Covariance Estimation in Two-Level Regression

Nicholas Moehle and Dimitry Gorinevsky

***Abstract*— This paper considers estimation of covariance matrices in multivariate linear regression models for two-level data produced by a population of similar units (individuals). The proposed Bayesian formulation assumes that the covariances for different units are sampled from a common distribution. Assuming that this common distribution is Wishart, the optimal Bayesian estimation problem is shown to be convex. This paper proposes a specialized scalable algorithm for solving this two-level optimal Bayesian estimation problem. The algorithm scales to datasets with thousands of units and trillions of data points per unit, by solving the problem recursively, allowing new data to be quickly incorporated into the estimates. An example problem is used to show that the proposed approach improves over existing approaches to estimating covariance matrices in linear models for two-level data.**

I. Introduction

Consider a two-level dataset

$$\mathcal{D} = \left\{ \{x_i(t), y_i(t)\}_{t=1}^{T_i} \right\}_{i=1}^{N}, \tag{1}$$

where index i is the dataset number, index t is the sample number. Each $x_i(t) \in \mathbf{R}^n$ is a vector of independent variables (regressors), and each $y_i(t) \in \mathbf{R}^m$ is a vector of dependent variables. There are N datasets at all; dataset i includes a total of T_i samples. Each dataset describes a separate individual unit of the overall population.

This paper studies linear multivariate regression models of the two-level data (1). For each unit, the model is described by the mean (regression parameters) and covariance of the model residual. The means for different units are assumed to be different; this is known as a model with fixed effects. The covariances for different units are assumed to be different but related; this is a novel formulation.

In the motivating example considered below, the covariance estimates are used in statistical monitoring to find outliers, which are labeled as possibly anomalous data. Accurate covariance estimation is necessary for the monitoring to be accurate. The described two-level formulation can be useful in many other problems requiring accurate estimation of covariance matrices, such as classification, linear discriminant analysis, portfolio management, and others.

The special case $N = 1$ in (1) yields a standard multivariate regression. In that case, accurate estimation of the m^2 elements of the covariance matrix requires that $T_1 \gg m$ [7]. For large m this might not hold. For an ill-conditioned covariance matrix, even more data samples are required. Using data from multiple units in a two-level data set can improve the covariance estimation accuracy.

Nicholas Moehle is with the Department of Mechanical Engineering, Stanford University, moehle@stanford.edu

Dimitry Gorinevsky is with Mitek Analytics LLC, Palo Alto, CA, dimitry@mitekan.com, and the Department of Electrical Engineering, Stanford University, gorin@stanford.edu

Covariance matrix estimation for one-level datasets has attracted substantial attention earlier. Several approaches to the maximum likelihood estimation (MLE) with shrinkage (regularization) have been proposed, e.g., see [10], [11] where further references can be found. The approaches related to this paper add regularization by using a Bayesian prior in a maximum *a posteriori* probability (MAP) estimation, such as an inverse Wishart prior for the covariance matrix, see [13]. The inverse Wishart is the conjugate prior for the covariance matrix of a multivariate normal distribution. As such, the MAP solution is a generalization of the MLE solution and has the same attractive properties. In particular, the solution can be efficiently computed for large datasets using recursive least squares (RLS) or a related formulation.

Two-level datasets are considered in multivariate analysis of covariance (MANCOVA), e.g., see [17]. In MANCOVA, the null hypothesis is that all data follows the same distribution. The goal is to decide if inter-unit covariance and intra-unit covariance is compatible with this hypothesis.

Two-level regression is used in 'soft' applications such as social sciences, biology, economics, and medicine (drug testing) [1], [2], [9], [18]. The most established solution approaches are based on Gibbs sampling [8] and other approximate methods. More scalable methods have been developed to estimate covariance structure for two-level datasets using expectation-maximization (EM) methods as a heuristic to find the maximum-likelihood estimates [3], [12].

There appears to be little prior work on scalable two-level modeling applicable to machine data monitoring. One exception is [6], discussing scalable algorithms for estimating a two-level regression model of aircraft fleet data. In [6], the same covariances are assumed for all units.

The contributions of this paper are as follows. First, we propose a novel two-level linear regression formulation with fixed effects in the regression parameters and in the covariances of the random effects. In this hierarchical Bayesian formulation the unit covariances are realizations of the generating Wishart distribution. Second, we show that for the proposed model, Bayesian optimal estimation for data (1) yields a (nonlinear) convex optimization problem. Third, we formulate a specialized convex optimization algorithm that solves the estimation problem using recursive updates and is scalable to very large datasets. Finally, numerical examples demonstrate that the proposed formulation improves accuracy of the estimation compared to the known methods.

II. Baseline Problem

This section considers a single-level dataset with $N = 1$, (there is only one unit). In this section we drop the unit index $i = 1$. The dataset (1) becomes

$$\{x(t), y(t)\}_{t=1}^{T}. \tag{2}$$

As a baseline for introducing the main contribution in the next section, this section briefly recaps the known formulation of multivariate Bayesian linear regression for dataset (2). In what follows, we use $p(\cdot)$ for all probability density functions and $\ell(\cdot)$ for all log likelihood functions. The meaning should be clear from the context.

A. Multivariate Regression

Consider a linear regression model $y(t) = Bx(t) + v(t)$, where $B \in \mathbf{R}^{m\times n}$ is the regression parameter matrix and $v(t) \in \mathbf{R}^m$ is the residual. For data (2), the model can be compactly represented as

$$Y = BX + V, \tag{3}$$

$$\begin{aligned} Y &= \left[\, y(1) \;\cdots\; y(T) \,\right], \\ X &= \left[\, x(1) \;\cdots\; x(T) \,\right], \\ V &= \left[\, v(1) \;\cdots\; v(T) \,\right]. \end{aligned} \tag{4}$$

The Bayesian formulation assumes that the residuals $v_i(t)$ are independently generated by the normal distribution

$$v(t) \sim \mathcal{N}(0, \Sigma), \tag{5}$$

with covariance matrix $\Sigma \in \mathbf{S}^m$, where $\mathbf{S}^m$ is the cone of positive definite $m \times m$ matrices.

The log likelihood function for Σ and B is

$$\begin{aligned} &\ell(\Sigma, B \mid \mathcal{D}) \\ &\quad = -\frac{1}{2}T \log|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(Y - BX)(Y - BX)^T\right). \end{aligned} \tag{6}$$

The parameter matrix $B \in \mathbf{R}^{m\times n}$ in (3) is considered a nuisance parameter; no prior for B is specified. As discussed in [11], [19], a prior for the covariance matrix Σ is needed if the number of samples T is insufficiently large. A well established approach, which we take as a baseline, is to use an inverse Wishart prior

$$\Sigma \sim \mathcal{W}^{-1}(\Psi, \nu), \tag{7}$$

where $\Psi \in \mathbf{S}^m$ is the *scale matrix* and the positive integer ν is the number of degrees of freedom. The inverse Wishart is the conjugate prior for the covariance matrix in a multivariate normal distribution.

The log likelihood of the inverse Wishart prior is, up to additive constants,

$$\ell(\Sigma|\Psi, \nu) = -\frac{1}{2}(\nu + m + 1)\log|\Sigma| - \frac{1}{2}\operatorname{tr}\left(\Psi\Sigma^{-1}\right). \tag{8}$$

The change of variables $\Omega_1 = \frac{1}{\nu+m+1}\Psi$ and $\alpha_1 = \nu+m+1$, transforms (8) into the log prior

$$\ell(\Sigma|\Omega_1, \alpha_1) = -\alpha_1 \log|\Sigma| - \alpha_1 \operatorname{tr}\left(\Omega_1\Sigma^{-1}\right). \tag{9}$$

The resulting log posterior has the form

$$\begin{aligned} \ell(B, \Sigma|\mathcal{D}) = &-\frac{1}{2}(N + \alpha_1)\log|\Sigma| - \frac{1}{2}\alpha_1 \operatorname{tr}\left(\Omega_1\Sigma^{-1}\right) \\ &- \frac{1}{2}\operatorname{tr}\left((Y - BX)^T\Sigma^{-1}(Y - BX)\right). \end{aligned} \tag{10}$$

The Maximum A posteriori Probability (MAP) estimates of B and Σ minimize the negative log-posterior (10); this minimization problem is convex in the variables Σ^{-1} and $\Sigma^{-1}B$. The first-order optimality conditions can be solved analytically (e.g., see [17]), yielding

$$B = YX^T(XX^T)^{-1} \tag{11}$$

$$\Sigma = \frac{1}{N + \alpha_1}\left((Y - BX)(Y - BX)^T + \alpha_1\Omega_1\right). \tag{12}$$

In what follows we assume that the scatter matrix XX^T is invertible. In practice, selecting independent regressors leads to invertible XX^T.

The estimates (11), (12) can be computed recursively. These estimates can be written in terms of the scatter matrices XX^T, YX^T, and YY^T. For (11) this is obvious; for (12) this requires expanding the matrix product in the numerator. As additional data become available, the scatter matrices can be updated using rank-one recursion.

III. Two-level Multiple Regression Problem

The rest of the paper considers two-level dataset (1) and (3), (4) is replaced by $Y_i = B_iX_i + V_i$, $(i = 1, \ldots, N)$, where

$$\begin{aligned} Y_i &= \left[\, y_i(1) \;\cdots\; y_i(T) \,\right], \\ X_i &= \left[\, x_i(1) \;\cdots\; x_i(T) \,\right], \\ V_i &= \left[\, v_i(1) \;\cdots\; v_i(T) \,\right]. \end{aligned} \tag{13}$$

A. Baseline Approaches

The formulation of Section II can be applied to two-level covariance estimation in two ways described below. These are are later used to benchmark the proposed approach.

1) Pooled Regression: The multivariate linear regression of Section II can be used for the pooled data

$$X = \begin{bmatrix} X_1 & \cdots & X_N \end{bmatrix} \qquad Y = \begin{bmatrix} Y_1 & \cdots & Y_N \end{bmatrix}.$$

The obtained estimates of B and Σ are identical across the population. The obvious deficiency of the pooled formulation is that it ignores differences in the units. If the units are substantially different, the obtained model can be inaccurate. This is illustrated in the example of Section V.

2) Separate Regressions: The second approach is to ignore any similarity between the units. The approach of II is then applied separately to each pair of data matrices X_i and Y_i. For each unit i, separate and unrelated estimates of B_i and Σ_i are computed. The problem of this approach is that there may be insufficient data to accurately estimate the covariance for each unit separately.

B. Estimation of the Two-level Model

We propose a generalization of the one-level Bayesian optimal estimation formulation (3), (4), (11), (12), to the two-level data (1), (13). The main novelty of the proposed approach is in modeling the commonality of the covariance structures of the level-one models.

Consider the regression model

$$y_i(t) \sim \mathcal{N}\big(B_i x_i(t), \Sigma_i\big). \tag{14}$$

By analogy with (6), ignoring additive constants, the log likelihood of Σ_i and B_i is

$$\begin{aligned}\ell(\Sigma_i, B_i|\mathcal{D}) = &-\frac{1}{2}T_i \log|\Sigma_i| \\ &-\frac{1}{2}\operatorname{tr}\left(\Sigma_i^{-1}(Y_i - B_i X_i)(Y_i - B_i X_i)^T\right).\end{aligned} \tag{15}$$

The one-level baseline approach of Section II, uses the inverse-Wishart prior for the covariance matrix. The proposed two-level formulation assumes that each individual covariance matrix is sampled from a population. In our Bayesian formulation, all individual covariance matrices are related by having the same Wishart prior

$$\Sigma_i|S, \beta \sim \mathcal{W}(\beta^{-1}S, \beta + m + 1). \tag{16}$$

This is the sampling distribution for a scatter matrix with zero-mean normally distributed columns. It has log prior

$$\begin{aligned}\ell(\Sigma_i|S) = &-\frac{1}{2}(\beta + m + 1)\log|S| \\ &-\frac{1}{2}\beta\log|\Sigma_i| - \frac{1}{2}\beta\operatorname{tr}\left(S^{-1}\Sigma_i\right).\end{aligned} \tag{17}$$

The hyperparameter β sets the strength of the prior, the similarity between the estimated unit covariance matrices. The hyperparameter S is the generating covariance matrix for the population. We propose an inverse Wishart hyperprior for the hyperparameter S

$$S \sim \mathcal{W}^{-1}(\alpha\Omega, \alpha - m - 1). \tag{18}$$

Below, it is shown that with this hyperprior the MLE formulation is convex. Then, the log prior is similar to (8)

$$\ell(S|\Omega, \alpha) = -\frac{1}{2}\alpha\log|S| - \frac{1}{2}\alpha\operatorname{tr}\left(\Omega\Sigma^{-1}\right). \tag{19}$$

The hyperhyperparameter α sets the strength of this hyperprior, how similar S is to Ω. The proposed Bayesian model structure with the parameters, hyperparameters, and nuisance parameters is summarized in Figure 1.

The log posterior is the sum of the log likelihood (15), the log priors (17) for each unit, and the population log prior (19)

$$\begin{aligned}&\ell(\Sigma_1, ..., \Sigma_N, B_1, ..., B_N, S|\mathcal{D}) \\ &\quad = -\ell(S|\Omega, \alpha) - \sum_{i=1}^{N}\left(\ell(\Sigma_i|S, \beta) + \sum_{t=1}^{T}\ell(\Sigma_i, B_i|\mathcal{D})\right)\end{aligned} \tag{20}$$

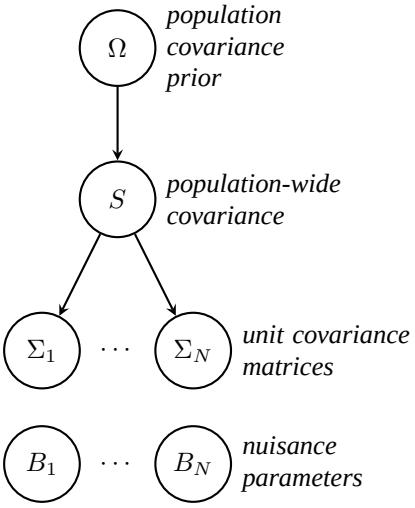

Fig. 1. The structure of the Bayesian priors.

The MAP estimates maximize (20). Substituting (15), (17), and (19) yields the optimization problem:

$$\begin{aligned}\text{minimize}\quad &\alpha\operatorname{tr}\left(\Omega\Sigma^{-1}\right) + (\alpha + N(\beta + m + 1))\log|S| \\ &+ \sum_{i=1}^{N}\Big((T_i - \beta)\log|\Sigma_i| + \beta\operatorname{tr}\left(S^{-1}\Sigma_i\right) \\ &+ \operatorname{tr}\left(\Sigma_i^{-1}(Y_i - B_i X_i)(Y_i - B_i X_i)^T\right)\Big)\end{aligned} \tag{21}$$

The decision variables in problem (21) are the matrices B_i and Σ_i, for $i = 1, \ldots, N$, and the matrix S. The hyperparameters α, β, and Ω are assumed to be given. These parameters can be used for tuning the estimator performance. In the absence of better information, a reasonable choice is $\Omega = \omega I$, where ω is a scalar and I is a unit matrix.

C. Convexity

Although the optimization problem (21) is not convex, a change of variables yields an equivalent problem that is convex. Define matrices $\Lambda_i = \Sigma_i^{-1}$, $M_i = \Sigma_i^{-1}B_i$, and $L^T L = S^{-1}$. In these new decision variables, (21) becomes

$$\begin{aligned}\text{minimize}\quad &\alpha\operatorname{tr}\left(L^T\Omega L\right) + (\alpha + N(\beta + m + 1))\log\left|L^T L\right| \\ &+ \sum_{i=1}^{N}\Big(-(T_i - \beta)\log|\Lambda_i| + \beta\operatorname{tr}\left(L^T\Lambda_i^{-1}L\right) \\ &+ \operatorname{tr}(Y_i^T\Lambda_i Y_i) - 2\operatorname{tr}(Y_i^T M_i X_i) \\ &+ \operatorname{tr}((M_i X_i)^T\Lambda_i^{-1}(M_i X_i))\Big)\end{aligned} \tag{22}$$

Convexity of follows from the fact that the log-det function

$$f : \mathbf{S}_+^n \to \mathbf{R} \quad f : X \mapsto \log|X|$$

and the matrix fractional function

$$g : \mathbf{S}_+^n \times \mathbf{R}^n \to \mathbf{R} \quad g : (X, y) \mapsto y^T X^{-1} y$$

are both convex [4]. In (22)

- $\log|L^T L| = \log(|L^T|\cdot|L|) = 2\log|L|$ is convex log-det function.

- $\mathrm{tr}(L^T\Lambda_i^{-1}L) = \sum_{j=1}^{n}(Le_j)^T\Lambda_i^{-1}(Le_j)$, where the e_j are the unit vectors is convex in both L and Λ_i as the sum of matrix fractional functions.
- $(T_i - \beta)\log|\Lambda_i|$ is convex for $T_i \geq \beta$ as a log-det function.

Other terms are trivially convex or are similar to the above.

IV. Solution of Two-Level Problem

A general-purpose convex solver could be used to compute a solution to (22). However, such solution will be inefficient or impossible for a large dataset. The number of decision variables in the problem is proportional to the number of units and can be very large. A large dataset may not even fit into computer memory. This section outlines a specialized method for computing the global optimum of (22). The method is efficient and scalable to very large problems.

A. First Order Optimality Conditions

The proposed approach is to compute the first order conditions for optimality to (22) and solve them iteratively. We will find these first order conditions using the transformed variables, but will iteratively solve them using the original, Bayesian model variables. This amounts to a block coordinate descent method on the transformed, convex problem.

Differentiating (22) with respect to L and then changing back to the original Bayesian model variables yields

$$S = \frac{1}{\alpha + N(\beta + m + 1)}\left(\alpha\Omega + \beta\sum_{i=1}^{N}\Sigma_i\right). \tag{23}$$

Differentiating (22) with respect to Λ_i gives in the original Bayesian model variables

$$0 = -(T_i - \beta)\Sigma_i - \beta\Sigma_i S\Sigma_i + Y_iY_i^T - B_iX_iX_i^TB_i^T. \tag{24}$$

This is an algebraic Riccati equation. It can be solved for Σ_i in cubic time using standard solvers. Differentiating (22) with respect to M_i yields

$$0 = 2\Lambda_i^{-1}M_iX_iX_i^T - 2Y_iX_i^T.$$

Changing back to the original Bayesian model variables gives

$$B_i = Y_iX_i^T(X_iX_i^T)^{-1}. \tag{25}$$

B. Discussion

The optimal estimate of S, given by (23), is a weighted sum of Ω and the individual covariance matrices. For $\alpha = 0$ and N large, S is the average of the unit covariance matrices. The optimal estimate (25) of the regression parameters B_i is the same as for the one-level regression for each unit.

For $\beta = 0$, the solutions to (24) and (25) are the MAP estimates to (see (11), (12)) for the one-level regression problems for each unit. To see this, substitute the solution for B_i in (25) into (24) and set $\beta = 0$ to obtain the sample covariance matrix

$$\Sigma_i = \tfrac{1}{T_i}\left(Y_iY_i^T - Y_iX_i^T(X_iX_i^T)^{-1}Y_i^TX_i^T\right),$$

which has the same form as (12) for $B = B_i$ and $\alpha_1 = 0$.

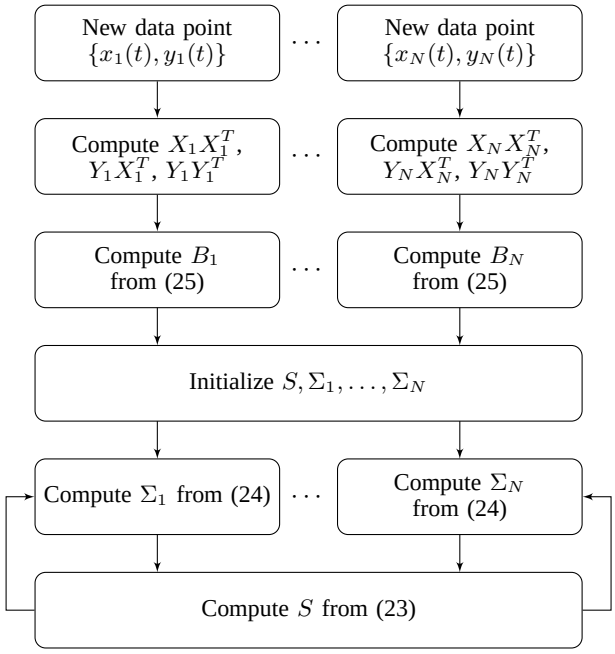

Fig. 2. The steps in the algorithm.

C. Algorithm

The solution for the two-level estimation problem of Section III can be computed as follows

1) Compute the scatter matrices $X_iX_i^T$, $Y_iX_i^T$, $Y_iY_i^T$, where X_i and Y_i are given by (13).
2) Compute B_i (25) for each unit i.
3) Initialize the matrices $\Sigma_1, \ldots, \Sigma_N$ and S.
4) For each i, solve (24) for Σ_i with S given.
5) Solve (23) for S, with $\Sigma_1, \ldots, \Sigma_N$ given.
6) Check for convergence. If the update have not converged, go to step 4.

The proposed algorithm is summarized in Figure IV-C. The algorithm performs block coordinate descent in the convex optimization problem (22). The coordinate blocks are given by matrices $\Sigma_1, \ldots, \Sigma_N$ and S. Such an algorithm is guaranteed to converge [16], [14].

The algorithm tuning parameters are the parameters of the two-level model explained in Section III, $\alpha > 0$, $\Omega \in \mathbf{S}_+^m$ (18), and $\beta > 0$ (16). The scalar α and matrix Ω in (18) should be chosen in the same way as the parameters α_1 and Ω_1 for the inverse Wishart prior in Section II-A. For large α, the population-level covariance S will be weighted heavily toward Ω. The scalar β affects the strength of the prior for the covariance matrices Σ_i. A larger value of β enforces greater similarity amongst the population of the covariances.

In the proposed algorithm, the optimal estimates depend on the data (1) through the scatter matrices X_iX_i, $Y_iX_i^T$, and $Y_iY_i^T$ only. This enables a recursive solution for a large (or growing) dataset (1). As new data is added, the scatter matrices can be recursively computed using rank-one update. Furthermore, for the proposed algorithm, the scatter matrices for each unit do not need to be loaded into memory simultaneously. This allows to parallelize computation of Σ_i.

D. Algorithmic Complexity

To analyze the computational complexity of the proposed algorithm, it is assumed that $O(n) = O(m)$, and that all units have approximately the same number of data points, and that there are many more data points per unit than there are units. i.e. $T_i \approx T \gg m, n$.

1) The cost of computing scatter matrices $X_i X_i^T$, $Y_i X_i$, and $Y_i Y_i$ for N units is approximately $O(n^2 NT)$.
2) Computing B_i (25) for each unit given the scatter matrices costs $O(n^3)$. This is done once for each unit, so the overall complexity is $O(n_x^3 N)$.
3) Computing coefficients of Riccati equation (24) costs $O(n^2)$ for each unit. Solving it costs $O(n^3)$ [5]. Over all units, the complexity is $O(n^3 N)$.
4) Coomputing Σ_{true} from (23) costs $O(n^2 N)$.

The global optimum is found by completing steps 1 and 2, then iterating over steps 3 and 4 until convergence. This process costs $O(n^2 NT + n^3 N N_{\text{iterations}})$, where $N_{\text{iterations}}$ is the number of iterations over steps 2 and 3 required for convergence. The recursive formulation allows for $\widetilde{T}$ new data points to be added for each unit, and the estimates updated, for a cost of $O(n^2 N\widetilde{T} + n^3 N N_{\text{iteration}})$. These computations can be fully parallelized over N processors at the cost $O(n^2\widetilde{T} + n^3 N_{\text{iterations}})$ for each.

V. Examples

The first of two examples in this section is a simple two-output dataset explain the use of the proposed two-level estimation approach. The second example describes a gas turbine fleet and illustrates the algorithm performance as compared to the baseline approaches.

A. Example 1: Two-output Dataset

We consider a dataset (1) with $m = 2$, $n = 1$, $N = 4$ and $T_i = 100$ for all valid i. We assume that $x_i(t) = 1$ for all i and t. Then the matrices $B_i \in \mathbf{R}^{2\times 1}$ are the means of the multivariate normal distributions for $y_i(t)$.

In this example, the elements of B_i (a column) are generated using a zero-mean normal distribution with covariance $3I_{2\times 2}$. The population covariance matrices $\Sigma_i \in \mathbf{S}^2$ are generated independently for each unit following $\Sigma_i \sim \mathcal{W}(S/100, 100 + m + 1)$ distribution with

$$S = \begin{bmatrix} 1 & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{10} \end{bmatrix}. \tag{26}$$

Matrices B_i, Σ_i specify the model for unit i. The output vectors $y_i(t)$ were generated independently from the distributions $y_i(t) \sim \mathcal{N}(B_i, \Sigma_i)$.

The generated dataset (1) was processed using the proposed method. The algorithm of Section IV-C was tuned by using the regularization parameters $\alpha = 0$, $\Omega = I_{2,2}$ in (18), and $\beta = 100$ in (16). The resulting covariance estimates for each unit are shown in Figure 3. Each ellipsoid is described by the center B_i and the covariance matrix Σ_i. The plotted ellipsoids are scaled up by a factor of 3 such that they contain most of the data points. The novel (proposed) algorithm implemented sequentially runs on a PC in 0.09 seconds.

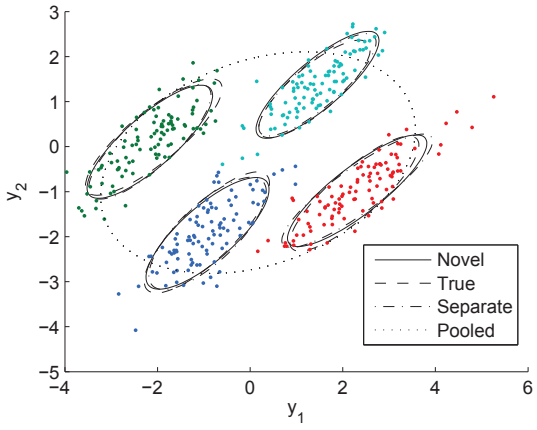

Fig. 3. Estimation results for Example 1 two-output dataset

TABLE I
Example 2 Gas Turbine Model Inputs and Outputs

Variable	Description	Unit
$x[1]$	inlet temp.	$^\circ F$
$x[2]$	inlet pressure	in. H_2O
$x[3]$	part load fraction	$nondimensional$
$x[4]$	steam injection flow	lb/hr
$y[1]$	heat rate	BTU/kWh
$y[2]$	efficiency	$nondimensional$

The ellipsoids corresponding to the ground truth covariance matrices Σ_i and the means B_i are plotted by the dashed lines, for all 4 units. The data $y_i(t)$ used in the estimation are plotted as dots. The ellipsoids produced by the proposed novel method are plotted by solid lines. They correspond to the estimated covariance matrices $\widehat{\Sigma}_i$ (enlarged by a factor of 3) and the estimated means $\widehat{B}_i$. As the figure shows, the ground truth and the estimated ellipsoids match closely.

B. Example 2: Gas Turbine Fleet

The next example uses simulated data from a fleet of gas turbines for power generation. The models of the form (14), estimated from the data, could be used for statistical monitoring of the performance of individual turbines.

The turbine model is based on [15]. The meanings of components $x[j]$ of the input vector x and the components $y[j]$ of the output vector y are explained in Table V-B. The model of [15] was linearized around the heat rate of $10,000$ BTH/kWh and efficiency of 80% (this is the operating point given in [15]). A central linear model of the form (14) is then described by a parameter matrix

$$B_{\text{ideal}} = \begin{bmatrix} -0.2962 & -0.2936 & 1 & 0.733 \\ 0.314 & 0.385 & 0.947 & -0.951 \end{bmatrix} \tag{27}$$

The linear input-output map B_i for each unit in the simulated fleet of the turbines was generated from the central model (27) as follows. Each column of B_i was normally distributed with mean given by the corresponding column of B_{ideal} and covariance matrix Σ_i. The covariances $\Sigma_i \in \mathbf{S}^2$ were generated for each unit, independently according to $\Sigma_i \sim \mathcal{W}(S/\beta, \beta + m + 1)$, following the assumptions

TABLE II

EXAMPLE 2 TURBINE FLEET RESULTS: 30 UNITS, 100 SAMPLES EA

Performance	$\left\|\Sigma_i - \widehat{\Sigma}_i\right\|_{\mathrm{F}}$	$\left\|\Sigma_i^{-1} - \widehat{\Sigma}_i^{-1}\right\|_{\mathrm{F}}$
Novel	0.0271	0.624
Separate	0.0536	1.65
Pooled	1.17	8.92

TABLE III

EXAMPLE 2 TURBINE FLEET RESULTS: 500 UNITS, 5000 SAMPLES EA

Performance	$\left\|\Sigma_i - \widehat{\Sigma}_i\right\|_{\mathrm{F}}$	$\left\|\Sigma_i^{-1} - \widehat{\Sigma}_i^{-1}\right\|_{\mathrm{F}}$
Novel	$4.38 \cdot 10^{-4}$	15.1
Separate	$5.23 \cdot 10^{-4}$	17.4
Pooled	0.129	794

made in Section III-B. The 2×2 population-wide generating covariance matrix S was taken to be the same as in (26).

Given B_i and Σ_i, the dataset was generated as follows. Input vectors $x_i(t) \in \mathbf{R}^4$ for turbine i at time t, were independently sampled from the distribution $x_i(t) \sim \mathcal{N}(0, \Sigma_i^x)$, with $\Sigma^x = I$. The output vectors $y_i(t) \in \mathbf{R}^2$ were then generated in accordance with $y_i(t) \sim \mathcal{N}(B_i x_i(t), \Sigma_i)$.

We considered a population of $N = 5000$ individual units with $T_i = T = 5000$ data points for each units. The data was processed according to the proposed two-level method, as outlined in Figure IV-C, and also according to the both baseline techniques discussed in Section III-A. The algorithm used regularization parameters $\alpha = 0$, $\Omega = I_{2,2}$ in (18) , and $\beta = 2000$ in (16). The time taken to run the algorithm sequentially on a PC was 8.1 seconds.

C. Algorithm Performance

Our motivation has been primarily to estimate the covariance matrices. One natural performance metric is the population average Frobenius distance between the true covariance Σ_i and the estimated covariance $\widehat{\Sigma}_i$ for each unit.

$$\|\Sigma_i - \widehat{\Sigma}_i\|_{\mathrm{F}} \tag{28}$$

A possible motivation for the two level regression modeling of the turbine fleet data is multivariate monitoring of anomalies. Such monitoring could be implemented using Hotelling-type statistics

$$T_2 = \left(y_i(t) - \widehat{B}_i x_i(t)\right)^T \widehat{\Sigma}_i^{-1} \left(y_i(t) - \widehat{B}_i x_i(t)\right)$$

where $\widehat{B}_i$ and $\widehat{\Sigma}_i$ are the estimates of the respective matrices obtained by the algorithm. The monitoring performance, then, depends on the estimation accuracy of the inverse covariance matrix. Therefore, the second performance metrics of the algorithm is the population-wide average of

$$\|\Sigma_i^{-1} - \widehat{\Sigma}_i^{-1}\|_{\mathrm{F}} \tag{29}$$

Monte Carlo simulations were completed to estimate (28) and (29) by repeatedly generating the data as described above. The average values obtained in the 100 completed simulations are summarized in Tables II and III. Both metrics (28) and (29) improve for the proposed method compared to the two baseline methods. The improvement is larger for the smaller dataset in Example 1. The model estimation errors $\|\widehat{B}_i - B_i\|_F$ are very small in all cases and are not shown.

VI. CONCLUSION

We have presented a technique for estimating linear models and covariances for two-level datasets. The proposed Bayesian approach assumes that the covariance matrices are sampled from a common population, It uses data from other units to improve the estimation of the individual covariances. The approach is scalable to very large datasets that may not fit into computer memory. In the provided simulation example, the technique shows an improvement over two versions of a standard one-level Bayesian approach to linear model and covariance estimation.

REFERENCES

[1] M. Arellano, "Computing robust standard errors for within-groups estimators," *Oxford Bulletin of Economics and Statistics*, Vol. 49, No. 4, 1987, pp.431–434.

[2] T. Asparouhov and B. Muthen, "Computationally efficient estimation of multilevel high-dimensional latent variable models," *Proc. Joint Statistical Meetings*, pages 2531–2535, July 2007, Salt Lake City, UT.

[3] P.J. Bickel and E. Levina, "Regularized estimation of large covariance matrices," *The Annals of Statistics*, Vol. 36, No. 1, 2008, pp. 199–227.

[4] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.

[5] A. Chinchuluun and P.M. Pardalos, *Optimization and optimal control*, volume 39. Springer, 2010.

[6] E. Chu, D. Gorinevsky, and S.P. Boyd, "Scalable statistical monitoring of fleet data," *18th IFAC World Congress*, volume 18, pages 13227–13232, Milano, Italy, August 2011.

[7] R. Couillet and M. Debbah, "Signal processing in large systems: a new paradigm," *IEEE Signal Processing Magazine*, Vol. 30, No. 1, 2003, pp. 24–39.

[8] A.E. Gelfand and A.F.M. Smith. "Sampling-based approaches to calculating marginal densities," *Journ. American Statistical Assoc.*, Vol, 85, No. 410, 1990, pp. 398–409.

[9] H. Goldstein. *Multilevel Statistical Models*, Wiley Series in Probability and Statistics, Wiley, 2010.

[10] W. James and C. Stein, "Estimation with quadratic loss," *Proc. 4th Berkeley Symp. on Math. Statistics and Probability*, volume 1, pages 1–379, 1961.

[11] O. Ledoit and M. Wolf, "A well-conditioned estimator for large-dimensional covariance matrices," *Journ. Multivariate Analysis*, Vol. 88, No. 2, 2004, pp. 365–411.

[12] S.Y. Lee and S.Y. Tsang, "Constrained maximum likelihood estimation of two-level covariance structure model via EM-type algorithms," *Psychometrika*, Vol. 64, No. 4, 1999, pp. 435–450.

[13] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979.

[14] Yu. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journ. Optimization*, Vol. 22, No. 2, 2012, pp. 341–362.

[15] J. Petek and P. Hamilton, "Performance monitoring of gas turbines," *Orbit*, Vol. 25, No. 1, 2005, pp. 64–74.

[16] P. Richtárik and M. Takáč, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Mathematical Programming, Ser. A*, 2012, pp. 1–38.

[17] B.G. Tabachnick, L.S. Fidell, and S.J. Osterlind, *Using multivariate statistics*, 4th Ed., New York: Allyn and Bacon, 2001.

[18] J. Teachman, G.J. Duncan, W.J. Yeung, and D. Levy, "Covariance structure models for fixed and random effects," *Sociological Methods & Research*, Vol 30, No. 2, 2001, pp. 271–288.

[19] J.H. Won and S.J. Kim, "Maximum likelihood covariance estimation with a condition number constraint," *Proc. Asilomar Conf. on Signals, Systems and Computers*, pages 1445–1449. Monterey, CA, Oct. 2006.