TWCCC ⋆ Texas – Wisconsin – California Control Consortium

Technical report number 2023-02

On multivariate linear regression with singular error covariances

Steven J. Kuntz[*] James B. Rawlings[*]

December 12, 2023

Abstract

Multivariate linear regression is a classic statistical method that has been used in a wide array of scientific and engineering fields, in some for over two centuries. While the maximum likelihood estimation problem is well-solved in the case of nonsingular data and error covariance matrices, the nonsingular case is less well understood, especially the singular error covariance case. The purpose of this report is to define and derive the maximum likelihood of the singular multivariate regression model, under no assumptions about the rank of the underlying data or parameters. We show that a naïve definition of the estimator has no solutions, almost surely, but it can be rigorously defined so that solutions exist and coincide with the nonsingular case. Illustrative examples of the technical results are included throughout, and applied examples in system identification are included after the technical results.

1 Introduction

Consider the multivariate linear regression model,

$$y_k = \Theta_0 x_k + e_k, \qquad\qquad e_k \overset{\text{i.i.d.}}{\sim} \mathrm{N}_p(0, \Sigma_0) \tag{1}$$

where $k = 1, \ldots, N$ is the sample index, $y_k, e_k \in \mathbb{R}^p$ are the measurements and measurement errors, $x_k \in \mathbb{R}^n$ are the predictors, N_p is the p-dimensional vector normal distribution (to

[*] University of California Santa Barbara (skuntz@ucsb.edu, jbraw@ucsb.edu).

be defined), and $\Theta_0 \in \mathbb{R}^{p\times n}$ and $\Sigma \succeq 0 \in \mathbb{R}^{p\times p}$ are the model parameters. It is convenient to express the linear model (1) in a compact matrix form,

$$Y_N = \Theta_0 X_N + E_N, \qquad\qquad E_N \sim \mathrm{N}_{p\times N}(0, \Sigma_0, I_N) \tag{2}$$

where $Y_N := \begin{bmatrix} y_1 & \dots & y_N \end{bmatrix}$, $E_N := \begin{bmatrix} e_1 & \dots & e_N \end{bmatrix}$, and $X_N := \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix}$ are matrices of measurements, measurement errors, and predictors, respectively, and $\mathrm{N}_{p\times N}$ is the $(p \times N)$-dimensional matrix normal distribution (to be defined). When the sample size N does not change, we suppress this notation and simply write $Y = Y_N$, $X = X_N$, and $E = E_N$.

A great deal of effort has been devoted to the study of estimators for the model (2), including maximum likelihood (ML) estimators [1, Ch. 8], maximum a posteriori (MAP) estimators [2, 3], and reduced-rank regression (RRR) estimators [4–8]. The vast majority of these results, however, assume the matrices $XX^\top$ and Σ_0 are nonsingular. As sensors have become cheaper and large systems rely more heavily on automation, singular and ill-conditioned problems may arise in a number of practical scenarios, including high-dimensional sensing (e.g., image processing, spectroscopy), systems with physical constraints or feedback (e.g., conservation laws, biological systems, controlled systems), and the analysis of happenstance data (e.g., process monitoring, state estimation). As such, a general theory for handling singular multivariate linear regression problems is needed.

A few papers have been devoted to ML estimators of the model (2) under the assumption that $XX^\top$ and Σ_0 may be singular [9, 10], but each of these papers assumes the rank of Σ_0 is known a priori, and none of these results are built upon a solid probabilistic basis for the definition of the ML estimator. The purpose of this report is to define and find the ML estimator of the parameters (Θ_0, Σ_0) of the model (2). We pose our definition in a way that makes it clear where the rank constraint on Σ_0 originates. In Section 2, we review the case of nonsingular $XX^\top$ and Σ_0. In Section 3 we provide a measure-theoretic definition of the singular normal probability density. In Section 4, we show that a naïve definition of the ML estimator has no solutions, almost surely. Finally, in Section 5, we provide a rank-constrained definition of the ML estimator that has solutions, almost surely, and show how the rank can be computed from the data. Many of the proofs and preliminary results are deferred to the appendices.

Notation. Throughout, we let $(\Omega, \mathcal{F}, \mathbb{P})$ denote a common probability space. A random variable Y on that probability space is a measurable function $Y : \Omega \to \mathcal{Y}$ from the probability space $(\Omega, \mathcal{F})$ to a measurable space $(\mathcal{Y}, \mathcal{F}_Y)$. We typically suppress this notation and simply call $Y \in \mathcal{Y}$ a random variable, with the σ-algebra $\mathcal{F}_Y$ implied from context. The random variable Y is completely described by its *probability distribution* $\mathbb{P}_Y := \mathbb{P} \circ Y^{-1}$ which is a probability measure on $(\mathcal{Y}, \mathcal{F}_Y)$. Unless otherwise specified, assume the σ-algebra of a random variable Y on a Banach space $\mathcal{Y}$ is the standard Borel algebra of $\mathcal{Y}$, denoted $\mathcal{B}(\mathcal{Y})$. As a shorthand, we let $\mathbb{P}[P] := \mathbb{P}(\{\, \omega \in \Omega : P(\omega) \,\})$ for any propositional function $P : \Omega \to \{\, \text{True}, \text{False} \,\}$. Let $\sim$ denote the phrase "is distributed as" and $\overset{\text{i.i.d.}}{\sim}$ denote the phrase "are independently and identically distributed as." We say $p : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is a *probability density function* (PDF) of the random variable $Y \in \mathcal{Y}$ with respect to a reference measure $\mu : \mathcal{F}_Y \to \mathbb{R}_{\geq 0}$ if $\mathbb{P}[Y \in \mathcal{A}] = \int_{\mathcal{A}} p d\mu$ for all $\mathcal{A}$ in the σ-algebra $\mathcal{F}_Y$. The

Radon-Nikodým theorem states a PDF of $Y \in \mathcal{Y}$ w.r.t. μ exists if and only if $\mu(\mathcal{A}) = 0$ implies $\mathbb{P}[Y \in \mathcal{A}] = 0$ for all $\mathcal{A} \in \mathcal{F}_Y$, and moreover, such a density is unique up to μ-null sets (c.f. [11, Thm. 32.2] or [12, Thm. 19.2]). Common reference measures include the n-dimensional Lebesgue measure λ^n and the $(m \times n)$-dimensional Lebesgue measure $\lambda^{m\times n}$.

2 Nonsingular Σ_0 case

Before we attempt the singular Σ_0 case, let us review the nonsingular Σ_0 case so they can be compared. Under the assumption that Σ_0 is nonsingular, the ML estimates of (Θ_0, Σ_0) for the model (2) are given by solving

$$\max_{\Theta \in \mathbb{R}^{p\times n}, \Sigma \succ 0 \in \mathbb{R}^{p\times p}} p(Y|X, \Theta, \Sigma) \tag{3}$$

By linearity, $Y|(X, \Theta, \Sigma) \sim \mathrm{N}_{p\times N}(\Theta X, \Sigma, I_N)$, and we have the conditional PDF (w.r.t. the Lebesgue measure $\lambda^{p\times N}$) [13, Thm. 2.2.1]:

$$p(Y|X, \Theta, \Sigma) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[\Sigma^{-1}(Y - \Theta X)(Y - \Theta X)^\top]\right)}{(2\pi)^{pN/2}|\Sigma|^{N/2}} \tag{4}$$

Taking the negative logarithm of (4) and dropping constants, we can equivalently write the maximization problem (3) as

$$\min_{\Theta \in \mathbb{R}^{p\times n}, \Sigma \succ 0 \in \mathbb{R}^{p\times p}} \phi(\Theta, \Sigma) := \frac{N}{2} \ln |\Sigma| + \frac{1}{2}\mathrm{tr}[\Sigma^{-1}(Y - \Theta X)(Y - \Theta X)^\top] \tag{5}$$

Solutions follow naturally from (5) as a staged minimization problem: first Θ is minimized as a function of Σ using convex programming theory, and then Σ is minimized after substituting back in the Θ solution. A general characterization of solutions to (3) are given in Proposition 1, for which a proof is supplied in Appendix A.

Proposition 1. The ML problem (3) has solutions if and only if $Y(I_N - X^+X)Y^\top$ is nonsingular. Moreover, if solutions exist, the pair $(\hat{\Theta}, \hat{\Sigma})$ solves (3) if and only if

$$\hat{\Theta} \in \{\, YX^+ + Q : \mathcal{R}(Q^\top) \subseteq \mathcal{N}(X^\top) \,\}, \qquad \hat{\Sigma} = \frac{1}{N} Y(I_N - X^+X)Y^\top \tag{6}$$

Proposition 1 reveals exactly when solutions to (3) exist and are unique. Specifically, existence of a solution depends on the data matrix $Y(I_N - X^+X)Y^\top$ being nonsingular. On the other hand, given a solution exists, uniqueness of the solution depends on the data matrix $XX^\top$ being nonsingular. To illustrate the importance of these data matrices, we consider the following elementary examples.

Example 2. In this example, we construct the simplest system where $XX^\top$ is singular. Suppose $n = p = 1$ and the true system (2) is generated by the parameters $\Theta_0 = \Sigma_0 = 1$.

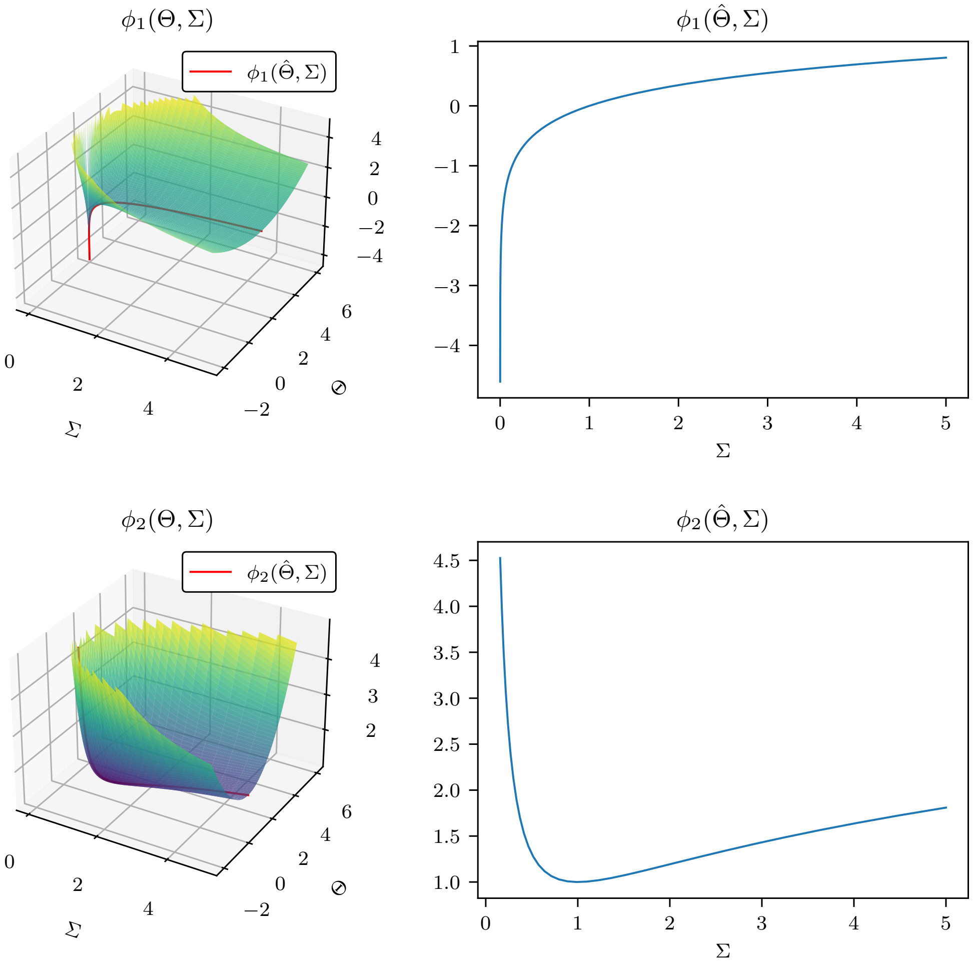

Figure 1: Plots of the objective functions (top) $\phi_1(\Theta, \Sigma)$ and (bottom) $\phi_2(\Theta, \Sigma)$ for Example 3.

We observe a sequence of samples $y_1, \dots, y_N$ corresponding to the predictors $x_1 = \dots = x_N = 0$. Clearly $XX^\top = 0$ is singular, and the objective is

$$\phi(\Theta, \Sigma) = \frac{N}{2} \ln |\Sigma| + \frac{1}{2\Sigma} \sum_{k=1}^{N} y_k^2$$

Since $\phi(\cdot, \Sigma)$ is a constant for each Σ, we can simply minimize Σ, which occurs at $\hat{\Sigma} = N^{-1} \sum_{k=1}^{N} y_k^2$, and set $\hat{\Theta} \in \mathbb{R}$. △

Example 3. In this example, we construct the simplest system where $Y(I_N - X^+ X)Y^\top$ is nonsingular. Suppose $n = p = 1$ and the true system (2) is generated by the parameters $\Theta_0 = \Sigma_0 = 1$. We observe a single sample $(x_1, y_1) = (1, 2)$. Then $y_1(1 - x_1^{-1} x_1) y_1 =$

$2(1-1)2=0$ and the likelihood function is

$$\phi_1(\Theta,\Sigma)=\frac{1}{2}\ln\Sigma+\frac{(2-\Theta)^2}{2\Sigma}$$

With $\Sigma>0$ fixed, we can minimize $\phi_1(\cdot,\Sigma)$ at $\hat{\Theta}=2$. However, evaluating the likelihood at $\hat{\Theta}=2$ gives

$$\phi_1(\hat{\Theta},\Sigma)=\frac{1}{2}\ln\Sigma$$

which is unbounded from below with $\Sigma\searrow 0$. Therefore no estimate of Σ can be obtained. However, this is not an unexpected result, as we have not collected enough data to estimate the parameters. Suppose a second sample $(x_2,y_2)=(1,0)$ is available. Then

$$\begin{bmatrix}y_1 & y_2\end{bmatrix}\left(I_2-\begin{bmatrix}x_1 & x_2\end{bmatrix}^+\begin{bmatrix}x_1 & x_2\end{bmatrix}\right)\begin{bmatrix}y_1\\y_2\end{bmatrix}=\begin{bmatrix}2 & 0\end{bmatrix}\left(I_2-\begin{bmatrix}1/2 & 1/2\\1/2 & 1/2\end{bmatrix}\right)\begin{bmatrix}2\\0\end{bmatrix}=1>0$$

and

$$\phi_2(\Theta,\Sigma)=\ln\Sigma+\frac{(2-\Theta)^2+(0-\Theta)^2}{2\Sigma}=\ln\Sigma+\frac{2(1-\Theta)^2+2}{2\Sigma}$$

With $\Sigma>0$ fixed, we can minimize $\phi_2(\cdot,\Sigma)$ at $\hat{\Theta}=1$. And evaluating the likelihood gives

$$\phi_2(\Theta,\Sigma)=\ln\Sigma+\frac{1}{\Sigma}$$

which has a minimum at $\hat{\Sigma}=1$. △

We conclude our discussion of the nonsingular Σ_0 case with the remark that Example 3 reveals the importance of acquiring sufficient data. It can be shown that $N=n+p$ is the minimum sample size for which we can design X to guarantee the existence and uniqueness of ML estimators of (Θ_0,Σ_0). However, the facts required to show this are easily subsumed into the singular Σ_0 case, so we defer the discussion until then.

3 Singular normal vectors and matrices

Before considering the singular version of (3), we take an aside to discuss the singular normal distribution. Recall, for any $m\in\mathbb{R}^n$ and positive definite $S\succ 0\in\mathbb{R}^{n\times n}$, the multivariate normal vector $z\sim\mathrm{N}_n(m,S)$ has the PDF,

$$p(z|m,S)=\frac{\exp\left(-\frac{1}{2}(z-m)^\top S^{-1}(z-m)\right)}{(2\pi)^{n/2}|S|^{1/2}} \tag{7}$$

with respect to the Lebesgue measure λ^n, and similarly for the conditional PDF.[1] As shown by Cramér [14, p. 290], Rao [15, pp. 527–528], and Srivastava and von Rosen [16, p. 4], if S is singular, then $z\sim\mathrm{N}_n(m,S)$ lies on a rank(S)-dimensional subspace of $\mathbb{R}^n$, almost surely.

[1] This PDF is defined with respect to the n-dimensional Lebesgue measure λ^n

As a result, $\lambda^n(\mathcal{A}) \leq \lambda^n(m + \mathcal{R}(S)) = 0$ for any measureable subset $\mathcal{A}$ of the affine space $m+\mathcal{R}(S)$. However, $z \in m+\mathcal{R}(S)$ almost surely, so by the Radon-Nikodým theorem z has no PDF with respect to λ^n. To define the multivariate normal for singular covariances, we need a definition that does not start with the PDF (7). We take the approach of Rao [15, p. 522] and fall back to the scalar normal distribution $z \sim \mathrm{N}(\mu, \sigma^2)$, defined in the classical sense for $\sigma^2 > 0$ (using, e.g., the Box-Muller transform), and defined as having a Dirac probability measure at $z = \mu$ for $\sigma^2 = 0$. A random vector must satisfy linearity, so we define it as satisfying linearity for all outer products of the form $a^\top z$.

Definition 4 ([15, p. 522]). The random vector z is normally distributed with mean $m \in \mathbb{R}^n$ and covariance $S \succeq 0 \in \mathbb{R}^{n\times n}$, denoted $z \sim \mathrm{N}_n(m, S)$, if $a^\top z \sim \mathrm{N}(a^\top m, a^\top Sa)$ for all $a \in \mathbb{R}^n$.

Using this definition, one can reverse the linear transformation over a basis of a vectors, and derive the following PDF of a singular normal vector.

Proposition 5 ([9]). Let $z \sim \mathrm{N}_n(m, S)$ and define the affine map $f(\cdot) := m + U_1(\cdot) : \mathbb{R}^r \to \mathbb{R}^n$ where $r := \mathrm{rank}(S)$ and $S = U_1\Sigma_1U_1^\top$ is the thin SVD of S. Then z has a PDF

$$p(z|m,S) = \begin{cases} \frac{\exp\left(-\frac{1}{2}(z-m)^\top S^+(z-m)\right)}{(2\pi)^{r/2}|S|_+^{1/2}}, & z \in m + \mathcal{R}(S) \\ 0, & z \notin m + \mathcal{R}(S) \end{cases} \tag{8}$$

with respect to the reference measure $\mu := \lambda^r \circ f^{-1} : \mathcal{B}(\mathbb{R}^n) \to \mathbb{R}_{\geq 0}$.

While the PDF (9) is *not* a valid density with respect to the n-dimensional Lebesgue measure λ^n, it is a valid density with respect to $\mu := \lambda^r \circ f^{-1}$, the r-dimensional measure supported on the affine space $m + \mathcal{R}(S)$:

$$\mathbb{P}[z \in \mathcal{A}] = \int_{\mathcal{A}} p(z|m,S)d\mu(z) \tag{9}$$

for any $\mathcal{A} \in \mathcal{B}(\mathbb{R}^n)$. In fact, the density outside of the affine space $m + \mathcal{R}(S)$ can be set to any value one wishes without changing (9), but to keep things simple and consistent we choose 0. If S is nonsingular, then $z \in m + \mathcal{R}(S) = m + \mathbb{R}^n = \mathbb{R}^n$ is always satisfied, $|S|_+ = |S|$, and $S^+ = S^{-1}$, and the nonsingular case (7) is recovered from Proposition 5.

Similarly, we define the matrix normal distribution without reference to an underlying PDF by considering the vectorization of the normally distributed matrix.

Definition 6 ([13, Defn. 2.2.1]). The random matrix Z is normally distributed with mean $M \in \mathbb{R}^{m\times n}$ and covariances $U \succeq 0 \in \mathbb{R}^{m\times m}$ and $V \succeq 0 \in \mathbb{R}^{n\times n}$, denoted $Z \sim \mathrm{N}_{m\times n}(M, U, V)$, if $\mathrm{vec}(Z) \sim \mathrm{N}_{mn}(\mathrm{vec}(M), U \otimes V)$.

Using Definition 6 in conjunction with Proposition 5, one can reverse the vectorization and derive the following PDF of a singular normal matrix.

Proposition 7 ([9]). Let $Z \sim \mathrm{N}_{m\times n}(M, U, V)$ and define the affine map $f(\cdot) = M + W_1(\cdot)Q_1^\top : \mathbb{R}^{r\times s} \to \mathbb{R}^{m\times n}$ where $r := \mathrm{rank}(U)$, $s := \mathrm{rank}(V)$, and $U = W_1\Sigma_1W_1^\top$ and $V = Q_1D_1Q_1^\top$ are the thin SVDs of U and V. Then Z has a PDF

$$p(Z|M,U,V) = \begin{cases} \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[U^+(Z-M)V^+(Z-M)^\top]\right)}{(2\pi)^{rs/2}|U|_+^{n/2}|V|_+^{m/2}}, & Z \in \{\, M + UQV : Q \in \mathbb{R}^{m\times n} \,\} \\ 0, & Z \notin \{\, M + UQV : Q \in \mathbb{R}^{m\times n} \,\} \end{cases} \tag{10}$$

with respect to the reference measure $\mu := \lambda^{r\times s} \circ f^{-1} : \mathcal{B}(\mathbb{R}^{m\times n}) \to \mathbb{R}_{\geq 0}$.

Again, the PDF (9) is *not* taken with respect to the matrix Lebesgue measure $\lambda^{m\times n}$, but an affine transformation of the lower-dimensional tranformation of it $\mu := \lambda^{r\times s} \circ f^{-1}$:

$$\mathbb{P}[Z \in \mathcal{A}] = \int_{\mathcal{A}} p(Z|M,U,V)d\mu(Z) \tag{11}$$

for any $\mathcal{A} \in \mathcal{B}(\mathbb{R}^{m\times n})$, and the density outside the space $\{\, M + UQV : Q \in \mathbb{R}^{m\times n} \,\}$ could have been chosen arbitrarily, albeit with greater complexity in the presentation.

4 Singular Σ_0 case

If Σ_0 is possibly singular, the maximum likelihood estimates of (Θ_0, Σ_0) for the model (2) are given by solving

$$\max_{\Theta\in\mathbb{R}^{p\times n}, \Sigma\succeq 0\in\mathbb{R}^{p\times p}} p(Y|X,\Theta,\Sigma) \quad \text{subject to} \quad p(Y|X,\Theta,\Sigma) > 0 \tag{12}$$

where

$$p(Y|X,\Theta,\Sigma) = \begin{cases} \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[\Sigma^+(Y-\Theta X)(Y-\Theta X)^\top]\right)}{(2\pi)^{\mathrm{rank}(\Sigma)N/2}|\Sigma|_+^{N/2}}, & Y \in \{\, \Theta X + \Sigma Z : Z \in \mathbb{R}^{p\times N} \,\} \\ 0, & Y \notin \{\, \Theta X + \Sigma Z : Z \in \mathbb{R}^{p\times N} \,\} \end{cases}$$

Here we have avoided the case where $p(Y|X,\Theta,\Sigma)$ is zero independently of (Θ, Σ) to rule out models that would suggest we have just observed zero-probability data. Since the objective is positive in the feasible region, we can take the negative logarithm to produce the equivalent minimization problem:

$$\min_{\Theta\in\mathbb{R}^{p\times n}, \Sigma\succeq 0\in\mathbb{R}^{p\times p}} \phi(\Theta,\Sigma) \quad \text{subject to} \quad Y \in \{\, \Theta X + \Sigma Z : Z \in \mathbb{R}^{p\times N} \,\} \tag{13a}$$

where

$$\phi(\Theta,\Sigma) := \frac{N}{2}[\ln(2\pi)\mathrm{rank}\Sigma + \ln|\Sigma|_+] + \frac{1}{2}\mathrm{tr}[\Sigma^+(Y-\Theta X)(Y-\Theta X)^\top] \tag{13b}$$

We fully characterize the solutions to (3) in Proposition 1. While the solutions are identical to that of (3), the methods required are substantially different, so the proof is included in Appendix C for completeness.

Proposition 8. The ML problem (12) has solutions if and only if $Y(I_N - X^+X)Y^\top$ is nonsingular. Moreover, the pair $(\hat{\Theta}, \hat{\Sigma})$ solves (12) if and only if (6).

Proposition 1 again claims solutions to (12) will only exist when $Y(I_N - X^+X)Y^\top$ is nonsingular. This poses an issue for the singular Σ_0 case because the errors should naturally be rank deficient almost surely. We explore this problem in the following example, which is an extension of Example 3.

Example 9. Suppose $n = p = 2$ and the true system (2) is generated by the parameters

$$\Theta = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

$\Theta = \Sigma = 1$. Suppose we observe the data[2]

$$X_4 = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, \qquad Y_4 = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

Then the objective is

$$\phi_4(\Theta, \Sigma) = 2 \ln \Sigma + \frac{1}{2} \mathrm{tr} \left[\Sigma^{-1} \left(\begin{bmatrix} 6 & 2 \\ 2 & 2 \end{bmatrix} - 2\Theta \begin{bmatrix} 2 & 0 \\ 2 & 2 \end{bmatrix} + 2\Theta\Theta^\top \right) \right]$$

With $\Sigma > 0$ fixed, we can minimize $\phi_1(\cdot, \Sigma)$ at

$$\hat{\Theta} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

Evaluating the objective at $\hat{\Theta} = 2$ gives

$$\phi_4(\hat{\Theta}, \Sigma) = 2 \ln \Sigma + \frac{1}{2} \mathrm{tr} \left[\Sigma^{-1} \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix} \right]$$

which is unbounded from below because we can take $\Sigma_{12} = \Sigma_{21} = 0$ and $\Sigma_{22} \searrow 0$ to get

$$\phi_4(\hat{\Theta}, \Sigma) = 2 \ln \Sigma_{11} + 2 \ln \Sigma_{22} + \Sigma_{11}^{-1} \to -\infty$$

Therefore ϕ_4 is unbounded from below and no estimate of Σ can be obtained. △

In fact, no sufficient sample number can be taken to produce an estimate if Σ_0 is singular. By linearity, have that $E := Y - \Theta_0 X = \Sigma_0 Q$ for some $Q \sim \mathrm{N}_{p \times N}(0, I_p, I_N)$. Therefore $\mathcal{R}(E) \subseteq \mathcal{R}(\Sigma_0)$. Moreover, $Y(I_N - X^+X) = \Sigma_0 Q(I_N - X^+X)$ so

$$\mathcal{R}(Y(I_N - X^+X)Y^\top) = \mathcal{R}(Y(I_N - X^+X)) \subseteq \mathcal{R}(\Sigma_0)$$

In other words, if Σ_0 is singular, then $Y(I_N - X^+X)Y^\top$ is singular, and the problem (12) has no solutions. As such, the naïve maximum likelihood formulation (12) appears incapable of handling models with structural rank deficiencies, motivating a reformulation.

[2] Although these matrices were hand-selected for illustrative purposes, the errors have the expected mean and standard deviation suggested by the system covariance matrix.

5 Rank-constrained multivariate linear regression

The ML estimator (12) is defined for a PDF that is taken with respect to a parameter-dependent measure. Most classic probability and statistics literture require a fixed reference measure with which to define the PDF [17, Theorems 7.49 and 7.54]. While [9] note this issue, they claim that the reference measure does not affect the results. While this claim is likely true (due to the prior knowledge of the rank), it benefits us to know how to formulate the ML estimator in a rigorous manner and provide a method by which the rank can be deduced from data. To this end, we consider the following measure-theoretic definition of a ML estimator.

Definition 10. Let $\theta : \Omega \to \Theta$ and $Y : \Omega \to \mathcal{Y}$ be random variables such that $\Theta \subseteq \mathbb{R}^p$ and $\mathcal{Y} \subseteq \mathbb{R}^n$. Suppose Y conditioned on any $\theta = \theta_1 \in \Theta$ has the conditional PDF $p(y|\theta_1)$ with respect to a common reference measure μ on a space $(\mathcal{Y}, \mathcal{F}_Y)$. Let y be an observation of Y conditioned on $\theta = \theta_0 \in \Theta$. We say $\hat{\theta}$ is a *μ-maximum likelihood* (ML) estimator of θ_0 if it solves

$$\max_{\theta \in \Theta} p(y|\theta) \tag{14}$$

It is fairly straightforward to show the ML estimator, if it exists, is independent up to equivalent reference measures.

Proposition 11. Let $\hat{\theta}_\mu$ and $\hat{\theta}_\nu$ be the μ- and ν-ML estimators of θ_0. If $\mu \equiv \nu$, then $\hat{\theta}_\mu = \hat{\theta}_\nu$ almost surely.

Proof. Let $p_\mu(\cdot|\theta_1)$ and $p_\nu(\cdot|\theta_1)$ be the PDFs of $Y|(\theta = \theta_1)$ with respect to the reference measure μ and ν, respectively. Then we have $p_\mu(\cdot|\theta_1) = p_\nu(\cdot|\theta_1)\frac{d\mu}{d\nu}(\cdot)$ almost surely, so the likelihood functions are equivalent up to a parameter-independent coefficient. □

Since we defined the reference measure in Section 3 by an affine transformation corresponding to the support of the singular normal random variable, the reference measure was parameter-dependent. Instead, we can consider an extension of that measure to the parameter-independent case. Let $m \in \mathbb{R}^n$ and $U_1 \in \mathbb{R}^{n \times r}$ such that $U_1^\top U_1 = I_r$, and define

$$\mu(A) := \lambda^r(\{\, x \in \mathbb{R}^r : m + U_1 x \in A \,\})$$

for each $A \in \mathcal{A}(m, U_1) := \{\, m + U_1 B : B \in \mathcal{B}(\mathbb{R}^n) \,\}$. Let

$$\mathcal{A} := \{\, A \in \mathcal{A}(m, U_1) : m \in \mathbb{R}^n, U_1 \in \mathbb{R}^{n \times r}, U_1^\top U_1 = I_r \,\}$$

It can be shown that $\mathcal{A}$ is a ring and μ is a pre-measure on it. Therefore, Carathéordy's extension theorem implies the existence of an extension that is a measure on $(\mathbb{R}^n, \sigma(\mathcal{A}))$. Therefore, the PDF (9) is also valid with respect to μ. A similar process can be used to extend the reference measure of Proposition 7 so that it is parameter-independent. However, we cannot avoid defining the rank r at the outset, so the parameter-independence of the reference measure is constraining the rank in our estimation problem.

Extending this idea to ML of (2), we have the rank-constrained ML problem

$$\max_{\Theta \in \mathbb{R}^{p\times n}, \Sigma \succeq 0 \in \mathbb{R}^{p\times p}} p(Y|X,\Theta,\Sigma) \quad \text{subject to} \quad p(Y|X,\Theta,\Sigma) > 0 \quad \text{and} \quad \operatorname{rank}(\Sigma) = r \qquad (15)$$

where the PDF is now taken with respect to μ^N as defined above. However, it is not clear how we may deduce the rank r. To see how r is readily apparent from the data, we require a preliminary fact about the rank of a zero-mean normally distributed matrix.

Proposition 12. If $Z \sim \mathrm{N}_{n\times p}(M,U,V)$ where $U \succeq 0 \in \mathbb{R}^{n\times n}$ and $V \succeq 0 \in \mathbb{R}^{p\times p}$, then $\operatorname{rank}(Z - M) = \min\{\operatorname{rank}(U), \operatorname{rank}(V)\}$ almost surely.

See Appendix B for a proof of Proposition 12. Consider the residual vector $R := Y(I_N - X^+X)$ for the standard solution. We have

$$R = Y(I_N - X^+X) = (\Theta_0 X + E)(I_N - X^+X) = E(I_N - X^+X)$$

and by linearity (Lemma 17),

$$R \sim \mathrm{N}_{p\times N}(0, \Sigma_0, I_N - X^+X)$$

By Proposition 12, we have $\operatorname{rank}(R) = \min\{\operatorname{rank}(\Sigma_0), \operatorname{rank}(I_N - X^+X)\}$ almost surely. Computationally it is simpler to check the rank of $RR^\top = Y(I_N - X^+X)Y^\top$ since $\operatorname{rank}(R) = \operatorname{rank}(RR^\top)$. Finally, we have by the rank-nullity theorem that

$$\operatorname{rank}(I_N - X^+X) = N - \operatorname{rank}(X) \geq N - \min\{n, N\}$$

so we can always choose the number of samples sufficient large ($N \geq n + p$) to guarantee that $\operatorname{rank}(RR^\top) = \operatorname{rank}(\Sigma_0)$. As a result, we can guess the rank from the data, and write the rank-constrained ML problem (15) with $r := \operatorname{rank}(Y(I_N - X^+X)Y^\top)$.

Solutions to (15) are characterized by the following proposition, with a proof included in Appendix C. Again, solutions, when they exist, are identical to that of (3). However, we now have a guarantee that solutions exist and are the correct rank, almost surely, and up to numerical precision of our computations.

Proposition 13. The pair $(\hat{\Theta}, \hat{\Sigma})$ solves (12) if and only if (6). Moreover, if $N \geq n + \operatorname{rank}(\Sigma_0)$, then $\operatorname{rank}(\hat{\Sigma}) = \operatorname{rank}(\Sigma_0)$ almost surely.

6 Applications

In this section, we consider examples in system identification. The examples are elementary, intended to demonstrate the practicality of Proposition 13. First, we consider the fully observed linear state-space model,

$$x_{k+1} = Ax_k + Bu_k + w_k, \qquad w_k \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, Q)$$

If we collect a finite trajectory of data $(x_0, \ldots, x_N, u_0, \ldots, u_{N-1})$, the ML estimates are

$$\begin{bmatrix} \hat{A} & \hat{B} \end{bmatrix} = YX^+, \qquad \hat{Q} = Y(I_N - X^+X)Y$$

according to Proposition 13, where $Y := \begin{bmatrix} x_1 & \dots & x_N \end{bmatrix}$ and $X := \begin{bmatrix} x_0 & \dots & x_{N-1} \\ u_0 & \dots & u_{N-1} \end{bmatrix}$.

Consider the system

$$x_{k+1} = \begin{bmatrix} 0.9 & 1 \\ 0 & 0 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k + w_k \qquad w_k \overset{\text{i.i.d.}}{\sim} \mathrm{N}\left(0, \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)$$

Notice that the rank deficiency of $Q := \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ necessitates the singular regression formulation. We collect a short trajectory ($N = 10$) of data and fit the parameters (A, B, Q),

$$\hat{A} = \begin{bmatrix} 0.661 & 0.903 \\ -0.239 & -0.097 \end{bmatrix}, \qquad \hat{B} = \begin{bmatrix} 0.011 \\ 1.011 \end{bmatrix}, \qquad \hat{Q} = \begin{bmatrix} 0.427 & 0.427 \\ 0.427 & 0.427 \end{bmatrix}$$

Notice that while $\hat{Q}$ is not an exact estimate, it has the same rank and range space as Q to within machine precision.

Next, we consider the partially-observed linear state-space model,

$$\begin{aligned} x_{k+1} &= Ax_k + Bu_k + Lv_k \\ y_k &= Cx_k + Du_k + v_k \end{aligned} \qquad \begin{bmatrix} Lv_k \\ v_k \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, S), \qquad S := \begin{bmatrix} LRL^\top & LR \\ RL^\top & R \end{bmatrix} \tag{16}$$

For simplicity, we put the system in innovations form, which can be done without loss of generality [18]. If we had access to the states $(x_0, \dots, x_N)$, inputs$(u_0, \dots, u_{N-1})$, and outputs $(y_0, \dots, y_{N-1})$, then we could straightforwardly estimate the parameters (A, B, C, D, S). Since we do not, we employ a common method of state approximation [19, Section 7.4]. If $A - LC$ is stable (i.e., the system is observable), we choose an integer n_p and there exists $\tilde{L} \in \mathbb{R}^{n \times \tilde{n}}$ such that

$$x_k \approx \tilde{L}\tilde{x}_k$$

where $n_z := m + p$, $\tilde{n} := n_z n_p$, and $\tilde{x}_k := \begin{bmatrix} y_{k-1}^\top & u_{k-1}^\top & \dots & y_{k-n_p}^\top & u_{k-n_p}^\top \end{bmatrix}^\top$, and the desired precision can be reached by choosing n_p sufficiently large. Suppose for simplicity that $A - LC$ is nilpotent of order $n_p + 1$ or less so that the system can be exactly rewritten

$$\begin{aligned} \tilde{x}_{k+1} &= \tilde{A}\tilde{x}_k + \tilde{B}u_k + \tilde{w}_k \\ y_k &= \tilde{C}\tilde{x}_k + Du_k + v_k \end{aligned} \qquad \begin{bmatrix} \tilde{w}_k \\ v_k \end{bmatrix} \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \tilde{S})$$

where $\tilde{n}' := \tilde{n} - n_z$,

$$\tilde{A} := \left[\begin{array}{c|c} \multicolumn{2}{c}{C\tilde{L}} \\ \hline \multicolumn{2}{c}{0_{m\times\tilde{n}}} \\ \hline I_{\tilde{n}'} & 0_{\tilde{n}'\times n_z} \end{array}\right], \quad \tilde{B} := \begin{bmatrix} D \\ I_m \\ 0_{\tilde{n}'\times m} \end{bmatrix}, \quad \tilde{C} := C\tilde{L}, \quad \tilde{S} := \begin{bmatrix} R & & \\ & 0_{\tilde{n}'\times\tilde{n}'} & \\ & & R \end{bmatrix}.$$

While this is a significant overparameterization of the system (16), it turns out to be well-posed and gives estimates that are no worse than standard ARX(n_p, n_p) estimates.

Consider the system

$$\begin{aligned} x_{k+1} &= \begin{bmatrix} a_1 & 1 \\ a_2 & 0 \end{bmatrix} x_k + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u_k + \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} v_k \qquad v_k \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1) \\ y_k &= \begin{bmatrix} 1 & 0 \end{bmatrix} x_k + v_k \end{aligned}$$

where $a_1 = 0.9$ and $a_2 = 0.1$. Clearly, $A - LC = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ is nilpotent (order 2), so we have an exact representation of the form $x_k = \tilde{L}\tilde{x}_k$ where $\tilde{x}_k := \begin{bmatrix} y_{k-1}^\top & u_{k-1}^\top \end{bmatrix}^\top$ and $\tilde{L} := \begin{bmatrix} 0.9 & 0 & 0.1 & 1 \\ 0.1 & 1 & 0 & 0 \end{bmatrix}$. Again, we collect a short trajectory ($N = 10$) of data and fit the parameters $(\tilde{A}, \tilde{B}, \tilde{C}, D, \tilde{S})$:

$$\hat{A} = \begin{bmatrix} \mathbf{0.2018} & \mathbf{-0.1585} & \mathbf{0.5085} & \mathbf{0.7528} \\ 5.015 \times 10^{-16} & -9.953 \times 10^{-16} & 5.085 \times 10^{-17} & 4.623 \times 10^{-16} \\ \mathbf{1.000} & -1.266 \times 10^{-16} & -2.842 \times 10^{-16} & 2.602 \times 10^{-16} \\ 3.642 \times 10^{-16} & \mathbf{1.000} & 1.404 \times 10^{-16} & 4.240 \times 10^{-16} \end{bmatrix},$$

$$\hat{B} = \begin{bmatrix} \mathbf{0.2321} \\ \mathbf{1.000} \\ 2.019 \times 10^{-16} \\ 1.778 \times 10^{-16} \end{bmatrix}, \qquad \hat{C} = \begin{bmatrix} \mathbf{0.202} & \mathbf{-0.159} & \mathbf{0.508} & \mathbf{0.753} \end{bmatrix}, \qquad \hat{D} = \mathbf{0.232},$$

$$\hat{S} = \begin{bmatrix} \mathbf{0.2073} & 8.105 \times 10^{-18} & 1.358 \times 10^{-17} & -2.111 \times 10^{-17} & \mathbf{0.2073} \\ 8.105 \times 10^{-18} & 1.084 \times 10^{-30} & -8.217 \times 10^{-32} & -3.746 \times 10^{-31} & 8.105 \times 10^{-18} \\ 1.358 \times 10^{-17} & -8.217 \times 10^{-32} & 1.679 \times 10^{-31} & -2.465 \times 10^{-32} & 1.358 \times 10^{-17} \\ -2.111 \times 10^{-17} & -3.746 \times 10^{-31} & -2.465 \times 10^{-32} & 5.382 \times 10^{-31} & -2.111 \times 10^{-17} \\ \mathbf{0.2073} & 8.105 \times 10^{-18} & 1.358 \times 10^{-17} & -2.111 \times 10^{-17} & \mathbf{0.2073} \end{bmatrix}$$

Again, $\hat{S}$ is of the correct rank and range space, but notice that a large number of entries in the other estimates $(\hat{A}, \hat{B})$ are near machine precision. This is because many data rows are duplicated across both Y and X, so some rows of Y can be computed exactly from X. The linear system solver takes case of this fact, meaning a structured ML estimation problem is not really necessary to most efficiently construct the state approximation $x_k = \tilde{L}\tilde{x}_k$.

7 Conclusions

Using a measure-theoretic definition of the ML estimator, we have shown that singular multivariate linear regression models can be estimated using only a minor modification of the nonsingular estimates. The importance of this theory was shown in numerical examples and applied to elementary problems in system identification. There are two areas of future research and applications of this work. First, Bayesian estimation, reduced-rank regression, and nonlinear regression problems can each be extended using the reference measure defined in Section 5. Second, more practical system identification problems, such as subspace identification and direct ML estimation of stochastic linear systems of the form (16) can be explored if some of the more exotic regression problems are addressed.

References

[1] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed. New York: John Wiley & Sons, 2003.

[2] G. E. P. Box and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, 1st ed. Reading, Massachusetts: Addison–Wesley, 1973.

[3] T. Minka, "Bayesian linear regression," MIT, Tech. Rep., 2000.

[4] A. J. Izenman, "Reduced-rank Regression for the Multivariate Linear Model: Its Relationship to Certain Classical Multivariate Techniques, and Its Application to the Analysis of Multivariate Data," PhD Thesis, University of California, Berkeley, 1972.

[5] ——, "Reduced-rank regression for the multivariate linear model," *J. Multivariate Analysis*, vol. 5, no. 2, pp. 248–264, 1975.

[6] M. K.-S. Tso, "Reduced-Rank Regression and Canonical Analysis," *J. Roy. Stat. Soc. Ser. B*, vol. 43, no. 2, pp. 183–189, 1981.

[7] P. Stoica and M. Viberg, "Maximum likelihood parameter and rank estimation in reduced-rank multivariate linear regressions," *IEEE Trans. Signal Process.*, vol. 44, no. 12, pp. 3069–3078, 1996.

[8] T. W. Anderson, "Asymptotic distribution of the reduced rank regression estimator under general conditions," *Ann. Stat.*, vol. 27, no. 4, pp. 1141–1154, Aug 1999.

[9] C. G. Khatri, "Some Results for the Singular Normal Multivariate Regression Models," *Sankhyā A*, vol. 30, no. 3, pp. 267–280, 1968.

[10] M. S. Srivastava and D. von Rosen, "Regression models with unknown singular covariance matrix," *Linear Algebra Appl.*, vol. 354, no. 1-3, pp. 255–273, 2002.

[11] P. Billingsley, *Probability and measure.* John Wiley & Sons, 2017.

[12] R. L. Schilling, *Measures, Integrals and Martingales.* Cambridge University Press, 2017.

[13] A. K. Gupta and D. K. Nagar, *Matrix Variate Distributions.* CRC Press, 1999.

[14] H. Cramér, *Mathematical methods of statistics.* Princeton University Press, 1946.

[15] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. New York: John Wiley and Sons, 1973.

[16] M. S. Srivastava and C. G. Khatri, *An Introduction to Multivariate Statistics.* North-Holland, 1979.

[17] M. J. Schervish, *Theory of Statistics*, ser. Springer Series in Statistics. New York, NY: Springer, 1995.

[18] J. V. Candy, T. E. Bullock, and M. E. Warren, “Invariant system description of the stochastic realization,” *Automatica*, vol. 15, no. 4, pp. 493–495, Jul 1979.

[19] L. Ljung, *System Identification: Theory for the User*, 2nd ed. New Jersey: Prentice Hall, 1999.

A Nonsingular multivariate linear regression

The approach to solving (5) is to solve an inner, convex problem in Θ, substitute that solution back into the objective, and then solve the outer problem in Σ. In Lemmas 14 and 15 we solve the inner and outer problems, respectively.

Lemma 14. Suppose $\Sigma \succ 0 \in \mathbb{R}^{p\times p}$ and consider the optimization problem

$$\min_{\Theta\in\mathbb{R}^{p\times n}} \phi(\Theta,\Sigma) := \frac{1}{2}\mathrm{tr}[\Sigma^{-1}(Y-\Theta X)(Y-\Theta X)^\top] \tag{17}$$

Then $\hat{\Theta}$ solves (17) if and only if

$$\hat{\Theta} \in \{\, YX^+ + Q : \mathcal{R}(Q) \subseteq \mathcal{N}(X)\,\} \tag{18}$$

Proof. Since the problem is convex and unconstrained, we can simply take the derivative and set it to zero:

$$\frac{\partial\phi}{\partial\Theta}(\hat{\Theta},\Sigma) = \Sigma^{-1}(Y-\hat{\Theta}X)X^\top = 0 \qquad \Leftrightarrow \qquad YX^\top = \hat{\Theta}XX^\top$$

which holds if and only if $\hat{\Theta} = YX(XX^\top)^+ + Q = YX^+ + Q$ for some $Q \in \mathbb{R}^{n\times p}$ such that $\mathcal{R}(Q) \subseteq \mathcal{N}(X)$, regardless of $\Sigma \succ 0$. □

Lemma 15. Let $R \in \mathbb{R}^{p\times N}$ and consider the optimization problem

$$\min_{\Sigma\succ 0\in\mathbb{R}^{p\times p}} \phi(\Sigma,R) := \frac{N}{2}\ln|\Sigma| + \frac{1}{2}\mathrm{tr}(\Sigma^{-1}RR^\top) \tag{19}$$

Then (19) has solutions if and only if $RR^\top$ is nonsingular. Moreover, if $RR^\top$ is nonsingular, then $\hat{\Sigma} = (1/N)RR^\top$ is the unique solution to (19).

Proof. Consider the singular value decomposition $RR^\top = USU^\top$, where $U \in \mathbb{R}^{p\times p}$ is unitary and $S \in \mathbb{R}^{p\times p}$ is diagonal with nonnegative entries.

Suppose $RR^\top$ is singular and consider the candidate estimate $\tilde{\Sigma} = U\tilde{S}U^\top$, where $\tilde{S} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with positive diagonal entries. Rewriting the objective in terms of $(U,S,\tilde{S})$:

$$\phi(U\tilde{S}U^\top,R) = \frac{N}{2}\ln|\tilde{S}| + \frac{1}{2}\mathrm{tr}(\tilde{S}^{-1}S) = \frac{1}{2}\sum_{i=1}^{p} N\ln\tilde{S}_{ii} + \frac{S_{ii}}{\tilde{S}_{ii}}$$

Since $RR^\top$ is singular, $S_{pp} = 0$ and $\tilde{S}_{pp} \searrow 0$ gives $\phi(U\tilde{S}U^\top, R) \to -\infty$. Therefore (17) has no solutions when $RR^\top$ is singular.

On the other hand, suppose $RR^\top$ is nonsingular. Then we can define its positive definite square root by $V := (RR^\top)^{1/2} = US^{1/2}U$, and the invertible transformation $\Omega = f(\Sigma) := V\Sigma^{-1}V$, where $\Sigma = f^{-1}(\Omega) = V\Omega^{-1}V$. Consider the singular value decomposition $\Omega = WDW^\top$. Rewriting the objective using properties of the determinant and trace:

$$\begin{aligned}
\phi(f^{-1}(\Omega), R) &= \frac{N}{2}\ln|V\Omega^{-1}V| + \frac{1}{2}\mathrm{tr}((V\Omega^{-1}V)^{-1}RR^\top) \\
&= N\ln|V| - \frac{N}{2}\ln|\Omega| + \frac{1}{2}\mathrm{tr}(\Omega) \\
&= N\ln|V| + \frac{1}{2}\sum_{i=1}^{p}(-N\ln D_{ii} + D_{ii})
\end{aligned}$$

The objective is minimized by $\hat{D}_{ii} = N$, independently of W. Then $\hat{\Omega} := W\hat{D}W^\top = W(NI_p)W^\top = NWW^\top = NI_p$ is the unique minimizer of $\phi(f^{-1}(\cdot), R)$. Taking the inverse transform gives that $\hat{\Sigma} := f^{-1}(\hat{\Omega}) = V(NI_p)^{-1}V = (1/N)V^2 = (1/N)RR^\top$ is the unique minimizer of $\phi(\cdot, R)$. □

Finally, we combine Lemmas 14 and 15 to solve (3) (equivalently, (5)).

Proof of Proposition 1. It suffices to work with the negative log-transformed problem (5). By Lemma 14, $\hat{\Theta}$ is a solution to the inner Θ optimization problem if and only if $\hat{\Theta} = YX^+ + Q$ for some $Q \in \mathbb{R}^{p\times n}$ such that $\mathcal{R}(Q) \subseteq \mathcal{N}(X)$. Substituting this back into ϕ, we get $R := Y - \hat{\Theta}X = Y - YX^+X = Y(I_N - X^+X)$ and the outer problem

$$\min_{\Sigma \succ 0 \in \mathbb{R}^{p\times p}} \phi(\hat{\Theta}, \Sigma) = \frac{N}{2}\ln|\Sigma| + \frac{1}{2}\mathrm{tr}(\Sigma^{-1}RR^\top) \tag{20}$$

By Lemma 15, the problem (20) has solutions if and only if $RR^\top = Y(I_N - X^+X)Y^\top$ is nonsingular, and moreover, if $RR^\top$ is nonsingular, then $\hat{\Sigma}$ uniquely solves (20). Therefore the pair $(\hat{\Theta}, \hat{\Sigma})$ solves (3) if and only if (6) hold. □

B Matrix normal properties

In this section we prove Proposition 12. To prove Proposition 12 we need some preliminary results.

Lemma 16. If $X \sim \mathrm{N}_{n\times p}(M, U, V)$ where $U \succeq 0 \in \mathbb{R}^{n\times n}$ and $V \succeq 0 \in \mathbb{R}^{p\times p}$, then $X^\top \sim \mathrm{N}_{p\times n}(M^\top, V, U)$.

Proof. The result follows from (4) and invariance of the trace under cyclic permutations. □

Lemma 17. If $X \sim \mathrm{N}_{n\times p}(M, U, V)$, then $AXB + C \sim \mathrm{N}_{n\times p}(AMB + C, AUA^\top, BVB^\top)$.

Proof. This follows from the fact that a matrix normal is fully defined by its mean matrix and covariance matrices. □

Lemma 18. If $X \sim \mathrm{N}_{n\times p}(M, FF^\top, GG^\top)$ where $F \in \mathbb{R}^{n\times r}$ and $G \in \mathbb{R}^{p\times s}$ are full column rank, then $X - M \sim FQG^\top$ where $Q \sim \mathrm{N}_{r\times s}(0, I_r, I_s)$.

Proof. This follows by linearity (Lemma 17). □

Lemma 19. If $X \sim \mathrm{N}_{n\times p}(0, I_n, I_p)$, then $\mathrm{rank}(X) = \min\{n, p\}$ with probability 1.

Proof. By Lemma 16, we can assume $n \geq p$ without loss of generality. For each $n \geq 1$ and $p = 1$, X is a vector which has rank $p = 1$ if and only if $X \neq 0$. Therefore

$$\mathbb{P}[\mathrm{rank}(X) = p] = \mathbb{P}[X \neq 0] = 1 - \mathbb{P}[X = 0] = 1 - \int_{\{\,0\,\}} \frac{\exp\left(-\frac{1}{2}x^\top x\right)}{(2\pi)^{n/2}} d\lambda^n(x) = 1$$

We complete the proof by induction. Assume the hypothesis holds for some $n > p \geq 1$. Let $X \sim \mathrm{N}_{n\times(p+1)}(0, I_n, I_{p+1})$ and consider the partition $X = \begin{bmatrix} X_1 & x_2 \end{bmatrix}$ where $X_1 \in \mathbb{R}^{n\times p}$ and $x_2 \in \mathbb{R}^n$. Then $X_1 \sim \mathrm{N}_{n\times p}(0, I_n, I_p)$ and $x_2 \sim \mathrm{N}_n(0, I_n)$ because all the entries of X are i.i.d. normals. For all $\tilde{X}_1 \in \mathbb{R}^{n\times p}$, $\mathcal{R}(\tilde{X}_1)$ is a subspace of $\mathbb{R}^n$ with dimension no greater than $p < n$ and $\lambda^n(\mathcal{R}(\tilde{X}_1)) = 0$. Therefore

$$\mathbb{P}[x_2 \in \mathcal{R}(X_1) | X_1 = \tilde{X}_1] = \int_{\mathcal{R}(\tilde{X}_1)} \frac{\exp\left(-\frac{1}{2}x^\top x\right)}{(2\pi)^{n/2}} d\lambda^n(x) \leq \frac{1}{(2\pi)^{n/2}} \lambda^n(\mathcal{R}(\tilde{X}_1)) = 0$$

regardless of the value of $\tilde{X}_1 \in \mathbb{R}^{n\times p}$, so $\mathbb{P}[x_2 \in \mathcal{R}(X_1)] = 0$. Moreover, $\mathbb{P}[\mathrm{rank}(X_1) < p] = 0$ by the assumption, so

$$\begin{aligned} \mathbb{P}[\mathrm{rank}(X) < p + 1] &= \mathbb{P}[(\mathrm{rank}(X_1) < p) \vee (x_2 \in \mathcal{R}(X_1))] \\ &\leq \mathbb{P}[\mathrm{rank}(X_1) < p] + \mathbb{P}[x_2 \in \mathcal{R}(X_1)] = 0 \end{aligned}$$

Finally, $\mathbb{P}[\mathrm{rank}(X) = p + 1] = 1 - \mathbb{P}[\mathrm{rank}(X) < p + 1] = 1$. □

Proof of Proposition 12.. Consider the thin SVDs $U = W_1\Sigma_1 W_1^\top$ and $V = Q_1 D_1 Q_1^\top$ and full-rank factors $F := W_1\Sigma_1^{1/2}$ and $G := Q_1 D_1^{1/2}$. By Lemma 18, we have $X - M = FQG^\top$ where $Q \sim \mathrm{N}_{r\times s}(0, I_r, I_s)$, $r := \mathrm{rank}(U)$, and $s := \mathrm{rank}(V)$. By Lemma 19, $\mathrm{rank}(X - M) = \mathrm{rank}(FQG^\top) = \mathrm{rank}(Q) = \min\{r, s\}$ (almost surely). □

C Singular multivariate linear regression

As in Appendix A, the solution approach will be to first solve an inner convex optimization problem in Θ as a function of Σ, and then to solve an outer nonconvex problem in Σ. Lemmas 20 and 21 solve these two problems. While the solutions are identical to that of Proposition 1, they require slightly different methods that may be of interest to the reader.

Lemma 20. Suppose $\Sigma \succeq 0 \in \mathbb{R}^{p\times p}$ and consider the optimization problem

$$\min_{\Theta\in\mathbb{R}^{p\times n}} \phi(\Theta, \Sigma) := \frac{1}{2}\mathrm{tr}[\Sigma^+(Y - \Theta X)(Y - \Theta X)^\top] \quad \text{subject to} \quad Y \in \mathcal{Z}(X, \Theta, \Sigma) \tag{21}$$

Then $\hat{\Theta}$ solves (21) if and only if

$$\hat{\Theta} \in \{\, YX^+ + Q : \mathcal{R}(Q) \subseteq \mathcal{N}(X)\,\} \tag{22}$$

Proof. We can relax the set constraint to a linear constraint by adding $Z \in \mathbb{R}^{p\times N}$ as an optimization variable,

$$\min_{\Theta\in\mathbb{R}^{p\times n}, Z\in\mathbb{R}^{p\times N}} \phi(\Theta,\Sigma) \quad \text{subject to} \quad Y = \Theta X + \Sigma Z \tag{23}$$

Substituting the linear constraint into the objective gives

$$\phi(\Theta,\Sigma) = \frac{1}{2}\text{tr}[\Sigma^+(Y-\Theta X)(Y-\Theta X)^\top] = \frac{1}{2}\text{tr}[\Sigma^+\Sigma Z Z^\top \Sigma] = \frac{1}{2}\text{tr}[\Sigma Z Z^\top]$$

for all (Θ,Σ) such that $Y = \Theta X + \Sigma Z$. Therefore (23) is equivalent to

$$\min_{\Theta\in\mathbb{R}^{p\times n}, Z\in\mathbb{R}^{p\times N}} \frac{1}{2}\text{tr}[\Sigma Z Z^\top] \quad \text{subject to} \quad Y = \Theta X + \Sigma Z \tag{24}$$

We can solve (24) with the method of Lagrange multipliers. Let

$$\mathcal{L}_\Sigma(\Theta, Z, \Lambda) := \frac{1}{2}\text{tr}[\Sigma Z Z^\top] + \text{tr}[\Lambda^\top(Y - \Theta X - \Sigma Z)]$$

where $\Lambda \in \mathbb{R}^{p\times N}$. Then $(\hat{\Theta}, \hat{Z})$ solve (24) (equivalently, (21)) if and only if

$$\frac{\partial \mathcal{L}_\Sigma}{\partial \Theta}(\hat{\Theta}, \hat{Z}, \hat{\Lambda}) = \hat{\Lambda} X^\top = 0 \tag{25a}$$

$$\frac{\partial \mathcal{L}_\Sigma}{\partial Z}(\hat{\Theta}, \hat{Z}, \hat{\Lambda}) = \Sigma\hat{Z} - \Sigma\hat{\Lambda} = 0 \tag{25b}$$

$$\frac{\partial \mathcal{L}_\Sigma}{\partial \Lambda}(\hat{\Theta}, \hat{Z}, \hat{\Lambda}) = Y - \hat{\Theta} X - \Sigma\hat{Z} = 0 \tag{25c}$$

for some $\hat{\Lambda} \in \mathbb{R}^{p\times N}$. Equation (25b) holds if and only if

$$\hat{Z} = \Sigma^+\Sigma\hat{\Lambda} + R$$

for some $R \in \mathbb{R}^{p\times N}$ such that $\mathcal{R}(R) \subseteq \mathcal{N}(\Sigma)$. Substituting this into (25c) gives the reduced system,

$$\hat{\Lambda} X^\top = 0, \qquad\qquad Y - \hat{\Theta} X - \Sigma\hat{\Lambda} = 0 \tag{26}$$

But (25a) implies $\mathcal{R}(\hat{\Lambda}^\top) \subseteq \mathcal{N}(X) = \mathcal{N}((X^+)^\top)$, so (26) implies $\hat{\Theta}$ must satisfy

$$\hat{\Theta} = YX^+ - \Sigma\hat{\Lambda}X^+ + Q = YX^+ + Q$$

for some $Q \in \mathbb{R}^{p\times n}$ such that $\mathcal{R}(Q) \subseteq \mathcal{N}(X^\top)$. Substituting this back into (26) gives

$$\hat{\Lambda} X^\top = 0, \qquad\qquad Y(I_N - X^+X) - \Sigma\hat{\Lambda} = 0$$

Moreover, we have that such a $\hat{\Lambda}$ exists if and only if $\mathcal{R}(Y(I_N - X^+X)) \subseteq \mathcal{R}(\Sigma)$, and the solution is given by

$$\hat{\Lambda} = \Sigma^+ Y(I_N - X^+X) + (I_p - \Sigma\Sigma^+)T(I_N - X^+X)$$

for some $T \in \mathbb{R}^{p\times N}$, and this implies $\hat{Z} = \Sigma^{+}Y(I_N - X^{+}X) + R$ for some $R \in \mathbb{R}^{p\times N}$ such that $\mathcal{R}(R) \subseteq \mathcal{N}(\Sigma)$. In summary, $\hat{\Lambda} \in \mathbb{R}^{p\times N}$ exists such that $(\hat{\Theta}, \hat{Z})$ satisfy (25) if and only if there exist $Q \in \mathbb{R}^{p\times n}$ and $R \in \mathbb{R}^{p\times N}$ such that

$$\begin{aligned} \hat{\Theta} &= YX^{+} + Q, && \mathcal{R}(Q) \subseteq \mathcal{N}(X^\top), \\ \hat{Z} &= \Sigma^{+}Y(I_N - X^{+}X) + R, && \mathcal{R}(R) \subseteq \mathcal{N}(\Sigma) \end{aligned}$$

In other words, we can solve (21) if and only if $\mathcal{R}(Y(I_N - X^{+}X)) \subseteq \mathcal{R}(\Sigma)$, and if this condition holds, then $\hat{\Theta}$ is a solution if and only if (22). □

Lemma 21. Let $R \in \mathbb{R}^{p\times N}$ and consider the optimization problem

$$\min_{\Sigma\succeq 0\in\mathbb{R}^{p\times p}} \phi(\Sigma, R) \quad \text{subject to} \quad \mathcal{R}(R) \subseteq \mathcal{R}(\Sigma) \tag{27a}$$

where

$$\phi(\Sigma, R) := \frac{N}{2}\ln(2\pi)\mathrm{rank}(\Sigma) + \frac{N}{2}\ln|\Sigma|_{+} + \frac{1}{2}\mathrm{tr}(\Sigma^{+}RR^\top) \tag{27b}$$

Then (27) has solutions if and only if $RR^\top$ is nonsingular. Moreover, if $RR^\top$ is nonsingular, then $\hat{\Sigma} = \frac{1}{N}RR^\top$ is the unique solution to (27).

Proof. To simplify the notation, let $r := \mathrm{rank}(\Sigma)$ and $\hat{r} := \mathrm{rank}(RR^\top)$ throughout. As in the proof of Lemma 15, we consider the singular value decomposition $RR^\top = USU^\top$, where $U \in \mathbb{R}^{p\times p}$ is unitary and $S \in \mathbb{R}^{p\times p}$ is diagonal with nonnegative entries.

Suppose $RR^\top$ is singular and consider the candidate estimate $\tilde{\Sigma} = U\tilde{S}U^\top$, where $\tilde{S} \in \mathbb{R}^{p\times p}$ is a diagonal matrix with $\tilde{S}_{ii} > 0$ for $i = 1, \dots, r$ and $\tilde{S}_{ii} \geq 0$ for $i = r+1, \dots, p$. Assume $\tilde{S}_{ii}$ are chosen so that $r \geq \hat{r}$ and the candidate $\tilde{\Sigma}$ is feasible. Rewriting the objective in terms of $(U, S, \tilde{S})$:

$$\begin{aligned} \phi(U\tilde{S}U^\top, R) &= \frac{N}{2}\ln(2\pi)r + \frac{N}{2}\ln|U\tilde{S}U^\top|_{+} + \frac{1}{2}\mathrm{tr}(U\tilde{S}^{+}U^\top USU^\top) \\ &= \frac{N}{2}\ln(2\pi)r + \frac{N}{2}\ln|\tilde{S}|_{+} + \frac{1}{2}\mathrm{tr}(\tilde{S}^{+}S) \\ &= \frac{N}{2}\ln(2\pi)r + \frac{N}{2}\sum_{i=1}^{r}\ln\tilde{S}_{ii} + \frac{1}{2}\sum_{i=1}^{\hat{r}}\frac{S_{ii}}{\tilde{S}_{ii}} \end{aligned}$$

Finally, we can choose $\tilde{S}$ such that $r = \hat{r}+1$ and take $\tilde{S}_{rr} \to 0$ to give $\phi(U\tilde{S}U^\top, R) \to -\infty$. Therefore (27) has no solutions when $RR^\top$ is singular.

On the other hand, suppose $RR^\top$ is nonsingular. Then $\mathcal{R}(\Sigma) \supseteq \mathcal{R}(R) = \mathbb{R}^p$, so Σ must also be nonsingular, and we can follow the proof of Lemma 15 to show that $\hat{\Sigma} = \frac{1}{N}RR^\top$ uniquely solves (27). □

We combine Lemmas 20 and 21 to solve (12).

Proof of Proposition 8. It suffices to work with the negative log-transformed problem (13). By Lemma 20, $\hat{\Theta}$ is a solution to the inner Θ optimization problem if and only if $\mathcal{R}(Y(I_N - X^+X)) \subseteq \mathcal{R}(\Sigma)$ and $\hat{\Theta} = YX^+ + Q$ for some $Q \in \mathbb{R}^{p\times n}$ such that $\mathcal{R}(Q) \subseteq \mathcal{N}(X)$. Substituting this back into ϕ, we get $R := Y - \hat{\Theta}X = Y - YX^+X = Y(I_N - X^+X)$ and the outer problem

$$\min_{\Sigma\succ 0\in\mathbb{R}^{p\times p}} \phi(\hat{\Theta}, \Sigma) = \frac{N}{2}\text{rank}(\Sigma) + \frac{N}{2}\ln|\Sigma|_+ + \frac{1}{2}\text{tr}(\Sigma^+ RR^\top) \tag{28}$$

By Lemma 21, the problem (28) has solutions if and only if $RR^\top = Y(I_N - X^+X)Y^\top$ is nonsingular, and moreover, if $RR^\top$ is nonsingular, then $\hat{\Sigma}$ uniquely solves (28). Therefore (12) has solutions if and only if $\mathcal{R}(Y(I_N - X^+X)) \subseteq \mathcal{R}(\Sigma)$ and $Y(I_N - X^+X)Y^\top$ is nonsingular, and the pair $(\hat{\Theta}, \hat{\Sigma})$ are solutions if and only if (6) hold. □

Finally, we modify the proofs of Proposition 8 and Lemmas 20 and 21 to accommodate the rank constraint in (15).

Proof of Proposition 13. It suffices to work with the negative log-transformed problem

$$\min_{\Theta\in\mathbb{R}^{p\times n}, \Sigma\succeq 0\in\mathbb{R}^{p\times p}} \phi(\Theta, \Sigma) \quad \text{subject to} \quad Y \in \mathcal{Z}(X, \Theta, \Sigma) \quad \text{and} \quad \text{rank}(\Sigma) = r \tag{29}$$

where $R := Y - \hat{\Theta}X = Y(I_N - X^+X)$ and we have rewritten the constraints $Y \in \mathcal{Z}(X, \hat{\Theta}, \Sigma)$ and $\text{rank}(\Sigma) = r := \text{rank}(RR^\top) = \text{rank}(R)$ as

$$\begin{aligned}
(Y \in \mathcal{Z}(X, \hat{\Theta}, \Sigma) := \{\, \hat{\Theta}X + \Sigma Z : Z \in \mathbb{R}^{p\times N} \,\} \wedge \text{rank}(\Sigma) = r) &\quad \Leftrightarrow \\
(R := Y - \hat{\Theta}X \in \{\, \Sigma Z : Z \in \mathbb{R}^{p\times N} \,\} \wedge \text{rank}(\Sigma) = r) &\quad \Leftrightarrow \\
(\mathcal{R}(R) \subseteq \mathcal{R}(\Sigma) \wedge \text{rank}(\Sigma) = r) &\quad \Leftrightarrow \\
\mathcal{R}(R) = \mathcal{R}(\Sigma) &
\end{aligned}$$

Following the proof of Lemma 20, we again get that $\hat{\Theta} \in \{\, YX^+ + Q : \mathcal{R}(Q^\top) \subseteq \mathcal{N}(X^\top) \,\}$ minimizes $\phi(\cdot, \Sigma)$ for any feasible Σ. Substituting this solution into the likelihood and dropping constants gives the following outer problem,

$$\min_{\Sigma\succeq 0\in\mathbb{R}^{p\times p}} \phi(\Sigma, R) := \frac{N}{2}\ln|\Sigma|_+ + \frac{1}{2}\text{tr}(\Sigma^+ RR^\top) \quad \text{subject to} \quad \mathcal{R}(R) = \mathcal{R}(\Sigma)$$

Consider the thin singular value decomposition $RR^\top = U_1 S_1 U_1^\top$. Each $\Sigma \succeq 0$ satisfies $\mathcal{R}(RR^\top) = \mathcal{R}(R) = \mathcal{R}(\Sigma)$ if and only if there exists a nonsingular matrix $M \succ 0 \in \mathbb{R}^{p\times p}$ such that $\Sigma = U_1 M U_1^\top$. Rewriting the objective in terms of M:

$$\begin{aligned}
\phi(U_1 M U_1^\top, R) &= \frac{N}{2}\ln|U_1 M U_1^\top|_+ + \frac{1}{2}\text{tr}((U_1 M U_1^\top)^+ U_1 S_1 U_1^\top) \\
&= \frac{N}{2}\ln|M| + \frac{1}{2}\text{tr}(M^{-1} S_1)
\end{aligned}$$

which, by Lemma 15, is minimized by $\hat{M} = (1/N)S_1$. Therefore $\hat{\Sigma} := U_1 \hat{M} U_1^\top = (1/N)U_1 S_1 U_1^\top = (1/N)RR^\top$ minimizes $\phi(\cdot, R)$. In summary, $(\hat{\Theta}, \hat{\Sigma})$ solves (15) if and only if (6) holds. □