Statistical Modeling of Fleet Data

Nicholas Moehle[1]

[1]Department of Mechanical Engineering
Stanford University

Advisor: Dimitry Gorinevsky
Information Systems Laboratory
Department of Electrical Engineering
Stanford University

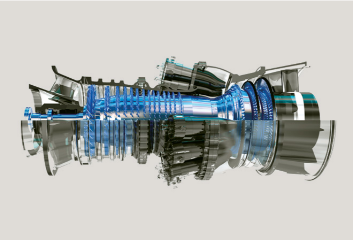
June 12, 2012

Introduction to *Fleet Data*

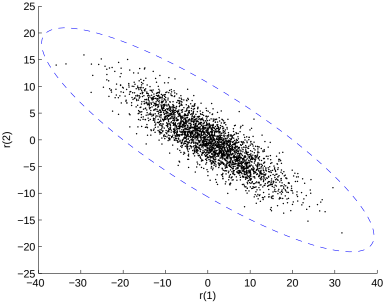
Consider the following problem:

- When GE sells a gas turbine, they offer a *Monitoring and Diagnostics* service
- GE has 1150 turbines currently being monitored, (and hundreds are the 7FA model) (Source: Orbit, Vol. 31)
- Data must be processed remotely (GE M&D, Atlanta)

How do we (tractably) generate a model? How do we use this model to detect anomalies?



Review: Confidence ellipsoid & exceedance monitoring



- Consider the ellipsoid $\{y \in \mathbb{R}^2 | (y - Bx)^T \Sigma^{-1} (y - Bx) \leq \alpha\}$.
- Represents a confidence bound.
- If we know $y \sim \mathcal{N}(Bx, \Sigma)$ in normal operation, then we use Bx and Σ for *anomaly detection*.
- Bx alone doesn't help us, need an accurate Σ.
- We will focus on finding B and Σ (x is assumed to be known).

What is fleet data?

Dataset is hierarchical:

- *Unit:* There exists N "units," so the data is divided into N subsets (indexed $i = 1, ..., N$).
- *Time:* Each of the N subsets contains T data points (indexed $t = 1, ..., T$).
  - t interpreted as sample time.
  - T is the same for each unit (i.e. each unit has the same number of data points).
- *Input/Output:* Each unit has an input output structure that is known *a priori* (input is $x_i(t)$, output is $y_i(t)$).
- *Multivariate data points* Each output data point $y_i(t)$ and input data point $x_i(t)$ is a vector in $\mathbb{R}^{n_y}$ and $\mathbb{R}^{n_x}$, respectively.

Therefore, we can summarize the data structure compactly by writing

$$\left\{ \{x_i(t), y_i(t)\}_{t=1}^{T} \right\}_{i=1}^{N}$$

Objectives

We propose a regression model of the following form:

$$y_i(t) = B_i x_i(t) + v_i(t)$$

with the following known variables:

- $y_i(t) \in \mathbb{R}^{n_y}$ is the (known) output of unit i at time t.
- $x_i(t) \in \mathbb{R}^{n_x}$ is the (known) exogenous input for unit i at time t.

We want to choose B_i such that:

- each $v_i(t)$ has low covariance.
- B_i is "similar" to B_j, $\forall\ i \neq j$
- The distribution of $v_i(t)$ is "similar" to that of $v_j(t)$.

We want reasonable computational complexity.

Regression approach

Recall the assumption:

$$y_i(t) = B_i x_i(t) + v_i(t)$$

For this specific approach, further define:

- The residual:
  - $v_i(t)|S \sim \mathcal{N}(0, S)$, i.i.d.
  - $S \in \mathbb{S}_+^{n_y}$ is the (unknown) residual covariance.
- The unit linear model:
  - $B_i|B_{\text{true}} \sim N_{n,k}(B_{\text{true}}, S/\alpha, I)$, i.i.d.
  - $N(\cdot,\cdot,\cdot)$ denotes the *matrix normal distribution*. For us, for each column of B_i, $b_i \sim \mathcal{N}(b_{\text{true}}, S/\alpha)$.
  - α is a weight, chosen *a priori*.
- No prior information is given about B_{true} or S.

MAP estimation of parameters

We find of $B_1, \dots, B_N$, B_{true}, and S which minimize the negative log likelihood function.
Using the law of total probability,

$$\ell(B_1, \dots, B_N, B_{\text{true}}, S|\mathcal{X}, \mathcal{Y}) = -\log f(\mathcal{X}, \mathcal{Y}|B_1, \dots, B_N, B_{\text{true}}, S)$$

$$= -\log \left(\prod_{i=1}^{N} \prod_{t=1}^{T} f(v_i(t)|B_1, \dots, B_N, B_{\text{true}}, S) \right)$$

$$= \sum_{i=1}^{N} \left(-\log f(B_i|B_{\text{true}}, S) - \sum_{t=1}^{T} \log f(v_i(t)|B_i, S) \right)$$

We have omitted the terms $f(B_{\text{true}})$ and $f(S)$, as there is no prior information about B_{true} or S (we can consider them to have improper priors).

MAP estimation of parameters

Plugging in the log-likelihoods and simplifying:

$$\ell = NT \log |S| + \sum_{i=1}^{N} \Bigg(\alpha \operatorname{tr}\left((B_i - B_{\text{true}})^T S^{-1} (B_i - B_{\text{true}}) \right) \\ + \operatorname{tr}\left((Y_i - B_i X_i)^T S^{-1} (Y_i - B_i X_i) \right) \Bigg)$$

- X_i and Y_i are data matrices
- constant terms are omitted
- Convex in the variables $S^{-1}B_i,\ S^{-1}B_{\text{true}},\ S^{-1}$

Normal equations

The normal equations are found by differentiating w.r.t. the (matrix) variables. They are:

- Unit Model:

$$Y_i X_i^T = \alpha(\widehat{B}_i - B_{\text{true}}) + \widehat{B}_i X_i X_i^T$$

- Average Model:

$$\widehat{B}_{\text{true}} = (1/N) \sum_{i=1}^{N} B_i$$

- Fleetwide Residual Covariance:

$$\widehat{S} = \frac{1}{NT} \sum_{i=1}^{N} \Big(\alpha(B_i - B_{\text{true}})(B_i - B_{\text{true}})^T + (Y_i - B_i X_i)(Y_i - B_i X_i)^T \Big)$$

Are we satisfied?

We chose B_i, B_{true} and S to minimize:

$$\ell = NT \log|S| + \sum_{i=1}^{N} \Bigg(\alpha \operatorname{tr}\left((B_i - B_{\text{true}})^T S^{-1} (B_i - B_{\text{true}}) \right) + \operatorname{tr}\left((Y_i - B_i X_i)^T S^{-1} (Y_i - B_i X_i) \right) \Bigg)$$

We wanted to encode the idea:

- each $v_i(t)$ should have low covariance. (Satisfied)
- B_i should be "similar" to B_j, $\forall\ i \neq j$ (Satisfied)
- The covariance of $v_i(t)$ should be "similar" to that of $v_j(t)$. (Unsatisfied)

Covariance approach

Recall the assumption:

$$y_i(t) = B_i x_i(t) + v_i(t)$$

For this specific approach, further define:

- The residual:
  - $v_i(t)|S \sim \mathcal{N}(0, S_i)$, i.i.d.
  - $S_i \in \mathbb{S}_+^{n_y}$ is the (unknown) residual covariance.
- The unit covariance:
  - $S_i \in \mathbb{S}_+^n$, where $S_i|S_{\text{true}} \sim \mathcal{W}(S_{\text{true}}/p, p)$, i.i.d.
  - $\mathcal{W}(\cdot,\cdot)$ is the Wishart distribution (for random matrix $Z \in \mathbb{R}^{n\times\nu}$, with columns of Z normally distributed, zero-mean i.i.d. random vectors with covariance Σ, then $ZZ^T \sim \mathcal{W}(\Sigma, \nu)$).
  - $p \in \mathbb{R}$ is a (known) weight parameter
- We have no prior information about B_i or S_{true}.

MAP estimation of parameters

We find the MAP estimates of S_{true}, and B_i and S_i, for all $i = 1, ..., N$.
Maximize the log likelihood function:

$$
\begin{aligned}
&\ell(S_1, ..., S_N, B_1, ..., B_N, S_{\text{true}} | \mathcal{X}, \mathcal{Y}) \\
&\quad = -\log \left(\prod_{i=1}^{N} \prod_{t=1}^{T} f(v_i(t) | S_1, ..., S_N, B_1, ..., B_N, S_{\text{true}}) \right) \\
&\quad = \sum_{i=1}^{N} \left(-\log f(S_i | S_{\text{true}}) - \sum_{t=1}^{T} \log f(v_i(t) | S_i, B_i) \right)
\end{aligned}
$$

As before, we ignore the terms $f(S_{\text{true}})$ and $f(B_i)$.

MAP estimation of parameters

Plugging in the log-likelihoods and simplifying:

$$= Np \log |S_{\text{true}}| + \sum_{i=1}^{N} \Bigg(p \operatorname{tr}\left(S_{\text{true}}^{-1} S_i\right) + (T + n + 1 - p) \log |S_i|$$

$$+ \operatorname{tr}\Big((Y_i - B_i X_i)^T S_i^{-1} (Y_i - B_i X_i)\Big)\Bigg)$$

With change of variables, $S_i^{-1} = P_i$, $L^T L = S_{\text{true}}^{-1}$, and $\widetilde{B}_i = S_i^{-1} B_i$, this is convex for large T:

$$= -Np \log \left| L^T L \right| + \sum_{i=1}^{N} \Bigg(p \operatorname{tr}\left(L^T P_i^{-1} L\right) - (T + n + 1 - p) \log |P_i|$$

$$+ \operatorname{tr}(Y_i^T P_i Y_i) - 2 \operatorname{tr}(Y_i^T \widetilde{B}_i X_i) + \operatorname{tr}((\widetilde{B}_i X_i)^T P_i^{-1} (\widetilde{B}_i X_i))\Bigg)$$

Normal equations

Two of the normal equations are found by differentiating w.r.t. the new variables, then substituting back into the natural variables:

- Unit Model:

$$Y_i X_i^T = \widehat{B}_i X_i X_i^T$$

- Average Covariance:

$$\widehat{S}_{\text{true}} = 1/N \sum_{i=1}^{N} S_i$$

- Unit Covariance:

$$0 = \widehat{S}_i(-pS_{\text{true}}^{-1})\widehat{S}_i - (1 + n + T - p)\widehat{S}_i + Y_i Y_i^T - B_i X_i X_i^T B_i^T$$

Normal equations

$$0 = \widehat{S}_i(-pS_{\text{true}}^{-1})\widehat{S}_i - (1 + n + T - p)\widehat{S}_i + Y_iY_i^T - B_iX_iX_i^TB_i^T$$

If we consider another change of variables:

- $Q^{(r)} = Y_iY_i^T - B_iX_iX_i^TB_i^T$
- $A^{(r)} = -(1/2)(1 + n + T - p)I$
- $B^{(r)} = I$
- $R^{(r)} = (1/p)S_{\text{true}}$

This can be solved easily as an algebraic Riccati equation:

$$A^TX + XA - XBR^{-1}B^TX + Q = 0$$

Normal equations

Summarizing:

$$Y_i X_i^T = \widehat{B}_i X_i X_i^T$$

$$0 = \widehat{S}_i(-p S_{\text{true}}^{-1})\widehat{S}_i - (1 + n + T - p)\widehat{S}_i + Y_i Y_i^T - B_i X_i X_i^T B_i^T$$

$$\widehat{S}_{\text{true}} = 1/N \sum_{i=1}^{N} S_i$$

The following algorithm can be used to obtain S_i, S_{true}, and B_i for all i

1. Compute $\widehat{B}_i$
2. Initialize $\widehat{S}_i$ and $\widehat{S}_{\text{true}}$.
3. Compute $\widehat{S}_i$
4. Compute $\widehat{S}_{\text{true}}$
5. Check the variables have converged. If so, stop. If not, go to step 3.

Convergence is guaranteed, by convexity

Now are we satisfied?

We chose B_i, S_{true} and S_i to minimize:

$$\ell = Np \log |S_{\text{true}}| + \sum_{i=1}^{N} \Big(p \operatorname{tr}\left(S_{\text{true}}^{-1} S_i\right) + (T + n + 1 - p) \log |S_i| + \operatorname{tr}\left((Y_i - B_i X_i)^T S_i^{-1} (Y_i - B_i X_i)\right) \Big)$$

We wanted to encode the idea:

- each $v_i(t)$ should have low covariance (satisfied)
- B_i should be "similar" to B_j, $\forall\ i \neq j$ (unsatisfied)
- The covariance of $v_i(t)$ should be "similar" to that of $v_j(t)$. (satisfied, though not obvious)

Brief simulation example

$$y_i(t) = B_i x_i(t) + v_i(t)$$

Random variables were generated according to:

- $x_i(t) \in \mathbb{R}^n$ $x_i(t) \sim \mathcal{N}(0, \Sigma_x)$, i.i.d. the (known) input of unit i at time t.
- $v_i(t) \in \mathbb{R}^n$, $v_i(t) \sim \mathcal{N}(0, S_i)$, i.i.d., is the residual for unit i at time t.
- $B_i \in \mathbb{R}^{n \times k}$, $B_i \sim N_{n,k}(B_{\text{true}}, S_i, I)$ is the static linear map for turbine i.
- $S_i \sim \mathcal{W}(S_{\text{true}}/p, p)$ is the covariance of the residual, with p degrees of freedom.

Constants for simulation

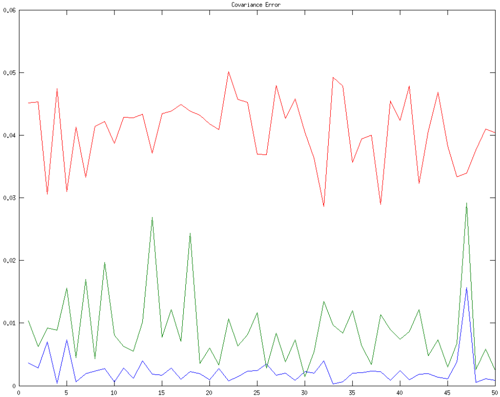
$$\Sigma_x = \begin{bmatrix} 1 & 0.5 & 0.25 & 0 \\ 0.5 & 1 & 0.5 & 0 \\ 0.25 & 0.25 & 1 & 0 \\ 0 & 0 & 0 & 0.001 \end{bmatrix} \qquad S_{\text{true}} = \begin{bmatrix} 0.1 & 0.01 \\ 0.01 & 0.001 \end{bmatrix}$$

$$B_{\text{true}} = \begin{bmatrix} -29.62 & -29.36 & 1 & 0.0733 \\ 0.0314 & 0.0385 & 0.947 & -9.51 \times 10^{-5} \end{bmatrix}$$

$$T = 100 \qquad N = 50 \qquad p = 10 \qquad \alpha = 1$$

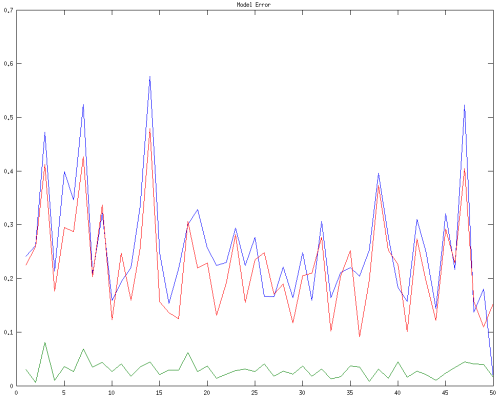
B_{true} obtained from "Performance monitoring of gas turbines," *Journal of Orbit*, Vol. 25, 2005

Results (unit covariance Error)



The error $\|S_i - \widehat{S}_i\|$ vs. i, for the regression model (green), the covariance model (blue), and the naive model (red)

Results (unit model error)



The error $\|B_i - \widehat{B}_i\|$ vs. i, for the regression model (green), the covariance model (blue), and the naive model (red)

Future work

- Using these approaches on real data will prove their efficacy. We are actively seeking such (real) data.
- When are these formulations better than naive approaches? When are they not?
- Is there a formulation that will acheive our original objectives? As a reminder:
  - each $v_i(t)$ should have low covariance
  - B_i should be "similar" to B_j, $\forall\ i \neq j$
  - The covariance of $v_i(t)$ should be "similar" to that of $v_j(t)$.
- Is the Wishart distribution the best prior for the unit covariances?

References

1. E. Chu, D. Gorinevsky, and S. Boyd. Detecting aircraft performance anomalies from cruise flight data. In AIAA Infotech Aerospace Conference, Atlanta, GA, 2010.
2. E. Chu, D. Gorinevsky, and S.P. Boyd. Scalable statistical monitoring of fleet data. In World Congress, volume 18, pages 1322713232, 2011.
3. M. Devaney and B. Cheetham. Case-based reasoning for gas turbine diagnostics. In AAAI 2005, 2005.
4. D. Gorinevsky. Bayesian fault isolation in multivariate statistical process monitoring dimitry gorinevsky. In American Control Conference (ACC), 2011, pages 19631968. IEEE, 2011.
5. P. D. hoff. Separable covariance arrays via the Tucker product, with applications to multivariate relational data. eprint arXiv:1008.2169, August 2010.
6. J. Petek and P. Hamilton. Performance monitoring of gas turbines. Journal of Orbit, 25(1):6474, 2005.
7. K.B. Petersen and M.S. Pedersen. The matrix cookbook. Technical University of Denmark, pages 715, 2008.
8. R. Rosipal and N. Kramer. Overview and recent advances in partial least squares. Subspace, a Latent Structure and Feature Selection, pages 3451, 2006.
9. AJ Volponi, H. DePold, R. Ganguli, and C. Daguang. The use of kalman filter and neural network methodologies in gas turbine performance diagnostics: a comparative study. Journal of engineering for gas turbines and power, 125:917, 2003.
10. Y. Zhang and J. Schneider. Learning multiple tasks with a sparse matrix-normal penalty. Advances in Neural Information Processing Systems, 2010.