$

A similar technique can be used when measurements are affected by zero-mean Gaussian noise in a *nonlinear* manner, i.e., $z_i = h_i(x_{\text{true}}, \epsilon_i)$ (in that case, the Jacobians of h_i's with respect to noise will in general depend on x).

Algorithm 2 Variable Elimination for Joint MAP

procedure VARIABLEELIMINATION
  Use a local optimization method to find

$$x^\star \in \operatorname*{argmin}_{x \in \mathcal{M}} -\log\det P^\star(x) + \langle M(x), P^\star(x) \rangle$$

  return $\left(x^\star, P^\star(x^\star)\right)$
end procedure

V. Algorithms for Joint Estimation

In principle, one can employ existing constrained optimization methods to directly solve (locally) Problems 1-4. However, this requires substantial modification of existing highly optimized solvers such as [3, 12, 19] and thus may not be ideal. In this section, present two types of algorithms that leverage Theorem 1 for solving the joint estimation problems.

A. Variable Elimination

We can eliminate P from the joint problems (13) by plugging in the optimal information matrix (as a function of x) $P^\star(x)$

for the inner subproblem (14) provided in Theorem 1. This leads to the following *reduced* optimization problem in x:

$$\underset{x\in\mathcal{M}}{\text{minimize}} \quad -\log\det P^\star(x) + \langle M(x), P^\star(x)\rangle. \tag{23}$$

Therefore, if $x^\star \in \mathcal{M}$ is an optimal solution for the above problem, then $x^\star$ and $P^\star(x^\star)$ are MAP estimates for x_{true} and P_{true}, respectively. This suggests the simple procedure outlined in Algorithm 2. Note that the reduced problem (23) (like Problems 1-4) is a *non*-convex optimization problem, and thus the first step in Algorithm 2 is subject to local minima.

Remark 6 (Reduced Problem for Problems 1 and 2)**.** The objective function of the reduced problem (23) further simplifies in the case of Problems 1 and 2. Specifically, for any $x \in \mathcal{M}$, the linear term in the objective (i.e., $\langle M(x), P^\star(x)\rangle$) is constant for the values of $P^\star(x)$ given in (15) and (16);

$$P^\star(x) = M(x)^{-1} \Rightarrow \langle M(x), P^\star(x)\rangle = m, \tag{24}$$

$$P^\star(x) = \operatorname{Diag}\big(M(x)\big)^{-1} \Rightarrow \langle M(x), P^\star(x)\rangle = m. \tag{25}$$

Therefore, in these cases the reduced problem further simplifies to the following:

$$\underset{x\in\mathcal{M}}{\text{minimize}} \quad \log\det M(x), \tag{26}$$

$$\underset{x\in\mathcal{M}}{\text{minimize}} \quad \log\det \operatorname{Diag}(M(x)). \tag{27}$$

These simplified problems have an intuitive geometric interpretation (as noted in [32] for the unconstrained ML estimation case): the MAP estimate of x_{true} is the value of $x \in \mathcal{M}$ that minimizes the volume of confidence ellipsoid (in Problem 1) and the volume of the projected (onto the standard basis) ellipsoid (in Problem 2) characterised by $M(x)$.

The variable elimination algorithm has the following practical drawbacks:

1) In many real-world estimation problems, the residuals r_i are typically sparse, meaning each measurement depends on only a small subset of elements in x. Exploiting this sparsity is essential for solving large-scale problems efficiently. However, this sparse structure is generally lost after eliminating P in the reduced problem (23).
2) Popular solvers in robotics and computer vision such as [3, 12, 19] are primarily nonlinear least squares solvers that assume Σ_{true} is known and focus solely on optimizing x. Consequently, these highly optimized tools cannot be directly applied to solve the reduced problem in (23).

B. Block-Coordinate Descent

In this section we show how block-coordinate descent (BCD) methods can be used to solve the problem of interest. A BCD-type algorithm alternates between the following two steps until a stopping condition (e.g., convergence) is satisfied:

1) Fix P to its most recent value and minimize the joint MAP objective function in Problems 1-4 with respect to $x \in \mathcal{M}$. This results in a standard nonlinear least squares problem (where residuals are weighted by P) over $\mathcal{M}$, which can be (locally) solved using existing solvers such as [3, 12, 19] (see Remark 2).
2) Fix x to its most recent value and minimize the joint MAP objective function with respect to $P \succeq 0$, subject to the constraints in (14). This step reduces to solving the inner subproblem for which we have analytical optimal solutions $P^\star(x)$ provided by Theorem 1.

Two variants of this procedure are shown in Algorithms 3 and 4. In Step 1 of Algorithm 3, $R(x,P) \triangleq \langle M(x), P\rangle$ denotes the component of the joint cost function $F(x,P)$ that depends on x. As demonstrated in Remark 2, minimizing this function with respect to x for a fixed P is equivalent to minimizing the associated weighted nonlinear least squares objective. Moreover, while Algorithm 3 uses Riemannian gradient descent to update x in Step 1, in principle one can use any (trust-region or line-search) optimization method that produces a descent iteration such as those already implemented in [3, 12, 19].[2]. Algorithm 4 is applicable to problems where in Step 1 of BCD one can exactly minimize $F(x,P)$ over x and find the (unique) minimizer x^t for the current value of P.

The BCD algorithms of the type considered here address the limitations of Algorithm 2. Specifically, the problem in the first step can be readily solved using standard solvers widely used in robotics and computer vision, which are also capable of exploiting sparsity in residuals. The second step is highly efficient, as it only requires computing the $m \times m$ matrix $M(x)$ as defined in (5), which can be done in $O(k\cdot m^2)$ time. In the case of Problem 3, one must also compute the eigendecomposition of $M(x)$. In practice, m (the dimension of the residuals) is typically a small constant (i.e., $m = O(1)$; e.g., in 3D PGO, $m = 6$), and therefore the overall time complexity of the second step of BCD is $O(k)$, i.e., linear in the number of measurements. In a 2D PGO problem (i.e., $m = 3$) with $k = 5{,}598$ measurements, updating the information matrix in Step 2 takes about one millisecond on a laptop CPU.

Before studying the convergence properties of these algorithms we introduce the necessary assumption below. See Appendix E for background information.

Assumption 1. Let $\mathcal{P}$ denote the constraint set for the noise information matrix P. The sets $\mathcal{M}$ and $\mathcal{P}$ are closed and nonempty and the function F is differentiable and its level set $\{(x,P) : F(x,P) \le \gamma\}$ is bounded for every scalar γ.

Assumption 2 (Lipschitz Smoothness in x)**.** Function $R(x,P)$ is continuously differentiable and Lipschitz smooth in x, i.e., there exists a positive constant L such that for all $(\xi,\Pi) \in \mathcal{M}\times\mathcal{P}$ and all $\zeta \in \mathcal{M}$:

$$\|\nabla_x R(\xi,\Pi) - \nabla_x R(\zeta,\Pi)\| \le L\|\xi-\zeta\|. \tag{28}$$

Next, we state the convergence result for Algorithm 3 when applied to Problems 3 and 4.

Theorem 3. *Let $\mathcal{M}$ and $\mathcal{P}$ be compact submanifolds of the Euclidean space. Under Assumptions 1 and 2 and setting*

[2]We analyze the convergence properties of Algorithm 3 in Theorem 3 when Riemannian gradient descent is used in Step 1.

Algorithm 3 Hybrid Block-Coordinate Descent for Joint MAP

procedure BCD(x_{init})
  $t \leftarrow 0$
  $x^0 \leftarrow x_{\text{init}}$
  // Initialize the information matrix
  $P^0 \leftarrow P^\star(x_{\text{init}})$
  while $t \leq \tau$ **do**
    $t \leftarrow t + 1$
    // Step 1: Update x^t using a descent step using retraction $\text{Retr}_{x^{t-1}}(\cdot)$, Riemannian gradient $\text{grad}_x R(\cdot,\cdot)$ with respect to x, and step-size η where $R(x,P) \triangleq \langle M(x), P\rangle$; see Theorem 3 and Appendix E.
    $x^t \leftarrow \text{Retr}_{x^{t-1}}\left(-\eta \,\text{grad}_x R(x^{t-1}, P^{t-1})\right)$
    // Step 2: optimize P (Theorem 1)
    $P^t \leftarrow P^\star(x^t)$
  end while
  return (x^t, P^t)
end procedure

Algorithm 4 Block-Exact BCD for Joint MAP

procedure BCD(x_{init})
  $t \leftarrow 0$
  // Initialize the information matrix
  $P^0 \leftarrow P^\star(x_{\text{init}})$
  while not converged **do**
    $t \leftarrow t + 1$
    // Step 1: optimize x
    $x^t \in \text{argmin}_{x\in\mathcal{M}} \ \frac{1}{2}\sum_{i=1}^k \|r_i(x)\|^2_{P^{t-1}}$
    // Step 2: optimize P (Theorem 1)
    $P^t \leftarrow P^\star(x^t)$
  end while
  return (x^t, P^t)
end procedure

$\eta = 1/\widetilde{L}$ with $\widetilde{L}$ defined in Lemma 1 in Appendix E, for the sequence $\{(x^t, P^t)\}$ generated by Algorithm 3 we have

$$\min_{t\in[\tau]} \|\text{grad}\, F(x^t, P^t)\| \leq C\sqrt{\frac{F(x_{\text{init}}, P_{\text{init}}) - F(x^\star, P^\star)}{\tau}}, \tag{29}$$

where $C = \sqrt{2\widetilde{L}}(1+\sqrt{2}\alpha)$ where α is given in Lemma 1.

Proof: See Appendix F. ∎

As can be seen in Algorithm 4, one might be able to solve the optimization problems associated with each of the coordinate blocks uniquely and exactly. For example, for the case where the residuals $r_i(x)$ are affine functions of x and $\mathcal{M}$ is convex, the optimization problem associated with x has a unique solution (assuming a non-singular Hessian) and can be solved exactly. This case satisfies the following assumption.

Assumption 3. For all $\Pi \in \mathcal{P}$ and all $\xi \in \mathcal{M}$, the following problems have unique solutions:

$$\min_{x\in\mathcal{M}} F(x, \Pi), \quad \min_{P\in\mathcal{P}} F(\xi, P). \tag{30}$$

Theorem 4. *Under Assumptions 1 and 3, the sequence $\{(x^t, P^t)\}$ generated by Algorithm 4 is bounded and has limit points. Moreover, every limit point $(x^\star, P^\star)$ is a local minimum of the optimization problem.*

Proof: The proof of the theorem follows directly from [24, Theorem 1]. ∎

VI. Experiments

A. Linear Measurement Model

We first evaluate the algorithms on a simple linear measurement model. Although the joint problem is still *non*-convex in this scenario, the problem is (strictly) convex in x and P separately, enabling exact minimization in both steps of BCD (Algorithm 4) with analytical solutions (i.e., linear least squares and (14)).

Setup: In these experiments, $\mathcal{M} = \mathbb{R}^{20}$ and x_{true} is set to the all-ones vector. We generated $k = 50$ measurements, where the dimension of each measurement is $m = 5$. Measurement z_i is generated according to

$$z_i = H_i x_{\text{true}} + \epsilon_i, \quad i \in [k]. \tag{31}$$

Each measurement function $H_i \in \mathbb{R}^{5\times 20}$ is a random matrix drawn from the standard normal distribution. The Measurement noise $\epsilon_i \sim \mathcal{N}(0, \Sigma_{\text{true}})$ where

$$\Sigma_{\text{true}} = \Sigma^{\text{base}} + \sigma^2 I, \tag{32}$$

in which $\Sigma^{\text{base}} \succeq 0$ is a fixed random covariance matrix, and $\sigma^2 \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ is a variable that controls the noise level in our experiments. We did not impose a prior on P, leading to unconstrained joint ML estimation of x_{true} and Σ_{true}.

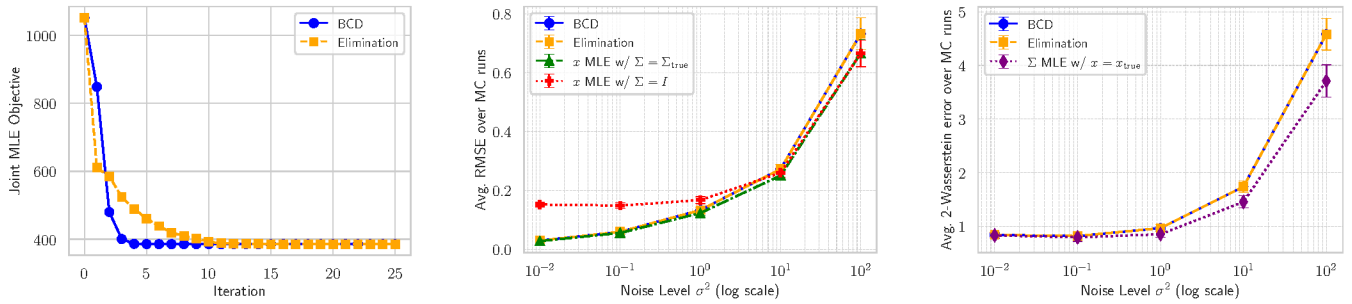
We conducted 50 Monte Carlo simulations per noise level, each with a different noise realization. In each trial, we generated measurement noise according to the model described above and applied Elimination and BCD (Algorithms 2 and 4) to estimate x_{true} and Σ_{true} under the Problem 1 formulation. Both algorithms were initialized with $x_{\text{init}} = 0$ and executed for up to 25 iterations. The Elimination algorithm uses the Limited-memory BFGS solver from SciPy [30].

Metrics: We use the root mean square error (RMSE) to evaluate the accuracy of estimating x_{true}. For covariance estimation accuracy, we use the 2-Wasserstein distance between the true noise distribution $\mathcal{N}(0, \Sigma_{\text{true}})$ and the estimated noise distribution $\mathcal{N}(0, \Sigma^\star)$:

$$\mathfrak{W}_2(\mathcal{N}_{\text{true}}, \mathcal{N}^\star) = \sqrt{\text{trace}\Big(\Sigma_{\text{true}} + \Sigma^\star - 2\big(\Sigma_{\text{true}}^{\frac{1}{2}}\Sigma^\star\Sigma_{\text{true}}^{\frac{1}{2}}\big)^{\frac{1}{2}}\Big)}. \tag{33}$$

Results: The results are shown in Figure 3.

Figure 3a illustrates the value of the objective function in Problem 1 during one of the Monte Carlo simulations ($\sigma = 0.1$). The objective value for BCD is recorded after updating x (Step 1 of Algorithm 4). This figure demonstrates that both methods eventually converge to the same objective value, although BCD exhibits faster convergence.



(a) Objective value over iterations

(b) Average RMSE in estimating x_{true}

(c) Average 2-Wasserstein error for noise covariance

Fig. 3: Results of experiments with linear measurement models (Section VI-A). The results shown in Figures 3b and 3c are averaged over 50 Monte Carlo (MC) runs. The error bars in these figures represent the 95% confidence intervals.

Figure 3b presents the average RMSE for the solution $x^\star$, obtained by Elimination and BCD (at their final iteration), across the Monte Carlo simulations for various noise levels. The figure also includes average RMSEs for estimates obtained using fixed covariance matrices: $\Sigma = \Sigma_{\text{true}}$ (i.e., the true noise covariance) and $\Sigma = I$ (i.e., an arbitrary identity matrix often used in practice when Σ_{true} is unknown). The results show that Elimination and BCD achieve identical RMSEs across all noise levels, with RMSE increasing as the noise level rises. Notably, under low noise, the accuracy of the solutions produced by these algorithms matches that achieved when the true covariance matrix Σ_{true} is known. This indicates that the algorithms can accurately estimate x_{true} without prior knowledge of Σ_{true}. The gap between the RMSEs widens as the noise level increases, which is expected because the algorithms must jointly estimate x_{true} and Σ_{true} under a low signal-to-noise ratio. Nonetheless, the RMSE trends consistently across noise levels.

Furthermore, the results highlight that using a naïve approximation of the noise covariance matrix (i.e., fixing $\Sigma = I$) leads to a poor estimate $x^\star$. However, as the noise level increases, the RMSE for this approximation eventually approaches that of the case where Σ_{true} is perfectly known. This behavior is partly due to the setup in (32): as σ^2 grows, the diagonal components of the covariance matrix dominate, making the true covariance approximately isotropic. Since the estimation of x_{true} (given a fixed noise covariance matrix) is invariant to the scaling of the covariance matrix, the performance of $\Sigma = I$ aligns with that of $\Sigma = \Sigma_{\text{true}}$ despite the scaling discrepancy.

Finally, Figure 3c shows the 2-Wasserstein distance averaged over the Monte Carlo simulations. Similar to the previous figure, the covariance estimates obtained by Elimination and BCD are consistently close to those derived using $x = x_{\text{true}}$. As the noise level increases, covariance estimation error also rises, widening the gap, as expected.

B. Pose-Graph Optimization Ablations

Dataset: We used a popular synthetic PGO benchmark, the Manhattan dataset [23], and generated new measurement realizations with varying values of the actual noise covariance matrix. The dataset consists of 3,500 poses and $k = 5{,}598$ relative-pose measurements. This dataset is notoriously poorly connected [18]. Therefore, to analyze the effect of connectivity on covariance estimation, we also performed experiments on modified versions of this dataset, where additional loop closures were introduced by connecting pose i to poses $i+2$ and $i+3$. This modification increased the total number of measurements to $k = 12{,}593$.

We generated zero-mean Gaussian noise in the Lie algebra $\mathfrak{se}(2) \cong \mathbb{R}^3$ for the following models:

1) **Homoscedastic Measurements:** In these experiments, all measurements share the same information matrix, given by $\alpha \times \text{diag}(20, 40, 30)$, where α is a scaling factor that controls the noise level (hereafter referred to as the "information level").
2) **Heteroscedastic Measurements:** We introduce two distinct noise models for odometry and loop-closure edges. The true information matrix for odometry noise is fixed at $\text{diag}(1000, 1000, 800)$, while the loop-closure noise information matrix is varied as $\alpha \times \text{diag}(20, 40, 30)$.

For each value of the information level $\alpha \in \{5, 10, 20, 30, 40\}$, we conducted 50 Monte Carlo simulations with different noise realizations.

Algorithms: We implemented a variant of Algorithm 3 in C++, based on g2o [19], for PGO problems. Instead of using Riemannian gradient descent, we used g2o's implementation of Powell's Dog-Leg method [25] on $\text{SE}(2)$ to optimize x. In each outer iteration, our implementation retrieves the residuals for each measurement, computes $M(x)$ as defined in (5), and updates the noise covariance estimate using Theorem 1. We set $\lambda_{\text{min}} = 10^{-4}$ and $\lambda_{\text{max}} = 10^4$ in all experiments to handle cases where the smallest eigenvalue of $S(x)$ is (approximately) zero. We then perform a single iteration of Powell's Dog-Leg method [25] to update the primary parameters x, initializing the solver at the latest estimate of x. To ensure convergence across all Monte Carlo trials, we ran 13 outer iterations. In this dataset, the per-iteration computational overhead of our algorithm (i.e., time spent updating the covariance matrix in Step 2 of Algorithm 3 relative to g2o with a given covariance)

ranged from 0.9 to 1.5 milliseconds on an Intel i7-6820HQ CPU, which is negligible.

We report the results for the following methods:

1) **BCD**: In line 11 of Algorithm 3, the noise information matrix is estimated using the ML estimate (i.e., no prior) subject to eigenvalue constraints. This is equivalent to (18) after replacing $M(x)$ with the sample covariance $S(x)$ at the current value for x.
2) **BCD (diag)**: In line 11 of Algorithm 3, the noise information matrix is estimated using the ML estimate (as above), subject to both diagonal and eigenvalue constraints. This is equivalent to (20) after replacing $M(x)$ with the sample covariance $S(x)$ at the current value for x.
3) **BCD (Wishart)**: In line 11 of Algorithm 3, the information matrix is updated according to (18); i.e., using the MAP estimate with a Wishart prior and under the eigenvalue constraints.
4) **BCD (diag+Wishart)**: In line 11 of Algorithm 3, the noise information matrix is updated according to (20); i.e., using the MAP estimate with a Wishart prior and under the diagonal and eigenvalue constraints.

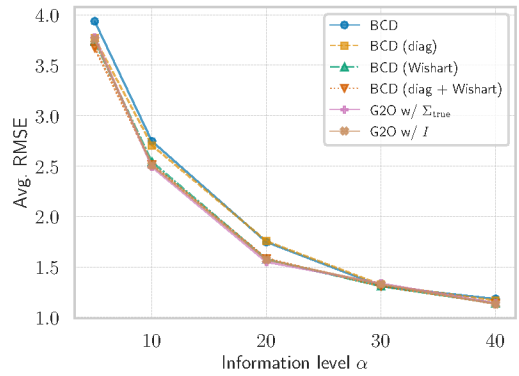
For BCD (Wishart) and BCD (diag+Wishart) where a Wishart prior was used, we applied our Algorithm 1 to set the prior parameters (i.e., V and ν), for a prior weight of $w_{\text{prior}} = 0.1$ and a prior estimate of $\Sigma_0 = 0.002I$ for both odometry and loop-closure edges. Note that this prior estimate is far from the true noise covariance value.

Additionally, we report results for estimating x using Powell's Dog-Leg solver in g2o under two fixed noise covariance settings: (i) the true covariance matrix (as a reference) and (ii) the identity matrix (a common ad hoc approximation used in practice). To ensure convergence across all trials, we performed eight iterations for these methods. All algorithms were initialized using a spanning tree to compute x_{init} [19].

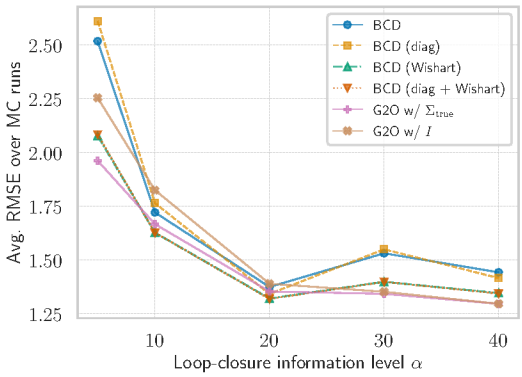
Metrics: We use RMSE to measure error in the estimated robot trajectory (positions). In all cases, the first pose is fixed to the origin and therefore aligning the estimates with the ground truth is not needed. We also use the 2-Wasserstein distance (33) to measure the covariance estimation error.

Results: The results are shown in Figures 4 and 5.

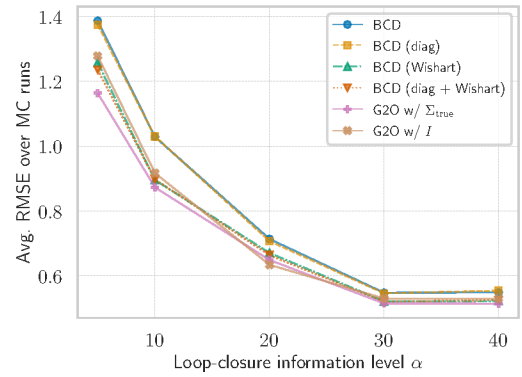
Figure 4 shows the average RMSE over Monte Carlo trials for different values of the information level α. Figure 4a presents the results for the homoscedastic case, while Figures 4b and 4c show the results for the heteroscedastic case before and after additional loop closures, respectively. The results show that our framework, across all variants, succesfully produced solutions with an RMSE close to the reference setting where the true noise covariance matrix was available (g2o with Σ_{true}). For almost all values of α and all experiments, BCD (Wishart) and BCD (diag+Wishart) achieved a lower average RMSE compared to the other variants, highlighting the importance of incorporating a prior. Notably, this is despite the prior being assigned a small weight w_{prior} and the fact that the prior estimate Σ_0 is not close to Σ_{true}. Incorporating a prior is particularly crucial in PGO, as we observed that the eigenvalues of $S(x)$ can become severely small in practice. In such cases,



(a) Homoscedastic Scenario



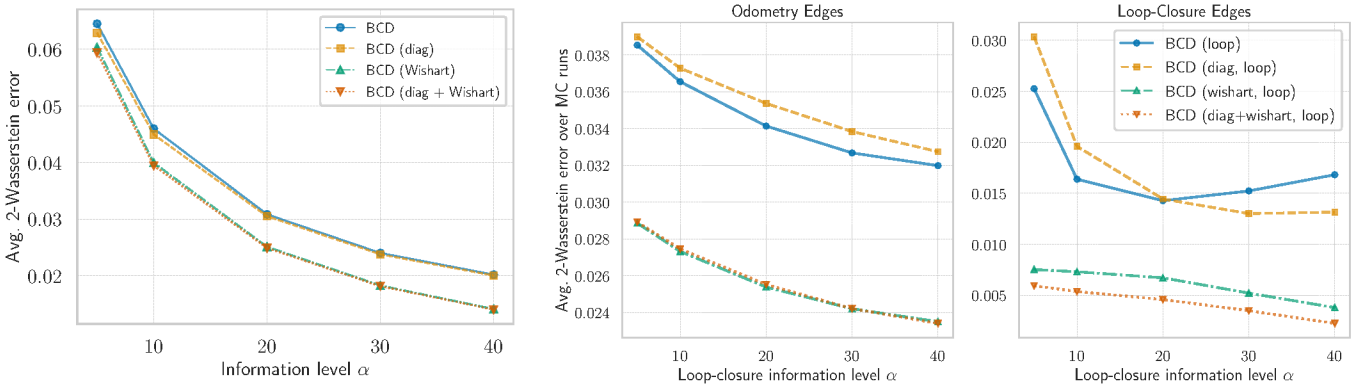
(b) Heteroscedastic Scenario



(c) Heteroscedastic Scenario with extra loop closures

Fig. 4: Average RMSE obtained by variants of BCD and g2o (with fixed true and identity covariances) as a function of information level α.

without a prior and without enforcing a minimum eigenvalue constraint $\Sigma \succeq \lambda_{\min} I$, the ML estimate for the noise covariance can collapse to zero, leading to invalid results and excessive overconfidence. The average RMSE achieved by those variants with diagonal and without constraints are generally similar. Interestingly, the RMSE achieved by the diagonal variants appears to be close to that of their counterparts without this constraint, despite the true noise covariance being diagonal.



(a) Homoscedastic Scenario

(b) Heteroscedastic Scenario – odometry (middle) and loop closure (right)

Fig. 5: Average 2-Wasserstein error achieved by variants of BCD and as a function of information level α.

For some values of α in Figure 4b, the BCD variants with the Wishart prior achieved a lower MSE than the reference solution. However, we expect that, on average, the reference solution will perform better with a larger number of Monte Carlo simulations. The RMSE trends look similar in all experiments, with the exception of a slight increase for $\alpha = 30$ for BCD and BCD (diag) in Figure 4b. In terms of RMSE, BCD variants with a Wishart prior outperform or perform comparably to the baseline with a fixed identity covariance in most cases. That said, especially under larger information levels (i.e., lower noise), this naïve baseline estimates x quite accurately.

Figure 5 presents the average 2-Wasserstein distance (33) between the noise covariance matrices estimated by various BCD variants and the true covariance in both homoscedastic and heteroscedastic scenarios. In all cases, BCD variants achieved a small 2-Wasserstein error. For reference, the 2-Wasserstein error between $\mathcal{N}(0, \Sigma_{\text{true}})$ and the baseline $\mathcal{N}(0, I)$ exceeds 1, which is more than 20 times the errors attained by our algorithms. As noted in Section I, an incorrect noise covariance leads to severe overconfidence or underconfidence in the estimated state x (e.g., the estimated trajectory in SLAM), potentially resulting in poor or catastrophic decisions.

The results indicate that, in most cases, diagonal BCD variants yield lower errors. This is expected, as the true noise covariance matrices are diagonal, and enforcing this prior information enhances estimation accuracy. Additionally, the MAP estimates obtained by BCD (Wishart) and BCD (diag+Wishart) outperform their ML-based counterparts. This is due to the fact that the eigenvalues of the sample covariance matrix $S(x)$ can be very small, indicating insufficient information for accurate noise covariance estimation. In such cases, the eigenvalue constraint effectively prevents a complete collapse of the estimated covariance. Interestingly, this issue was not encountered in [32] for ML estimation of the noise covariance matrix in PnP. This discrepancy may suggest additional challenges in estimating noise covariance in SLAM (and related) problems, which typically involve a larger number of variables and sparse, graph-structured relative measurements. In such cases, incorporating a prior estimate may be necessary for accurate identification of noise covariance matrices. Finally, Figure 5b illustrates that the MAP estimation error is lower for the noise covariance of loop closures compared to odometry edges. This difference is partly because the prior value based on Σ_0 is closer to the true noise covariance matrix for loop closures.

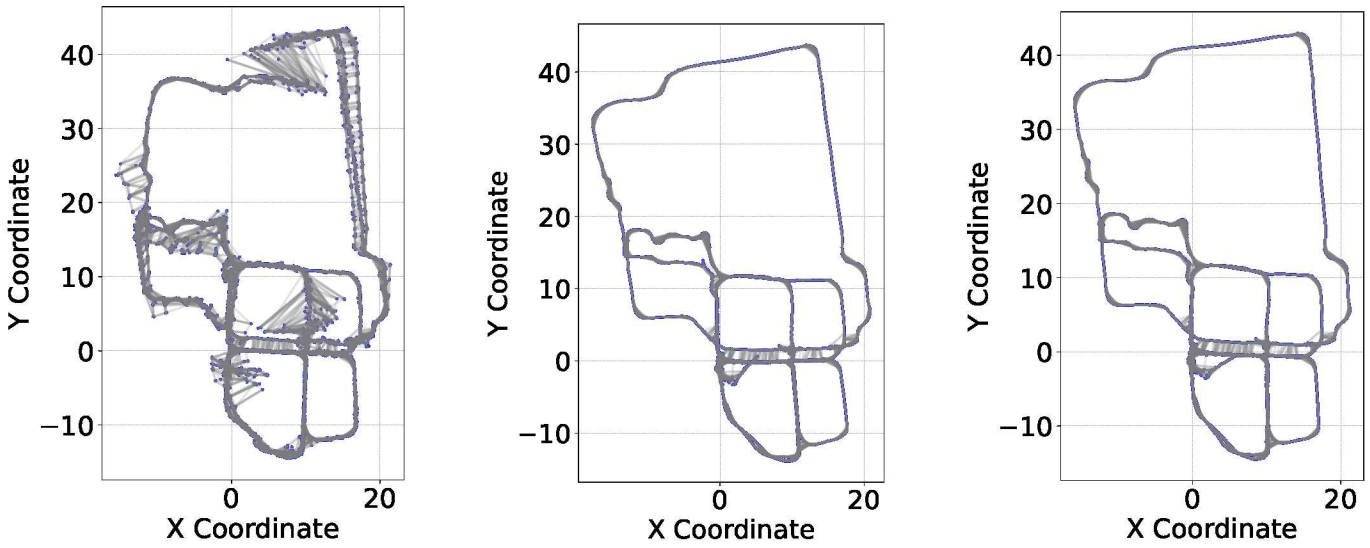
C. RIM Dataset

In this section, we present qualitative results on RIM, a large real-world 3D PGO dataset collected at Georgia Tech [10]. This dataset consists of 10,195 poses and 29,743 measurements. The g2o dataset includes a default noise covariance matrix, but with these default values, g2o struggles to solve the problem. Specifically, the Gauss-Newton and Dog-Leg solvers terminate after a single iteration, while the Levenberg-Marquardt solver completes 10 iterations but produces the trajectory estimate shown in Figure 6a.

We ran BCD without and with the Wishart prior (i.e., ML and MAP estimation) for 10 outer iterations, using a single Dog-Leg iteration to optimize x in each round. We set the prior parameter based on $w_{\text{prior}} = 0.1$ and $\Sigma_0 = 0.01I$. The results are displayed in Figures 6b and 6c, respectively. It is evident that the trajectory estimates obtained by BCD are significantly more accurate than those obtained using the original noise covariance values.

A closer visual inspection reveals that the trajectory estimated using the Wishart prior (right) appears more precise than the ML estimate (middle). This is expected, as without imposing a prior, some eigenvalues of $S(x)$ collapsed until they reached the eigenvalue constraint $\Sigma \succeq \lambda_{\min} I$. In contrast, in the MAP case, the estimated covariance did not reach this lower bound.

Interestingly, despite Σ_0 being isotropic, the estimate obtained by BCD with the Wishart prior was not. In particular, the information matrix component associated with the z coordinate was almost twice as large as those of x and y. This likely reflects the fact that the trajectory in this dataset is relatively flat. This finding highlights that in this case, in addition to the



(a) g2o w/ original covariance (b) BCD – joint ML estimate (w/o prior) (c) BCD – joint MAP estimate w/ prior

Fig. 6: RIM Dataset

prior, the measurements also contributed information to the estimation of the noise covariance matrix.

VII. Limitations

Algorithm 3, in its current form, is not robust to outliers, which limits its applicability in real-world scenarios. However, existing robust M-estimation techniques can be readily adapted for use with BCD. These methods iteratively reweight residuals based on a chosen robust cost function. Our preliminary results indicate that incorporating these weights into Step 1 of BCD (as is standard) and also into Step 2 (by replacing the sample covariance matrix $S(x)$ with a weighted average of residuals) yields an outlier-robust variant of BCD for the joint estimation problem. We are currently evaluating this approach and aim to implement it in real-time LiDAR and visual SLAM and odometry systems such as [9, 28].

Additionally, Theorem 3 requires $\mathcal{M}$ to be compact. This result therefore does not immediately apply to problems where x contains translational components. However, we believe we can address this issue in SLAM and many other geometric estimation problems by confining translational components to potentially large but bounded subsets of $\mathbb{R}^d$ (where $d \in \{2, 3\}$ is the problem dimension).

Finally, we acknowledge the extensive research on IGLS and Feasible GLS (FGLS) conducted as early as 1970s in econometrics and statistics, which remain largely unknown in engineering; see relevant references in [27, Chapter 12.5] and [32]. Exploring this rich literature may result in new insights and methods that can further improve noise covariance estimation in SLAM and computer vision applications.

VIII. Conclusion

This work presented a novel and rigorous framework for *joint* estimation of primary parameters and noise covariance matrices in SLAM and related computer vision estimation problems. We derived analytical expressions for (conditionally) optimal noise covariance matrix (under ML and MAP criteria) under various structural constraints on the true covariance matrix. Building on these solutions, we proposed two types of algorithms for finding the optimal estimates and theoretically analyzed their convergence properties. Our results and algorithms are quite general and can be readily applied to a broad range of estimation problems across various engineering disciplines. Our algorithms were validated through extensive experiments using linear measurement models and PGO problems. The results show that the state and the noise covariance matrix can be jointly estimated from the measurements (and, optionally, a prior on the covariance) with negligible computational overhead relative to standard solvers.

References

[1] Paul D Abramson. *Simultaneous estimation of the state and noise statistics in linear dynamical systems*, volume 332. National Aeronautics and Space Administration, 1970.

[2] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[3] Sameer Agarwal, Keir Mierle, and The Ceres Solver Team. Ceres Solver, 10 2023. URL https://github.com/ceres-solver/ceres-solver.

[4] Timothy D Barfoot, James R Forbes, and David J Yoon. Exactly sparse Gaussian variational inference with application to derivative-free batch nonlinear state estimation. *The International Journal of Robotics Research*, 39(13): 1473–1502, 2020.

[5] Christopher M Bishop and Nasser M Nasrabadi. *Pattern*

recognition and machine learning, volume 4. Springer, 2006.

[6] Nicolas Boumal. An introduction to optimization on smooth manifolds. Available online, Aug 2020. URL http://www.nicolasboumal.net/book.

[7] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, Jose Neira, Ian D Reid, and John J Leonard. Simultaneous localization and mapping: Present, future, and the robust-perception age. *arXiv preprint arXiv:1606.05830*, 2016.

[9] Carlos Campos, Richard Elvira, Juan J. Gómez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 37 (6):1874–1890, 2021.

[10] Luca Carlone, Roberto Tron, Kostas Daniilidis, and Frank Dellaert. Initialization techniques for 3d slam: A survey on rotation estimation and its use in pose graph optimization. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 4597–4604. IEEE, 2015.

[11] Zhaozhong Chen, Harel Biggie, Nisar Ahmed, Simon Julier, and Christoffer Heckman. Kalman filter autotuning through enforcing chi-squared normalized error distributions with bayesian optimization. *arXiv preprint arXiv:2306.07225*, 2023.

[12] Frank Dellaert and GTSAM Contributors. borglab/gtsam, May 2022. URL https://github.com/borglab/gtsam.

[13] Frank Dellaert, Michael Kaess, et al. Factor graphs for robot perception. *Foundations and Trends® in Robotics*, 6(1-2):1–139, 2017.

[14] Kamak Ebadi, Lukas Bernreiter, Harel Biggie, Gavin Catt, Yun Chang, Arghya Chatterjee, Christopher E Denniston, Simon-Pierre Deschênes, Kyle Harlow, Shehryar Khattak, et al. Present and future of SLAM in extreme environments: The DARPA SubT challenge. *IEEE Transactions on Robotics*, 2023.

[15] Alejandro Fontan, Javier Civera, Tobias Fischer, and Michael Milford. Look ma, no ground truth! ground-truth-free tuning of structure from motion and visual SLAM. *arXiv preprint arXiv:2412.01116*, 2024.

[16] Alejandro Fontan, Javier Civera, and Michael Milford. Anyfeature-vslam: Automating the usage of any chosen feature into visual SLAM. In *Robotics: Science and Systems*, volume 2, 2024.

[17] Robin Forsling, Simon J Julier, and Gustaf Hendeby. Matrix-valued measures and wishart statistics for target tracking applications. *arXiv preprint arXiv:2406.00861*, 2024.

[18] Kasra Khosoussi, Matthew Giamou, Gaurav S Sukhatme, Shoudong Huang, Gamini Dissanayake, and Jonathan P How. Reliable graphs for SLAM. *The International Journal of Robotics Research*, 38(2-3):260–298, 2019.

[19] Rainer Kuemmerle, Giorgio Grisetti, Hauke Strasdat, Kurt Konolige, and Wolfram Burgard. g2o: A general framework for graph optimization. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2011.

[20] Ziqi Lu, Yihao Zhang, Kevin Doherty, Odin Severinsen, Ethan Yang, and John Leonard. SLAM-supported self-training for 6d object pose estimation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2833–2840. IEEE, 2022.

[21] Edmond Malinvaud. *Statistical methods of econometrics*. 1980.

[22] Raman Mehra. Approaches to adaptive filtering. *IEEE Transactions on automatic control*, 17(5):693–698, 1972.

[23] E. Olson, J. Leonard, and S. Teller. Fast iterative alignment of pose graphs with poor initial estimates. In *Robotics and Automation, 2006. ICRA 2006. Proceedings 2006 IEEE International Conference on*, pages 2262–2269. Ieee, 2006.

[24] Liangzu Peng and René Vidal. Block coordinate descent on smooth manifolds: Convergence theory and twenty-one examples. *arXiv preprint arXiv:2305.14744*, 2023.

[25] MJD Powell. *A new algorithm for unconstrained optimization*. UKAEA, 1970.

[26] Mohamad Qadri, Zachary Manchester, and Michael Kaess. Learning covariances for estimation with constrained bilevel optimization. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15951–15957. IEEE, 2024.

[27] George A. F. Seber and C. J. Wild. *Nonlinear Regression*. Wiley-Interscience, 1989.

[28] Tixiao Shan, Brendan Englot, Drew Meyers, Wei Wang, Carlo Ratti, and Rus Daniela. Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5135–5142. IEEE, 2020.

[29] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustmenta modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999.

[30] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, and et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[31] Jeremy Nathan Wong, David Juny Yoon, Angela P Schoellig, and Timothy D Barfoot. Variational inference with parameter learning applied to vehicle trajectory estimation. *IEEE Robotics and Automation Letters*, 5 (4):5291–5298, 2020.

[32] Tian Zhan, Chunfeng Xu, Cheng Zhang, and Ke Zhu. Generalized maximum likelihood estimation for perspective-n-point problem. *IEEE Robotics and Automation Letters*, 2025.

[33] Lingyi Zhang, David Sidoti, Adam Bienkowski, Krishna R Pattipati, Yaakov Bar-Shalom, and David L Kleinman. On the identification of noise covariances and adaptive

kalman filtering: A new look at a 50 year-old problem. *IEEE Access*, 8:59362–59388, 2020.

APPENDIX A
WISHART PRIOR

Consider $\mathcal{W}(P; V, \nu)$ as a prior for the information matrix. Here we assume $\nu \geq m+1$ where m is the dimension of covariance matrix and $V \succ 0$. The mode of this distribution is given by

$$\underset{P \succeq 0}{\operatorname{argmax}} \quad \mathcal{W}(P; V, \nu) = (\nu - m - 1)V. \tag{34}$$

Let $\Sigma_0 \succ 0$ be our prior guess for the covariance matrix. We set the value of the scale matrix V by matching the mode with our prior estimate Σ_0^{-1}:

$$(\nu - m - 1)V = \Sigma_0^{-1} \Leftrightarrow V^{-1} = (\nu - m - 1)\Sigma_0 \tag{35}$$

Let w_{prior} be the following:

$$w_{\text{prior}} \triangleq \frac{\nu - m - 1}{k} \geq 0. \tag{36}$$

Then we can rewrite V^{-1} as

$$V^{-1} = w_{\text{prior}} k \Sigma_0. \tag{37}$$

In the unconstrained case (15), the optimal (MAP) estimate (conditioned on a fixed value of x) is given by:

$$\Sigma^\star(x) \triangleq P^\star(x)^{-1} = M(x) \tag{38}$$

$$= \frac{V^{-1} + kS(x)}{k + \nu - m - 1} \tag{39}$$

$$= \frac{w_{\text{prior}} k \Sigma_0 + kS(x)}{(w_{\text{prior}} + 1)k} \tag{40}$$

$$= \frac{w_{\text{prior}}}{w_{\text{prior}} + 1}\Sigma_0 + \frac{1}{w_{\text{prior}} + 1}S(x). \tag{41}$$

This shows that the (conditional) MAP estimator simply blends the sample covariance matrix $S(x)$ and prior Σ_0 based on the *prior weight* w_{prior}. We directly set w_{prior} (e.g., $w_{\text{prior}} = 0.1$ by default) based on our confidence in the prior relative to the likelihood. Large values of w_{prior} will result in stronger priors. For the chosen value of w_{prior},

$$\nu = w_{\text{prior}} k + m + 1, \tag{42}$$

$$V = w_{\text{prior}} k \Sigma_0. \tag{43}$$

This procedure is summarized in Algorithm 1.

Remark 7 (Wishart Parameters)**.** It is worth noting that Algorithm 1 represents just one simple approach to setting the prior parameters. Alternative strategies can also be employed, such as learning the parameters from an offline calibration dataset, or directly specifying the value of ν independently of k (with larger values of ν indicating greater confidence in the prior estimate Σ_0). By explicitly setting a constant value for w_{prior},[3] the relative influence of the prior versus the measurements remains fixed. We adopted this strategy in our experiments for simplicity. However, it is important to recognize that setting ν as a function of k violates the Bayesian principle that the prior should be specified before observing any data. Consequently, when w_{prior} is kept fixed, unlike in standard MAP estimation the measurements will never completely dominate the prior as k increases.

[3]That is, setting the prior parameter ν as a function of k in (36).

APPENDIX B
PROOF OF PROPOSITION 1

The objective function is strictly convex because log-determinant is strictly concave over $\mathbb{S}_{\succ 0}$ and the second term is linear in P. The equality (diagonal) and linear matrix inequality (eigenvalue) constraints are also linear and convex, respectively. Since the objective is strictly convex, there is at most one optimal solution (i.e., if a minimizer exists, it is unique).

APPENDIX C
PROOF OF THEOREM 1

In the following, we provide primal-dual pairs that satisfy the KKT conditions for each variant. Since the problem is differentiable and convex, these pairs must be primal-dual optimal. Moreover, these solutions are unique because the objective function is strictly convex.

1): For a given $x \in \mathcal{M}$, the Lagrangian of the inner subproblem in Problem 1 is given by

$$\mathcal{L}(P, Q) = -\log\det P + \langle M(x), P\rangle - \langle Q, P\rangle, \tag{44}$$

where $Q \succeq 0$ is the Lagrange multiplier corresponding to the semidefinite cone constraint. Recall that $M(x) \succ 0$ by construction. Verify that $P^\star(x) = M(x)^{-1}$ and $Q = 0$ trivially satisfy the KKT conditions, and thus $P^\star(x) = M(x)^{-1}$ is the optimal solution to the primal problem.

2): For the case of the inner subproblem in Problem 2, the problem can be rewritten as

$$\begin{aligned} \underset{\lambda_1,\ldots,\lambda_m}{\text{minimize}} \quad & -\sum_{i=1}^m \log\lambda_i + \sum_{i=1}^m \lambda_i M_{ii}(x) \\ \text{subject to} \quad & \lambda_i \geq 0, \quad i \in [m]. \end{aligned} \tag{45}$$

The Lagrangian of this problem is given by

$$\mathcal{L}(\lambda_1, \ldots, \lambda_m, q_1, \ldots, q_m) = \sum_{i=1}^m \left[-\log\lambda_i + \lambda_i M_{ii}(x) - \lambda_i q_i\right], \tag{46}$$

where q_i are the Lagrange multipliers corresponding to the constraints. The KKT conditions for this problem are

$$-1/\lambda_i + M_{ii}(x) - q_i = 0, \quad i \in [m] \tag{47}$$

$$q_i\lambda_i = 0, \quad q_i \geq 0, \quad \lambda_i \geq 0, \quad i \in [m]. \tag{48}$$

It can be observed that since $M_{ii}(x)$ is positive by the virtue of $M(x)$ being a positive definite matrix, $\lambda_i^\star = M_{ii}(x)^{-1}$ and $q_i^\star = 0$, $\forall i \in [m]$, is the solution to the above KKT system.

3): The Lagrangian of the inner subproblem of Problem 3 is given by

$$\mathcal{L}(P,\underline{Q},\overline{Q}) = -\log\det P + \langle M(x), P\rangle \tag{49}$$
$$+ \langle \underline{Q}, I/\lambda_{\max} - P\rangle + \langle \overline{Q}, P - I/\lambda_{\min}\rangle. \tag{50}$$

The KKT conditions corresponding to this problem read

$$\nabla_P \mathcal{L}(P,\underline{Q},\overline{Q}) = -P^{-1} + M(x) - \underline{Q} + \overline{Q} = 0 \tag{51}$$
$$\langle \underline{Q}, I/\lambda_{\max} - P\rangle = 0 \tag{52}$$
$$\langle \overline{Q}, P - I/\lambda_{\min}\rangle = 0 \tag{53}$$
$$\underline{Q} \succ 0,\ \overline{Q} \succ 0,\ P = P^\top \tag{54}$$
$$I/\lambda_{\max} \le P \le I/\lambda_{\min}. \tag{55}$$

Let $\underline{Q}^\star = U(x)\underline{\Lambda}(x)U(x)^\top$ and $\overline{Q}^\star = U(x)\overline{\Lambda}(x)U(x)^\top$ where the elements of diagonal matrices $\underline{\Lambda}(x)$ and $\overline{\Lambda}(x)$ are

$$\underline{\Lambda}_{ii}(x) = \begin{cases} 0 & D_{ii}(x) \in [0, \lambda_{\max}], \\ D_{ii}(x) - \lambda_{\max} & D_{ii}(x) \in (\lambda_{\max}, \infty), \end{cases} \tag{56}$$

and

$$\overline{\Lambda}_{ii}(x) = \begin{cases} \lambda_{\min} - D_{ii}(x) & D_{ii}(x) \in [0, \lambda_{\min}], \\ 0 & D_{ii}(x) \in (\lambda_{\min}, \infty). \end{cases} \tag{57}$$

It can be observed that the choice of $P^\star(x)$ as in (18) along with $\overline{Q}^\star(x)$ and $\underline{Q}^\star(x)$ defined above satisfy the KKT conditions and are the primal-dual optimal solutions to the problem. This completes the proof.

4): The solution to the inner subproblem of Problem 4 can be obtained by a reformulation similar to case 2) above and checking the conditions similar to case 3).

Appendix D
Proof of Theorem 2

Consider a strictly feasible pair (x_0, P_0) where $x_0 \in \mathcal{M}$ and $P_0 \succ 0$. Let $u_0 \in \mathbb{R}^m$ be any vector in the nullspace of $S(x_0)$. For any $c > 0$, the pair $(x_0, P_0 + c\,u_0u_0^\top)$ remains strictly feasible. Observe that

$$F_{\text{ML}}(x_0, P_0 + c\,u_0u_0^\top) =$$
$$-\log\det\left(P_0 + c\,u_0u_0^\top\right) + \langle S(x_0), P_0 + c\,u_0u_0^\top\rangle \tag{58}$$
$$= -\log\det\left(P_0 + c\,u_0u_0^\top\right) + \langle S(x_0), P_0\rangle \tag{59}$$
$$= -\log\left(1 + c\,u_0^\top P_0^{-1}u_0\right) - \log\det P_0 + \langle S(x_0), P_0\rangle \tag{60}$$
$$= -\log\left(1 + c\,u_0^\top P_0^{-1}u_0\right) + F_{\text{ML}}(x_0, P_0), \tag{61}$$

where we used the matrix determinant lemma in (60). Since $P_0 \succ 0$, we have that $u_0^\top P_0^{-1}u_0 > 0$. Hence, as $c \to \infty$, we have

$$F_{\text{ML}}(x_0, P_0 + c\,u_0u_0^\top) \to -\infty.$$

Therefore, the problem is unbounded below and does not admit a solution.

Appendix E
Background Information for Section V-B

Let $\mathcal{Y}$ be a smooth submanifold in Euclidean space. Each $y \in \mathcal{Y}$ is associated with a linear subspace, called the *tangent space*, denoted by $T_y\mathcal{Y}$. Informally, the tangent space contains all directions in which one can tangentially pass through y. For a formal definition see [2, Definition 3.5.1, p. 34]. A smooth function $F : \mathcal{Y} \to \mathbb{R}$ is associated with the (Euclidean) gradient $\nabla F(y)$ of F, as well as the orthogonal projection of $\nabla F(y)$ onto the tangent space $T_y\mathcal{Y}$, known as the *Riemannian gradient* of F at y, denoted by $\operatorname{grad} F(y)$. See [2, p. 46] for more detail. The manifold $\mathcal{Y}$ is also associated with a map $\operatorname{Retr}_y : T_y\mathcal{Y} \to \mathcal{Y}$, called a *retraction* that enables moving from y along the direction $v \in T_y\mathcal{Y}$ while remaining on the manifold – see [6, Definition 3.47] and [2, Definition 4.1.1] for formal definitions.

A sufficient condition for Assumption 2 to hold is for the manifolds to be compact subsets of the Euclidean space. Under this assumption we have the following lemma [24, Lemma 1].

Lemma 1. *Let $\mathcal{M}$ be a compact submanifold of the Euclidean space. Then we have,*

$$\|\operatorname{Retr}_\xi(v) - \xi\| \le \alpha\|v\|, \quad \forall \xi \in \mathcal{M}, \quad \forall v \in T_\xi\mathcal{M}$$

for some constant α. Moreover, if R satisfies Assumption (2) and $\mathcal{P}$ is compact, then there exists a positive $\widetilde{L}$ such that for all $(\xi, \Pi) \in \mathcal{M} \times \mathcal{P}$ and all $v \in T_\xi\mathcal{M}$, we have

$$R(\operatorname{Retr}_\xi(v), \Pi) \le R(\xi, \Pi) + \langle \operatorname{grad}_x R(\xi, \Pi), v\rangle + \frac{\widetilde{L}}{2}\|v\|. \tag{62}$$

Appendix F
Proof of Theorem 3

The proof follows from specializing the proof of [24, Theorem 4] to the problem considered in this paper. Specifically, the compactness of manifolds along with Assumptions 1 and 2 lead to the existence of of $\widetilde{L}$ as described in Lemma 1. Then following the steps of the proof of [24, Theorem 4] and setting $b = 2$ completes the proof.