TWCCC ⋆ Texas – Wisconsin – California Control Consortium

Technical report number 2023-01

On the unified theory of linear Gaussian estimation: solution methods, applications, and extensions*

Steven J. Kuntz† James B. Rawlings†

December 12, 2023

*The title is in reference to Rao's 1971 paper "Unified Theory of Linear Estimation".

†University of California Santa Barbara (skuntz@ucsb.edu, jbraw@ucsb.edu).

Abstract

Linear Gaussian estimation, i.e., estimation of β (or a linear function of β) in the model $y = X\beta + e$ where $e \sim \mathrm{N}(0, V)$, is a classic and ubiquitous problem in statistics. Linear Gaussian estimation under the most restrictive assumptions (X full column rank, $V = \sigma^2 I$) dates back to the late 18th century. Estimates without assumptions on the rank of X or V were stated in closed-form in the early 1970s. Recently, linear estimation has taken many new and non-Gaussian formulations with the popularity of Bayesian regression priors (e.g., Tikhonov regularization, ridge and LASSO regression, sparse modeling). Given its distinguished history and prominent role in the fields of statistics, optimization, and optimal estimation and control, results on linear estimation (and least squares) are extensive and widely scattered in the literature, often with strikingly *different* but nevertheless equivalent closed-form solutions appearing in different fields.

This review is intended to serve as a self-contained and compact resource for these many definitions and closed-form solutions of estimators of the linear model. We survey a wide variety of estimator definitions, including ordinary/generalized least squares estimators, maximum likelihood estimators (MLEs), maximum a posteriori (MAP) estimators, and best linear unbiased estimators (BLUEs), and derive closed-form solutions to these estimation problems. While solutions to the BLUE problem are available in the literature under our assumptions, we know of no other literature that has closed-form solutions of the MLE and MAP problems under these assumptions. Despite the breadth of estimator definitions available, we show that all of these can be formulated as an equivalent equality constrained generalized least squares (ECGLS) problem, i.e., minimization of a quadratic objective subject to linear equality constraints. Moreover, we show that the estimator of a perturbed linear model with nonsingular variance ($y = X\beta + e$ where $e \sim \mathrm{N}(0, V + \rho I)$) is a stable approximation of the estimator with singular variance. In this review, we also discuss many of the applications (semidefinite Kalman filtering and optimal control, saddle point systems, linear inverse problems) and extensions (nonlinear regression and inverse problems, Bayesian regression, sparse modeling) of the classical and generalized linear estimation problem, and show how they arise from specific solution methods in linear estimation.

Contents

1 Introduction **5**
1.1 History . . . 5
1.2 Summary of results . . . 8
1.3 Applications and modern extensions . . . 12
1.4 Notation . . . 13
1.5 Summary . . . 15

2 Background, definitions, and preliminary results **15**
2.1 Background and definitions . . . 15
2.2 Preliminary results . . . 20

3 Estimators and their solutions **23**
3.1 Generalized least squares . . . 24
3.2 Tikhonov generalized least squares . . . 24
3.3 Equality constrained generalized least squares . . . 25
3.4 Maximum likelihood estimator . . . 26
3.5 Maximum a posteriori estimator . . . 28
3.6 Best affine unbiased estimator . . . 30

4 Computing the estimators **31**
4.1 Convex optimization formulations . . . 32
4.2 Linear system formations . . . 32
4.3 Gradient methods and early stopping . . . 33

5 Applications in control and estimation **35**
5.1 Linear quadratic regulator . . . 36
5.2 Kalman filter . . . 39
5.3 Sparse control and estimation reformulations . . . 40

6 Modern extensions **43**
6.1 A generalized perturbation method for degenerate distributions . . . 43
6.2 Nonlinear regression . . . 44
6.3 Bayesian regression . . . 44
6.4 Sparse estimators . . . 45

A Least squares proofs **48**
A.1 Generalized least squares proofs . . . 48
A.2 Tikhonov generalized least squares proofs . . . 49
A.3 Equality constrained generalized least squares proofs . . . 49

B Maximum likelihood proofs **51**
B.1 ECGLS equivalence . . . 52
B.2 Saddle point equivalence . . . 53
B.3 Barrier function method . . . 55

C Maximum a posteriori estimator proofs **57**

D Best affine unbiased estimator proofs **60**

E Background **63**
E.1 Linear algebra . . . 63
E.2 Projectors . . . 64
E.3 Linear equations . . . 64
E.4 Singular value decomposition . . . 65
E.5 The matrix 2-norm . . . 65
E.6 Matrix limits . . . 66
E.7 Probability . . . 66
E.8 Optimization . . . 67

F Block matrix pseudoinversion proof **68**

G Pseudoinverse of sum of positive semidefinite matrices **70**

H Global bounds on the perturbed problem **74**

I Proof of the limit of the perturbed problem solution **78**

J Miscellaneous results **83**

1 Introduction

Consider the following linear regression model

$$y = X\beta + e, \qquad e \sim \mathrm{N}(0, V) \tag{LGM}$$

in which $y \in \mathbb{R}^n$ are the observations, $\beta \in \mathbb{R}^p$ are the model parameters, $X \in \mathbb{R}^{n\times p}$ is a (known) predictor matrix, $e \in \mathbb{R}^n$ are the model errors, and $V \in \mathbb{R}^{n\times n}$ is a (known) positive semidefinite error covariance matrix. We seek estimators of the parameters β of the model (LGM), as a function of the observations y and known matrices X, V. While these estimators are defined later in this section, let us first conceptually describe them as they are useful in the following historical discussion.

- The ordinary least squares (OLS) estimator $\hat{\beta}$ minimizes the squared 2-norm of the errors $\|y - X\beta\|^2$.
- The generalized least squares (GLS) estimator $\hat{\beta}$ minimizes the squared H-(semi)norm of the errors $\|y - X\beta\|_H^2$.
- The Tikhonov generalized least squares (TGLS) estimator $\hat{\beta}$ minimizes the regularized objective $\|y - X\beta\|_H^2 + \|\beta - \beta_0\|_\Gamma^2$, where β_0 is an initial guess and Γ is a positive semidefinite regularization weighting matrix.
- The equality constrained generalized least squares (ECGLS) estimator $\hat{\beta}$ minimizes the squared H-(semi)norm of the errors $\|y - X\beta\|_H^2$ subject to a linear equality constraint $w = Z\beta$.
- The maximum likelihood estimator (MLE) $\hat{\beta}$ maximizes the probability density of the observations $f(y; \beta)$.
- The maximum a posteriori (MAP) estimator $\hat{\beta}$ maximizes the conditional probability density of the parameters given the observations $f(\beta|y)$, where β and e are independent with prior distribution $\beta \sim \mathrm{N}(\beta_0, \Sigma)$.
- The minimum variance unbiased estimator (MVUE) $\hat{\beta}$ has the minimum variance among all unbiased estimators $\theta = f(y)$.
- The best linear unbiased estimator (BLUE) $\hat{\beta}$ has minimum variance among all *linear* unbiased estimators $\theta = Ay$.
- The best affine unbiased estimator (BAUE) $\hat{\beta}$ has minimum variance among all *affine* unbiased estimators $\theta = Ay + c$.

1.1 History

Estimates of the parameters β of the model (LGM) were first defined and derived by Gauss [38].[1] Gauss' assumptions are stated, in modern matrix notation, as $\mathrm{rank}(X) = p$

[1] Although Laplace [61] and Legendre [64] have each been debated as possibly responsible for inventing what is commonly called the *method of least squares*, Gauss [38] is generally given credit. Some also credit

and $V = \sigma^2 I$ for some $\sigma > 0$. Under these assumptions, the *unique* BLUE of the parameters β is

$$\hat{\beta}_{\text{Gauss}} := (X'X)^{-1}X'y \tag{1}$$

Aitken [1] considered more general assumptions: $\text{rank}(X) = p$ and V is positive definite. Under these assumptions, the *unique* BLUE of the parameters β is

$$\hat{\beta}_{\text{Aitken}} := (X'V^{-1}X)^{-1}X'V^{-1}y \tag{2}$$

Gauss' and Aitken's estimators (1) and (2) are often referred to as the OLS and GLS estimators. However, we wish to consider X with dependent columns and semidefinite weighting matrices, so we reserve the OLS and GLS labels for a more general class of estimators. Nonetheless, if $\text{rank}(X) = p$, then (1) and (2) are equivalent to the OLS estimator and the GLS estimator with weighting matrix $H = V^{-1}$, respectively. Moreover, our OLS and GLS objectives are equivalent (up to constant scaling and shifting) to the negative log-likelihood function, making both MLEs under their respective assumptions. According to Kagan and Salaevskii [55], both (1) and (2) are not only BLUE, but also MVUE, as all the unbiased estimators of (LGM) are also linear estimators. Alternatively, one can use the Lehmann-Scheffé theorem [65, 66] to show that (1) and (2) are MVUE. It is also easily shown, by differentiating the log-likelihood function, that (1) and (2) achieve the Cramér-Rao lower bound [16, 24, 87] and are therefore MVUE.

Many statistical treatments of the model (LGM) stop with Aitken's estimator (2). However, X with dependent columns and singular V may occur in practice. In systems with more predictors than measurements ($n < p$), X always has dependent columns. Singularity in V occurs commonly in economic data with budget constraints [107, Chapter 6.7] or in physical data with conservation laws.

When X has dependent columns, it is no longer possible to find an unbiased estimator of β. However, it is still possible to find an unbiased estimator of the *parameteric function* $W\beta$, whenever $\mathcal{R}(W') \subseteq \mathcal{R}(X')$. Goldman and Zelen [40] first considered the estimation of $W\beta$ given the model (LGM) with any $X \in \mathbb{R}^{n\times p}$ and positive semidefinite $V \in \mathbb{R}^{n\times n}$. They found the following closed-form expression for the BLUE of $W\beta$ when $\mathcal{R}(X) \subseteq \mathcal{R}(V)$:

$$\widehat{W\beta}_{\text{GZ}} := W(X'V^{+}X)^{+}X'V^{+}y \tag{3}$$

where $(\cdot)^{+}$ is the (Moore-Penrose) pseudoinverse [73, 81]. Zyskind and Martin [118] derived normal equations for the purpose of computing the BLUE, but they did not state a closed-form expression. Rao [88] derived the following closed-form expression for the *unique* BLUE under no assumptions on X and V:

$$\widehat{W\beta}_{\text{Rao}} := W(X'V_0^{+}X)^{+}X'V_0^{+}y \tag{4}$$

where $V_0 := V + XEX'$ and $E \in \mathbb{R}^{p\times p}$ is any positive semidefinite matrix that satisfies $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$. The matrix E arises as a free parameter in the pseudoinversion of the

Markov [68] for pointing out that Gauss' estimator is the BLUE even for non-Gaussian errors, which is why (LGM) is sometimes called the Gauss-Markov model, but Gauss' proof does not require a Gaussian error assumption [82]. For more information on the early history of least squares, see [82, 104].

block matrix,

$$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}$$

which is a key step in Rao's derivation.[2,3] An alternative expression for the BLUE comes from Albert [3],

$$\widehat{W\beta}_{\text{Albert}} := WX^+(I - VS(SVS)^+S)y \tag{5}$$

where $S := I - XX^+$. Albert's expression (5) shows how the BLUE deviates from the OLS estimator, $W\hat{\beta}_{\text{OLS}} := WX^+y$. Note that the BLUE is unique, so Rao's and Albert's estimators are equal (i.e., $\widehat{W\beta}_{\text{Rao}} = \widehat{W\beta}_{\text{Albert}}$).

As with (1) and (2), the estimators (3)–(5) are closely related to other estimators. Zyskind and Martin [118] showed that any MLE $\hat{\beta}_{\text{MLE}}$ is equivalent to the BLUE up to left-multiplication by W (i.e., $\widehat{W\beta}_{\text{Rao}} = W\hat{\beta}_{\text{MLE}}$). Seely [96] and Drygas [30] claim that the Lehmann-Scheffé theorem applies to the estimators (3)–(5), and therefore they are not only BLUEs, but also MVUEs. However, this does not rule out nonlinear MVUEs, so (3)–(5) are not necessarily unique MVUEs.[4] We refer the reader to Magnus and Neudecker [67, Chapters 11, 13] for a modern treatment of Rao's results that includes equivalences between the BLUE and least squares-type estimators,[5] and to [44] for a modern review of the results leading to and following from (3)–(5).

It is also worth pointing out the estimators (1)–(5) are still the unique BLUE under non-Gaussian error distributions (with zero mean and covariance V). That is, if the errors are not Gaussian, and the linear model is more generally stated

$$y = X\beta + e, \qquad \mathbb{E}[e] = 0, \qquad \text{var}[e] = V \tag{LM}$$

then the estimators (1)–(5) are still the unique BLUE under their relevant assumptions about X and V. However, they are not necessarily MLE or MVUE. A general treatment of non-Gaussian models is outside of the scope of this paper, so we refer to the model (LGM) throughout. However, our derivation of Rao's estimator (4) does not rely on the probability density of the errors, so the results are valid for the model (LM). We refer the reader to [39, 58, 86] for examples of nonlinear unbiased estimators with lower variance than (1) and (2), and to [47, 84, 86] for recent results on finding MVUEs in the context of (LM).

[2] Rao used a g-inverse rather than the pseudoinverse, but since the pseudoinverse is a g-inverse, all results hold when restricted to the pseudoinverse. See [72, 89, 90, 91, 92] for a further discussion of Rao's solution.

[3] While Pringle and Rayner [85] first derived an expression for the pseudoinverse of this block matrix, they did not include the free parameter E.

[4] Other methods of showing the BLUE is an MVUE do not work in the general case. Kagan and Salaevskii's result [55] requires that X has independent columns, so it cannot be applied to (3)–(5). The Cramér-Rao theorem requires regularity conditions on the probability density function [16, Proposition 3.4.4] that are not satisfied by degenerate normal distributions, but it may be possible to use a constrained version of the Cramér-Rao theorem instead [42, 70].

[5] They show $\widehat{W\beta}_{\text{Rao}} = W\hat{\beta}_{\text{GLS}}$ where $\hat{\beta}_{\text{GLS}}$ is the GLS estimator with weighting matrix $H = V_0^+$ [67, Theorem 13.15]. They also mention an equivalence between $\hat{\beta}_{\text{GLS}}$ and the ECGLS estimator with weighting matrix $H = V^+$ and linear constraint $(I - VV^+)(y - X\beta) = 0$, but this is left as an exercise and not proven [67, p. 318].

1.2 Summary of results

For the purposes of summarizing the main results on estimators of (LGM), we collect acronyms of the relevant models, estimators, and optimization problems, and define their solution sets in table 1.

Table 1: Summary of acronyms.

(a) Model acronyms and references.

Acronym	Model	Reference
LGM	Linear Gaussian model	(LGM)
p-LGM	Perturbed linear Gaussian model	(p-LGM)
B-LGM	Bayesian linear Gaussian model	(B-LGM)

(b) Estimator acronyms and solution sets.

Acronym	Estimator	Solution set
OLS	Ordinary least squares estimator	–
GLS	Generalized least squares estimator	$\hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, H)$
TGLS	Tikhonov generalized least squares estimator	$\hat{\mathbb{B}}_{\mathrm{TGLS}}(y, X, H, \beta_0, \Gamma)$
ECGLS	Equality constrained generalized least squares estimator	$\hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, H, w, Z)$
MLE	Maximum likelihood estimator	$\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$
MAP	Maximum a posteriori estimator	$\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$
MVUE	Minimum variance unbiased estimator	–
BLUE	Best linear unbiased estimator	–
BAUE	Best affine unbiased estimator	$\widehat{\mathbb{WB}}_{\mathrm{BAUE}}(y, X, V, W)$

(c) Special problem acronyms and references.

Acronym	Problem	Reference
SPP	Saddle point problem	(SPP)
MTP	Minimum trace problem	(MTP)

The estimator definitions, problem reformulations, and closed-form solutions are diagrammed in figures 1 to 4. Figure 1 summarizes the equivalences relating to GLS, TGLS, and ECGLS estimators. Figure 2 summarizes the equivalences relating to MLEs. Figure 3 summarizes the equivalences relating to MAP estimators. Figure 4 summarizes the equivalences relating to BAUEs. Rao's estimator (4) appears in figures 2 and 4, Albert's estimator (5) appears in figure 2, and the results of Magnus and Neudecker [67] appear in figures 1a, 1c, 2, and 4.

Rao's estimator (4) is derived, in the MLE context, by writing an equivalent saddle point problem,

$$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{SPP}$$

Albert's estimator (5) is derived as a perturbation result, i.e., as the limit of the MLE of the *perturbed model*,

$$y = X\beta + e, \qquad e \sim \mathrm{N}(0, V + \rho I) \tag{p-LGM}$$

$$\boxed{\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, H)} \xleftrightarrow{\text{Defn. 24}} \boxed{\hat{\beta} \text{ solves } \min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_H^2}$$

$$\Updownarrow \text{Thm. 25,[67,Ch. 13]} \qquad \Updownarrow \text{Lem. 10}$$

$$\boxed{\hat{\beta} \in (X'HX)^+ X'Hy + \mathcal{N}(X'HX)} \xleftrightarrow{\text{Lem. 3}} \boxed{X'HX\hat{\beta} = X'Hy}$$

(a) GLS estimator equivalences and solutions.

$$\boxed{\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{TGLS}}(y, X, H, \beta_0, \Gamma)} \xleftrightarrow{\text{Defn. 27}} \boxed{\hat{\beta} \text{ solves } \min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_H^2 + \frac{1}{2}\|\beta - \beta_0\|_\Gamma^2}$$

$$\Updownarrow \text{Thm. 28} \qquad \Updownarrow \text{Thm. 28}$$

$$\boxed{\hat{\beta} \in \Gamma_0\Gamma_0^+\beta_0 + \Gamma_0^+ X'H(y - X\beta_0) + \mathcal{N}(\Gamma_0)} \xleftrightarrow{\text{Thm. 25}} \boxed{\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{GLS}}\left(\begin{bmatrix} y \\ \beta_0 \end{bmatrix}, \begin{bmatrix} X \\ I \end{bmatrix}, \begin{bmatrix} H & 0 \\ 0 & \Gamma \end{bmatrix}\right)}$$

(b) TGLS estimator equivalences and solutions, where $\Gamma_0 := X'HX + \Gamma$.

$$\boxed{\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, H, w, Z)} \xleftrightarrow{\text{Defn. 30}} \boxed{\hat{\beta} \text{ solves } \min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_H^2 \text{ s.t. } w = Z\beta}$$

$$\Updownarrow \text{Thm. 31,[67,Ch. 13]} \qquad \Updownarrow \text{Lem. 10}$$

$$\boxed{\begin{array}{c}\hat{\beta} \in \beta_0 + G^+Z'F^+(w - Z\beta_0) \\ +\mathcal{N}(G), w \in \mathcal{R}(Z)\end{array}} \xleftrightarrow{\text{Lem. 3}} \boxed{\exists \hat{\lambda} \text{ s.t. } \begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'Hy \\ w \end{bmatrix}}$$

$$\text{Thm. 31} \qquad \text{Thm. 31}$$

$$\boxed{\begin{array}{c}\hat{\beta} \in Z^+w + B(BX'HXB)^+BX'Hz \\ +B\mathcal{N}(BX'HXB), w \in \mathcal{R}(Z)\end{array}} \xleftrightarrow{\text{Thm. 25}} \boxed{\begin{array}{c}\hat{\beta} \in Z^+w + B\hat{\mathbb{B}}_{\mathrm{GLS}}(z, XB, H), \\ w \in \mathcal{R}(Z)\end{array}}$$

(c) ECGLS estimator equivalences and solutions, where $G := X'HX + Z'Z$, $F := ZG^+Z'$, $\beta_0 := G^+X'Hy$, $B := I - Z^+Z$, and $z := y - Z^+w$.

Figure 1: Problem equivalences and solutions for least-squares-type estimators.

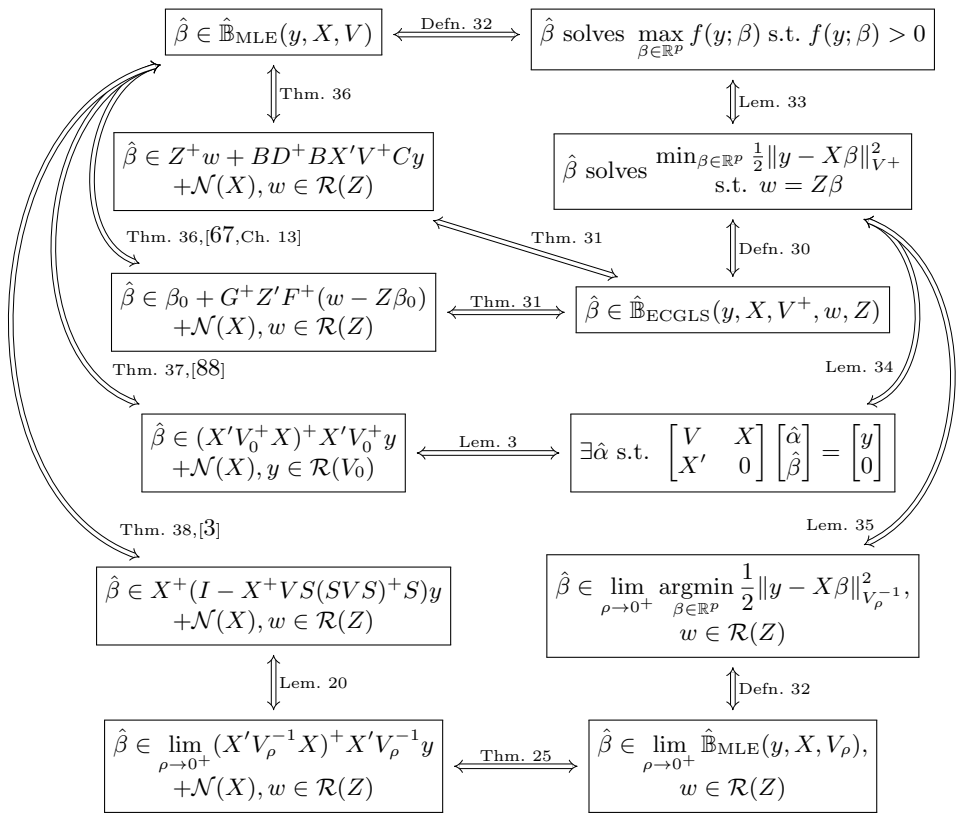

Figure 2: MLE equivalences and solutions, where $T := I - VV^+$, $w := Ty$, $Z := TX$, $B := I - Z^+Z$, $C := I - XZ^+$, $D := BX'V^+XB$, $G := X'V^+X + Z'Z$, $F := ZG^+Z'$, $\beta_0 := G^+X'V^+y$, $V_0 := V + XEX'$ for any positive definite E such that $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, $S := I - XX^+$, and $V_\rho := V + \rho I$ for each $\rho > 0$.

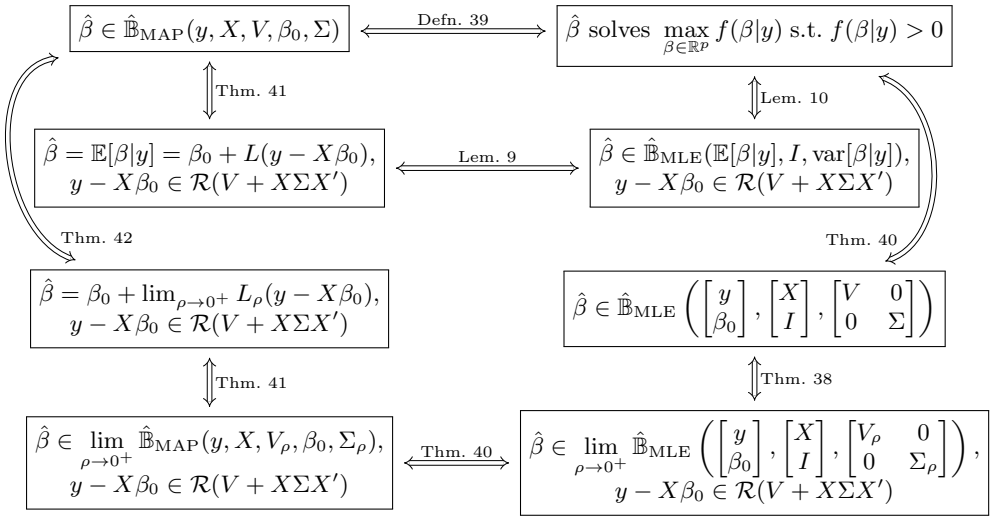

Figure 3: MAP estimator equivalences and solutions, where $V_\rho := V + \rho I$, $\Sigma_\rho := \Sigma + \rho I$, $L := \Sigma X'(V + X\Sigma X')^+$, and $L_\rho := \Sigma_\rho X'(V_\rho + X\Sigma_\rho X')^{-1}$ for each $\rho > 0$.

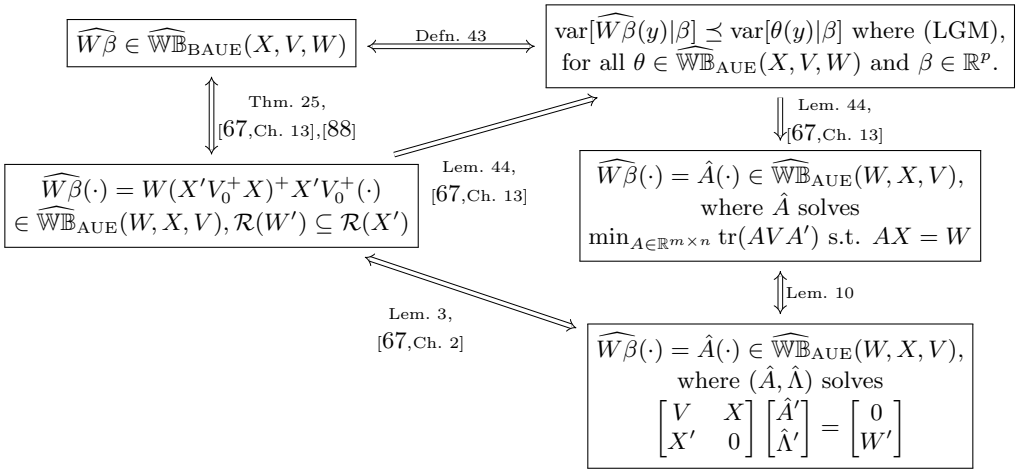

Figure 4: BAUE equivalences and solutions, where $V_0 := V + XEX'$ for any positive semidefinite E such that $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, and $\widehat{\mathbb{WB}}_{\mathrm{AUE}}(X, V, W) := \{\, \theta(\cdot) = A(\cdot) + c : \mathcal{R}(V_0) \to \mathbb{R}^m \mid \mathbb{E}[\theta(y)|\beta] = W\beta \text{ where (LGM)}, \forall \beta \in \mathbb{R}^p \,\}$.

which we justify using methods in variational analysis. Using the method of Magnus and Neudecker [67, pp. 286–287], we derive Rao's estimator through the following minimum trace problem

$$\min_{A \in \mathbb{R}^{m \times p}} \frac{1}{2}\mathrm{tr}(AVA') \quad \text{subject to} \quad AX = W \tag{MTP}$$

1.3 Applications and modern extensions

Linear Gaussian estimation is related to a wide variety of problems in control, optimization, and estimation. We discuss some of these applications and extensions in this section, going into more detail in sections 5 and 6.

1.3.1 Kalman filtering

The Kalman filter (KF) and extended Kalman filter (EKF) equations can be derived as (generalized) least squares estimates [13, 53, 56, 100]. More specifically, we show in section 5.2 that the full information estimation (FIE) and one-step ahead KF estimates are MAP estimates using the method outlined in [2, Chapter IX]. These formulations admit singular noise covariance matrices, which are used when the states have linear equality constraints [5, 83, 99, 112]. Stability of the KF has been discussed in both time-invariant [98] and time-varying [93] contexts.

1.3.2 Linear quadratic regulation

The linear quadratic regulator (LQR) has a Lagrangian dual relationship with the KF. Stability results are often established first for the LQR before being translated to the KF via duality [93, 98]. In the singular case (singular input penalties or singular measurement noise covariance), this duality is subtle and leads to the following situation. While the singular KF solution is always unique, it may not exist (due to an implied restriction on the measurements). Conversely, while the singular LQR always exists, it may not be unique (due to a free parameter corresponding to semidefinite directions of the objective). In section 5.1 we derive closed-form expressions that cover the complete set of LQR solutions.

1.3.3 Gradient descent

Gradient descent, gradient flow, and their variants are fundamental numerical algorithms in a wide variety of applied fields. When gradient descent is used on the OLS problem with $\mathrm{rank}(X) = p$, both the gradient descent and gradient flow algorithms produce iterates that solve certain TGLS problems [4]. When gradient descent is used on the OLS problem with $\mathrm{rank}(X) < p$, the iterates converge to the OLS solution that is closest in 2-norm distance to the initial guess. These facts are further discussed in section 4.3. This analysis also gives insight to the behavior of gradient descent for nonlinear regression algorithms, which is further explored in section 6.

1.3.4 Bayesian regression

The MAP estimator is closely related to the topic of regularized regression [21, 22]. With nonsingular V, the MAP estimator corresponds to Tikhonov regularization [109, 110], and with $V = \sigma^2 I$ and $\Sigma = \lambda I$ corresponds to ridge regression ($\|\cdot\|_2$ regularization) [51, 52]. With other prior distributions on β we can produce other regularized regression models. For example, if we assume the elements of β are independent and identically distributed with a Laplace distribution, the MAP estimation problem corresponds to LASSO regression (i.e., $\|\cdot\|_1$ regularization) [49, 108, 117]. These regularization techniques are commonly used in machine learning to reduce the variance of estimates [45, 78], as well as in algorithms for solving ill-conditioned problems [77, 111] and inverse problems [21, 22].

1.3.5 Sparse modeling

Sparse modeling, or finding estimates of β with the fewest nonzero elements, is an important problem in signal and image processing [19]. While formulations of this problem with and without noise are NP-hard [79], small to moderately sized problems can be solved using mixed integer programming [15, 17].

LASSO regression is particularly useful for sparse modeling problems, as it is equivalent to the $\|\cdot\|_0$-regularized problem for a wide class of X matrices [19, 23, 29].[6] This is because $\|\cdot\|_1$ is the best convex approximation of $\|\cdot\|_0$. Nonconvex approximations of $\|\cdot\|_0$ also make useful priors for sparse modeling, with some even producing exact solutions to the $\|\cdot\|_0$-regularized problem [37, 63, 101, 102]. For example, Fung and Mangasarian [37] demonstrate that the solution to the $\|\cdot\|_0$-regularized regression problem is the limit of the solution to the $\|\cdot\|_q$-regularized regression problem as $q \to 0^+$.[7] We show this fact in section 6 in a similar manner to our derivation of Albert's estimator (5).

1.4 Notation

Table 2 lists the notation and definitions that are used in the subsequent sections. It is worth pointing out some facts relating to table 2. First, note that $\|\cdot\|_W$ with W positive *semi*definite is a *semi*norm because it does not have the positive definiteness property. When $0 \le q < 1$, we refer to $\|\cdot\|_q$ as a pseudonorm because, for $0 < q < 1$, the triangle inequality does not hold, and for $q = 0$ absolute homogeneity does not hold. Moreover, we have the limit $\|x\|_0 = \lim_{q\to 0^+} \|x\|_q^q$, pointwise in $x \in \mathbb{R}^n$. Throughout, we may use $A \succ 0$ ($A \succeq 0$) as a shorthand to denote that A is positive definite (semidefinite). We use the Painlevé–Kuratowski notion of set convergence [59], as it is suitable for analysis of optimization problems [95, pp. 108–110].

[6] Define the 0-pseudonorm $\|\cdot\|_0$ as an operator that returns the number of nonzero elements of its argument.

[7] Fung and Mangasarian [37] showed a stronger result, that there is a positive constant $\overline{q} \in \mathbb{R}_{>0}$ such that the $\|\cdot\|_q$-regularized regression problem produces the same solution for all $q \in [0, \overline{q}]$. We demonstrate only the weaker limit result.

Table 2: Summary of notation.

Notation	Definition
$\mathbb{I}, \mathbb{I}_{\geq 0}, \mathbb{I}_{>0}$	Set of integers, nonnegative integers, and positive integers.
$\mathbb{R}, \mathbb{R}_{\geq 0}, \mathbb{R}_{>0}, \mathbb{R}^n$, $\mathbb{R}^{m\times n}$	Set of real numbers, nonnegative reals, positive reals, real n-vectors, and real $n \times m$ matrices.
I	Identity matrix, dimensions implied by context.
A^{-1}	Inverse of $A \in \mathbb{R}^{n\times n}$ (when it exists).
A', A^+	Transpose and pseudoinverse of $A \in \mathbb{R}^{m\times n}$.
$\mathcal{R}(A), \mathcal{N}(A)$	Range and null space of $A \in \mathbb{R}^{m\times n}$.
$\operatorname{rank}(A)$	Rank of $A \in \mathbb{R}^{m\times n}$.
$\overline{\sigma}(A), \underline{\sigma}(A)$	Largest and smallest singular values, respectively, of a matrix $A \in \mathbb{R}^{m\times n}$.
A is positive definite	if $A = A'$ and $x'Ax > 0$ for all $x \neq 0 \in \mathbb{R}^n$.
A is positive semidefinite	if $A = A'$ and $x'Ax \geq 0$ for all $x \in \mathbb{R}^n$.
$\mathbb{S}^n_{++}$ ($\mathbb{S}^n_{+}$)	Set of positive definite (semidefinite) $n \times n$ matrices.
$\succ$ ($\succeq$)	Loewner partial order on $\mathbb{S}^n_{++}$ ($\mathbb{S}^n_{+}$), defined as $A \succ B$ ($A \succeq B$) if $A - B$ is positive definite (semidefinite).
$\mathbb{D}^n, \mathbb{D}^n_{>0}, \mathbb{D}^n_{\geq 0}$	Set of diagonal, positive definite diagonal, and positive semidefinite diagonal $n \times n$ matrices.
$\operatorname{diag}(A_1, \ldots, A_n)$	$:= \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_n \end{bmatrix}$, the diagonalization of matrices $A_i \in \mathbb{R}^{n_i \times m_i}$.
$\lVert x \rVert$	$:= \sqrt{x'x}$, the 2-norm of $x \in \mathbb{R}^n$.
$\lVert x \rVert_W$	$:= \sqrt{x'Wx}$, the W-norm (W-seminorm) of $x \in \mathbb{R}^n$ for $W \in \mathbb{S}^n_{++}$ ($W \in \mathbb{S}^n_{+}$).
$\lVert x \rVert_q$	$:= (\sum_{i=1}^n \lvert x_i \rvert^q)^{1/q}$, the q-norm (q-pseudonorm) of $x \in \mathbb{R}^n$, where $q \geq 1$ ($0 < q < 1$).
$\lVert x \rVert_0$	$:= \#\{\, x_i \neq 0 \,\} = \lim_{q\to 0^+} \lVert x \rVert_q^q$, the 0-pseudonorm of $x \in \mathbb{R}^n$.
$\lVert A \rVert$	$:= \max_{\lVert x \rVert = 1} \lVert Ax \rVert$, the induced 2-norm of $A \in \mathbb{R}^{m\times n}$.
$A + \mathbb{B}$	$:= \{\, A + B \mid B \in \mathbb{B} \,\}$, the matrix-set sum of $A \in \mathbb{R}^{m\times n}$ and $\mathbb{B} \subseteq \mathbb{R}^{m\times n}$.
$A\mathbb{B}$	$:= \{\, AB \mid B \in \mathbb{B} \,\}$, the matrix-set product of $A \in \mathbb{R}^{m\times n}$ and $\mathbb{B} \subseteq \mathbb{R}^{n\times m}$.
$\mathbb{A}B$	$:= \{\, AB \mid A \in \mathbb{A} \,\}$, the matrix-set product of $\mathbb{A} \subseteq \mathbb{R}^{n\times m}$ and $B \in \mathbb{R}^{m\times n}$.
$d(A, \mathbb{B})$	$:= \inf_{B\in\mathbb{B}} \lVert A - B \rVert$, the point-to-set distance between $A \in \mathbb{R}^{m\times n}$ and $\mathbb{B} \subseteq \mathbb{R}^{m\times n}$.
$\lim_{\alpha\to\alpha_0} \mathbb{A}(\alpha)$	$:= \{\, A \in \mathbb{R}^{m\times n} \mid \lim_{\alpha\to\alpha_0} d(A, \mathbb{A}(\alpha)) = 0 \,\}$, the limit of the set-valued function $\mathbb{A}(\alpha)$ as $\alpha \to \alpha_0$.
$\lim_{\alpha\to 0^+} \mathbb{A}(\alpha)$	$:= \{\, A \in \mathbb{R}^{m\times n} \mid \lim_{\alpha\to 0^+} d(A, \mathbb{A}(\alpha)) = 0 \,\}$, the limit of the set-valued function $\mathbb{A}(\alpha)$ as $\alpha \to 0^+$.
$\mathbb{E}[x], \operatorname{var}[x]$	Expectation and variance of a random variable x.
$\mathbb{E}_x[f], \operatorname{var}_x[f]$	Expectation and variance of a function $f(x)$ with respect to the random variable x.

1.5 Summary

The rest of this paper is outlined as follows. In section 2, we present a basic toolset of linear algebra, probability, and convex optimization with which one can derive all of the relevant estimators. Section 3 contains all the estimator definitions, statements on problem reformulations and closed-form solutions, and any corollaries that are useful for computations, applications, and extensions. Problem formulation is analyzed from a computational perspective in section section 4, including relating the output of gradient methods with early termination to the TGLS problem. Applications in optimal control and estimation are discussed in section 5. Finally, in section 6, we show how the tools of linear estimation can be applied to modern extensions of the linear estimation problem.

Proofs of many of the results in the body are delayed until the appendix. Appendices A to D contain proofs of the results stated in section 3. Appendices F and I contain proofs of some preliminary results in section 2 that are required in the proofs of the results stated in section 3.

2 Background, definitions, and preliminary results

In this section we present definitions and results that are helpful to derive the results in figures 1 to 4. We defer to appendix E the basics of matrix analysis, probability, and convex optimization that are necessary to follow the proofs in the subsequent sections. Proofs of the results in this section are deferred to the appendices. Proofs of the results in this section are deferred to the appendices or left to external references.

2.1 Background and definitions

First, we present background and definitions that can be found in textbooks on linear algebra [41], probability theory [90], and convex optimization [18].

2.1.1 Pseudoinverse

Thus far, we have avoided defining two important concepts. First, the pseudoinverse $A^+ \in \mathbb{R}^{m\times n}$ solves the system of equations (6). Existence and uniqueness of the pseudoinverse were established by Moore [73] and later rediscovered by Penrose [81], which is restated (for real matrices) in the following theorem.

Theorem 1 ([73, 81])**.** *For any matrix $A \in \mathbb{R}^{n\times m}$, the system of equations*

$$AXA = A, \qquad XAX = X, \qquad (XA)' = XA, \qquad (AX)' = AX \tag{6}$$

has a unique solution, called the pseudoinverse *of A, denoted $X = A^+$.*

Some basic corollaries to theorem 1 are stated below for completeness. They are used throughout, without reference.

$$\begin{aligned} &(A^+)^+ = A, && A^+ = (A'A)^+A' = A'(AA')^+, && \mathcal{R}(A^+) = \mathcal{R}(A') = \mathcal{R}(A'A), \\ &(A')^+ = (A^+)', && A = A' \Rightarrow AA^+ = A^+A, && \mathcal{N}(A^+) = \mathcal{N}(A') = \mathcal{N}(AA') \end{aligned}$$

It was first shown by Albert [2, pp. 19–23] that the pseudoinverse of any $A \in \mathbb{R}^{m\times n}$ could be defined as the limit of the approximation $(A'A + \alpha I)^{-1}A'$ as $\alpha \to 0^+$. Later, Golub and Van Loan [41, pp. 296–297] proposed an exact form for the norm of the residual $R(\alpha) := A^+ - (A'A + \alpha I)^{-1}A'$ in terms of α and $\underline{\sigma}(A)$. These results are summarized in the following lemma.

Lemma 2 ([41])**.** *For any $A \in \mathbb{R}^{m\times n}$ and $\alpha > 0$ let $R(\alpha) := (A'A + \alpha I)^{-1}A' - A^+$. Then*

$$R(\alpha) = -\alpha V_1(\alpha I + \Sigma_1^2)^{-1}\Sigma_1^{-1}U_1' \tag{7a}$$

$$\|R(\alpha)\| = \frac{\alpha}{\underline{\sigma}(A)(\underline{\sigma}^2(A)) + \alpha} \tag{7b}$$

$$A^+ = \lim_{\alpha\to 0^+}(A'A + \alpha I)^{-1}A' \tag{7c}$$

given the SVD (8)*.*

2.1.2 Linear equations

lemma 3 provides necessary and sufficient conditions for the existence of solutions to linear *vector* equations, as well as closed-form solutions to those problems. This lemma is useful in solving least squares problems. corollary 4 extends the lemma 3 to linear *matrix* equations, which is used to solve constrained minimum trace problems (MTP). The proofs of lemma 3 and corollary 4 can be found in appendix E.3. Lemma 3 was adapted from [67, Theorems 2.11–12]. Corollary 4 was adapted from [67, Theorems 2.13], which characterizes the solutions to the more general linear matrix equation $AXB = C$.

Lemma 3 ([67])**.** *Let $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. The following statements are equivalent.*

1. *$\mathcal{S} := \{\, x \in \mathbb{R}^p \mid Ax = b \,\}$ is nonempty.*
2. *$b \in \mathcal{R}(A)$.*
3. *$AA^+b = b$.*

If the above statements hold, then $\mathcal{S} = A^+b + \mathcal{N}(A)$.

Corollary 4 ([67])**.** *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{m\times p}$. The following statements are equivalent.*

1. *$\mathcal{S} := \{\, X \in \mathbb{R}^{n\times p} \mid AX = B \,\}$ is nonempty.*
2. *$\mathcal{R}(B) \subseteq \mathcal{R}(A)$.*
3. *$AA^+B = B$.*

If the above statements hold, then $\mathcal{S} = A^+B + \{\, (I - A^+A)Q \mid Q \in \mathbb{R}^{n\times p} \,\}$.

2.1.3 Projectors

We say a matrix $P \in \mathbb{R}^{n\times n}$ is a projector if $P^2 = P$. We say it is an orthogonal projector if additionally $P = P'$. It follows from theorem 1 that the following functions of the pseudoinverse are orthogonal projectors,

$$A^+A, \qquad AA^+, \qquad I - A^+A, \qquad I - AA^+$$

Moreover, we can write equivalent expressions for $\mathcal{R}(A)$ and $\mathcal{N}(A)$ in terms of these projectors,

$$\begin{aligned}\mathcal{R}(A) &= \{\, AA^+q \mid q \in \mathbb{R}^n \,\} = \mathcal{R}(AA^+),\\ \mathcal{N}(A) &= \{\, (I - A^+A)q \mid q \in \mathbb{R}^m \,\} = \mathcal{R}(I - A^+A)\end{aligned}$$

and likewise for $\mathcal{R}(A')$ and $\mathcal{N}(A')$. Finally, we have the following results about projectors, the proofs of which can be found in appendix E.2. Finally, we have the following result about projectors, which is easily demonstrated using theorem 1.

Lemma 5. *If $P \in \mathbb{R}^{n\times n}$ is an orthogonal projector, then $P^+ = P$ and $(PA)^+ = (PA)^+P$ for all $A \in \mathbb{R}^{n\times m}$.*

2.1.4 Singular value decomposition

We say a matrix $Q \in \mathbb{R}^{n\times r}$ is orthogonal if $Q'Q = I$. Notice that if Q is orthogonal, then QQ' is an orthogonal projector. The following lemma states some properties of the pseudoinverse of orthogonal matrices.

Lemma 6. *If $Q \in \mathbb{R}^{n\times m}$ is orthogonal, then $Q^+ = Q'$ and $(QA)^+ = A^+Q'$ for all $A \in \mathbb{R}^{m\times n}$.*

We say a matrix $U \in \mathbb{R}^{n\times n}$ is unitary if $U'U = UU' = I$. If a matrix $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{n\times n}$ is unitary, then $U_1'U_1 = I$, $U_2'U_2 = I$, and $U_1U_1' + U_2U_2' = I$, where r is the number of columns of U_1. Clearly, U_1 and U_2 are orthogonal matrices, and U_1U_1' and U_2U_2' are orthogonal projectors.

These definitions allow us to state the existence of the singular value decomposition (SVD) [41, Theorem 2.4.1].

Theorem 7. *For any $A \in \mathbb{R}^{n\times m}$, there exist unitary matrices $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{n\times n}$ and $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \in \mathbb{R}^{m\times m}$ and constants $\sigma_1 \geq \ldots \geq \sigma_r > 0$ such that*

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix} = U_1\Sigma_1V_1' \tag{8}$$

where $r := \operatorname{rank}(A)$ and $\Sigma_1 := \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r\times r}$.

The columns of U and V are called the left and right singular vectors, and the diagonal entries of Σ_1 are called the singular values. The singular vectors contain important information about the range and null spaces of A and A',

$$\mathcal{R}(A) = \mathcal{R}(U_1), \qquad \mathcal{N}(A) = \mathcal{R}(V_2), \qquad \mathcal{R}(A') = \mathcal{R}(V_1), \qquad \mathcal{N}(A') = \mathcal{R}(U_2)$$

given the SVD (8). Moreover, it can easily be shown that

$$A^+ = V_1 \Sigma_1^{-1} U_1'$$

by checking the conditions of theorem 1, given the SVD (8).

A special case of the SVD arises when $A \in \mathbb{S}_+^n$, in which case the left and right singular values are equal [41], i.e.

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1' \\ U_2' \end{bmatrix} = U_1 \Sigma_1 U_1' \in \mathbb{S}_+^n \tag{9}$$

is a SVD. A consequence of this special case is that, given the SVD (9), we can define a matrix square root for each $A \in \mathbb{S}_+^n$ as $A^{1/2} := U_1 \Sigma_1^{1/2} U_1'$ for which

$$A^{1/2} A^{1/2} = U_1 \Sigma_1^{1/2} U_1' U_1 \Sigma_1^{1/2} U_1' = U_1 \Sigma_1 U_1' = A$$

and

$$\mathcal{R}(A^{1/2}) = \mathcal{R}(U_1) = \mathcal{R}(A), \qquad \mathcal{N}(A^{1/2}) = \mathcal{R}(U_2) = \mathcal{N}(A)$$

Notice also that for every $A \in \mathbb{R}^{n \times m}$, $AA' \in \mathbb{S}_+^n$ because

$$AA' = U_1 \Sigma_1^2 U_1'$$

has the form of the SVD (9). Likewise, for any $A \in \mathbb{S}_+^n$ and $B \in \mathbb{R}^{n \times m}$, with the SVD (9) $B'AB = (B'U_1\Sigma_1^{1/2})(B'U_1\Sigma_1^{1/2})'$ and therefore $B'AB \in \mathbb{S}_+^m$. Another consequence of (9) is that every $A \in \mathbb{S}_{++}^n$ is nonsingular since the inverse can be defined as $A^{-1} = U\Sigma^{-1}U'$, given the SVD $A = U\Sigma U'$.

2.1.5 Degenerate Gaussian

Second, we define the multivariate Gaussian distribution for positive semidefinite covariance matrices. A derivation of this likelihood function (10) can be found in [90, pp. 527–528]. Below, we state a more parsimonious definition of the positive semidefinite covariance multivariate Gaussian, sometimes called the *degenerate* or *singular* multivariate Gaussian or normal.

Definition 8 ([90])**.** *A random variable $x \in \mathbb{R}^n$ has a Gaussian distribution with mean $\mu \in \mathbb{R}^n$ and (possibly singular) covariance matrix $\Sigma \in \mathbb{S}_+^n$, denoted $x \sim \mathrm{N}(\mu, \Sigma)$, if the probability density function is*

$$f(x) := \frac{1}{(2\pi)^{n/2} |\Sigma|_+^{1/2}} \exp\left(-\frac{1}{2} \|x - \mu\|_{\Sigma^+}^2 \right) \tag{10}$$

on the support $x \in \mu + \mathcal{R}(\Sigma)$, and is zero elsewhere, where $|\cdot|_+ : \mathbb{R}^{n \times n} \to \mathbb{R}$ is called the pseudodeterminant *and is defined for all $A \in \mathbb{R}^{n \times n}$ as the product of the nonzero singular values of A.*[8,9]

[8] When $\Sigma \in \mathbb{S}_{++}^n$, the singular values of Σ are strictly positive and the inverse exists, so we have $|\Sigma|_+ = |\Sigma|$ and $\Sigma^+ = \Sigma^{-1}$. Moreover, $\mathcal{R}(\Sigma) = \mathbb{R}^n$ so the support is nondegenerate.

[9] To integrate over this density, one must define, using the disintegration theorem, a Lebesgue measure over the rank(Σ)-dimensional subspace $\mu + \mathcal{R}(\Sigma)$ [90].

The definition of the singular values is presented later in this section. We refer the reader to theorem 7 for the definition of the singular values. To ensure the likelihood functions in (MLE) and (MAP) produce estimates consistent with the models (LGM) and (B-LGM), we set the likelihood function to zero for impossible values of β. In a typical MLE or MAP problem, the likelihood function is equal to the probability density function for all parameter values, but since we are performing optimization over a *degenerate* distribution, we require constraints on the feasible values of β.

Some properties of non-degenerate Gaussians carry over to the degenerate case with minor modifications. For example, if $x \sim \mathrm{N}(\mu, \Sigma)$, then $\mathbb{E}[Ax] = A\mu$ and $\mathrm{var}[Ax] = A\Sigma A'$. Moreover, we have the following lemma due to Marsaglia [69].

Lemma 9 ([69])**.** *Let $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ be random variables such that*

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_x & \Sigma_{xy}' \\ \Sigma_{xy}' & \Sigma_y \end{bmatrix} \right)$$

Then $x|y \sim \mathrm{N}(\mu_x + \Sigma_{xy}\Sigma_y^+(y - \mu_y), \Sigma_x - \Sigma_{xy}\Sigma_y^+\Sigma_{xy}')$ for all $y \in \mu_y + \mathcal{R}(\Sigma_y)$.

A proof of lemma 9 can be found in appendix E.7. It is worth pointing out that lemma 9 leaves $x|y$ (and its density function) undefined when y leaves the support of $f(\cdot; \mu_y, \Sigma_y)$. This is desirable because it does not make sense to condition x on an impossible realization of y.

2.1.6 Convex optimization

We say a set $\mathcal{S} \subseteq \mathbb{R}^n$ is convex if $tx + (1-t)y \in \mathcal{S}$ for all $x, y \in \mathcal{S}$ and $0 \leq t \leq 1$. We say a function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

for all $x, y \in \mathbb{R}^n$ and $0 < t < 1$. If a function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then it is convex if and only if

$$f(y) \geq f(x) + (y - x)' \frac{df}{dx}(x)$$

for all $x, y \in \mathbb{R}^n$ [18, p. 69]. An immediate consequence of this fact is that, if f is differentiable and convex and $\mathcal{S}$ is convex, then x^0 solves $\min_{x \in \mathcal{S}} f(x)$ if and only if $x^0 \in \mathcal{S}$ and $(x - x^0)'(df/dx)(x^0) \geq 0$ for all $x \in \mathcal{S}$. A further corollary is that, if f is differentiable and convex, then $x^0 \in \mathbb{R}^n$ solves $\min_{x \in \mathbb{R}^n} f(x)$ if and only if $(df/dx)(x^0) = 0$. Lemma 10 provides necessary and sufficient conditions for the solutions to convex optimization problems

$$\min_{x \in \mathbb{R}^n} f(x) \qquad \text{subject to} \qquad Ax = b \tag{11}$$

using the method of Lagrange multipliers. A proof of lemma 10 can be found in [18, pp. 141–142]. The proof (adapted from [18, pp. 141–142]) can be found in appendix E.8. We use the necessary and sufficient conditions (13) to solve the convex optimization problems that we encounter.

Lemma 10 ([18])**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathcal{R}(A)$. Define the Lagrangian*

$$\mathcal{L}(x, \lambda) = f(x) + \lambda'(Ax - b) \tag{12}$$

Then $x^0 \in \mathbb{R}^n$ solves (11) if and only if there exists $\lambda^0 \in \mathbb{R}^m$ such that

$$\frac{\partial \mathcal{L}}{\partial x}(x^0, \lambda^0) = 0, \qquad \frac{\partial \mathcal{L}}{\partial \lambda}(x^0, \lambda^0) = 0 \tag{13}$$

2.1.7 Miscellaneous facts

We present Woodbury's matrix identity below [97, 114].

Theorem 11 ([97, 114])**.** *For any $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times m}$, and $D \in \mathbb{R}^{m\times n}$,*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

subject to existence of the inverses.

All positive definite matrices $A \in \mathbb{S}^n_{++}$ are invertible, i.e., $A^{-1} \in \mathbb{S}^n_{++}$ exists. For any positive semidefinite matrix $A \in \mathbb{S}^n_+$, there exists a positive semidefinite matrix $A^{1/2} \in \mathbb{S}^n_+$ such that $A = A^{1/2}A^{1/2}$. Moreover, $A \in \mathbb{S}^n_+$ if and only if there exists a matrix $B \in \mathbb{R}^{n\times r}$ such that $A = BB'$ where $r = \operatorname{rank}(A)$.

2.2 Preliminary results

Next, we present some preliminary results, most of which are from the linear algebra literature and the others are corollaries unique to this work.

2.2.1 Block matrix pseudoinversion

As alluded to in section 1, the pseudoinverse of the block matrix

$$M = \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \tag{14}$$

is essential in deriving results about the estimators of (LGM). First, it is useful to state some consequences of corollary 4 that are not only used in deriving M^+, but are also useful in the estimator derivations of appendices A, B, and D.

Lemma 12 ([67])**.** *For any $V \in \mathbb{S}^n_+$, $X \in \mathbb{R}^{n\times p}$, and $E \in \mathbb{S}^p_+$, let $V_0 := V + XEX'$ and $W_0 := X'V_0^+X$. If $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, then*

1. *$\mathcal{R}(V) \subseteq \mathcal{R}(V_0)$,*
2. *$V_0V_0^+X = X$, $V_0V_0^+V = V$,*
3. *$\mathcal{R}(X') = \mathcal{R}(W_0)$, and $W_0W_0^+X' = X'$.*

The formula for M, MM^+, and M^+M are given in the following lemma.

Lemma 13 ([67, 85])**.** *For any $V \in \mathbb{S}^n_+$, $X \in \mathbb{R}^{n\times p}$, and $E \in \mathbb{S}^p_+$, consider* (14) *and let $V_0 := V + XEX'$ and $W_0 := X'V_0^+X$. If $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, then*

$$M^+ = \begin{bmatrix} V_0^+ - V_0^+XW_0^+X'V_0^+ & V_0^+XW_0^+ \\ W_0^+X'V_0^+ & W_0W_0^+EW_0W_0^+ - W_0^+ \end{bmatrix}$$
$$MM^+ = M^+M = \begin{bmatrix} V_0V_0^+ & 0 \\ 0 & W_0W_0^+ \end{bmatrix}$$

Lemmas 12 and 13 were adapted from [67, Theorems 3.20–21]. In appendix F we prove the generalized versions of these lemmas. as well as an immediate corollary to lemma 13 which is stated below.

Corollary 14. *For any $V \in \mathbb{S}^n_+$, $X \in \mathbb{R}^{n\times p}$, and $E \in \mathbb{S}^p_+$, let $V_0 := V + XEX'$ and $V_1 := V + XX'$. If $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, then*

$$(X'V_0^+X)^+X'V_0^+ = (X'V_1^+X)^+X'V_1^+$$

2.2.2 Pseudoinverse of sum of positive semidefinite matrices

Another important result is a formula, proposed by Minamide [71], for the pseudoinverse of $V_1 = V + XX'$. Minamide's lemma is restated below and proven[10] in appendix G for completeness.

Lemma 15 (Minamide [71])**.** *For any $V \in \mathbb{S}^n_+$ and $X \in \mathbb{R}^{n\times p}$, let $Z = (I - VV^+)X$, $B = I - Z^+Z$, $C = I - XZ^+$, and $D = I + BX'V^+XB$. Then,*

$$(V + XX')^+ = C'V^+C + Z'^+Z^+ - C'V^+XBD^{-1}BX'V^+C$$

Notice that we have lost dependence on E, but can regain it using corollary 14. A direct consequence of Minamide's lemma is stated in the following lemma, which can be used to show (31c).

Lemma 16. *For any $V \in \mathbb{S}^n_+$ and $X \in \mathbb{R}^{n\times p}$, let $V_1 = V + XX'$, $Z = (I - VV^+)X$, $B = I - Z^+Z$, and $C = I - XZ^+$. Then,*

$$(X'V_1^+X)^+X'V_1^+ = Z^+ + (BX'V^+XB)^+BX'V^+C$$

The proof of the above lemma can be found in appendix G.

2.2.3 Bounds on oblique projectors and weighted pseudoinverses

In order to take the limit of the perturbed solution, we require some bounds on the norms of so-called *oblique* pseudoinverses and projectors. Stewart [103] first proposed an upper bound on these matrices and Vavasis [111] later proved that the upper bound is exact. These results are summarized in theorem 17.[11]

[10] Minamide proved lemma 15 using a combination of nullspace and projector arguments, but in appendix G we show the formula directly using the definition of the pseudoinverse (theorem 1).

[11] We take a slight liberty in the statements of theorem 17 and lemma 19 in that we do not require A to be full column rank. However, the proofs in [80, 103] can easily be extended to A of arbitrary rank using the *thin* SVD.

Theorem 17 ([103, 111])**.** *For any nonzero matrix $A \in \mathbb{R}^{m\times n}$,*

$$\sup_{D\in\mathbb{D}^m_{>0}} \|(A'DA)^+A'D\| \leq \frac{1}{\underline{\sigma}(A)\chi(A)} \tag{15a}$$

$$\sup_{D\in\mathbb{D}^m_{>0}} \|A(A'DA)^+A'D\| = \frac{1}{\chi(A)} \tag{15b}$$

where

$$\chi(A) := \inf_{x\in\mathbb{X}(A), y\in\mathbb{Y}(A)} \|x-y\| \tag{16a}$$

$$\mathbb{X}(A) := \{\, x \in \mathcal{R}(A) \mid \|x\| = 1 \,\} \tag{16b}$$

$$\mathbb{Y}(A) := \{\, y \mid \exists D \in \mathbb{D}^n_{>0} \text{ s.t. } A'Dy = 0 \,\} \tag{16c}$$

The above theorem implies a more general upper bound on approximations to the linear estimation problem.

Corollary 18. *For any $X \in \mathbb{R}^{n\times p}$ and $V \in \mathbb{S}^n_{++}$,*

$$\sup_{D\in\mathbb{D}^m_{>0}} \|(X'V_D^{-1}X)^+X'V_D^{-1}\| \leq \frac{1}{\underline{\sigma}(X)\chi(Q'U_1)} \tag{17a}$$

$$\sup_{D\in\mathbb{D}^m_{>0}} \|X(X'V_D^{-1}X)^+X'V_D^{-1}\| \leq \frac{1}{\chi(Q'U_1)} \tag{17b}$$

where $V_D := V + QDQ'$, $V = QSQ'$ is the SVD of V, $X = U_1\Sigma_1V_1'$ is the economic SVD of X, and $\chi(\cdot)$ is defined in (16a)*.*

Finally, we state an equivalent form of the infimum (16a) that can be directly computed from the SVD of A. Note that the computational cost of this problem is exponential in m, so it may not be desirable to actually compute this bound.

Lemma 19 (Generalized from [80, 103])**.** *For any nonzero matrix $A \in \mathbb{R}^{m\times n}$,*

$$\chi(A) = \min_{U\in\mathbb{U}(A)} \underline{\sigma}(U) \tag{18}$$

where $\chi(\cdot)$ is defined in (16a) *and $\mathbb{U}(A)$ is the set of all submatrices formed by the rows of U_1, where $A = U_1\Sigma_1V_1'$ is the economic SVD of A.*

2.2.4 Limit of the perturbed problem

One approach to solving (MLE) is to add a perturbation ρI to the positive semidefinite covariance V so that it is positive definite, find the MLE with the perturbed covariance $V + \rho I$, and take the limit of that MLE as the perturbation goes to zero. This approach is inspired by Bellman's application to extending the symmetric matrix eigenvalue decomposition from matrices with distinct roots to any symmetric matrix [9, p. 40], and Albert's solution to the ECGLS problem (ECGLS) with $H \in \mathbb{S}^n_{++}$ [2, pp. 119–121]. The following lemma characterizes the error of this approximation and establishes that the limit of the perturbed solution coincides with the exact solution.

Lemma 20. *For any $V \in \mathbb{S}^n_+$ and $X \in \mathbb{R}^{n\times p}$, there exist constants $\alpha_i, \beta_i > 0$ for $i = 1, 2, 3$ such that*

$$\|(X'V_\rho^{-1}X)^+X'V_\rho^{-1} - X^+ + X^+VS(SVS)^+S\| \leq \sum_{i=1}^{3} \frac{\alpha_i\rho}{\beta_i + \rho} \qquad \forall \rho > 0 \tag{19}$$

and moreover,

$$\lim_{\rho\to 0^+} (X'V_\rho^{-1}X)^+X'V_\rho^{-1} = X^+ - X^+VS(SVS)^+S \tag{20}$$

where $S := I - XX^+$ and $V_\rho := V + \rho I$ for all $\rho > 0$.

While proving lemma 20 in appendix I, we derive expressions for the constants α_i and β_i in terms of the singular values of X, V, and $SV^{1/2}$, and the constant $\chi(Q'U_1)$ defined in theorem 17.

2.2.5 Miscellaneous results

A miscellaneous result concerns the connection between the limiting result of lemma 20 and Rao's estimator (4), which we show in appendix J.

Lemma 21. *For any $V \in \mathbb{S}^n_+$, $X \in \mathbb{R}^{n\times p}$, and $E \in \mathbb{S}^p_+$, if $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, then*

$$(X'V_0^+X)^+X'V_0^+ = X^+ - X^+V(SVS)^+$$

where $V_0 = V + XEX'$ and $S = I - XX^+$.

As a corollary to the above lemma, we can show Rao's estimator (4) is independent of the choice of E (subject to the relevant range constraint). This corollary may be seen as an alternative, yet equivalent statement of corollary 14.

Corollary 22. *For any $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, $E_0 \in \mathbb{S}^p_+$, and $E_1 \in \mathbb{S}^p_+$, let $V_0 = V + XE_0X'$ and $V_1 = V + XE_1X'$. If $\mathcal{R}(X) \subseteq \mathcal{R}(V + XE_0X')$ and $\mathcal{R}(X) \subseteq \mathcal{R}(V + XE_1X')$, then*

$$(X'V_0^+X)^+X'V_0^+ = (X'V_1^+X)^+X'V_1^+$$

Finally, we state necessary and sufficient conditions for which (LGM) is a consistent model. This lemma is significant because it indicates the set of values where the observations y can be found.

Lemma 23. *Let $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, $E \in \mathbb{S}^p_+$, and $V_0 = V + XEX'$, and assume $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$. There exists $\beta \in \mathbb{R}^p$ such that* (LGM) *has nonzero probability if and only if $y \in \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) = \mathcal{R}(V_0)$, almost surely.*

Proofs of the above three results can be found in appendix J.

3 Estimators and their solutions

In this section we define and state results pertaining to estimators for the model (LGM). Proofs of the results stated in this section can be found in appendices A to D.

3.1 Generalized least squares

Below, we define the GLS estimator of the model (LGM).

Definition 24. *Define* $\hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, H)$ *as the set of solutions to*

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_H^2 \tag{GLS}$$

where $H \in \mathbb{S}_+^n$ *is a positive semidefinite weighting matrix. We say* $\hat{\beta}$ *is a GLS estimator (of the model* (LGM)*, with weighting matrix* H*) if* $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, H)$*.*

It is straightforward to solve (GLS) using the fact that unconstrained minimization of a convex objective is equivalent to finding a stationary point of the objective function [14, 18]. Solutions to (GLS) are stated in the following theorem.

Theorem 25. *For any* $y \in \mathbb{R}^n$*,* $X \in \mathbb{R}^{n \times p}$*, and* $H \in \mathbb{S}_+^n$*,*

$$\hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, H) = (X'HX)^+ X'Hy + \mathcal{N}(X'HX) \tag{21}$$

A corollary to theorem 25 is stated below. corollary 26 extends the standard completion of squares results for positive definite weighting matrices to the semidefinite case and is used in section 5 to derive the LQR solution set.

Corollary 26. *For any* $y \in \mathbb{R}^n$*,* $X \in \mathbb{R}^{n \times p}$*, and* $H \in \mathbb{S}_+^n$*, if* $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, H)$*, then*

$$V(\beta) := \frac{1}{2}\|y - X\beta\|_H^2 = \frac{1}{2}\|\beta - \hat{\beta}\|_{X'HX}^2 + \frac{1}{2}\|y\|_{H_0}^2 \tag{22}$$

for all $\beta \in \mathbb{R}^p$ *where* $H_0 := H - HX(X'HX)^+X'H$*.*

Proofs of theorem 25 and corollary 26 can be found in appendix A. Clearly, if $\mathrm{rank}(X) = p$ and $H = \sigma^2 I$, the unique solution is Gauss' estimator (1), and if $\mathrm{rank}(X) = p$ and $H \in \mathbb{S}_{++}^n$, the unique solution is Aitken's estimator (2).

3.2 Tikhonov generalized least squares

We can add the regularization term $\frac{1}{2}\|\beta - \beta_0\|_\Gamma^2$ to the objective of (GLS) to give the following TGLS problem.

Definition 27. *Define* $\hat{\mathbb{B}}_{\mathrm{TGLS}}(y, X, H, \beta_0, \Gamma)$ *as the set of solutions to*

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_H^2 + \frac{1}{2}\|\beta - \beta_0\|_\Gamma^2 \tag{TGLS}$$

where $H \in \mathbb{S}_+^n$ *and* $\Gamma \in \mathbb{S}_+^p$ *are positive semidefinite weighting matrices and* $\beta_0 \in \mathbb{R}^p$ *is a bias parameter. We say* $\hat{\beta}$ *is a TGLS estimator (of the model* (LGM)*, with weighting matrices* H, Γ *and bias* β_0*) if* $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, H, \beta_0, \Gamma)$*.*

It is straightforward to solve (TGLS) by rewriting the objective in the form of (GLS). Solutions to (TGLS) are stated in the following theorem.

Theorem 28. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $H \in \mathbb{S}^n_+$, $\Gamma \in \mathbb{S}^p_+$, and $\beta_0 \in \mathbb{R}^p$,*

$$\hat{\mathbb{B}}_{\text{TGLS}}(y, X, H, \beta_0, \Gamma) = \hat{\mathbb{B}}_{\text{GLS}}\left(\begin{bmatrix} y \\ \beta_0 \end{bmatrix}, \begin{bmatrix} X \\ I \end{bmatrix}, \begin{bmatrix} H & 0 \\ 0 & \Gamma \end{bmatrix}\right) \tag{23a}$$

$$= \Gamma_0^+\Gamma_0\beta_0 + L(y - X\beta_0) + \mathcal{N}(\Gamma_0) \tag{23b}$$

where $\Gamma_0 := X'HX + \Gamma$ and $L := \Gamma_0^+ X'H$.

A corollary to theorem 28 is stated below. corollary 29 extends the completion-of-squares result in corollary 26 to include regularization terms and is also used in section 5 to derive the LQR solution set.

Corollary 29. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $H \in \mathbb{S}^n_+$, $\beta_0 \in \mathbb{R}^p$, and $\Gamma \in \mathbb{S}^p_+$, if $\hat{\beta} \in \hat{\mathbb{B}}_{\text{TGLS}}(y, X, H, \beta_0, \Gamma)$, then*

$$V(\beta) := \frac{1}{2}\|y - X\beta\|_H^2 + \frac{1}{2}\|\beta - \beta_0\|_\Gamma^2 = \frac{1}{2}\|\beta - \hat{\beta}\|_{\Gamma_0}^2 + \frac{1}{2}\|y - X\beta_0\|_{\Gamma_1}^2 \tag{24}$$

for all $\beta \in \mathbb{R}^p$, where $\Gamma_0 := X'HX + \Gamma$ and $\Gamma_1 := H - HX\Gamma_0^+X'H$.

Proofs of theorem 28 and corollary 29 can be found in appendix A. Clearly, if $\text{rank}(X) = p$, $H = \sigma^2 I$, $\Gamma \in \mathbb{S}^p_{++}$, and $\beta_0 = 0$, then the TGLS estimator is Tikhonov's estimator, and if additionally $\Gamma = \lambda I$ for some $\lambda > 0$, the TGLS estimator is a ridge estimator.

3.3 Equality constrained generalized least squares

Some problems require the imposition of a linear constraint $Z\beta = w$ with constraint parameters $Z \in \mathbb{R}^{c\times n}$ and $w \in \mathbb{R}^c$, where c is the dimension of the constraint. Given this constraint, the ECGLS estimator is defined as follows.

Definition 30. *Define $\hat{\mathbb{B}}_{\text{ECGLS}}(y, X, H, w, Z)$ as the set of solutions to*

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_H^2 \quad \text{subject to} \quad w = Z\beta \tag{ECGLS}$$

where $H \in \mathbb{S}^n_+$ is a positive semidefinite weighting matrix and $Z \in \mathbb{R}^{c\times n}$ and $w \in \mathbb{R}^c$ are the constraint parameters. We say $\hat{\beta}$ is a ECGLS estimator (of the model (LGM)*, with weighting matrix H and constraint parameters $Z \in \mathbb{R}^{c\times n}$ and $w \in \mathbb{R}^c$) if $\hat{\beta} \in \hat{\mathbb{B}}_{\text{ECGLS}}(y, X, H, w, Z)$.*

There are two ways to solve (ECGLS), both of which are captured in the following theorem. First, the constraint equation $Z\beta = w$ is solved up to a free parameter α, eliminating the constraints and reparameterizing the problem to the new parameters α, which can then be solved using theorem 25. Second, the method of Lagrange multipliers are applied to generate an augmented saddle point system [14, 18]. Each method produces a different closed-form solution to the ECGLS problem.

Theorem 31. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $H \in \mathbb{S}^n_+$, $w \in \mathbb{R}^c$, and $Z \in \mathbb{R}^{c\times n}$, $\hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, H, w, Z)$ is nonempty if and only if $w \in \mathcal{R}(Z)$. Moreover, if $w \in \mathcal{R}(Z)$ then*

$$\hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, H, w, Z) = Z^+ w + B\hat{\mathbb{B}}_{\mathrm{GLS}}(y, XB, H) \tag{25a}$$

$$= Z^+ w + B(BX'HXB)^+ BX'Hz + B\mathcal{N}(BX'HXB) \tag{25b}$$

where $B := I - Z^+ Z$ and $z := y - Z^+ w$, and

$$\hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, H, w, Z) = \beta_0 + G^+ Z' F^+ (w - Z\beta_0) + \mathcal{N}(G) \tag{26}$$

where $G := X'HX + Z'Z$, $F := ZG^+Z'$, and $\beta_0 := G^+ X'Hy$.

The proof of theorem 31 can be found in appendix A. The reparameterization method produces the solution set (25). The method of Lagrange multipliers produces the solution set (26). We refer the reader to [67, Chapter 11] for a modern treatment of the GLS and ECGLS problems. Two expressions (21) and (26) can be found in [67, Chapter 11], whereas the remaining expression (25) is unique to this work.

3.4 Maximum likelihood estimator

The MLE is applicable to a wide variety of estimation problems. As the namesake suggests, the MLE maximizes the likelihood of y. For certain likelihood functions (e.g., convex and log-convex) the MLE is easily solved numerically using a wide variety of optimization software.

The MLE of the model (LGM) is defined in definition 32 below. We define the Gaussian distribution in the case where the covariance matrix is possibly singular in section 2.

Definition 32. *Define $\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$ as the set of solutions to*

$$\max_{\beta\in\mathbb{R}^p} f(y;\beta) \quad \text{subject to} \quad f(y;\beta) > 0 \tag{MLE}$$

where $f(\cdot;\beta)$ is the probability density of y in the model (LGM)*, given the parameters $\beta \in \mathbb{R}^p$.*[12] *We say that $\hat{\beta}$ is a MLE of the model* (LGM) *if $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$.*

We present three methods for deriving solutions to (MLE). The first method is based on rewriting (MLE) in the form of (ECGLS) and invoking theorem 31 to obtain two expressions for the solution. The second method equates (MLE) to the saddle point system (SPP). The third method is based on finding a sequence of MLEs $\hat{\beta}_\rho$ of the perturbed model (p-LGM), and taking the limit as $\rho \to 0^+$. The problem reformulations in these methods are state in lemmas 33 to 35 below.

Lemma 33. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $V \in \mathbb{S}^n_+$,*

$$\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V) = \hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, V^+, w, Z) \tag{27}$$

and $\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$ is nonempty if and only if $w \in \mathcal{R}(Z)$, where $T := I - VV^+$, $w := Ty$, and $Z := TX$.

[12] Using properties of the Gaussian distribution, $y \sim \mathrm{N}(X\beta, V)$. See definition 8 for a restatement of the probability density function of a Gaussian random variable given in [90, pp. 527–528]. Note the probability *measure* takes on a degenerate character.

Lemma 34. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $V \in \mathbb{S}^n_+$, $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$ if and only if there exists $\hat{\alpha} \in \mathbb{R}^n$ such that* (SPP).

Lemma 35. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $V \in \mathbb{S}^n_+$, if $w \in \mathcal{R}(Z)$, then*

$$\lim_{\rho\to 0^+} \left\{ \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|^2_{V_\rho^{-1}} \right\} = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \left\{ \lim_{\rho\to 0^+} \frac{1}{2}\|y - X\beta\|^2_{V_\rho^{-1}} \right\} \tag{28}$$

where $T := I - VV^+$, $Z := TX$, $w := Ty$, and $V_\rho := V + \rho I$ for all $\rho > 0$.

The closed-form solutions that follow from the problem reformulations in lemmas 33 and 34 are stated in theorems 36 to 38 below.

Theorem 36. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $V \in \mathbb{S}^n_+$, if $\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$ is nonempty, then*

$$\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V) = Z^+y + B(BX'V^+XB)^+BX'V^+Cy + \mathcal{N}(X) \tag{29a}$$

$$= \beta_0 + G^+Z'F^+(w - Z\beta_0) + \mathcal{N}(X) \tag{29b}$$

where $T := I - VV^+$, $w := Ty$, $Z := TX$, $B := I - Z^+Z$, $C := I - XZ^+$, $G := X'V^+X + Z'Z$, $F := ZG^+Z'$, and $\beta_0 := G^+X'V^+y$.

Theorem 37. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, let $V_0 := V + XEX'$ for some $E \in \mathbb{S}^p_+$ such that $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$. Then $\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$ is nonempty if and only if $y \in \mathcal{R}(V_0)$. Moreover, if $y \in \mathcal{R}(V_0)$, then*

$$\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V) = \hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, V_0^+) \tag{30a}$$

$$= (X'V_0^+X)^+X'V_0^+y + \mathcal{N}(X) \tag{30b}$$

Theorem 38. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $V \in \mathbb{S}^n_+$, if $\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$ is nonempty, then*

$$\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V) = \lim_{\rho\to 0^+} \hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V_\rho) \tag{31a}$$

$$= \lim_{\rho\to 0^+} (X'V_\rho^{-1}X)^+X'V_\rho^{-1}y + \mathcal{N}(X) \tag{31b}$$

$$= X^+(I - VS(SVS)^+S)y + \mathcal{N}(X) \tag{31c}$$

where $S := I - XX^+$ and $V_\rho := V + \rho I$ for all $\rho > 0$.

Proofs of theorems 36 to 38 and lemmas 33 and 34 can be found in appendix B. To our knowledge, no one has solved, let alone rigorously defined (MLE) for the model (LGM) in the case where $V \in \mathbb{S}^n_+$. However, there are connections to other literature that are present in the proofs of theorems 36 and 38. theorem 36 follows straightforwardly from the definition of the degenerate normal [90, pp. 527–528] and the ECGLS problem [67, Theorems 11.36], and theorem 38 is similar to a barrier function method originally stated by Albert [2, Chapter VII]. It appears that theorem 37 uses a novel method to write (MLE) as the saddle point system (SPP). Solutions to the saddle point system (SPP) follow straightforwardly from [67, Theorems 3.20–21].

We also provide indirect proofs of theorems 36 and 37 based on equating the expressions (29a), (30b), and (31c).

3.5 Maximum a posteriori estimator

Suppose that $\beta \sim \mathrm{N}(\beta_0, \Sigma)$ independently of the errors e, i.e., we have the Bayesian linear Gaussian model,

$$y = X\beta + e, \qquad e \sim \mathrm{N}(0, V), \qquad \beta \sim \mathrm{N}(\beta_0, \Sigma), \qquad e, \beta \text{ independent} \tag{B-LGM}$$

The distribution of β is called the *prior* distribution. We can form the *posterior* likelihood (i.e., the likelihood of $\beta|y$) by using Bayes' theorem. The MAP estimator maximizes the posterior distribution, which is defined as follows.

Definition 39. *Define* $\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$ *as the set of solutions to*

$$\max_{\beta \in \mathbb{R}^p} f(\beta|y) \quad \text{subject to} \quad f(\beta|y) > 0 \tag{MAP}$$

where $f(\cdot|y)$ *is the conditional probability density of* β *given* y.[13] *We say that* $\hat{\beta}$ *is a MAP of the model* (B-LGM) *if* $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$.

We present three methods for deriving solutions to (MAP). The first is based on rewriting the model (B-LGM) in the form of (LGM). The second is an application of the conditioning of joint (degenerate) Gaussian random variables, proven by Marsaglia [69] and restated in lemma 9. The third method is based on finding a sequence of MAP estimators $\hat{\beta}_\rho$ of the following *perturbed models*,

$$y = X\beta + e, \qquad e \sim \mathrm{N}(0, V + \rho I), \qquad \beta \sim \mathrm{N}(\beta_0, \Sigma + \rho I), \qquad e, \beta \text{ independent}$$

where $\rho > 0$, and taking the limit as $\rho \to 0^+$. The results from these methods are stated in theorems 40 to 42 below.

Theorem 40. *For any* $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $V \in \mathbb{S}^n_+$, $\beta_0 \in \mathbb{R}^p$, *and* $\Sigma \in \mathbb{S}^p_+$,

$$\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma) = \hat{\mathbb{B}}_{\mathrm{MLE}}\left(\begin{bmatrix} y \\ \beta_0 \end{bmatrix}, \begin{bmatrix} X \\ I \end{bmatrix}, \begin{bmatrix} V & 0 \\ 0 & \Sigma \end{bmatrix} \right) \tag{32}$$

Moreover, $\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$ *is nonempty if and only if* $y - X\beta_0 \in \mathcal{R}(V + X\Sigma X')$.

Theorem 41. *For any* $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $V \in \mathbb{S}^n_+$, $\beta_0 \in \mathbb{R}^p$, *and* $\Sigma \in \mathbb{S}^p_+$, *if* $\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$ *is nonempty, then*

$$\begin{aligned} \hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma) &= \hat{\mathbb{B}}_{\mathrm{MLE}}(\mathbb{E}[\beta|y], I, \mathrm{var}[\beta|y]) && \text{(33a)} \\ &= \{\, \mathbb{E}[\beta \mid y] \,\} && \text{(33b)} \\ &= \{\, \beta_0 + L(y - X\beta_0) \,\} && \text{(33c)} \end{aligned}$$

where $L := \Sigma X'(V + X\Sigma X')^+$.

[13] See definition 8 for a restatement of the probability density function of a Gaussian random variable given in [90, pp. 527–528].

Theorem 42. *Let $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, $\beta_0 \in \mathbb{R}^p$, and $\Sigma \in \mathbb{S}^p_+$. If $\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$ is nonempty, then*

$$\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma) = \{\, \beta_0 + L(y - X\beta_0) \,\} \tag{34a}$$

$$= \lim_{\rho\to 0^+} \{\, \beta_0 + L_\rho(y - X\beta_0) \,\} \tag{34b}$$

$$= \lim_{\rho\to 0^+} \hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V_\rho, \beta_0, \Sigma_\rho) \tag{34c}$$

where $L := \Sigma X'(V + X\Sigma X')^+$, $V_\rho = V + \rho I$, $\Sigma_\rho = \Sigma + \rho I$, and $L_\rho := \Sigma_\rho X'(V_\rho + X\Sigma_\rho X')^{-1}$ for each $\rho > 0$.

Proofs of theorems 40 to 42 can be found in appendix C. Theorem 40 is a consequence of Bayes' theorem. Theorem 41 follows straightforwardly by generalizing common techniques in the Bayesian and regularized regression literature. Then theorem 42 is a corollary to theorems 38 and 41, and is therefore inspired by Albert [2, Chapter VII]. Those familiar with state estimation of discrete linear dynamical systems may notice the formulae (34a) and (34b) are similar to the Kalman filter formula. Indeed, the L and L_ρ matrices defined above are analogs to the Kalman gain of problems with semidefinite and positive definite measurement noise covariance. Moreover, theorem 41 shows that the Kalman filtered state estimates are MAP estimates of the state, as we show in section 5.

It is worth pointing out that the equations (32) do *not* hold when $y - X\beta_0 \notin \mathcal{R}(V + X\Sigma X')$, because $\hat{\mathbb{B}}_{\mathrm{MLE}}(\overline{y}, \overline{X}, \overline{V}) = \{\, \beta_0 + L(y - X\beta_0) \,\}$ is nonempty despite the fact that $\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$ is empty. Care should be taken to use the formula (32) only when the solution actually exists. Theorem 42 does *not* imply that $\lim_{\rho\to 0^+} L_\rho = L$. Instead, we have the weaker result $\lim_{\rho\to 0^+} L_\rho e_0 = Le_0$ for all $e_0 := y - X\beta_0 \in \mathcal{R}(V + X\Sigma X')$. For example, consider the model parameters

$$V = \Sigma = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \qquad X = I$$

For the above parameters, L is computed as

$$L = \Sigma X'(V + X\Sigma X')^+ = \begin{bmatrix} 0 & 0 \\ 0 & 1/2 \end{bmatrix}$$

and L_ρ is computed as

$$L_\rho = (\Sigma + \rho I)X'(V + \rho I + X(\Sigma + \rho I)X')^{-1} = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

for all $\rho > 0$. While $\lim_{\rho\to 0^+} L_\rho$ does exist, it is not equal to L,

$$\lim_{\rho\to 0^+} L_\rho = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix} \neq L = \begin{bmatrix} 0 & 0 \\ 0 & 1/2 \end{bmatrix}$$

However, we can take any

$$e_0 = \begin{bmatrix} 0 \\ \alpha_0 \end{bmatrix} \in \mathcal{R}(V + X\Sigma X') = \mathcal{R}\left(\begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}\right) = \left\{ \begin{bmatrix} 0 \\ \alpha \end{bmatrix} \,\middle|\, \alpha \in \mathbb{R} \right\}$$

to give $L_\rho e_0 = \alpha_0/2$, $Le_0 = \alpha_0/2$, and therefore $\lim_{\rho\to 0^+} L_\rho e_0 = Le_0$.

3.6 Best affine unbiased estimator

To find the MVUE, one must find a *function* of the observations y that is unbiased and has minimum variance among the set of unbiased functions of y. Optimizing over the space of functions is a complex problem compared to optimizing over a parameterized space of functions, so much of the study on the model (LGM) has been concerned with finding an estimator with minimum variance among *linear* unbiased estimators (i.e., a BLUE). It turns out the BAUE can capture a wider class of constrained estimators [67, Chapter 13], so we address the BAUE problem. Below, we define (uniformly) unbiased estimators, affine estimators, and the BAUE.

Definition 43. *Let $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, and $W \in \mathbb{R}^{m\times n}$, and consider the model* (LGM)*. Denote the set of (uniformly) unbiased estimators of the parametric function $W\beta$ (given the model* (LGM)*) as*[14]

$$\widehat{\mathbb{WB}}_{\mathrm{UE}}(X,V,W) := \left\{ \theta : \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) \to \mathbb{R}^m \;\middle|\; \begin{array}{c} \mathbb{E}[\theta(y)|\beta] = W\beta \text{ where (LGM)}, \\ \forall \beta \in \mathbb{R}^p \end{array} \right\}$$

Denote the set of affine estimators as

$$\widehat{\mathbb{WB}}_{\mathrm{AE}} := \{\, \theta(\cdot) = A(\cdot) + c : \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) \to \mathbb{R}^m \mid A \in \mathbb{R}^{m\times p}, c \in \mathbb{R}^m \,\}$$

Denote the set of (uniformly) unbiased affine estimators of the parametric function $W\beta$ (given the model (LGM)*) as*

$$\widehat{\mathbb{WB}}_{\mathrm{AUE}}(X,V,W) := \widehat{\mathbb{WB}}_{\mathrm{AE}} \cap \widehat{\mathbb{WB}}_{\mathrm{UE}}(X,V,W)$$

Define the estimators that minimize variance over the set of (uniformly) unbiased affine estimators as[15]

$$\widehat{\mathbb{WB}}_{\mathrm{BAUE}}(X,V,W) := \left\{ \theta \in \widehat{\mathbb{WB}}_{\mathrm{AUE}}(X,V,W) \;\middle|\; \begin{array}{c} \mathrm{var}[\theta(y)|\beta] \preceq \mathrm{var}[\tilde{\theta}(y)|\beta] \\ \text{where (LGM)}, \\ \forall \tilde{\theta} \in \widehat{\mathbb{WB}}_{\mathrm{AUE}}(X,V,W), \\ \forall \beta \in \mathbb{R}^p \end{array} \right\} \qquad \text{(BAUE)}$$

We say $\widehat{W\beta}$ is a BAUE (of the parametric function $W\beta$ given the model (LGM)*) if $\widehat{W\beta} \in \widehat{\mathbb{WB}}_{\mathrm{BAUE}}(X,V,W)$.*

In solving (BAUE) we solve, as an intermediate step, the linear equality constrained minimum trace problem (MTP), which, by the connection between the BAUE and the MLE, is closely related to the least squares problem. This technique is used extensively by Magnus and Neudecker [67, Chapter 13]. We use it to derive Rao's result [88] under an optimization framework via the following lemma.

[14] We restrict the domain of all estimators to $\mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$. This is because $e \in \mathcal{R}(V)$ (almost surely) implies $y \in \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$ (almost surely).

[15] While (BAUE) is not defined in the language of optimization, one can view (BAUE) as equivalent to minimizing a positive semidefinite matrix-valued function over the Loewner partial order $\preceq$.

Lemma 44. *Let $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, $W \in \mathbb{R}^{m\times n}$, and $\hat{\mathbb{A}}(X,V,W)$ be the set of solutions to* (MTP)*. If $\widehat{\mathbb{WB}}_{\text{BAUE}}(X,V,W)$ is nonempty and*

$$\hat{A}_1 y = \hat{A}_2 y \qquad \forall \hat{A}_1, \hat{A}_2 \in \hat{\mathbb{A}}(X,V,W), y \in \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) \tag{35}$$

then, for any $\hat{A} \in \hat{\mathbb{A}}(X,V,W)$,

$$\widehat{\mathbb{WB}}_{\text{BAUE}}(X,V,W) = \{\, \hat{A}(\cdot) : \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) \to \mathbb{R}^m \,\} \tag{36}$$

Using lemma 44, we can solve (MTP) to show that (BAUE) is a singleton, where the unique solution is that proposed by Rao [88]. This result is stated in the following theorem.

Theorem 45. *Let $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, $W \in \mathbb{R}^{m\times n}$, and $V_0 := V + XEX'$ for any $E \in \mathbb{S}^p_+$ such that $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$. Then $\widehat{\mathbb{WB}}_{\text{BAUE}}(X,V,W)$ is nonempty if and only if $\mathcal{R}(W') \subseteq \mathcal{R}(X')$. Moreover, if $\widehat{\mathbb{WB}}_{\text{BAUE}}(X,V,W)$ is nonempty, then*

$$\widehat{\mathbb{WB}}_{\text{BAUE}}(X,V,W) = \{\, W(X'V_0^+X)^+X'V_0^+(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m \,\} \tag{37}$$

It is also clear that if a BAUE exists, it is unique and linear, so it is also the *unique* best *linear* unbiased estimator (BLUE). Moreover, there is a clear connection between (30b) and (37), which is stated in the following corollary.

Corollary 46. *For any $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, and $W \in \mathbb{R}^{m\times n}$, if $\widehat{W\beta} \in \widehat{\mathbb{WB}}_{\text{BAUE}}(X,V,W)$, then*

$$\{\, \widehat{W\beta}(y) \,\} = W\hat{\mathbb{B}}_{\text{MLE}}(y,X,V) \qquad \forall y \in \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$$

Proofs of lemma 44, theorem 45, , and corollary 46 can be found in appendix D.

As previously stated, the results also apply to non-Gaussian error distributions with $\mathbb{E}[e] = 0$ and $\text{var}[e] = V$ because the derivation of the BAUE depends only on the first and second order statistics of the underlying distribution. In appendix D, we prove the following theorem and thus derive the unique BAUE solution proposed by Rao [88].

4 Computing the estimators

Computing closed-form solution of the estimators defined in section 3 is intractable for large and ill-conditioned systems, as computing the relevant pseudoinverses requires taking singular value decompositions or solving many linear system subproblems, which compounds errors and amplifies conditioning issues. However, as illustrated in figures 1 to 4 and stated in the lemmas and theorems of section 3, all of our problems can be reformulated as either a convex optimization problem (specifically, minimizing a convex quadratic objective subject to linear equality constraints), or a linear system, each without requiring the computation of any pseudoinverses to pose the problem. Convex optimization problems and linear systems are well-studied problems that can be solved by a wide variety of high-quality, publicly available numerical algorithms. Moreover, gradient descent forms the basis for many of these numerical algorithms, and it has been shown that early-stopping

of these algorithms implicitly regularizes the solution [4]. In this section, we discuss the most numerically attractive reformulations of the estimation problems defined in section 3 and show how gradient methods with early stopping can be related to (TGLS), where the initial guess becomes the bias parameter.

4.1 Convex optimization formulations

Each estimation problem can be written as a convex optimization problem. The estimation problems (GLS), (TGLS), and (ECGLS) are convex by definition. While lemmas 33 and 35 reformulate (MLE) as a convex optimization problem, they do *not* do so without requiring the computation of a (pseudo)inverse to pose the problem.[16] Instead, one can solve the following problem,

$$\min_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^p} \frac{1}{2}\|\alpha\|_V^2 \qquad \text{subject to} \qquad y = X\beta + V\alpha \tag{38}$$

which is shown to be equivalent to (MLE) in the proof of lemma 34. Equation (38) has an increased number of optimization variables compared to (MLE) but avoids the need for pseudoinverse or projector matrix computations. A similar strategy can be taken for (MAP),

$$\min_{(\alpha_1,\alpha_2)\in\mathbb{R}^{n+p},\beta\in\mathbb{R}^p} \frac{1}{2}\|\alpha_1\|_V^2 + \frac{1}{2}\|\alpha_2\|_\Sigma^2 \qquad \text{subject to} \qquad \begin{array}{c} y = X\beta + V\alpha_1, \\ \beta_0 = \beta + V\alpha_2 \end{array} \tag{39}$$

using theorem 40. The other reformulations in theorems 41 and 42 still require pseudoinverse calculations, making them poor candidates for optimization. Finally, as shown in lemma 44 and theorem 45, the problem (BAUE) can be solved by setting $\widehat{W\beta} = \hat{A}y$ where $\hat{A}$ solves (MTP).

While each of these convex optimization problems can be solved as a linear system using lemma 10, it is worth stating the original optimization problem because it allows us to put convex constraints on the argument and retain numerical tractability.

4.2 Linear system formations

Because the convex optimization problems discussed in section 4.1 have quadratic objectives and only linear equality constraints, we can use lemma 10 to rewrite them as linear systems. In the statistics literature, these linear systems are often called the *normal equations* for the problem. From the proofs of theorems 25, 28, and 31, the normal equations for the problems (GLS), (TGLS), and (ECGLS) are

$$X'HX\beta = X'Hy \tag{40a}$$

$$(X'HX + \Gamma)\beta = X'Hy + \Gamma\beta_0 \tag{40b}$$

$$\begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} = \begin{bmatrix} X'Hy \\ w \end{bmatrix} \tag{40c}$$

[16] The inverse computation in lemma 35 also becomes ill-conditioned as $\rho \to 0^+$, so to get an accurate solution, one must necessarily make the problem ill-conditioned.

where $\lambda \in \mathbb{R}^c$ is a Lagrange multiplier corresponding to the constraint $w = Z\beta$. It is clear that, in (40), the structure of the coefficient matrix can be exploited. For (40a) and (40b), the coefficient matrices are positive semidefinite. Solution methods for solving positive semidefinite linear systems are well studied [36, 57, 75, 76]. For (40c), the coefficient matrix corresponds to that of a *saddle point system*, which is well-studied [11, 12, 28, 31, 32, 43, 62, 111, 116]. We already showed in lemma 34 that (38) is equivalent to (SPP). Likewise, it can be shown that (39) is equivalent to

$$\begin{bmatrix} V & 0 & X \\ 0 & \Sigma & I \\ X' & I & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} = \begin{bmatrix} y \\ \beta_0 \\ 0 \end{bmatrix} \tag{41}$$

and (MTP) is equivalent to,

$$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} A' \\ \Lambda' \end{bmatrix} = \begin{bmatrix} 0 \\ W' \end{bmatrix} \tag{42}$$

As with (40c), the fact that (41) and (42) are saddle point systems may be exploited.

4.3 Gradient methods and early stopping

Gradient descent is the bedrock on which many modern optimization algorithms and linear system solvers are built. All of the optimization problems described herein can be posed as solving a linear system. In this section, we consider applying gradient descent and gradient flow to the OLS problem (i.e., computing elements of $\hat{\mathbb{B}}_{\text{GLS}}(y, X, I)$). Specifically, we consider the following two algorithms,

$$\beta^{(k)} = \beta^{(k-1)} + \varepsilon X'(y - X\beta^{(k-1)}), \qquad \beta^{(0)} = \beta_0 \tag{43}$$

$$\dot{\beta}(t) = \varepsilon X'(y - X\beta(t)), \qquad \beta(0) = \beta_0 \tag{44}$$

where, for each algorithm, $\beta_0 \in \mathbb{R}^p$ is the initial guess and $\varepsilon > 0$ is a tunable rate parameter. It is worth pointing out that the algorithms (43) and (44) can be extended to handle the GLS problem by using the equivalence $\hat{\mathbb{B}}_{\text{GLS}}(y, X, V) = \hat{\mathbb{B}}_{\text{GLS}}(H^{1/2}y, H^{1/2}X, I)$. Moreover, they can solve (feasible) linear systems by noting that, for any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $H \in \mathbb{S}^p_+$, $y \in \mathcal{R}(X)$ implies $\{\, \beta \mid y = X\beta \,\} = \hat{\mathbb{B}}_{\text{GLS}}(y, X, I)$. Therefore, we can solve (ECGLS), (MLE), (MAP), and (BAUE) using any of the normal equation formulations in figures 1c and 2 to 4.

Ali et al. [4] showed that the algorithms (43) and (44), with early stopping and a zero initial guess $\beta_0 = 0$, implicitly regularize the solution. Theorems 47 and 48 generalize these results to the early stopping and convergence behavior of the general gradient algorithms (43) and (44). The early stopping algorithm implicitly regularizes the solution based on an equivalent TGLS problem with a positive definite regularization parameter and bias parameter equal to the initial guess β_0. Moreover, the part of the regularization penalty with columns in the range space of X drops out as the algorithms converge, while the remaining part of the regularization penalty is always a projector into the nullspace of X. As a result, both algorithms converge to the element of $\hat{\mathbb{B}}_{\text{GLS}}(y, X, I)$ that is closest in 2-norm to the initial guess β_0.

Theorem 47 (Generalized from [4])**.** *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $\beta_0 \in \mathbb{R}^p$, and $\varepsilon \in (0, 1/\|X\|^2)$, let*

$$X = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} S_1 & \\ & 0 \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix} = U_1 S_1 V_1' \tag{45}$$

denote the SVD of X. Then the gradient descent iterates $\beta^{(k)} \in \mathbb{R}^p$, given by (43)*, satisfy*

$$\hat{\mathbb{B}}_{\text{TGLS}}(y, X, I, \beta_0, \Gamma^{(k)}) = \{\, \beta^{(k)} \,\}, \qquad \forall k > 0 \tag{46}$$

where $\Gamma^{(k)} := V_1[(I - \varepsilon S_1^2)^{-k} - I]^{-1} S_1^2 V_1' + V_2 V_2'$. Moreover,

$$\lim_{k\to\infty} \beta^{(k)} = X^+ y + V_2 V_2' \beta_0 \tag{47}$$

Theorem 48 (Generalized from [4])**.** *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, and $\beta_0 \in \mathbb{R}^p$, let* (45) *denote the SVD of X. Then the gradient flow trajectory $\beta(\cdot) : \mathbb{R}_{\geq 0} \to \mathbb{R}^p$, given by* (44)*, satisfies*

$$\hat{\mathbb{B}}_{\text{TGLS}}(y, X, I, \beta_0, \Gamma(t)) = \{\, \beta(t) \,\}, \qquad \forall t > 0 \tag{48}$$

where $\Gamma(t) := V_1[\exp(\varepsilon S_1^2 t) - I]^{-1} S_1^2 V_1' + V_2 V_2'$. Moreover,

$$\lim_{t\to\infty} \beta(t) = X^+ y + V_2 V_2' \beta_0 \tag{49}$$

Theorem 47 is proven below. The proof of theorem 48 is nearly identical, but with matrix powers and power series replaced by matrix exponentials and integrals, so the proof is omitted.

Proof of theorem 47. Let $k > 0$. We can rewrite the gradient descent iterate $\beta^{(k)}$ in terms of the initial guess β_0 using a common linear systems formula,

$$\beta^{(k)} = (I - \varepsilon X'X)^k \beta_0 + \varepsilon \sum_{j=0}^{k-1} (I - \varepsilon X'X)^j X'y \tag{50}$$

Next, we rewrite the formula (50) in terms of the SVD (45),

$$\begin{aligned}
\beta^{(k)} &= [V_1(I - \varepsilon S_1^2)^k V_1' + V_2 V_2']\beta_0 + \varepsilon V_1 \left[\sum_{j=0}^{k-1} (I - \varepsilon S_1^2)^j \right] S_1 U_1' y \\
&= [V_1(I - \varepsilon S_1^2)^k V_1' + V_2 V_2']\beta_0 + V_1[I - (I - \varepsilon S_1^2)^k] S_1^{-1} U_1' y \\
\beta^{(k)} &= \beta_0 + V_1[I - (I - \varepsilon S_1^2)^k] S_1^{-1} U_1'(y - X\beta_0) \qquad (51)
\end{aligned}$$

where the second equality follows from the partial geometric series formula $\sum_{j=0}^{k-1} r^j = \frac{1-r^k}{1-r}$ for all scalars $r \neq 1$.[17] Using theorem 28, we have

$$\begin{aligned}
\hat{\mathbb{B}}_{\text{TGLS}}(y, X, I, \beta_0, \Gamma^{(k)}) &= \{\, \beta_0 + (X'X + \Gamma^{(k)})^{-1} X'(y - X\beta_0) \,\} \\
&= \{\, \beta_0 + V_1\{I + [(I - \varepsilon S_1^2)^{-k} - I]^{-1}\}^{-1} S_1^{-1} U_1'(y - X\beta_0) \,\} \\
&= \{\, \beta_0 + V_1[I - (I - \varepsilon S_1^2)^k] S_1^{-1} U_1'(y - X\beta_0) \,\} = \{\, \beta^{(k)} \,\}
\end{aligned}$$

[17] We can use the scalar formula here because the matrices in the geometric series are diagonal.

where the third and fourth equalities follows from theorem 60 and (51), respectively. Since $\|\varepsilon X'X\| = \varepsilon\|X\|^2 < 1$, all the diagonal entries of $I - \varepsilon S_1^2$ are positive and less than 1. Therefore $\lim_{k\to\infty}(I - \varepsilon S_1^2)^k = 0$ and

$$\lim_{k\to\infty} \beta^{(k)} = \beta_0 + V_1 S_1^{-1} U_1'(y - X\beta_0) = X^+ y + V_2 V_2' \beta_0$$

where the first equality follows by taking the limit of (51) and the second equality follows by the SVD (45). □

Proof of theorem 48. Let $t > 0$. First, we can rewrite the gradient flow at time t in terms of the initial guess β_0 using a common linear systems formula,

$$\beta(t) = \exp(-\varepsilon X'Xt)\beta_0 + \int_0^t \exp(-\varepsilon X'X(t-\tau))X'y d\tau \tag{52}$$

Next, we rewrite the formula (52) in terms of the SVD (45),

$$\begin{aligned}
\beta(t) &= V_1 \exp(-\varepsilon S_1^2 t) V_1' \beta_0 + V_1 \left[\int_0^t \exp(-\varepsilon S_1^2 (t-\tau)) d\tau\right] S_1 U_1' y \\
&= V_1 \exp(-\varepsilon S_1^2 t) V_1' \beta_0 + V_1 [I - \exp(-\varepsilon S_1^2 t)] S_1^{-1} U_1' y \\
&= \beta_0 + V_1 [I - \exp(-\varepsilon S_1^2 t)] S_1^{-1} U_1'(y - X\beta_0) \qquad (53)
\end{aligned}$$

where the second equality follows from the integral formula $\int_0^t \exp(-a(t-\tau))d\tau = \frac{1-\exp(-at)}{a}$ for all scalars $a > 0$.[18] Using theorem 28, we have

$$\begin{aligned}
\hat{\mathbb{B}}_{\mathrm{TGLS}}(y, X, I, \beta_0, \Gamma(t)) &= \{\, \beta_0 + (X'X + \Gamma(t))^{-1} X'(y - X\beta_0) \,\} \\
&= \{\, \beta_0 + V_1 \{I + [\exp(\varepsilon S_1^2 t) - I]^{-1}\}^{-1} S_1^{-1} U_1'(y - X\beta_0) \,\} \\
&= \{\, \beta_0 + V_1 [I - \exp(-\varepsilon S_1^2 t)] S_1^{-1} U_1'(y - X\beta_0) \,\} = \{\, \beta(t) \,\}
\end{aligned}$$

where the third and fourth equalities follows from theorem 60 and (53), respectively. Since S_1 has positive diagonal entries, we have $\lim_{t\to\infty} \exp(-\varepsilon S_1^2 t) = 0$ and

$$\lim_{t\to\infty} \beta(t) = \beta_0 + V_1 S_1^{-1} U_1'(y - X\beta_0) = X^+ y + V_2 V_2' \beta_0$$

where the first equality follows by taking the limit of (53) and the second equality follows by the SVD (45). □

5 Applications in control and estimation

Consider the following linear time-varying system,

$$x_{k+1} = A_k x_k + B_k u_k \tag{54a}$$

$$y_k = C_k x_k + D_k u_k \tag{54b}$$

[18] We can use the scalar formula here because the matrices in the geometric series are diagonal.

where $k \in \mathbb{I}_{\geq 0}$ is the time index, $x_k \in \mathbb{R}^{n_x}$ is the state, $y_k \in \mathbb{R}^{n_y}$ is the measurement, and $u_k \in \mathbb{R}^{n_u}$ is the exogenous input. We assume that rank $\begin{bmatrix} B_k' & D_k' \end{bmatrix} = n_u$ for each $k \in \mathbb{I}_{\geq 0}$. If this were not the case (i.e., rank $\begin{bmatrix} B_k' & D_k' \end{bmatrix} < n_u$ for some $k \in \mathbb{I}_{\geq 0}$), then there would be inputs that affect neither the state evolution nor the stage cost.

In the LQR problem, we seek an input sequence $\mathbf{u}_{N-1} := (u_0, u_1, \ldots, u_{N-1})$ that minimizes the quadratic objective

$$V_N(x; \mathbf{u}_{N-1}) = \sum_{j=0}^{N-1} \|y_j\|^2 + \|x_N\|_{P_N}^2$$

subject to (54) and $x_0 = x$, for some user-defined terminal penalty $P_N \in \mathbb{S}_+^{n_x}$. We solve this problem via a dynamic programming approach, with the current optimal input being a linear function of the current state. Each stage in the dynamic program is solved by completing the squares via corollary 26 or corollary 29. An alternative derivation can be found in [33, 34], which also involves completing the squares, but relies on index shifting to rewrite $V_N(x; \mathbf{u}_{N-1})$ in a form where squares can be completed for every $j = 0, 1, \ldots, N-1$ simultaneously.

In the KF problem, we seek the probability densities of the random variables $x_k|\mathbf{y}_k$ and $x_k|\mathbf{y}_{k-1}$ under the assumption that

$$x_0 \sim \mathrm{N}(\hat{x}_0^-, \hat{P}_0^-), \qquad u_k \sim \mathrm{N}(0, I), \qquad \text{independently} \tag{55}$$

where $\mathbf{y}_k := (y_0, y_1, \ldots, y_k)$, for each $k \in \mathbb{I}_{\geq 0}$. With a slight abuse of notation, we let $\mathbf{y}_{-1}$ denote an empty vector. Conditioning a random variable on $\mathbf{y}_{-1}$ returns the same random variable. This notation lets us include the initial distribution (55) in our KF definition and use the notation $x_0 = x_0|\mathbf{y}_{-1}$. Since $x_k|\mathbf{y}_k$ has a Gaussian distribution, it suffices to find a recursion for the mean and covariance. To do this, we use Albert's method [2, Chapter IX], which is a natural extension of Marsaglia's lemma (lemma 9) [69].

Stability of the LQR and KF for the system (54) is an important property for successful applications. Stability results, under the assumption of positive definiteness in either $D_k' D_k$ (for the LQR) or $D_k D_k'$ (for the KF), have appeared in many textbooks [6, 7, 50, 60]. In the general (semidefinite) case, Silverman [98] provides necessary and sufficient conditions for which the time-invariant LQR and KF are stable, and Rappaport and Silverman [93] provide sufficient conditions for the time-varying LQR and KF.[19] Further discussion of stability is outside of the scope of this work. For a modern treatment of the semidefinite case, including solution methods, algorithms, and stability theory, we refer the reader to the work of Ferrante and Ntogramatzidis [33, 34, 35].

5.1 Linear quadratic regulator

For the regulation problem, it is convenient to also define the block matrix

$$\begin{bmatrix} Q_k & S_k \\ S_k' & R_k \end{bmatrix} = \begin{bmatrix} C_k' \\ D_k' \end{bmatrix} \begin{bmatrix} C_k & D_k \end{bmatrix} \in \mathbb{S}_+^{n_x + n_y}$$

[19] These works require stability to hold independently of the choice of $P_N \in \mathbb{S}_+^{n_x}$ (for the LQR) or $\hat{P}_0^- \in \mathbb{S}_+^{n_x}$ (for the KF). For stability conditions that depend on $P_N \in \mathbb{S}_+^{n_x}$ or $\hat{P}_0^- \in \mathbb{S}_+^{n_x}$, see [20, 25, 26].

With this definition, it is easy to see that $\|y_k\|^2 = \|x_k\|_{Q_k}^2 + \|u_k\|_{R_k}^2 + 2x_k' S_k u_k$ for each $k \in \mathbb{I}_{\geq 0}$, which is the traditional form of the stage cost.

Theorem 49. *Consider the system* (54) *and suppose* $S_k = 0$ *for each* $k = 0, 1, \ldots, N-1$. *For each initial state* $x \in \mathbb{R}^{n_x}$ *and terminal weight* $P_N \in \mathbb{S}_+^{n_x}$, *the input sequence* $\mathbf{u}_{N-1}^0 = (u_0^0, u_1^0, \ldots, u_{N-1}^0)$ *solves the following LQR problem,*

$$V_N^0(x) := \min_{\mathbf{u}_{N-1} \in \mathbb{R}^{n_u N}} \sum_{j=0}^{N-1} \|y_j\|^2 + \|x_N\|_{P_N}^2 \tag{56a}$$

$$\text{subject to (54) and } x_0 = x \tag{56b}$$

if and only if

$$u_k^0 \in -(B_k' P_{k+1} B_k + R_k)^+ B_k' P_{k+1} A_k x_k + \mathcal{N}(B_k' P_{k+1} B_k + R_k) \tag{57a}$$

$$P_k := A_k' P_{k+1} A_k + Q_k - A_k' P_{k+1} B_k (B_k' P_{k+1} B_k + R_k)^+ B_k' P_{k+1} A_k \tag{57b}$$

for each $k = 0, 1, \ldots, N-1$. *Moreover,* $V_N^0(x) = \|x\|_{P_0}^2$.

Proof. To shorten the notation, let

$$\hat{u}_k := -\mathcal{R}_k^+ B_k' P_{k+1} A_k x_k, \qquad \mathcal{R}_k := B_k' P_{k+1} B_k + R_k$$

for each $k = 0, 1, \ldots, N-1$, and let

$$V_k(x; \mathbf{u}_{k-1}) := \sum_{j=0}^{k-1} \|y_j\|^2 + \|x_k\|_{P_k}^2$$

for each $k = 0, 1, \ldots, N$. Since $V_N(x; \mathbf{u}_{N-1})$ is the objective function for (56), it suffices to minimize it subject to (54) and $x_0 = x$.

Using corollary 29, we have

$$\|x_k\|_{Q_k}^2 + \|u_k\|_{R_k}^2 + \|A_k x_k + B_k u_k\|_{P_{k+1}}^2 = \|x_k\|_{P_k}^2 + \|u_k - \hat{u}_k\|_{\mathcal{R}_k}^2$$

and

$$V_{k+1}(x; \mathbf{u}_k) = V_k(x; \mathbf{u}_{k-1}) + \|u_k - \hat{u}_k\|_{\mathcal{R}_k}^2 \tag{58}$$

for each $k = 0, 1, \ldots, N-1$. Applying (58) recursively gives

$$V_N(x; \mathbf{u}_{N-1}) = \|x\|_{P_0}^2 + \sum_{j=1}^{N-1} \|u_j - \hat{u}_j\|_{\mathcal{R}_j}^2$$

where we use the terminal identity $V_0(x) = \|x\|_{P_0}^2$. In this form, it is clear from theorem 25 that $\mathbf{u}_{N-1}^0$ minimizes $V_N(x; \mathbf{u}_{N-1})$ if and only if

$$u_k^0 \in \hat{u}_k + \mathcal{N}(\mathcal{R}_k) = -\mathcal{R}_k^+ B_k' P_{k+1} A_k x_k + \mathcal{N}(\mathcal{R}_k)$$

for each $k = 0, 1, \ldots, N-1$. Moreover, $V_N^0(x) = V_N(x; \mathbf{u}_{N-1}^0) = \|x\|_{P_0}^2$. □

Theorem 50. *Consider the system* (54)*. For each initial state* $x \in \mathbb{R}^{n_x}$ *and terminal weight* $P_N \in \mathbb{S}_+^{n_x}$*, the input sequence* $\mathbf{u}_{N-1}^0 := (u_0^0, u_1^0, \ldots, u_{N-1}^0)$ *solves the following LQR problem,*

$$V_N^0(x) = \min_{\mathbf{u}_{N-1} \in \mathbb{R}^{n_x N}} \sum_{j=0}^{N-1} \|y_j\|^2 + \|x_N\|_{P_N}^2 \tag{59a}$$

$$\text{subject to (54) and } x_0 = x \tag{59b}$$

if and only if

$$u_k^0 \in -(B_k' P_{k+1} B_k + R_k)^+ (B_k' P_{k+1} A_k + S_k') x_k + \mathcal{N}(\mathcal{R}_k) \tag{60a}$$

$$\begin{aligned} P_k := {} & A_k' P_{k+1} A_k + Q_k \\ & - (A_k' P_{k+1} B_k + S_k)(B_k' P_{k+1} B_k + R_k)^+ (B_k' P_{k+1} A_k + S_k') \end{aligned} \tag{60b}$$

for each $k = 0, 1, \ldots, N-1$*. Moreover,* $V_N^0(x) = \|x\|_{P_0}^2$*.*

Proof. For this proof, we use theorem 49 as an intermediate step to solve the more general case. To shorten the notation, let

$$\begin{aligned} \mathcal{R}_k &:= B_k' P_{k+1} B_k + R_k, & \mathcal{S}_k &:= A_k' P_{k+1} B_k + S_k, \\ \hat{u}_k^- &:= -\mathcal{R}_k^+ B_k' P_{k+1} A_k x_k, & \hat{u}_k &:= \hat{u}_k^- - \mathcal{R}_k^+ S_k' x_k = -\mathcal{R}_k^+ \mathcal{S}_k' x_k, \end{aligned}$$

$$P_k^- := A_k' P_{k+1} A_k + Q_k - A_k' P_{k+1} B_k \mathcal{R}_k^+ B_k' P_{k+1} A_k$$

for each $k = 0, 1, \ldots, N-1$, and let

$$V_k(x; \mathbf{u}_{k-1}) := \sum_{j=0}^{k-1} \|y_j\|^2 + \|x_k\|_{P_k}^2$$

for each $k = 0, 1, \ldots, N$. Since $V_N(x; \mathbf{u}_{N-1})$ is the objective function for (59), it suffices to minimize it subject to (54) and $x_0 = x$.

In a similar manner to the proof of theorem 49, we use corollary 29 to give

$$\|x_k\|_{Q_k}^2 + \|u_k\|_{R_k}^2 + \|A_k x_k + B_k u_k\|_{P_{k+1}}^2 = \|x_k\|_{P_k^-}^2 + \|u_k - \hat{u}_k^-\|_{\mathcal{R}_k}^2 \tag{61}$$

Next, since $S_k = C_k' D_k$, $R_k = D_k' D_k$, and $\mathcal{R}_k = B_k' P_{k+1} B_k + R_k$, we have

$$\mathcal{R}(S_k') \subseteq \mathcal{R}(D_k') = \mathcal{R}(R_k) \subseteq \mathcal{R}(\mathcal{R}_k) \tag{62}$$

using basic properties of $\mathcal{R}(\cdot)$. Expanding the square $\|u_k - \hat{u}_k\|_{\mathcal{R}_k}^2 = \|(u_k - \hat{u}_k^-) + \mathcal{R}_k^+ S_k' x_k\|_{\mathcal{R}_k}^2$ gives

$$\begin{aligned} \|u_k - \hat{u}_k\|_{\mathcal{R}_k}^2 &= \|u_k - \hat{u}_k^-\|_{\mathcal{R}_k}^2 + \|\mathcal{R}_k^+ S_k' x_k\|_{\mathcal{R}_k}^2 + 2x_k' S_k \mathcal{R}_k^+ \mathcal{R}_k (u_k - \hat{u}_k^-) \\ &= \|u_k - \hat{u}_k^-\|_{\mathcal{R}_k}^2 + \|x_k\|_{S_k \mathcal{R}_k^+ S_k'}^2 + 2x_k' S_k u_k \\ &\quad + 2x_k' S_k \mathcal{R}_k^+ A_k' P_{k+1} B_k x_k \\ \|u_k - \hat{u}_k\|_{\mathcal{R}_k}^2 &= \|u_k - \hat{u}_k^-\|_{\mathcal{R}_k}^2 + 2x_k' S_k u_k + \|x_k\|_{P_k^-}^2 - \|x_k\|_{P_k}^2 \end{aligned} \tag{63}$$

where the second equality follows from lemma 12 and (62). Combining (61) and (63) gives

$$\|x_k\|_{Q_k}^2 + \|u_k\|_{R_k}^2 + 2x_k' S_k u_k + \|A_k x_k + B_k u_k\|_{P_{k+1}}^2 = \|x_k\|_{P_k}^2 + \|u_k - \hat{u}_k\|_{\mathcal{R}_k}^2$$

and therefore

$$V_{k+1}(x; \mathbf{u}_k) = V_k(x; \mathbf{u}_{k-1}) + \|u_k - \hat{u}_k\|_{\mathcal{R}_k}^2 \tag{64}$$

for each $k = 0, 1, \ldots, N-1$. Given the recursion (64), the result follows (almost) identically to the proof of theorem 49, starting from (58). □

5.2 Kalman filter

Similarly to the regulation problem, we define the following block matrix (with a slight abuse of notation),

$$\begin{bmatrix} Q_k & S_k \\ S_k' & R_k \end{bmatrix} := \begin{bmatrix} B_k \\ D_k \end{bmatrix} \begin{bmatrix} B_k' & D_k' \end{bmatrix} \in \mathbb{S}_+^{n+p}$$

With this definition, it is clear that

$$\begin{bmatrix} B_k \\ D_k \end{bmatrix} u_k \sim \mathrm{N}\left(0, \begin{bmatrix} Q_k & S_k \\ S_k' & R_k \end{bmatrix}\right)$$

independently, for each $k \in \mathbb{I}_{\geq 0}$. We can obtain a recursion on $\mathbb{E}[x_k|\mathbf{y}_{k-1}]$ and $\mathrm{var}[x_k|\mathbf{y}_{k-1}]$ through Marsaglia's lemma [69], in a similar manner to Albert [2, Chapter IX].

Theorem 51 ([2, Chapter IX])**.** *Consider the system* (54) *and assume* (55)*. Then,*

$$x_k|\mathbf{y}_{k-1} \sim \mathrm{N}(\hat{x}_k^-, \hat{P}_k^-) \tag{65}$$

where

$$\begin{aligned} \hat{x}_{k+1}^- &:= A_k \hat{x}_k^- + (A_k \hat{P}_k^- C_k' + S_k)(C_k P_k^- C_k' + R_k)^+ (y_k - C_k \hat{x}_k^-) \\ \hat{P}_{k+1}^- &:= A_k \hat{P}_k^- A_k + Q_k - (A_k \hat{P}_k^- C_k' + S_k)(C_k P_k^- C_k' + R_k)^+ (C_k \hat{P}_k^- A_k' + S_k') \end{aligned}$$

for each $k \in \mathbb{I}_{\geq 0}$. Moreover,

$$x_k|\mathbf{y}_k \sim \mathrm{N}(\hat{x}_k, \hat{P}_k) \tag{66}$$

where

$$\begin{aligned} \hat{x}_k &:= \hat{x}_k^- + \hat{P}_k^- C_k' (C_k \hat{P}_k^- C_k' + R_k)^+ (y_k - C_k \hat{x}_k^-) \\ \hat{P}_k &:= \hat{P}_k^- - \hat{P}_k^- C_k' (C_k P_k^- C_k' + R_k)^+ C_k \hat{P}_k^- \end{aligned}$$

for each $k \in \mathbb{I}_{\geq 0}$.

Proof. We prove (65) by induction, and (66) is established as an intermediate step towards proving (65). The base case $x_0|\mathbf{y}_{-1} \sim \mathrm{N}(\hat{x}_0^-, \hat{P}_0^-)$ is satisfied by assumption (recall that $x_0 = x_0|\mathbf{y}_{-1}$). Therefore we only need to show $x_k|\mathbf{y}_{k-1} \sim \mathrm{N}(\hat{x}_k^-, \hat{P}_k^-)$ implies $x_k|\mathbf{y}_k \sim \mathrm{N}(\hat{x}_k, \hat{P}_k)$ and $x_{k+1}|\mathbf{y}_k \sim \mathrm{N}(\hat{x}_{k+1}^-, \hat{P}_{k+1}^-)$.

Suppose $x_k|\mathbf{y}_{k-1} \sim \mathrm{N}(\hat{x}_k^-, \hat{P}_k^-)$. To simplify the notation we let

$$\begin{aligned} e_k &:= y_k - C_k\hat{x}_k^-, & L_k^f &:= \hat{P}_k^- C_k'(C_k\hat{P}_k^- C_k' + R_k)^+, & \hat{Q}_k &:= Q_k - L_k^p S_k', \\ \hat{w}_k &:= L_k^p e_k, & L_k^p &:= S_k(C_k\hat{P}_k^- C_k' + R_k)^+ \end{aligned}$$

for each $k \in \mathbb{I}_{\geq 0}$. By independence of u_k and $(x_0, u_0, \ldots, u_{k-1})$,

$$\begin{bmatrix} x_k|\mathbf{y}_{k-1} \\ u_k \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \hat{x}_k^- \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{P}_k^- & 0 \\ 0 & I \end{bmatrix}\right)$$

Using the linearity property of Gaussian random variables,

$$\begin{aligned} \begin{bmatrix} x_k \\ B_k u_k \\ y_k \end{bmatrix} |\mathbf{y}_{k-1} &= \begin{bmatrix} I & 0 \\ 0 & B_k \\ C_k & D_k \end{bmatrix} \begin{bmatrix} x_k|\mathbf{y}_{k-1} \\ u_k \end{bmatrix} \\ &\sim \mathrm{N}\left(\begin{bmatrix} \hat{x}_k^- \\ 0 \\ C_k\hat{x}_k^- \end{bmatrix}, \begin{bmatrix} \hat{P}_k^- & 0 & \hat{P}_k^- C_k' \\ 0 & Q_k & S_k \\ C_k\hat{P}_k^- & S_k' & C_k\hat{P}_k^- C_k' + R_k \end{bmatrix}\right) \end{aligned}$$

Using to condition on y_k,

$$\begin{bmatrix} x_k \\ B_k u_k \end{bmatrix} |(\mathbf{y}_{k-1}, y_k) = \begin{bmatrix} x_k \\ B_k u_k \end{bmatrix} |\mathbf{y}_k \sim \mathrm{N}\left(\begin{bmatrix} \hat{x}_k \\ \hat{w}_k \end{bmatrix}, \begin{bmatrix} \hat{P}_k & -L_k^f S_k' \\ -S_k(L_k^f)' & \hat{Q}_k \end{bmatrix}\right)$$

Taking the marginal distribution in $x_k|\mathbf{y}_k$ gives (66). Using the linearity property again gives

$$x_{k+1}|\mathbf{y}_k = (A_k x_k + B_k u_k)|\mathbf{y}_k \sim \mathrm{N}(\hat{x}_{k+1}^-, \hat{P}_{k+1}^-)$$

□

5.3 Sparse control and estimation reformulations

While the closed-form expressions (60) and (65) can be readily used on small-scale control and estimation problems, large-scale problems may not be amenable to these expressions. When the system dimensions are large (n for the LQR and p for the KF), the frequent pseudoinverse computations may cause numerical issues that compromise closed-loop stability. In the following discussion, we pose the LQR and KF problems as estimation problems. These problems can be solved using optimization or linear systems algorithms, as discussed in section 4. For both problems, when N is large, it is beneficial to consider *sparse* formulations of the optimization problems and linear systems. We first illustrate this fact with the LQR.

5.3.1 Linear quadratic regulator

The naïve way to solve (59) (without recursion) is to rewrite the objective in terms of only the initial state x and control actions $\mathbf{u}_{N-1}$,

$$V_N(x; \mathbf{u}_{N-1}) = \left\| \begin{bmatrix} \mathcal{O}_N \\ A^N \end{bmatrix} x + \begin{bmatrix} \mathcal{G}_N \\ \mathcal{C}_N \end{bmatrix} \mathbf{u}_{N-1} \right\|^2_{\left[\begin{smallmatrix} I & 0 \\ 0 & P_N \end{smallmatrix}\right]}$$

where

$$\mathcal{G}_N := \begin{bmatrix} D_0 & & & & \\ C_1B_0 & D_1 & & & \\ \vdots & \ddots & \ddots & & \\ C_{N-2}\mathcal{A}_{1:N-3}B_0 & \dots & C_{N-2}B_{N-3} & D_{N-2} & \\ C_{N-1}\mathcal{A}_{1:N-2}B_0 & \dots & C_{N-1}A_{N-2}B_{N-3} & C_{N-1}B_{N-2} & D_{N-1} \end{bmatrix},$$

$$\mathcal{O}_N := \begin{bmatrix} C_0 \\ C_1A_0 \\ \vdots \\ C_{N-2}\mathcal{A}_{0:N-3} \\ C_{N-1}\mathcal{A}_{0:N-2} \end{bmatrix},$$

$$\mathcal{C}_N := \begin{bmatrix} \mathcal{A}_{1:N-1}B_0 & \mathcal{A}_{2:N-1}B_1 & \dots & A_{N-1}B_{N-2} & B_{N-1} \end{bmatrix},$$

$$\mathcal{A}_{i:i+j} := A_{i+j} \times A_{i+j-1} \times \ldots \times A_i \qquad \forall i, j \in \mathbb{I}_{\geq 0}$$

This reformulation makes it clear that (59) is a GLS problem, i.e.,

$$\hat{\mathbb{B}}_{\text{GLS}}\left(\begin{bmatrix}\mathcal{O}_N \\ A^N\end{bmatrix} x, -\begin{bmatrix}\mathcal{G}_N \\ \mathcal{C}_N\end{bmatrix}, \begin{bmatrix} I & 0 \\ 0 & P_N\end{bmatrix}\right) = \underset{\mathbf{u}_{N-1}\in\mathbb{R}^{n_u N}}{\operatorname{argmin}} V_N(x; \mathbf{u}_{N-1})$$

Taking the derivative of the objective, we can use lemma 10 to get that $\mathbf{u}^0_{N-1}$ minimizes $V_N(x; \mathbf{u}_{N-1})$ if and only if

$$(\mathcal{G}_N'\mathcal{G}_N + \mathcal{C}_N'P_N\mathcal{C}_N)\mathbf{u}^0_{N-1} = -(\mathcal{G}_N'\mathcal{O}_N + \mathcal{C}_N'P_NA^N)x \tag{67}$$

But (67) is a *dense* problem that scales *cubically* with N, making it numerically challenging to compute solutions to this problem for large N.

When N is large, it is better to solve for both $\mathbf{u}_{N-1}$ and $\mathbf{x}_{N+1} := (x_0, x_1, \ldots, x_N)$ simultaneously, enforcing the dynamics (54) through constraints,

$$\min_{\mathbf{x}_N\in\mathbb{R}^{n_x(N+1)}, \mathbf{u}_{N-1}\in\mathbb{R}^{n_u N}} V_N(\mathbf{x}_N, \mathbf{u}_{N-1}) \quad \text{subject to} \quad c_N(\mathbf{x}_N, \mathbf{u}_{N-1}) = 0 \tag{68}$$

where

$$V_N(\mathbf{x}_N, \mathbf{u}_{N-1}) := \sum_{j=0}^{N-1} \|y_j\|^2 + \|x_N\|^2_{P_N}$$

$$c_N(\mathbf{x}_N, \mathbf{u}_{N-1}) := \begin{bmatrix} x_0 - x \\ x_1 - A_0x_0 - B_0u_0 \\ x_2 - A_1x_1 - B_1u_1 \\ \vdots \\ x_N - A_{N-1}x_{N-1} - B_{N-1}u_{N-1} \end{bmatrix}$$

We can rewrite this as a ECGLS problem,

$$\min_{\mathbf{x}_N \in \mathbb{R}^{n_x(N+1)}, \mathbf{u}_{N-1} \in \mathbb{R}^{n_u N}} \left\| \begin{bmatrix} \mathbf{x}_N \\ \mathbf{u}_{N-1} \end{bmatrix} \right\|_{\mathbf{V}_N}^2 \quad \text{subject to} \quad \mathbf{Z}_N \begin{bmatrix} \mathbf{x}_N \\ \mathbf{u}_{N-1} \end{bmatrix} = \mathbf{w}_N \tag{69}$$

where[20]

$$\begin{aligned}
\mathbf{V}_N &:= \left[\begin{array}{cc|c} \bigoplus_{j=0}^{N-1} C_j' C_j & 0 & \bigoplus_{j=0}^{N-1} C_j' D_j \\ 0 & P_N & 0 \\ \hline \bigoplus_{j=0}^{N-1} C_j' D_j & 0 & \bigoplus_{j=0}^{N-1} D_j' D_j \end{array}\right] \\
\mathbf{Z}_N &:= \left[\begin{array}{c|c} I - \begin{bmatrix} 0 & 0 \\ \bigoplus_{j=0}^{N-1} A_j & 0 \end{bmatrix} & \begin{matrix} 0 \\ -\bigoplus_{j=0}^{N-1} B_j \end{matrix} \end{array}\right] \\
\mathbf{w}_N &:= \begin{bmatrix} x \\ 0 \end{bmatrix}
\end{aligned}$$

which is clearly a *sparse* problem. Therefore $(\mathbf{x}_N^0, \mathbf{u}_{N-1}^0)$ solves (68) if and only if $(\mathbf{x}_N^0, \mathbf{u}_{N-1}^0) \in \hat{\mathbb{B}}_{\text{ECGLS}}(0, I, \mathbf{V}_N, \mathbf{w}_N, \mathbf{Z}_N)$. From the proof of theorem 31, we have that $(\mathbf{x}_N^0, \mathbf{u}_{N-1}^0)$ solves (68) if and only if there exists $\boldsymbol{\lambda}_N^0$ such that

$$\left[\begin{array}{c|c} \mathbf{V}_N & \mathbf{Z}_N' \\ \hline \mathbf{Z}_N & 0 \end{array}\right] \left[\begin{array}{c} \mathbf{x}_N^0 \\ \mathbf{u}_{N-1}^0 \\ \hline \boldsymbol{\lambda}_N^0 \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline \mathbf{w}_N \end{array}\right] \tag{70}$$

While the sparse normal equations (70) has $2n_x N$ more variables than their dense counterpart (67), sparse solvers can be used on (70), which have a far lower computational burden when N is large. It is also worth noting that the optimization problem (69) can be augmented with additional (linear equality or convex inequality) state and input constraints while preserving convexity of the problem, allowing the control actions of linear model predictive controllers to be computed with sparse convex optimization solvers for large N [54, 94, 113, 115].

5.3.2 Full information estimation

In some situations it is desirable to estimate the entire state sequence, $\mathbf{x}_N | \mathbf{y}_{N-1}$ where $\mathbf{x}_N := (x_0, x_1, \ldots, x_N)$. In this case, we can write the initial state prior as $x_0 = \hat{x}_0^- + e_0$ where $e_0 \sim \mathrm{N}(0, \hat{P}_0^-)$ and the model (54) and (55) can be written as

$$\tilde{y} = \tilde{X} \mathbf{x}_N + \tilde{e}$$

using theorem 40, where

$$\tilde{y} := \left[\begin{array}{c} \hat{x}_0^- \\ 0 \\ \hline \mathbf{y}_{N-1} \end{array}\right], \quad \tilde{X} := \left[\begin{array}{c} I - \begin{bmatrix} 0 & 0 \\ \bigoplus_{j=0}^{N-1} A_j & 0 \end{bmatrix} \\ \hline \bigoplus_{j=0}^{N-1} C_j \quad 0 \end{array}\right], \quad \tilde{e} := \left[\begin{array}{c|c} I & 0 \\ 0 & -\bigoplus_{j=0}^{N-1} B_j \\ \hline 0 & \bigoplus_{j=0}^{N-1} D_j \end{array}\right] \left[\begin{array}{c} e_0 \\ \hline \mathbf{u}_{N-1} \end{array}\right]$$

[20] The direct sum $\bigoplus$ is defined as $\bigoplus_{k=1}^{N} M_k := \begin{bmatrix} M_1 & & \\ & \ddots & \\ & & M_N \end{bmatrix}$ for any $M_i \in \mathbb{R}^{m_i \times n_i}$.

Therefore, by theorems 40 and 41, $\hat{\mathbf{x}}_N = \mathbb{E}[\mathbf{x}_N|\mathbf{y}_{N-1}]$ if and only if $\hat{\mathbf{x}}_N \in \hat{\mathbb{B}}_{\mathrm{MLE}}(\tilde{y}, \tilde{X}, \tilde{V})$, where

$$\tilde{V} := \mathrm{var}[\tilde{e}] = \left[\begin{array}{cc|c} \hat{P}_0^- & 0 & 0 \\ 0 & \bigoplus_{j=0}^{N-1} B_j B_j' & -\bigoplus_{j=0}^{N-1} B_j D_j' \\ \hline 0 & -\bigoplus_{j=0}^{N-1} D_j B_j' & \bigoplus_{j=0}^{N-1} D_j D_j' \end{array}\right]$$

From theorem 37, we have that $\hat{\mathbf{x}}_N = \mathbb{E}[\mathbf{x}_N|\mathbf{y}_{N-1}]$ if and only if there exists $\hat{\boldsymbol{\alpha}}_N \in \mathbb{R}^{n(N+1)+pN}$ such that $(\hat{\mathbf{x}}_N, \hat{\boldsymbol{\alpha}}_N)$ solve

$$\min_{\mathbf{x}_N \in \mathbb{R}^{n(N+1)}, \boldsymbol{\alpha}_N \in \mathbb{R}^{n(N+1)+pN}} \|\boldsymbol{\alpha}_N\|_{\tilde{V}}^2 \quad \text{subject to} \quad \tilde{y} = \tilde{X}\mathbf{x}_N + \tilde{V}\boldsymbol{\alpha}_N \tag{71}$$

or equivalently, $\hat{\mathbf{x}}_N = \mathbb{E}[\mathbf{x}_N|\mathbf{y}_{N-1}]$ if and only if there exists $\hat{\boldsymbol{\alpha}}_N \in \mathbb{R}^{n(N+1)+pN}$ such that

$$\left[\begin{array}{c|c} \tilde{V} & \tilde{X} \\ \hline \tilde{X}' & 0 \end{array}\right] \left[\begin{array}{c} \hat{\boldsymbol{\alpha}}_N \\ \hline \hat{\mathbf{x}}_N \end{array}\right] = \left[\begin{array}{c} \tilde{y} \\ \hline 0 \end{array}\right] \tag{72}$$

The formulations (71) and (72) are again amenable to sparse convex optimization solvers and sparse linear solvers. If we wish to add (linear equality or convex inequality) constraints to the estimates $\mathbf{x}_N$, we can add them to (71). However, in both problems, we do not estimate the error covariance for this computation, so the computational tractability of the large-scale problem comes at the cost of additional information about the error statistics.

6 Modern extensions

We conclude this paper with a summary of related areas of research that might benefit from methods discussed herein.

6.1 A generalized perturbation method for degenerate distributions

A generalized proof of theorem 38 could utilize [95, Theorems 7.17, 7.33] to show that

$$\operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left\{ \lim_{\rho \to 0^+} \phi_\rho(\beta) \right\} \supseteq \lim_{\rho \to 0^+} \left\{ \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \phi_\rho(\beta) \right\} \tag{73}$$

holds for a more general class of objectives.[21] This approach of exchanging the limit and minimizer has applications in the estimation of parameters of other degenerate distributions (e.g., singular elliptical distributions [8, 27]). If the inclusion (73) holds, one can solve the (easier) non-degenerate problem and take the limit of the solution as $\rho \to 0^+$ to obtain a solution to the (harder) degenerate problem. Moreover, the non-degenerate problem with arbitrarily small ρ is arbitrarily close to a solution of the degenerate problem, so a limit need not be taken if the problem can be solved with small enough ρ to meet the desired tolerance.

[21] This result holds, for example, when the limit $\phi = \lim_{\rho \to 0^+} \phi_\rho$ exists (pointwise); ϕ_ρ, ϕ are convex, lower semicontinuous, and proper for all $\rho > 0$; and the right-hand side limit of (73) exists.

6.2 Nonlinear regression

Consider the following nonlinear regression model,

$$y = f(\beta) + e, \qquad e \sim \mathrm{N}(0, V) \tag{74}$$

The problem of finding an estimate of β given the observations y and function $f : \mathbb{R}^p \to \mathbb{R}^n$ is often called the *nonlinear inverse problem* [22, 46, 48, 74, 105, 106]. Development of algorithms for solving the problem (74) is a current and ongoing research topic [10, 22, 46]. While deriving estimates of (74) is outside of the scope of this paper, concepts from linear estimation can be generalized to the nonlinear setting. In particular, the gradient methods discussed in section 4 can be extended to compute (locally) optimal estimators.

Consider the nonlinear least squares problem,

$$\min_{\beta \in \mathbb{R}^p} \phi(\beta) := \frac{1}{2}\|y - f(\beta)\|^2 \tag{75}$$

The nonlinear map from parameters to observations complicates computation of the solutions to (75), as we no longer have convexity. However, gradient methods can still be used to compute *locally* optimal solutions to (75). The gradient flow algorithm for (75) is

$$\dot{\beta}(t) = -\left[\frac{df}{d\beta}(\beta(t))\right]' [y - f(\beta(t))] \tag{76}$$

It can be shown that, for any stationary point $\beta_0 \in \mathbb{R}^p$ of (76), the function $V(\beta) := \phi(\beta) - \phi(\beta_0)$ is a *local* Lyapunov function and therefore the gradient flow is *locally* stable about β_0. If we also assume that $(df/d\beta)(\beta)$ is full column rank in a neighborhood of β_0, then we can also guarantee local *asymptotic* stability about β_0. Existence of the asymptotic Lyapunov function also implies local optimality of β_0 because $V(\beta) > 0 = V(\beta_0)$ for all $\beta \neq \beta_0$ in a neighborhood of β_0. In other words, the gradient flow algorithm can converge only to locally optimal points.

6.3 Bayesian regression

In section 3.5, we considered a model with a Gaussian prior on β. However, more general priors can be used, which produce different properties of the estimator. Suppose the parameters β have a prior distribution of the following form,

$$f(\beta) := c_1 \exp\left(-c_2\|\beta - \beta_0\|_\gamma^{c_3}\right) \tag{77}$$

where $\|\cdot\|_\gamma : \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ is a norm and $c_1, c_2, c_3 > 0$ are chosen such that f is a probability distribution.[22] Then the MAP estimator for the model

$$y = X\beta + e, \qquad e \sim \mathrm{N}(0, V), \qquad \beta \sim f, \qquad \beta, e \text{ independent}$$

[22] We could also choose any distribution such that $f(\beta)$ is log-convex, but the form (77) makes it clear how the estimation problem relates to commonly used regularization methods.

is given by

$$\max_{\beta \in \mathbb{R}^p} f(y|\beta) f(\beta) \qquad \text{subject to} \qquad f(y|\beta) f(\beta) > 0 \tag{78}$$

Similarly to the proof of lemma 34, we can reformulate (78) as a convex optimization problem,

$$\min_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^p} \frac{1}{2} \|\alpha\|_V^2 + c_2 \|\beta - \beta_0\|_\gamma^{c_3} \qquad \text{subject to} \qquad y = X\beta + V\alpha$$

When $\|\cdot\|_\gamma^{c_3} = \|\cdot\|_2^2$, (77) and (78) corresponds to classic 2-norm regularization [51, 52], which is widely used in machine learning [45, 78], algorithms for ill-conditioned linear systems [77, 111], and algorithms for inverse problems [21, 22]. When $\|\cdot\|_\gamma^{c_3} = \|\cdot\|_1$, (77) and (78) corresponds to 1-norm regularization, or LASSO regression, which is often used to promote the sparsity of estimates [49, 108, 117].

6.4 Sparse estimators

A common situation in signal processing, image processing, nonlinear system identification, and machine learning arises when there are many more parameters than observations. In these situations we wish to find a *sparse* estimator, or more specifically, one which minimizes the 0-pseudonorm $\|\beta\|_0 := \#\{\beta_i \neq 0\}$. We can define these estimators as an augmentation of the estimators discussed herein by augmenting the estimator objective with an additive penalty $\lambda\|\beta\|_0$ or multiplicative penalty $\exp(-\lambda\|\beta\|_0)$. The sparsity (hyper)parameter $\lambda > 0$ requires tuning to achieve the desired level of sparsity in the estimate. For example, we can define sparse versions of the ECGLS and MLE problems as follows.[23]

$$\min_{\beta \in \mathbb{R}^p} \|y - X\beta\|_H^2 + \lambda\|\beta\|_0 \quad \text{subject to} \quad Z\beta = w \tag{79}$$

$$\max_{\beta \in \mathbb{R}^p} f(y|\beta) \exp(-\lambda\|\beta\|_0) \quad \text{subject to} \quad f(y|\beta) > 0 \tag{80}$$

The sparse MLE problem (86) can be reformulated as

$$\min_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^p} \frac{1}{2} \|\alpha\|_V^2 + \lambda\|\beta\|_0 \quad \text{subject to} \quad y = X\beta + V\alpha \tag{81}$$

Due to the nonconvex and discontinuous objective functions of problems (85) and (86), estimators cannot be computed exactly in polynomial time [79]. However, small to moderately sized problems can be solved using mixed integer programming methods [15, 17].

Alternatively, we can approximate the sparse solution with a (possibly nonconvex) shrinkage penalty or prior. Recall that the 0-pseudonorm is the limit of the q-(pseudo)norm as $q \to 0^+$, i.e. $\|\beta\|_0 = \lim_{q \to 0^+} \|\beta\|_q^q$, pointwise in $\beta \in \mathbb{R}^p$. For some $q, \lambda > 0$, consider the following generalized normal prior on β,

$$\beta \sim f(\beta; q, \lambda) := \frac{\lambda^{1/q}}{[2\Gamma(1 + q^{-1})]^p} \exp\left(-\lambda\|\beta\|_q^q\right)$$

[23] Note that the sparse MLE problem is *not* a MAP problem because the multiplicative penalty $\exp(-\lambda\|\beta\|_0)$ cannot be used to formulate a probability density over $\beta \in \mathbb{R}^p$. Instead, the sparse MLE problem should be viewed as a *modified* MLE problem.

and define the q-pseudonorm MAP (qMAP) with shrinkage penalty λ as follows,[24]

$$\max_{\beta\in\mathbb{R}^p} f(y|\beta)f(\beta;q,\lambda) \quad \text{subject to} \quad f(y|\beta)>0 \tag{82}$$

Again, (87) can be reformulated as a convex optimization problem,

$$\min_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^p} \frac{1}{2}\|\alpha\|_V^2+\lambda\|\beta\|_q^q \quad \text{subject to} \quad y=X\beta+V\alpha \tag{83}$$

Moreover, the objective of (81) converges uniformly to the objective (83). Therefore, taking the limit of the set of solutions to (83) as $q\to 0^+$ gives a limit set that is a subset of the set of solutions to (81) [95, Theorems 7.15, 7.33].[25] In other words,

$$\lim_{q\to 0^+}\left\{\begin{matrix}\operatorname*{argmin}_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^p} \frac{1}{2}\|\alpha\|_V^2+\lambda\|\beta\|_q^q \\ \text{subject to } y=X\beta+V\alpha\end{matrix}\right\} \subseteq \left\{\begin{matrix}\operatorname*{argmin}_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^p} \frac{1}{2}\|\alpha\|_V^2+\lambda\|\beta\|_0 \\ \text{subject to } y=X\beta+V\alpha\end{matrix}\right\} \tag{84}$$

Fung and Mangasarian [37] showed that there exists constant $\overline{q}>0$ such that (86) and (87) have the same solution sets for all $q\in[0,\overline{q}]$ [37]. Therefore, the limit may not need to be taken completely, and early stopping of the limit may produce exact solutions. In fact, for some cases of (y,X), solving (87) with $q=1$ can give the exact solution to (86) [19, 23, 29]. This convex relaxation of the problem (86) is equivalent to LASSO regression and is similar to compressed sensing.

Definition 52. *Let $\hat{\mathbb{B}}_{\text{sECGLS}}(y,X,H,w,Z,\lambda)$ be the set of solutions to the problem*

$$\min_{\beta\in\mathbb{R}^p} \|y-X\beta\|_H^2+\lambda\|\beta\|_0 \quad \text{subject to} \quad Z\beta=w \tag{85}$$

where $H\in\mathbb{S}_+^n$ is a positive semidefinite weighting matrix, $Z\in\mathbb{R}^{c\times n}$ and $w\in\mathbb{R}^c$ are the constraint parameters, and $\lambda>0$ is the sparsity parameter. We say $\hat{\beta}$ is a sECGLS estimator (of the model (LGM)*, with weighting matrix H, constraint parameters w,Z, and sparsity parameter λ) if $\hat{\beta}\in\hat{\mathbb{B}}_{\text{sECGLS}}(y,X,H,w,Z,\lambda)$.*

Definition 53. *Let $\hat{\mathbb{B}}_{\text{sMLE}}(y,X,V,\lambda)$ be the set of solutions to the problem*

$$\max_{\beta\in\mathbb{R}^p} L_y(\beta;y,X,V)\exp(-\lambda\|\beta\|_0) \tag{86}$$

where $L_y(\cdot;y,X,V)$ is the likelihood function of the observations.[26] *We say that $\hat{\beta}$ is a sMLE of the model* (LGM) *with sparsity λ if $\hat{\beta}\in\hat{\mathbb{B}}_{\text{sMLE}}(y,X,V,\lambda)$.*

[24] Bayes' rule justifies maximizing over $f(y|\beta)f(\beta;q,\lambda)$ rather than $f(\beta|y;q,\lambda)$, and using the constraint $f(y|\beta)>0$ rather than $f(\beta|y;q,\lambda)>0$.

[25] We have skipped some technical detail here for the sake of brevity. In particular, the following additional facts are required in the hypotheses of [95, Theorems 7.15, 7.33]. Let ϕ and ϕ_q denote the extended-value objective functions (i.e., let them be equal to ∞ when the constraint is violated) of (81) and (83), respectively. Denote the c-sublevel set of a function $f:\mathbb{R}^n\to\mathbb{R}$ as $\operatorname{lev}_c f := \{x\in\mathbb{R}^n \mid f(x)\le c\}$.

1. ϕ_ρ is *level-bounded* (i.e., $\operatorname{lev}_c\phi, \operatorname{lev}_c\phi_\rho$ are bounded for all $c\in\mathbb{R}$), for all $\rho>0$.
2. ϕ,ϕ_ρ are *lower semicontinuous* (i.e., $\operatorname{lev}_c\phi, \operatorname{lev}_c\phi_\rho$ are closed for all $c\in\mathbb{R}$ [95, Theorem 1.6]) and *proper* (i.e., ϕ,ϕ_ρ never equal $-\infty$ and do not always equal ∞), for all $\rho>0$.

[26] This is the same likelihood function as the one defined in definition 32.

Definition 54. *Define* $\hat{\mathbb{B}}_{q\mathrm{MAP}}(y, X, V, \lambda)$ *as the set of solutions to the problem*

$$\max_{\beta \in \mathbb{R}^p} L_{\beta|y}(\beta; y, X, V, q, \lambda) \tag{87}$$

where $L_{\beta|y}(\cdot; y, X, V, q, \lambda)$ *is the likelihood function of the parameters conditioned on the observations given that* $\beta \sim q\mathrm{N}(0, \lambda)$.[27] *We say that* $\hat{\beta}$ *is a qMAP of the model* (LGM) *with prior* $\beta \sim q\mathrm{N}(0, \lambda)$ *if* $\hat{\beta} \in \hat{\mathbb{B}}_{q\mathrm{MAP}}(y, X, V, \lambda)$.

Using the qMAP definition, we have a convenient interpretation of the sparse MLE as the limiting estimator of a class of non-Gaussian MAP estimators, that is the limit of any qMAP estimator as $q \to 0^+$ is a sparse MLE. The proof of this fact follows similarly to the argument (73) using [95, Theorems 7.15, 7.33].

Conjecture 55. *Let* $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $V \in \mathbb{S}^n_+$, *and* $T = I - VV^+$. *If* $y \in \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$, *then*

$$\hat{\mathbb{B}}_{\mathrm{sMLE}}(y, X, V, \lambda) = \hat{\mathbb{B}}_{\mathrm{sECGLS}}(y, X, V^+, TX, Ty, \lambda) \tag{88a}$$

$$\supseteq \lim_{q \to 0^+} \hat{\mathbb{B}}_{q\mathrm{MAP}}(y, X, V, \lambda, q) \tag{88b}$$

Proof. The equality (88a) follows in the same way as in the proof of lemma 33. We show the inclusion (88b) by rewriting the qMAP problem (87) as a minimization over the negative log-likelihood. Denote the feasible set as

$$\mathbb{B} = \{\, \beta \in \mathbb{R}^p \mid y - X\beta \in \mathcal{R}(V) \,\} = \{\, \beta \in \mathbb{R}^p \mid TX\beta = Ty \,\}$$

Using Bayes theorem and dropping constants, we have the negative log-likelihood

$$-\ln L_{\beta|y}(\beta|y; X, V, q, \lambda) \propto -\ln L_{y,\beta}(y, \beta; X, V, q, \lambda) \propto \|y - X\beta\|^2_{V^+} + \lambda\|\beta\|_q^q$$

for all $\beta \in \mathbb{B}$. Therefore we can optimize over the objectives

$$\phi_q(\beta) = \begin{cases} \|y - X\beta\|^2_{V^+} + \lambda\|\beta\|_q^q, & \beta \in \mathbb{B} \\ \infty, & \text{otherwise} \end{cases}$$

$$\phi(\beta) = \begin{cases} \|y - X\beta\|^2_{V^+} + \lambda\|\beta\|_0, & \beta \in \mathbb{B} \\ \infty, & \text{otherwise} \end{cases}$$

for the qMAP and sMLE problems, respectively. For every $q > 0$, ϕ_q is lower semi-continuous as it is the sum of lower semi-continuous and proper functions [95, Proposition 1.39].[28] As $q \to 0^+$, we have $\|\beta\|_q^q \to \|\beta\|_0$ for all $\beta \in \mathbb{R}^p$. Therefore, $\phi_q \to \phi$ *uniformly* as $q \to 0^+$. The result (88b) follows from [95, Theorems 7.15, 7.33]. □

As a result of conjecture 55, one can find an approximate sMLE by solving for the qMAP for some small $q > 0$.

[27] We can replace the posterior density function with the joint density function using Bayes' theorem $p_{\beta|y}(\beta|y; X, V, q, \lambda) \propto p_{y,\beta}(y, \beta; X, V, q, \lambda) = p_{y|\beta}(y|\beta; X, V)p_\beta(\beta; q, \lambda)$ where $y|\beta \sim \mathrm{N}(X\beta, V)$ and $\beta \sim q\mathrm{N}(0, \lambda)$. We do not need to consider the density of y because it is not a function of β, and it exists when $p_{y|\beta}(y|\beta; X, V)$ exists.

[28] We say a function $f : \mathbb{R}^p \to \mathbb{R}$ is proper if $f \not\equiv \infty$. We say a function $f : \mathbb{R}^p \to \mathbb{R}$ is lower semi-continuous if, for every $\alpha \in \mathbb{R}$, the sublevel set $\{\, x \in \mathbb{R}^p \mid f(x) \leq \alpha \,\}$ is closed [95, Theorem 1.6].

A Least squares proofs

In this appendix we prove theorems 25, 28, and 31 and corollaries 26 and 29 using the method of Lagrange multipliers [14, 18].

A.1 Generalized least squares proofs

Proof of theorem 25. If suffices to show $\hat{\beta} \in \hat{\mathbb{B}}_{\text{GLS}}(y, X, H)$ if and only if $\hat{\beta} \in (X'HX)^+X'Hy + \mathcal{N}(X'HX)$. By definition, $\hat{\beta} \in \hat{\mathbb{B}}_{\text{GLS}}(y, X, H)$ if and only if it solves (GLS). Denote the objective as

$$\phi(\beta) = \frac{1}{2}\|y - X\beta\|_H^2$$

Using lemma 10, we have $\hat{\beta}$ is a solution to (GLS) if and only if

$$\frac{\partial \phi}{\partial \beta}(\hat{\beta}) = 2X'HX\hat{\beta} - 2X'Hy = 0$$

By lemma 3, solutions to the above equation exist because $X'Hy \in \mathcal{R}(X'H^{1/2}) = \mathcal{R}(X'HX)$, and $\hat{\beta}$ is a solution if and only if $\hat{\beta} \in (X'HX)^+X'Hy + \mathcal{N}(X'HX)$. □

Proof of corollary 26. First, we expand the objective function,

$$\begin{aligned} V(\beta) &= \frac{1}{2}\|y - X(\beta - \hat{\beta}) - X\hat{\beta}\|_H^2 \\ &= \frac{1}{2}\|\beta - \hat{\beta}\|_{X'HX}^2 + \frac{1}{2}\|y - X\hat{\beta}\|_H^2 - (\beta - \hat{\beta})'X'H(y - X\hat{\beta}) \end{aligned} \tag{89}$$

Next, recall that $\hat{\beta} \in \hat{\mathbb{B}}_{\text{GLS}}(y, X, V)$ if and only if $\hat{\beta} = (X'HX)^+X'Hy + \hat{\alpha}$ for some $\hat{\alpha} \in \mathcal{N}(X'HX)$. Then $HX\hat{\alpha} = 0$ and

$$H(y - X\hat{\beta}) = Hy - HX(X'HX)^+X'Hy = H(I - X(X'HX)^+X'H)y$$

Moreover, the cross term in (89) is zero,

$$X'H(y - X\hat{\beta}) = (X'H - X'HX(X'HX)^+X'H)y = (X'H - X'H)y = 0 \tag{90}$$

and we can rewrite the normed error in (89) as follows,

$$\begin{aligned} \|y - X\hat{\beta}\|_H^2 &= y'(I - HX(X'HX)^+X')H(I - X(X'HX)^+X'H)y \\ &= y'(H - HX(X'HX)^+X'H)y = \|y\|_{H_0}^2 \end{aligned} \tag{91}$$

Finally, combining (89)–(91) gives (22). □

A.2 Tikhonov generalized least squares proofs

Proof of theorem 28. It is clear that the objective of (TGLS) is equal to $\frac{1}{2}\|\tilde{y}-\tilde{X}\beta\|^2_{\tilde{H}}$ where

$$\tilde{y} = \begin{bmatrix} y \\ \beta_0 \end{bmatrix}, \qquad \tilde{X} = \begin{bmatrix} X \\ I \end{bmatrix}, \qquad \tilde{H} = \begin{bmatrix} H & 0 \\ 0 & \Gamma \end{bmatrix} \tag{92}$$

This shows (23a). To show (23b), we use (23a) and theorem 25:

$$\begin{aligned}
\hat{\mathbb{B}}_{\mathrm{TGLS}}(y, X, H, \beta_0, \Gamma) &= \hat{\mathbb{B}}_{\mathrm{GLS}}(\tilde{y}, \tilde{X}, \tilde{H}) \\
&= (\tilde{X}'\tilde{H}\tilde{X})^+\tilde{X}'\tilde{H}\tilde{y} + \mathcal{N}(\tilde{X}'\tilde{H}\tilde{X}) \\
&= (X'HX + \Gamma)^+(X'Hy + \Gamma\beta_0) + \mathcal{N}(X'HX + \Gamma) \\
&= (X'HX + \Gamma)^+(X'HX + \Gamma)\beta_0 \\
&\qquad + (X'HX + \Gamma)^+X'H(y - X\beta_0) + \mathcal{N}(X'HX + \Gamma) \\
&= \Gamma_0^+\Gamma_0\beta_0 + L(y - X\beta_0) + \mathcal{N}(\Gamma_0)
\end{aligned}$$

□

Proof of corollary 29. Using the definitions (92) and corollary 26, we have

$$V(\beta) = \frac{1}{2}\|\tilde{y} - \tilde{X}\beta\|^2_{\tilde{H}} = \frac{1}{2}\|\beta - \hat{\beta}\|^2_{\tilde{X}'\tilde{H}\tilde{X}} + \frac{1}{2}\|\tilde{y}\|^2_{\tilde{H}_0} \tag{93}$$

where $\tilde{H}_0 = \tilde{H} - \tilde{H}\tilde{X}(\tilde{X}'\tilde{H}\tilde{X})^+\tilde{X}'\tilde{H}$. The first terms in the right-hand sides of (24) and (93) are clearly equivalent because

$$\tilde{X}'\tilde{H}\tilde{X} = X'HX + \Gamma = \Gamma_0 \tag{94}$$

Moreover, by lemma 12, we have the following identities,

$$\begin{aligned}
\Gamma\Gamma_0^+X'H &= \Gamma_0\Gamma_0^+X'H - X'HX\Gamma_0^+X'H = X'(H - HX\Gamma_0^+X'H) = X'\Gamma_1 \\
\Gamma - \Gamma\Gamma_0^+\Gamma &= \Gamma - \Gamma\Gamma_0^+\Gamma_0 + \Gamma\Gamma_0^+X'HX = X'(H - HX\Gamma_0^+X'H)X = X'\Gamma_1X
\end{aligned}$$

which imply

$$\begin{aligned}
\|\tilde{y}\|^2_{\tilde{H}_0} &= y'Hy + \beta_0'\Gamma\beta_0 - (y'HX + \beta_0'\Gamma)\Gamma_0^+(X'Hy + \Gamma\beta_0) \\
&= y'(H - HX\Gamma_0^+X'H)y + \beta_0'(\Gamma - \Gamma\Gamma_0^+\Gamma)\beta_0 - 2\beta_0'\Gamma\Gamma_0^+X'Hy \\
&= y'\Gamma_1y + \beta_0'X'\Gamma_1X\beta_0 - 2\beta_0'X'\Gamma_1y = \|y - X\beta_0\|^2_{\Gamma_1}
\end{aligned} \tag{95}$$

Combining (93)–(95) gives (24). □

A.3 Equality constrained generalized least squares proofs

Proof of theorem 31. If $w \notin \mathcal{R}(Z)$, the feasible set is empty (lemma 3) and therefore $\hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, H, w, Z)$ must be empty. It suffices to assume $w \in \mathcal{R}(Z)$ and show (25) and (26).

To show (25), we eliminate the constraint and reparameterize the optimization problem. By lemma 3, β satisfies the constraint $w = Z\beta$ if and only if

$$\begin{aligned}\beta \in Z^+w + \mathcal{N}(Z) &= Z^+w + \{\, (I - Z^+Z)\alpha \mid \alpha \in \mathbb{R}^p \,\} \\ &= Z^+w + \{\, B\alpha \mid \alpha \in \mathbb{R}^p \,\}\end{aligned}$$

In other words, $\beta = Z^+w + B\alpha$ for some $\alpha \in \mathbb{R}^p$. Under this parameterization the constraint is eliminated, and the error is $y - X\beta = z - XB\alpha$. The objective function can be written

$$\phi(Z^+w + B\alpha) = \frac{1}{2}\|y - X\beta\|_H^2 = \frac{1}{2}\|z - XB\alpha\|_H^2$$

and therefore $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, H, w, Z)$ if and only if $\hat{\beta} = Z^+w + B\hat{\alpha}$ for some $\hat{\alpha} \in \hat{\mathbb{B}}_{\mathrm{GLS}}(z, XB, H)$. By theorem 25, $\hat{\alpha} \in \hat{\mathbb{B}}_{\mathrm{GLS}}(z, XB, H)$ if and only if

$$\hat{\alpha} \in (BX'V^+XB)^+BX'Hz + \mathcal{N}(BX'V^+XB)$$

Therefore $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, H, w, Z)$ if and only if

$$\hat{\beta} \in Z^+w + B(BX'V^+XB)^+BX'Hz + B\mathcal{N}(BX'V^+XB)$$

To show (26), we use the method of Lagrange multipliers. The Lagrangian is defined as

$$\mathcal{L}(\beta, \lambda) = \frac{1}{2}\|y - X\beta\|_H^2 + \lambda'(Z\beta - w)$$

Since the ECGLS objective is convex and constraint is linear, $\hat{\beta}$ is a solution if and only if there exists $\hat{\lambda} \in \mathbb{R}^c$ such that $(\partial\mathcal{L}/\partial\beta)(\hat{\beta}, \hat{\lambda}) = 0$ and $(\partial\mathcal{L}/\partial\lambda)(\hat{\beta}, \hat{\lambda}) = 0$ (lemma 10). Taking the derivative, we have

$$\frac{\partial\mathcal{L}}{\partial\beta}(\hat{\beta}, \hat{\lambda}) = X'HX\hat{\beta} - X'Hy + Z'\hat{\lambda} = 0, \qquad \frac{\partial\mathcal{L}}{\partial\lambda}(\hat{\beta}, \hat{\lambda}) = Z\hat{\beta} - w = 0$$

which are equivalent to the linear system

$$\begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} X'Hy \\ w \end{bmatrix} \tag{96}$$

Using properties of the range space, $X'Hy \in \mathcal{R}(X'H^{1/2}) = \mathcal{R}(X'HX) \subseteq \mathcal{R}(X'HX + Z'Z) = \mathcal{R}(G)$ and by lemma 3, $GG^+X'Hy = X'Hy$. Moreover, by lemma 12 and lemma 3, $w \in \mathcal{R}(Z) = \mathcal{R}(F)$ and $FF^+w = w$. Then by lemma 13,

$$\begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix}^+ \begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix} \begin{bmatrix} X'Hy \\ w \end{bmatrix} = \begin{bmatrix} GG^+ & 0 \\ 0 & FF^+ \end{bmatrix} \begin{bmatrix} X'Hy \\ w \end{bmatrix} = \begin{bmatrix} X'Hy \\ w \end{bmatrix}$$

and by lemma 3, (96) has solutions and $(\hat{\beta}, \hat{\lambda})$ are solutions if and only if

$$\begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} \in \begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix}^+ \begin{bmatrix} X'Hy \\ w \end{bmatrix} + \mathcal{N}\left(\begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix}\right)$$

In other words, there exists $\hat{\lambda} \in \mathbb{R}^c$ such that $\hat{\beta}$ solves (96) if and only if

$$\hat{\beta} \in \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix}^+ \begin{bmatrix} X'Hy \\ w \end{bmatrix} + \begin{bmatrix} I & 0 \end{bmatrix} \mathcal{N}\left(\begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix} \right)$$

Simplifying the constant term, we have by lemma 13,

$$\begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix}^+ \begin{bmatrix} X'Hy \\ w \end{bmatrix} = \beta_0 + G^+ Z' F^+ (w - Z\beta_0)$$

Likewise for the null space term, we have by lemma 13,

$$\begin{aligned} &\begin{bmatrix} I & 0 \end{bmatrix} \mathcal{N}\left(\begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix} \right) \\ &\quad = \begin{bmatrix} I & 0 \end{bmatrix} \left\{ \left(I - \begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix}^+ \begin{bmatrix} X'HX & Z' \\ Z & 0 \end{bmatrix} \right) q \;\middle|\; q \in \mathbb{R}^{n+c} \right\} \\ &\quad = \{ (I - GG^+) q_1 \mid q_1 \in \mathbb{R}^p \} = \mathcal{N}(G) \end{aligned}$$

Combining these results, we have that there exists $\hat{\lambda} \in \mathbb{R}^c$ such that $\hat{\beta}$ solves (96) (and equivalently $\hat{\beta} \in \hat{\mathbb{B}}_{\text{ECGLS}}(y, X, H, w, Z)$) if and only if

$$\hat{\beta} \in \beta_0 + G^+ Z' F^+ (w - Z\beta_0) + \mathcal{N}(G)$$

□

An immediate corollary to theorem 31 is that both of the solutions have equivalent minimum norm and null space components. corollary 56 is inconsequential to the subsequent sections and therefore the proof is omitted. At the time of this writing, we do not know of a direct (algebraic) proof of the following result.

Corollary 56. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, $H \in \mathbb{S}^n_+$, $Z \in \mathbb{R}^{c \times n}$, and $w \in \mathcal{R}(Z)$,*

$$\begin{aligned} Z^+ w + (BX'V^+XB)^+ BX'Hz &= \beta_0 + G^+ Z' F^+ (w - Z\beta_0) \\ B\mathcal{N}(BX'V^+XB) &= \mathcal{N}(G) \end{aligned}$$

where $B = I - Z^+Z$, $z = y - XZ^+w$, $G = X'HX + Z'Z$, $F = ZG^+Z'$, and $\beta_0 = G^+X'Hy$.

B Maximum likelihood proofs

In this section we examine several methods for solving the MLE problem (MLE). There are three methods to do this, and all are shown to give distinct but equivalent closed-form solutions. The first method is based on solving an equivalent ECGLS problem, the second is based on solving an equivalent saddle point system, and the third is based on taking the limit of the perturbed MLE.

B.1 ECGLS equivalence

Proof of lemma 33. By theorem 31, $\hat{\mathbb{B}}_{\text{ECGLS}}(y, X, V^+, w, Z)$ is nonempty if and only if $w \in \mathcal{R}(Z)$. Therefore it suffices to show (27). We start with the probability density of y. Since $y \sim \mathrm{N}(X\beta, V)$, we have, by definition 8,

$$f(y;\beta) = (2\pi)^{-\frac{p}{2}} |V|_+^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|y - X\beta\|_{V^+}^2\right) > 0$$

for all $y - X\beta \in \mathcal{R}(V)$ and $f(y;\beta) = 0$ otherwise. Therefore the constraint $f(y;\beta) > 0$ is equivalent to $y - X\beta \in \mathcal{R}(V)$. The range constraint can be rewritten as the following linear equality constraint,

$$y - X\beta \in \mathcal{R}(V) \quad \Leftrightarrow \quad (I - VV^+)(y - X\beta) = w - Z\beta = 0$$

Maximizing the likelihood (subject to $w = Z\beta$) is equivalent to minimizing the negative log-likelihood (subject to $w = Z\beta$), which is given by

$$-\ln f(y;\beta) = \frac{p}{2}\ln(2\pi) + \frac{1}{2}\ln|V|_+ + \frac{1}{2}\|y - X\beta\|_{V^+}^2 \tag{97}$$

for all β such that $w = Z\beta$, and is undefined otherwise. After dropping the constant terms in (97), it is clear that minimizing the negative log-likelihood (subject to $w = Z\beta$) is equivalent to (ECGLS) with the stated weighting matrix and constraint parameter definitions, which demonstrates (27). □

To prove (29), we require the following preliminary lemma.

Lemma 57. *For any orthogonal projectors $A, B \in \mathbb{R}^{n\times n}$, if $AB = A$ and $BA = B$, then $A = B$.*

Proof. Using the definition of orthogonal projectors and the hypotheses, we can check the conditions in theorem 1 to show $A^+ = B$. But by lemma 5, we have $A^+ = A$, so $A = B$. □

Proof of theorem 36. First, note that, by lemma 33 and theorem 31,

$$\begin{aligned}\hat{\mathbb{B}}_{\text{MLE}}(y, X, V) &= \hat{\mathbb{B}}_{\text{ECGLS}}(y, X, V^+, w, Z)\\ &= Z^+w + B(BX'V^+XB)^+BX'V^+(y - XZ^+w) + B\mathcal{N}(BX'V^+XB)\end{aligned}$$

Using lemma 5, we have $Z^+w = (TX)^+Ty = (TX)^+y = Z^+y$ and moreover, $y - XZ^+w = (I - XZ^+)y = Cy$. This effectively demonstrates equivalence of the minimum norm component,

$$\hat{\mathbb{B}}_{\text{MLE}}(y, X, V) = Z^+y + B(BX'V^+XB)^+BX'V^+Cy + B\mathcal{N}(BX'V^+XB) \tag{98}$$

Next, we show $X^+X = M := Z^+Z + (BX'V^+XB)^+BX'V^+XB$ by lemma 57. Noting that M is the sum of orthogonal projectors, it is symmetric, and moreover $M^2 = M$ because the cross terms of M^2 are zero,

$$Z^+Z(BX'V^+XB)^+BX'V^+XB = Z^+ZB(BX'V^+XB)^+BX'V^+XB) = 0$$

by lemma 5. Therefore M is an orthogonal projector. Next, we have $MX^+X = M$ using the facts where $Z^+ZX^+X = Z^+TXX^+X = Z^+Z$ and $XBX^+X = XX^+X - XZ^+ZX^+X = X - XZ^+Z = XB$. Finally, we have $X^+XM = X^+X$ using the facts

$$\begin{aligned} XB = (I - VV^+ + VV^+)XB &= ZB + VV^+XB = VV^+XB \\ X^+X(BX'V^+XB)^+BX'V^+XB &= X^+XB(BX'V^+XB)^+BX'V^+XB \\ &= X^+VV^+XB(BX'V^+XB)^+BX'V^+XB \\ &= X^+VV^+XB = X^+XB \end{aligned}$$

which follow from lemmas 5 and 12 and the fact

$$\mathcal{R}(BX'V^+) \subseteq \mathcal{R}(BX'(V^+)^{1/2}) = \mathcal{R}(BX'V^+XB)$$

Therefore $M = X^+X$ by lemma 57, and

$$\begin{aligned} B\mathcal{N}(BX'V^+XB) &= B(I - (BX'V^+XB)^+BX'V^+XB)\mathbb{R}^p \\ &= (I - M)\mathbb{R}^p = (I - X^+X)\mathbb{R}^p = \mathcal{N}(X) \end{aligned} \tag{99}$$

by lemma 5. Finally (29a) follows from (98) and (99).

For (29b), note that by lemma 33 and theorem 31,

$$\begin{aligned} \hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V) &= \hat{\mathbb{B}}_{\mathrm{ECGLS}}(y, X, V^+, w, Z) \\ &= \beta_0 + G^+Z'F^+(w - Z\beta_0) + \mathcal{N}(G) \end{aligned} \tag{100}$$

Rewriting G, we have $G = X'V^+X + Z'Z = X'(V^+ + T)X$. It is easy to see from the SVD

$$V = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} S_1 & \\ & 0 \end{bmatrix} \begin{bmatrix} Q_1' \\ Q_2' \end{bmatrix} = Q_1S_1Q_1'$$

that

$$V^+ + T = Q_1S_1^{-1}Q_1' + Q_2Q_2' = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} S_1^{-1} & \\ & I \end{bmatrix} \begin{bmatrix} Q_1' \\ Q_2' \end{bmatrix}$$

and $(V^+ + T)^{1/2}$ must be invertible. Therefore

$$\mathcal{N}(G) = \mathcal{N}(X'(V^+ + T)X) = \mathcal{N}((V^+ + T)^{1/2}X) = \mathcal{N}(X) \tag{101}$$

Finally, (29b) follows from (100) and (101). □

B.2 Saddle point equivalence

Proof of lemma 34. From the proof of lemma 33, we have $\hat{\beta} \in \hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$ if and only if $\hat{\beta}$ solves

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_{V^+}^2 \qquad \text{subject to} \qquad y - X\beta \in \mathcal{R}(V) \tag{102}$$

Using the definition of the range space $\mathcal{R}(V) = \{\, V\alpha \mid \alpha \in \mathbb{R}^n \,\}$, it is clear that $y - X\beta \in \mathcal{R}(V)$ if and only if $y - X\beta = V\alpha$ for some $\alpha \in \mathbb{R}^n$. Therefore $\hat{\beta}$ solves (102) if and only if there exists $\hat{\alpha} \in \mathbb{R}^n$ such that $(\hat{\alpha}, \hat{\beta})$ solves

$$\min_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_{V^+}^2 \qquad \text{subject to} \qquad y = X\beta + V\alpha$$

or equivalently,

$$\min_{\alpha\in\mathbb{R}^n,\beta\in\mathbb{R}^p} \frac{1}{2}\|\alpha\|_V^2 \qquad \text{subject to} \qquad y = X\beta + V\alpha \tag{103}$$

The Lagrangian is defined as

$$\mathcal{L}(\alpha, \beta, \lambda) := \frac{1}{2}\|\alpha\|_V^2 + \lambda'(X\beta + V\alpha - y)$$

Since the objective is convex and constraint is linear, by lemma 10, $(\hat{\alpha}, \hat{\beta})$ is a solution to (103) if and only if there exists $\hat{\lambda} \in \mathbb{R}^n$ such that

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \alpha}(\hat{\alpha}, \hat{\beta}, \hat{\lambda}) &= V\hat{\alpha} + V\hat{\lambda} = 0 \\
\frac{\partial \mathcal{L}}{\partial \beta}(\hat{\alpha}, \hat{\beta}, \hat{\lambda}) &= X'\hat{\lambda} = 0 \\
\frac{\partial \mathcal{L}}{\partial \lambda}(\hat{\alpha}, \hat{\beta}, \hat{\lambda}) &= X\hat{\beta} + V\hat{\alpha} - y = 0
\end{aligned}$$

which, after making the substitution $V\hat{\alpha} = -V\hat{\lambda}$, is equivalent to the linear vector equation,

$$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} -\hat{\lambda} \\ \hat{\beta} \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{104}$$

But we can always choose $\hat{\lambda} = -\hat{\alpha}$, so $\hat{\lambda} \in \mathbb{R}^n$ satisfying (104) exists if and only if $\hat{\alpha} \in \mathbb{R}^n$ satisfying (SPP) exists. □

Proof of theorem 37. We have that $\hat{\beta} \in \hat{\mathbb{B}}_{\text{MLE}}(y, X, V)$ if and only if there exists $\hat{\alpha} \in \mathbb{R}^n$ such that (SPP) by lemma 34. By lemma 3, solutions to (SPP) exist if and only if

$$\begin{bmatrix} V & X' \\ X & 0 \end{bmatrix} \begin{bmatrix} V & X' \\ X & 0 \end{bmatrix}^+ \begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} y \\ 0 \end{bmatrix} \tag{105}$$

Using lemma 13, we have

$$\begin{bmatrix} V & X' \\ X & 0 \end{bmatrix} \begin{bmatrix} V & X' \\ X & 0 \end{bmatrix}^+ \begin{bmatrix} y \\ 0 \end{bmatrix} = \begin{bmatrix} V_0 V_0^+ y \\ 0 \end{bmatrix}$$

and therefore (105) if and only if $y \in \mathcal{R}(V_0)$ (lemma 12). In other words, the system (SPP) has a solution (and, by lemma 34, $\hat{\mathbb{B}}_{\text{MLE}}(y, X, V)$ is nonempty) if and only if $y \in \mathcal{R}(V_0)$. Moreover, if $y \in \mathcal{R}(V_0)$, then $(\hat{\alpha}, \hat{\beta})$ solves (SPP) if and only if

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \in \begin{bmatrix} (V_0^+ - V_0^+ X W_0^+ X' V_0^+) y \\ W_0^+ X' V_0^+ y \end{bmatrix} + \begin{bmatrix} I - V_0 V_0^+ & 0 \\ 0 & I - W_0 W_0^+ \end{bmatrix} \mathbb{R}^{n+p}$$

by lemmas 3 and 13. Clearly, there exists $\hat{\alpha} \in \mathbb{R}^n$ such that (SPP) if and only if $\hat{\beta} \in W_0^+ X' V_0^+ y + \mathcal{N}(W_0) = \hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, V_0^+)$. The result follows by noting that $\mathcal{R}(W_0) = \mathcal{R}(X')$ (lemma 12) is equivalent to $\mathcal{N}(W_0) = \mathcal{N}(X)$. □

Another proof of theorem 37 follows indirectly from theorem 36, where the expressions (29a) and (30b) are equated using facts from section 2.

Proof of theorem 37 *(indirect).* Using theorem 36, lemma 15, , and corollary 14,

$$\begin{aligned}\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V) &= Z^+ y + (BX'V^+XB)^+ BX'V^+Cy + \mathcal{N}(X) \\ &= (X'(V + XX')^+X)^+ X'(V + XX')^+ y + \mathcal{N}(X) \\ &= (X'V_0^+X)^+ X'V_0^+ y + \mathcal{N}(X)\end{aligned}$$

From lemma 12, we have $\mathcal{R}(X') = \mathcal{R}(X'V_0^+X)$ and thus $\mathcal{N}(X) = \mathcal{N}(X'V_0^+X)$. Using theorem 25,

$$\begin{aligned}\hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, V_0^+) &= (X'V_0^+X)^+ X'V_0^+ y + \mathcal{N}(X'V_0^+X) \\ &= (X'V_0^+X)^+ X'V_0^+ y + \mathcal{N}(X) \\ &= \hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)\end{aligned}$$

□

B.3 Barrier function method

Proof of lemma 35. Let $\phi_\rho(\beta) := \frac{1}{2}\|y - X\beta\|^2_{V_\rho^{-1}}$ and $\phi(\beta) := \lim_{\rho \to 0^+} \phi_\rho(\beta)$, and denote the SVD of V as

$$V = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1' \\ Q_2' \end{bmatrix}$$

Taking the limit as $\rho \to 0^+$ gives

$$\begin{aligned}\phi(\beta) &= \lim_{\rho \to 0^+} \frac{1}{2}\|y - X\beta\|^2_{V_\rho^{-1}} \\ &= \lim_{\rho \to 0^+} \frac{1}{2}\|y - X\beta\|^2_{Q_1(S_1+\rho I)^{-1}Q_1'} + \frac{1}{2\rho}\|y - X\beta\|^2_{Q_2 Q_2'} \\ &= \begin{cases} \frac{1}{2}\|y - X\beta\|^2_{V^+} & y - X\beta \in \mathcal{R}(V) \\ \infty & y - X\beta \notin \mathcal{R}(V) \end{cases}\end{aligned}$$

Because $\phi_\rho \to \phi$ and ϕ, ϕ_ρ are convex, we can use [95, Theorem 7.33] to show[29]

$$\lim_{\rho\to 0^+} \left\{ \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \phi_\rho(\beta) \right\} \subseteq \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \phi(\beta) \tag{106}$$

To show equality it suffices to show that each side of (106) are affine sets of equal dimension. Starting with the left-hand side, we have, by theorem 25,

$$\lim_{\rho\to 0^+} \left\{ \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \phi_\rho(\beta) \right\} = \lim_{\rho\to 0^+} \hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V_\rho) \tag{107a}$$

$$= \left\{ \lim_{\rho\to 0^+} (X'V_\rho^{-1}X)^+ X'V_\rho^{-1} y \right\} + \mathcal{N}(X) \tag{107b}$$

which is clearly affine with dimension $\dim(\mathcal{N}(X))$ when the limit exists, and it does by lemma 20. For the right-hand side, we note that

$$\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V) = \operatorname*{argmin}_{\beta\in\mathbb{R}^p} \phi(\beta) \tag{108}$$

Therefore, since $w \in \mathcal{R}(Z)$, the right-hand side is affine with dimension $\dim(\mathcal{N}(X))$ (theorem 36), and (106) must hold with equality. □

Proof of theorem 38. First, we have $w \in \mathcal{R}(Z)$ by lemma 33. Therefore, lemma 35 and (107) and (108) imply (31a) and (31b). The remaining equality (31c) follows by lemma 20. □

Finally, we can indirectly prove theorem 38 by equating the expressions (30b) and (31c).

Proof of theorem 38 *(indirect).* By lemma 33 and theorem 25,[30]

$$\begin{aligned}
\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V_\rho) &= \hat{\mathbb{B}}_{\mathrm{GLS}}(y, X, V_\rho^{-1}) \\
&= (X'V_\rho^{-1}X)^+ X'V_\rho^{-1} y + \mathcal{N}(X'V_\rho^{-1}X) \\
&= (X'V_\rho^{-1}X)^+ X'V_\rho^{-1} y + \mathcal{N}(X)
\end{aligned}$$

where we have used the fact $\mathcal{N}(X'V_\rho^{-1}X) = \mathcal{N}(V_\rho^{-1/2}X) = \mathcal{N}(X)$. Taking the limit of both sides completes the second part of the proof,

$$\begin{aligned}
\lim_{\rho\to 0^+} \hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V_\rho) &= \lim_{\rho\to 0^+} (X'V_\rho^{-1}X)^+ X'V_\rho^{-1} y + \mathcal{N}(X) \\
&= X^+(I - V(SVS)^+)y + \mathcal{N}(X)
\end{aligned}$$

[29] We have skipped some technical detail here for the sake of brevity. In particular, the following facts are required in the hypothesis of [95, Theorem 7.33]. Denote the c-sublevel set of a function $f : \mathbb{R}^n \to \mathbb{R}$ as $\operatorname{lev}_c f := \{ x \in \mathbb{R}^n \mid f(x) \le c \}$.

1. ϕ, ϕ_ρ are *level-bounded* (i.e., $\operatorname{lev}_c\phi, \operatorname{lev}_c\phi_\rho$ are bounded for all $c \in \mathbb{R}$), for all $\rho > 0$.
2. ϕ, ϕ_ρ are *lower semicontinuous* (i.e., $\operatorname{lev}_c\phi, \operatorname{lev}_c\phi_\rho$ are closed for all $c \in \mathbb{R}$ [95, Theorem 1.6]) and *proper* (i.e., ϕ, ϕ_ρ never equal $-\infty$ and do not always equal ∞), for all $\rho > 0$.

Both of these facts can easily be shown by observing that ϕ_ρ is strictly convex on $\mathbb{R}^p$, and ϕ is strictly convex on the feasible set $\{ \beta \in \mathbb{R}^p \mid y - X\beta \in \mathcal{R}(V) \}$.

[30] The constraints for this problem are trivial since V_ρ is nonsingular, so the ECGLS problem reduces to a GLS problem.

where the second equality follows from lemma 20. Finally, the proof is completed using lemma 21 and theorem 37,

$$X^+(I - V(SVS)^+)y + \mathcal{N}(X) = (X'V_0^+X)^+X'V_0^+y + \mathcal{N}(X) = \hat{\mathbb{B}}_{\text{MLE}}(y, X, V)$$

□

C Maximum a posteriori estimator proofs

To show theorems 40 to 42, we first require the following preliminary lemma.

Lemma 58. *For any $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, $\beta, \beta_0 \in \mathbb{R}^p$, and $\Sigma \in \mathbb{S}^p_+$, let*

$$\tilde{y} := \begin{bmatrix} y \\ \beta_0 \end{bmatrix}, \qquad \tilde{X} := \begin{bmatrix} X \\ I \end{bmatrix}, \qquad \tilde{V} := \begin{bmatrix} V & 0 \\ 0 & \Sigma \end{bmatrix} \tag{109}$$

and $L := \Sigma X'(V + X\Sigma X')^+$. Then the following statements are equivalent.

1. *$y - X\beta \in \mathcal{R}(V)$ and $\beta - \beta_0 \in \mathcal{R}(\Sigma)$.*
2. *$\tilde{y} - \tilde{X}\beta \in \mathcal{R}(\tilde{V})$.*
3. *$\beta - \beta_0 - L(y - X\beta_0) \in \mathcal{R}(\Sigma - LX\Sigma)$ and $y - X\beta_0 \in \mathcal{R}(V + X\Sigma X')$.*

Proof. (1. ⇔ 2.) The first two statements are easily shown to be equivalent via the definition of $\mathcal{R}(\cdot)$.

(1. ⇔ 3.) First, let $V_0 := V + X\Sigma X'$ and $\Sigma_0 := (\Sigma - LX\Sigma)$, and note the following identities due to corollary 4 and the fact that $\mathcal{R}(V) \subseteq \mathcal{R}(V_0)$ and $\mathcal{R}(X\Sigma^{1/2}) = \mathcal{R}(X\Sigma X') \subseteq \mathcal{R}(V_0)$,

$$\begin{aligned} LV &= \Sigma X'V_0^+V = \Sigma X'V_0^+V_0 - \Sigma X'V_0^+X\Sigma X' = \Sigma_0 X' \\ (I - XL)V_0 &= V + X\Sigma X' - X\Sigma X'V_0^+V_0 = V \\ X\Sigma_0 &= X\Sigma - V_0V_0^+X\Sigma + VV_0^+X\Sigma = VV_0^+X\Sigma \end{aligned}$$

(⇒) Suppose $y - X\beta \in \mathcal{R}(V)$ and $\beta - \beta_0 \in \mathcal{R}(\Sigma)$. Then there exist $\alpha_1 \in \mathbb{R}^n$ and $\alpha_2 \in \mathbb{R}^p$ such that $y - X\beta = V\alpha_1$ and $\beta - \beta_0 = \Sigma\alpha_2$. Moreover,

$$\begin{aligned} y - X\beta_0 = y - X\beta + X(\beta - \beta_0) = V\alpha_1 + X\Sigma\alpha_2 &= \begin{bmatrix} V^{1/2} & X\Sigma^{1/2} \end{bmatrix} \begin{bmatrix} V^{1/2}\alpha_1 \\ \Sigma^{1/2}\alpha_2 \end{bmatrix} \\ &\in \mathcal{R}\left(\begin{bmatrix} V^{1/2} & X\Sigma^{1/2} \end{bmatrix}\right) = \mathcal{R}\left(\begin{bmatrix} V^{1/2} & X\Sigma^{1/2} \end{bmatrix} \begin{bmatrix} V^{1/2} \\ \Sigma^{1/2}X' \end{bmatrix}\right) = \mathcal{R}(V_0) \end{aligned}$$

and

$$\begin{aligned} \beta - \beta_0 - L(y - X\beta_0) &= (I - LX)(\beta - \beta_0) - L(y - X\beta) \\ &= (I - LX)\Sigma\alpha_2 - LV\alpha_1 \\ &= \Sigma_0\alpha_2 - \Sigma_0X'\alpha_1 \in \mathcal{R}(\Sigma_0) \end{aligned}$$

($\Leftarrow$) Suppose $y - X\beta_0 \in \mathcal{R}(V_0)$ and $\beta - \beta_0 - L(y - X\beta_0) \in \mathcal{R}(\Sigma_0)$. Then there exist $\alpha_1 \in \mathbb{R}^n$ and $\alpha_2 \in \mathbb{R}^p$ such that $y - X\beta_0 = V_0\alpha_1$ and $\beta - \beta_0 - L(y - X\beta_0) = \Sigma_0\alpha_2$. Moreover,

$$\begin{aligned}\beta - \beta_0 &= \beta - \beta_0 - L(y - X\beta_0) + L(y - X\beta_0) = \Sigma_0\alpha_2 + L(y - X\beta_0)\\ &= \Sigma(I - X'V_0^+X\Sigma)\alpha_2 + \Sigma X'V_0(y - X\beta_0) \in \mathcal{R}(\Sigma)\end{aligned}$$

and

$$\begin{aligned}y - X\beta &= y - X\beta_0 - X(\beta - \beta_0 - L(y - X\beta_0) + L(y - X\beta_0))\\ &= (I - XL)(y - X\beta_0) - X(\beta - \beta_0 - L(y - X\beta_0))\\ &= (I - XL)V_0\alpha_1 - X\Sigma_0\alpha_2 = V\alpha_1 - VV_0^+X\Sigma\alpha_2 \in \mathcal{R}(V)\end{aligned}$$

□

Proof of theorem 40. To show (32), we show that the MAP and MLE problems have the same feasible set their objectives are proportional for all β in that feasible set.

Consider the shorthand notation (109) and the joint density of y and β,

$$\begin{bmatrix} y \\ \beta \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} X\beta_0 \\ \beta_0 \end{bmatrix}, \begin{bmatrix} V + X\Sigma X' & X\Sigma \\ X'\Sigma & \Sigma \end{bmatrix}\right)$$

By lemma 9, we have,

$$\beta|y \sim \mathrm{N}(\beta_0 + L(y - X\beta_0), \Sigma - LX\Sigma) \tag{110}$$

for all $y - X\beta_0 \in \mathcal{R}(V + X\Sigma X')$. Therefore, the probability density corresponding to (MAP) is given by

$$f(\beta|y) = (2\pi)^{-\frac{n}{2}}|\Sigma - LX\Sigma|_+^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\beta - \beta_0 - L(y - X\beta_0)\|^2_{(\Sigma - LX\Sigma)^+}\right) > 0 \tag{111}$$

whenever $\beta - \beta_0 - L(y - X\beta_0) \in \mathcal{R}(\Sigma - LX\Sigma)$ and $y - X\beta_0 \in \mathcal{R}(V + X\Sigma X')$, and $f(\beta|y) = 0$ otherwise. In other words, the MAP problem has the feasible set

$$\left\{\beta \in \mathbb{R}^n \;\middle|\; \begin{matrix} \beta - \beta_0 - L(y - X\beta_0) \in \mathcal{R}(\Sigma - LX\Sigma), \\ y - X\beta_0 \in \mathcal{R}(V + X\Sigma X') \end{matrix}\right\} \tag{112}$$

Note the second constraint $y - X\beta_0 \in \mathcal{R}(V + X\Sigma X')$ only serves to make the feasible set empty when the conditional density function is ill-defined. We can always choose $\beta = \beta_0 + L(y - X\beta_0)$ to satisfy the first constraint, so the feasible set is empty (and by implication, $\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$ is empty) if and only if $y - X\beta_0 \notin \mathcal{R}(V + X\Sigma X')$.

For the MLE problem, we have the probability density

$$f(\tilde{y}; \beta) = (2\pi)^{-\frac{n+p}{2}}|\tilde{V}|_+^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\tilde{y} - \tilde{X}\beta\|^2_{\tilde{V}^+}\right) > 0 \tag{113}$$

whenever $\tilde{y} - \tilde{X}\beta \in \mathcal{R}(\tilde{V})$ and $f(\tilde{y}; \beta) = 0$ otherwise. In other words, the MLE problem has the feasible set

$$\{\, \beta \in \mathbb{R}^p \mid \tilde{y} - \tilde{X}\beta \in \mathcal{R}(\tilde{V}) \,\} \tag{114}$$

By lemma 58, the feasible sets (112) and (114) are equivalent. Let β be in the feasible set. Then, by lemma 58, we also have $y - X\beta \in \mathcal{R}(V)$ and $\beta - \beta_0 \in \mathcal{R}(\Sigma)$. Using basic facts about the pseudoinverse and pseudodeterminant of block diagonal matrices,[31] we can rewrite (113) as

$$\begin{aligned} f(\tilde{y};\beta) &= (2\pi)^{-\frac{p}{2}}|V|_+^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|y - X\beta\|_{V^+}^2\right)(2\pi)^{-\frac{n}{2}}|\Sigma|_+^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\|\beta_0 - \beta\|_{\Sigma^+}^2\right) \\ &= f(y|\beta)f(\beta) \end{aligned}$$

which is clearly proportional to (111) by Bayes' theorem. □

Proof of theorem 41. To simplify the notation, let $\overline{y} := \mathbb{E}[\beta|y] = \beta_0 + L(y - X\beta_0)$, $\overline{X} := I$, $\overline{S} := I - \overline{X}\overline{X}^+ = 0$, and $\overline{V} := \mathrm{var}[\beta|y] = \Sigma - LX\Sigma$, where the formula for the expectation and variance follow from (110). By theorem 40, $\hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma)$ being nonempty implies that $y - X\beta_0 \in \mathcal{R}(V + X\Sigma X')$, and the probability density corresponding to (MAP) is given by (111) whenever

$$\overline{y} - \overline{X}\beta = \beta_0 + L(y - X\beta_0) \in \mathcal{R}(\Sigma - LX\Sigma) = \mathcal{R}(\overline{V})$$

and $f(\beta|y) = 0$ otherwise. But the probability density (111) is clearly equivalent to the probability density corresponding to the MLE problem

$$\hat{\mathbb{B}}_{\mathrm{MLE}}(\beta_0 + L(y - X\beta_0), I, \Sigma - LX\Sigma) = \hat{\mathbb{B}}_{\mathrm{MLE}}(\overline{y}, \overline{X}, \overline{V})$$

and by theorem 38, we have

$$\begin{aligned} \hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma) &= \hat{\mathbb{B}}_{\mathrm{MLE}}(\overline{y}, \overline{X}, \overline{V}) \\ &= \overline{X}^+(I - \overline{V}\overline{S}(\overline{S}\overline{V}\overline{S})^+\overline{S})\overline{y} + \mathcal{N}(\overline{X}) \\ &= \{\, \overline{y} \,\} = \{\, \mathbb{E}[\beta \mid y] \,\} = \{\, \beta_0 + L(y - X\beta_0) \,\} \end{aligned}$$

where the third equality follows from the fact that $\overline{S} = 0$ and $\mathcal{N}(\overline{X}) = \{\, 0 \in \mathbb{R}^p \,\}$. □

Proof of theorem 42. Equations (34a) and (34c) follow from theorem 41. To finish the proof we show that the first and last expressions are equivalent,

$$\begin{aligned} \hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V, \beta_0, \Sigma) &= \hat{\mathbb{B}}_{\mathrm{MLE}}\left(\begin{bmatrix} y \\ \beta_0 \end{bmatrix}, \begin{bmatrix} X \\ I \end{bmatrix}, \begin{bmatrix} V & 0 \\ 0 & \Sigma \end{bmatrix}\right) \\ &= \lim_{\rho \to 0^+} \hat{\mathbb{B}}_{\mathrm{MLE}}\left(\begin{bmatrix} y \\ \beta_0 \end{bmatrix}, \begin{bmatrix} X \\ I \end{bmatrix}, \begin{bmatrix} V & 0 \\ 0 & \Sigma \end{bmatrix} + \rho I\right) \\ &= \lim_{\rho \to 0^+} \hat{\mathbb{B}}_{\mathrm{MAP}}(y, X, V + \rho I, \beta_0, \Sigma + \rho I) \end{aligned}$$

where the first and last equalities follow by theorem 41 and the second equality follows by theorem 38. □

[31] For all $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{p\times p}$, $\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}^+ = \begin{bmatrix} A^+ & 0 \\ 0 & B^+ \end{bmatrix}$ and $\left|\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right|_+ = |A|_+ \cdot |B|_+$.

D Best affine unbiased estimator proofs

In order to prove lemma 44, theorem 45, , and corollary 46, we first require the following preliminary lemma.

Lemma 59. *Let $X \in \mathbb{R}^{n\times p}$, $V \in \mathbb{S}^n_+$, $W \in \mathbb{R}^{m\times n}$, $V_0 := V + XEX'$ for any $E \in \mathbb{S}^p_+$ such that $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, and $\hat{\mathbb{A}}(X, V, W)$ be the set of solutions to* (MTP)*. Then the following statements hold.*

1. $\mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) = \mathcal{R}(V_0)$.
2. $\widehat{\mathbb{WB}}_{\text{AUE}}(X, V, W) = \{\, A(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m \mid AX = W \,\}$.
3. $\widehat{\mathbb{WB}}_{\text{BAUE}}(X, V, W)$ *is nonempty if and only if* $\mathcal{R}(W') \subseteq \mathcal{R}(X')$.
4. $\hat{\mathbb{A}}(X, V, W)$ *is nonempty if and only if* $\mathcal{R}(W') \subseteq \mathcal{R}(X')$.
5. $\widehat{\mathbb{WB}}_{\text{BAUE}}(X, V, W) \subseteq \{\, \hat{A}(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m \mid \hat{A} \in \hat{\mathbb{A}}(X, V, W) \,\}$.

Proof. (1.) Noting that $\mathcal{R}(V) \subseteq \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$, $\mathcal{R}(X) \subseteq \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$, $\mathcal{R}(V) \subseteq \mathcal{R}(V_0)$, and $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, we have

$$V_0 V_0^+ \begin{bmatrix} V & X \end{bmatrix} = \begin{bmatrix} V_0 V_0^+ V & V_0 V_0^+ X \end{bmatrix} = \begin{bmatrix} V & X \end{bmatrix}$$

and

$$\begin{aligned} \begin{bmatrix} V & X \end{bmatrix} \begin{bmatrix} V & X \end{bmatrix}^+ V_0 &= \begin{bmatrix} V & X \end{bmatrix} \begin{bmatrix} V & X \end{bmatrix}^+ V + \begin{bmatrix} V & X \end{bmatrix} \begin{bmatrix} V & X \end{bmatrix}^+ XEX' \\ &= V + XEX' = V_0 \end{aligned}$$

which implies that $\mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) = \mathcal{R}(V_0)$ (corollary 4).

(2.) Let $\widehat{\mathbb{WB}}(X, W) := \{\, A(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m \mid AX = W \,\}$. Note that all functions in $\widehat{\mathbb{WB}}_{\text{AUE}}(X, V, W)$ and $\widehat{\mathbb{WB}}(X, W)$ have the same domain because $\mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) = \mathcal{R}(V_0)$.

($\subseteq$) Suppose $\theta(\cdot) = \hat{A}(\cdot) + \hat{c} \in \widehat{\mathbb{WB}}_{\text{AUE}}(X, V, W)$. Then

$$W\beta = \mathbb{E}[\theta(y)|\beta] = \mathbb{E}[\hat{A}y + \hat{c}|\beta] = \hat{A}X\beta + \hat{c}$$

for all $\beta \in \mathbb{R}^p$. For this to be true, we must have $\hat{c} = 0$ and $\hat{A}X = W$. Therefore, $\theta(\cdot) = \hat{A}(\cdot) \in \widehat{\mathbb{WB}}(X, W)$.

($\supseteq$) Suppose $\theta(\cdot) = \hat{A}(\cdot) \in \widehat{\mathbb{WB}}(X, W)$. Then

$$\mathbb{E}[\theta(y)|\beta] = \mathbb{E}[\hat{A}y|\beta] = \hat{A}X\beta = W\beta$$

for any $\beta \in \mathbb{R}^p$, and therefore $\theta \in \widehat{\mathbb{WB}}_{\text{AUE}}(X, V, W)$.

(3–4.) According to the second part of this lemma, the feasible set of the BAUE "optimization" problem $\widehat{\mathbb{WB}}_{\text{AUE}}(X, V, W) = \widehat{\mathbb{WB}}(X, W)$ is nonempty if and only if there exists $A \in \mathbb{R}^{m\times n}$ such that $AX = W$. Likewise, the feasible set of (MTP) is nonempty if

and only if there exists $A \in \mathbb{R}^{m\times n}$ such that $AX = W$. But $AX = W$ has solutions if and only if $\mathcal{R}(W') \subseteq \mathcal{R}(X')$ by corollary 4, so the third and fourth statements are true.

(5.) Suppose $\theta \in \widehat{\mathbb{WB}}_{\text{BAUE}}(X, V, W) \subseteq \widehat{\mathbb{WB}}(X, W)$. Then $\theta(\cdot) = \hat{A}(\cdot)$ for some $\hat{A} \in \mathbb{R}^{m\times n}$. It suffices to show $\hat{A}$ is a solution to (MTP). By (BAUE) and (2.) of this lemma, we have

$$\text{var}[\theta(y)|\beta] \preceq \text{var}[\tilde{\theta}(y)|\beta] \qquad \forall \beta \in \mathbb{R}^p, \tilde{\theta} \in \widehat{\mathbb{WB}}_{\text{AUE}}(X, V, W) = \widehat{\mathbb{WB}}(X, W) \tag{115}$$

But $\theta, \tilde{\theta} \in \widehat{\mathbb{WB}}_{\text{BAUE}}(X, V, W)$ implies that $\theta(\cdot) = \hat{A}(\cdot)$, $\tilde{\theta}(\cdot) = \tilde{A}(\cdot)$, $\hat{A}X = W$ and $\tilde{A}X = W$ for some $\hat{A}, \tilde{A} \in \mathbb{R}^{m\times n}$. Rewriting (115) in terms of $\hat{A}, \tilde{A}$,

$$\hat{A}V\hat{A}' \preceq \tilde{A}V\tilde{A}' \qquad \forall \tilde{A} \in \{\, A \in \mathbb{R}^{m\times n} \mid AX = W \,\} \tag{116}$$

The Loewner is preserved under trace[32] so (116) implies

$$\text{tr}(\hat{A}V\hat{A}') \leq \text{tr}(\tilde{A}V\tilde{A}') \qquad \forall \tilde{A} \in \{\, A \in \mathbb{R}^{m\times n} \mid AX = W \,\}$$

and therefore $\hat{A}$ is a solution to (MTP). □

Proof of lemma 44. Let $V_0 := V + XEX'$ for any $E \in \mathbb{S}^p_+$ such that $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$. Then $\mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) = \mathcal{R}(V_0)$ by lemma 59. Suppose (35) holds and $\widehat{\mathbb{WB}}_{\text{BAUE}}(X, V, W)$ is nonempty. Then $\hat{\mathbb{A}}(X, V, W)$ is nonempty (lemma 59) and the functions $\hat{A}_1(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m$ and $\hat{A}_2(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m$ are equivalent for any $\hat{A}_1, \hat{A}_2 \in \hat{\mathbb{A}}(X, V, W)$. We can rewrite the set of so-called *minimum trace estimators* as a singleton,

$$\{\, \hat{A}(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m \mid \hat{A} \in \hat{\mathbb{A}}(X, V, W) \,\} = \{\, \hat{A}_1(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m \,\} \tag{117}$$

for any $\hat{A}_1 \in \hat{\mathbb{A}}(X, V, W)$. By lemma 59, we have

$$\begin{aligned} \widehat{\mathbb{WB}}_{\text{BAUE}}(X, V, W) &\subseteq \{\, \hat{A}(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m \mid \hat{A} \in \hat{\mathbb{A}}(X, V, W) \,\} \\ &= \{\, \hat{A}_1(\cdot) : \mathcal{R}(V_0) \to \mathbb{R}^m \,\} \end{aligned} \tag{118}$$

But $\widehat{\mathbb{WB}}_{\text{BAUE}}(X, V, W)$ is nonempty, so (118) must hold with equality, which demonstrates (36). □

Proof of theorem 45. We aim to use lemma 44 to derive the BAUE. Let $W_0 := X'V_0^+X$ and suppose $\widehat{\mathbb{WB}}_{\text{BAUE}}(X, V, W)$ is nonempty. Then $\hat{\mathbb{A}}(X, V, W)$ is nonempty by lemma 59. It suffices to show that, $WW_0^+X'V_0^+ \in \hat{\mathbb{A}}(X, V, W)$ and (35) holds.

We first solve (MTP) by the method of Lagrange multipliers. The Lagrangian is defined as

$$\mathcal{L}(A, \Lambda) = \frac{1}{2}\text{tr}(AVA') + \text{tr}(\Lambda'(AX - W))$$

[32] That is, $A \preceq B$ implies $\text{tr}A \leq \text{tr}B$ for all A, B of suitable dimensions.

Since (MTP) has a convex objective and linear constraints, lemma 10 implies that $\hat{A} \in \hat{\mathbb{A}}(X, V, W)$ if and only if there exits $\hat{\Lambda} \in \mathbb{R}^{m\times n}$ such that[33]

$$\frac{\partial \mathcal{L}}{\partial A}(\hat{A}, \hat{\Lambda}) = V\hat{A}' + X\hat{\Lambda}' = 0, \qquad \frac{\partial \mathcal{L}}{\partial \Lambda}(\hat{A}, \hat{\Lambda}) = \hat{A}X - W = 0$$

or equivalently

$$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} \hat{A}' \\ \hat{\Lambda}' \end{bmatrix} = \begin{bmatrix} 0 \\ W' \end{bmatrix} \tag{119}$$

By corollary 4 and lemma 12, $\mathcal{R}(W') \subseteq \mathcal{R}(X') = \mathcal{R}(W_0)$ and $W_0 W_0^+ W' = W'$. Using lemma 13,

$$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}^+ \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} 0 \\ W' \end{bmatrix} = \begin{bmatrix} V_0 V_0^+ & 0 \\ 0 & W_0 W_0^+ \end{bmatrix} \begin{bmatrix} 0 \\ W' \end{bmatrix} = \begin{bmatrix} 0 \\ W' \end{bmatrix}$$

By corollary 4, $(\hat{A}, \hat{\Lambda})$ is a solution to (119) if and only if

$$\begin{aligned} \begin{bmatrix} \hat{A}' \\ -\hat{\Lambda}' \end{bmatrix} &\in \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}^+ \begin{bmatrix} 0 \\ W' \end{bmatrix} + \left\{ \left(I - \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}^+ \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \right) Q \;\middle|\; Q \in \mathbb{R}^{n+p\times m} \right\} \\ &= \begin{bmatrix} V_0^+ X W_0^+ W' \\ W_0 W_0^+ E W_0 W_0^+ W' - W_0^+ W' \end{bmatrix} + \left\{ \begin{bmatrix} (I - V_0 V_0^+) Q_1 \\ (I - W_0 W_0^+) Q_2 \end{bmatrix} \;\middle|\; \begin{bmatrix} Q_1 \\ Q_2 \end{bmatrix} \in \mathbb{R}^{n+p\times m} \right\} \end{aligned}$$

where the equality follows from lemma 13. In other words, there exists $\hat{\Lambda} \in \mathbb{R}^{m\times n}$ such that $\hat{A}$ solves (119) if and only if

$$\hat{A} \in \hat{\mathbb{A}}(X, V, W) = W W_0^+ X' V_0^+ + \{\, Q(I - V_0 V_0^+) \mid Q \in \mathbb{R}^{m\times n} \,\}$$

Choosing $Q = 0$ shows that $W W_0^+ X' V_0^+ \in \hat{\mathbb{A}}(X, V, W)$. Let $\hat{A}_1, \hat{A}_2 \in \hat{\mathbb{A}}(X, V, W)$ and $y \in \mathcal{R}(V_0)$. Then there exists $Q_1, Q_2 \in \mathbb{R}^{m\times n}$ such that

$$\hat{A}_i = W W_0^+ X' V_0^+ + Q_i (I - V_0 V_0^+)$$

for $i = 1, 2$. Moreover, $(I - V_0 V_0^+) y = 0$ and $\hat{A}_1 y = \hat{A}_2 y = W W_0^+ X' V_0^+ y$, which demonstrates (35). □

Proof of corollary 46. Let $V_0 := V + XEX'$ for any $E \in \mathbb{S}_+^p$ such that $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$. Then $\mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) = \mathcal{R}(V_0)$ by lemma 59. Suppose $y \in \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) = \mathcal{R}(V_0)$ and $\widehat{W\beta} \in \widehat{\mathbb{WB}}_{\mathrm{BAUE}}(X, V, W)$. Then $\widehat{W\beta}(y) = W(X' V_0^+ X)^+ X' V_0^+ y$ by theorem 45. Since $\widehat{\mathbb{WB}}_{\mathrm{BAUE}}(X, V, W)$ was assumed nonempty, we have $\mathcal{R}(W') \subseteq \mathcal{R}(X')$ by lemma 59. Then theorem 37 gives

$$W\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V) = W(X' V_0^+ X)^+ X' V_0^+ y + W\mathcal{N}(X) = \{\, W(X' V_0^+ X)^+ X' V_0^+ y \,\}$$

where $W\mathcal{N}(X) = \{\, W(I - X^+ X) q \mid q \in \mathbb{R}^p \,\} = \{\, 0 \in \mathbb{R}^p \,\}$ by corollary 4. Finally, we have that $\{\, \widehat{W\beta}(y) = W(X' V_0^+ X)^+ X' V_0^+ y \,\} = W\hat{\mathbb{B}}_{\mathrm{MLE}}(y, X, V)$. □

[33] While we have shifted to matrix arguments from vector arguments, lemma 10 still applies. To see this, consider vectorization of the constraint $AX = W$ and note the identity $\mathrm{tr}(A'B) = [\mathrm{vec}(A)]' \mathrm{vec}(B)$. Since vectorization is bijective and differentiable, the unvectorized derivatives are zero if and only if the vectorized derivatives are zero.

E Background

In this appendix we collect background results which are referenced in the main text and subsequent appendices.

E.1 Linear algebra

Throughout we use the following properties of the range and null spaces,

$$\mathcal{R}(A) \subseteq \mathcal{R}(\begin{bmatrix} A & B \end{bmatrix}), \qquad \mathcal{R}(A) \supseteq \mathcal{R}(AB)$$

and

$$\mathcal{R}(A) \subseteq \mathcal{R}(B) \qquad \Leftrightarrow \qquad \mathcal{N}(A') \supseteq \mathcal{N}(B')$$

for A and B of suitable dimension. We present Woodbury's matrix identity and two corollaries to it [41].

Theorem 60. *For any $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{m\times m}$, and $D \in \mathbb{R}^{m\times n}$,*

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

subject to existence of the inverses.

Proof of theorem 60. First note the following identities,

$$\begin{aligned} BC(C^{-1} + DA^{-1}B) &= B + BCDA^{-1}B = (A + BCD)A^{-1}B \\ (A + BCD)^{-1}BC &= A^{-1}B(C^{-1} + DA^{-1}B)^{-1} \end{aligned}$$

where the second follows from the first by multiplying $(A + BCD)^{-1}$ on the left and $(C^{-1} + DA^{-1}B)^{-1}$ on the right. Then

$$\begin{aligned} A^{-1} &= (A + BCD)^{-1}(A + BCD)A^{-1} \\ &= (A + BCD)^{-1} + (A + BCD)^{-1}BCDA^{-1} \\ &= (A + BCD)^{-1} + A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \end{aligned}$$

□

Below, we prove theorem 1.

Proof of theorem 1. Existence follows from the fact that $X = V_1\Sigma_1^{-1}U_1'$ is a solution to (6), where $A = U_1\Sigma_1V_1'$ is the economic SVD of A.

To show uniqueness, suppose both $\tilde{X}, X \in \mathbb{R}^{m\times n}$ are solutions to (6). Then by the third equality of (6), we have $A\tilde{X} = AXA\tilde{X}$, and by symmetry of AX and $A\tilde{X}$, we can take the transpose to show $A\tilde{X} = A\tilde{X}AX$. We also have $\tilde{X}A = \tilde{X}AXA = XA\tilde{X}A$ by a similar argument using the last equality of (6). Combining these results gives

$$\tilde{X} = \tilde{X}A\tilde{X} = \tilde{X}A\tilde{X}AX = \tilde{X}AX = XA\tilde{X}AX = XAX = X$$

and clearly $A^+ = X = \tilde{X}$ is the unique solution to (6). □

E.2 Projectors

Below, we prove lemmas 5 and 61 which are stated but not proven in section 2.

Proof of lemma 5. The first identity is shown by noting that $PP = P$ is symmetric and $PPP = P$, which covers all conditions in (6). The second identity is also shown by substitution into (6),

$$\begin{aligned}
(PA)^+PPA &= (PA)^+PA \\
PA(PA)^+P &= (PA(PA)^+)'P' = (PPA(PA)^+)' \\
&= (PA(PA)^+)' = PA(PA)^+ \\
PA(PA)^+PPA &= PA(PA)^+PA = PA \\
(PA)^+PPA(PA)^+P &= (PA)^+PA(PA)^+P = (PA)^+P
\end{aligned}$$

□

Lemma 61. *If $P \in \mathbb{R}^{n\times n}$ is a projector such that $0 < \operatorname{rank}(P) < n$, then $\|P\| \geq 1$ and $\|P\| = \|I - P\|$.*

Proof of lemma 61. For the first part, we have $\|P\| = \|P^2\| \leq \|P\|^2$ which can only be true if $\|P\| \geq 1$ or $\|P\| = 0$. However, $\|P\| \neq 0$ because $\operatorname{rank}(P) > 0$, so $\|P\| \geq 1$.

For the second part, it suffices to show $\|P\| \leq \|I - P\|$ since $I - P$ is a projector and therefore $\|I - P\| \leq \|I - (I - P)\| = \|P\|$. Let $u \in \mathbb{R}^p$ such that $\|u\| = 1$. Denote $x = Pu$ and $y = u - x = (I - P)u$. If $x = 0$, we have $\|Pu\| = 0$. If $y = 0$, then $\|Pu\| = \|u\| = 1 \leq \|I - P\|$. If $x \neq 0$ and $y \neq 0$ (which is guaranteed to happen by the rank constraint), then let $w = \tilde{x} + \tilde{y}$ where

$$\tilde{x} = \frac{\|y\|}{\|x\|}x, \qquad \tilde{y} = \frac{\|x\|}{\|y\|}y$$

Then $\|w\|^2 = \|y\|^2 + \|x\|^2 + 2x'y = \|u\|^2 = 1$, and we have the equivalence

$$\|Pu\| = \|x\| = \|\tilde{y}\| = \|(I - P)w\| \leq \|I - P\|$$

Taking the maximum of both sides over all $\|u\| = 1$ gives the result. □

E.3 Linear equations

Below, we prove lemma 3 and corollary 4 which was stated but not proven in section 2.

Proof of lemma 3. (1. $\Leftrightarrow$ 2.) By definition, $Ax = b$ has solutions if and only if $b \in \mathcal{R}(A)$.

(1. $\Rightarrow$ 3.) Suppose there exists $x \in \mathbb{R}^p$ such that $Ax = b$. Then $b = Ax = AA^+Ax = AA^+b$.

(1. $\Leftarrow$ 3.) Suppose $AA^+b = b$. Then with $x = A^+b$, we have $Ax = AA^+b = b$.

Suppose $b \in \mathcal{R}(A)$. Then with $x_0 = x - A^+b$, we can rewrite the (nonempty) solution set as desired,

$$\begin{aligned}\mathcal{S} = \{\, x \in \mathbb{R}^p \mid Ax = b \,\} &= \{\, A^+b + x_0 \in \mathbb{R}^p \mid A(A^+b + x_0) = b \,\} \\ &= \{\, A^+b + x_0 \in \mathbb{R}^p \mid Ax_0 = 0 \,\} \\ &= A^+b + \{\, x_0 \in \mathbb{R}^p \mid Ax_0 = 0 \,\} \\ &= A^+b + \mathcal{N}(A)\end{aligned}$$

□

Proof of corollary 4. The above statements can each be rewritten,

1. the linear vector equation $Ax_i = b_i$ has a solution for x_i for $i = 1, \ldots, p$,
2. $b_i \in \mathcal{R}(A)$ for $i = 1, \ldots, p$,
3. $AA^+b_i = b_i$ for $i = 1, \ldots, p$,

where x_i and b_i are the i-th columns of X and B. The result follows by lemma 3. □

E.4 Singular value decomposition

Proof of lemma 6. These identities follow directly from the definition of orthogonal matrices ($Q'Q = I$) and substitution into (6). □

E.5 The matrix 2-norm

The 2-norm is convex and submultiplicative, that is

$$\|A + B\| \leq \|A\| + \|B\|, \qquad \|AC\| \leq \|A\|\|C\|$$

for any $A, B \in \mathbb{R}^{m\times n}$ and $C \in \mathbb{R}^{n\times p}$. Under an orthogonal transformation, the vector 2-norm is preserved,

$$\|Ux\| = \sqrt{x'U'Ux} = \sqrt{x'x} = \|x\|$$

for any orthogonal matrix $U \in \mathbb{R}^{m\times n}$ and vector $x \in \mathbb{R}^p$. Therefore the matrix 2-norm is also preserved,

$$\|UAV'\| = \max_{x\in\mathbb{R}^q} \frac{\|UAV'x\|}{\|x\|} = \max_{x\in\mathbb{R}^q} \frac{\|AV'x\|}{\|x\|} = \max_{z\in\mathbb{R}^n} \frac{\|AV'Vz\|}{\|Vz\|} = \max_{z\in\mathbb{R}^n} \frac{\|Az\|}{\|z\|} = \|A\|$$

for any orthogonal matrices $U \in \mathbb{R}^{m\times n}$, $V \in \mathbb{R}^{q\times p}$ and matrix $A \in \mathbb{R}^{n\times p}$. From this result, we can write the 2-norm of a matrix and its pseudoinverse in terms of the singular values,

$$\|A\| = \|U_1\Sigma_1V_1'\| = \|\Sigma_1\| = \sigma_1, \qquad \|A^+\| = \|U_1\Sigma_1^{-1}V_1'\| = \|\Sigma_1^{-1}\| = \sigma_r^{-1}$$

given the SVD (8). It is also clear that $\|U\| = 1$ for any orthogonal $U \in \mathbb{R}^{n\times m}$, and $\|AA'\| = \|U_1\Sigma_1^2U_1'\| = \|\Sigma_1^2\| = \|A\|^2$ for any $A \in \mathbb{R}^{n\times m}$.

E.6 Matrix limits

Proof of (7). Rewriting $R(\alpha)$ in terms of the SVD matrices,

$$\begin{aligned} R(\alpha) &= (A'A + \alpha I)^{-1}(A' - (A'A + \alpha I)A^+) \\ &= (A'A + \alpha I)^{-1}(A' - A'AA^+ - \alpha A^+) \\ &= -\alpha(A'A + \alpha I)^{-1}A^+ \\ &= -\alpha(V_1\Sigma_1^2V_1' + \alpha I)^{-1}V_1\Sigma_1^{-1}U_1' \end{aligned}$$

and using theorem 60,

$$\begin{aligned} R(\alpha) &= -\alpha(\alpha^{-1}I - \alpha^{-2}V_1(\alpha^{-1}I + \Sigma_1^{-2})^{-1}V_1')V_1\Sigma_1^{-1}U_1' \\ &= -V_1[\alpha(\alpha^{-1}I - \alpha^{-2}(\alpha^{-1}I + \Sigma_1^{-2})^{-1})\Sigma_1^{-1}]U_1' \\ &= -V_1[\alpha(\alpha I + \Sigma_1^2)^{-1}\Sigma_1^{-1}]U_1' \end{aligned}$$

where $r = \text{rank}(A)$. Noting that the final expression of $R(\alpha)$ is a SVD (up to the ordering of the singular values) with maximum singular value $\frac{\alpha}{\underline{\sigma}(A)(\underline{\sigma}^2(A))+\alpha}$, we have

$$\|R(\alpha)\| = \frac{\alpha}{\underline{\sigma}(A)(\underline{\sigma}^2(A)) + \alpha}$$

and therefore $\lim_{\alpha\to 0^+}\|R(\alpha)\| = \lim_{\alpha\to 0^+}\frac{\alpha}{\underline{\sigma}(A)(\underline{\sigma}^2(A))+\alpha} = 0$ and $\lim_{\alpha\to 0^+} R(\alpha) = 0$. □

E.7 Probability

Below, we prove lemma 9 using the method outlined by Marsaglia [69].

Proof of lemma 9. First note that since the joint covariance

$$\Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}' & \Sigma_y \end{bmatrix}$$

is positive semidefinite, there exist $U \in \mathbb{R}^{n\times r}$ and $V \in \mathbb{R}^{m\times r}$, where $r = \text{rank}(\Sigma)$, such that

$$\Sigma = \begin{bmatrix} U \\ V \end{bmatrix}\begin{bmatrix} U \\ V \end{bmatrix}' = \begin{bmatrix} UU' & UV' \\ VU' & VV' \end{bmatrix}$$

and therefore $\mathcal{R}(\Sigma_{xy}') = \mathcal{R}(VU') \subseteq \mathcal{R}(V) = \mathcal{R}(VV') = \mathcal{R}(\Sigma_y)$, and by corollary 4, we have $\Sigma_{xy}\Sigma_y^+\Sigma_y = \Sigma_{xy}$. Let $z = x - \Sigma_{xy}\Sigma_y^+ y$. Using the linearity of Gaussians, we have the joint distribution of

$$\begin{bmatrix} z \\ y \end{bmatrix} = \begin{bmatrix} I & -\Sigma_{xy}\Sigma_y^+ \\ 0 & I \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}$$

is Gaussian with mean

$$\mathbb{E}\left[\begin{bmatrix} z \\ y \end{bmatrix}\right] = \begin{bmatrix} I & -\Sigma_{xy}\Sigma_y^+ \\ 0 & I \end{bmatrix}\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = \begin{bmatrix} \mu_x - \Sigma_{xy}\Sigma_y^+\mu_y \\ \mu_y \end{bmatrix}$$

and covariance

$$\begin{aligned}\operatorname{var}\left(\begin{bmatrix} z \\ y \end{bmatrix}\right) &= \begin{bmatrix} I & -\Sigma_{xy}\Sigma_y^+ \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}' & \Sigma_y \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_y^+\Sigma_{xy}' & I \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_x - \Sigma_{xy}\Sigma_y^+\Sigma_{xy}' & 0 \\ \Sigma_{xy}' & \Sigma_y \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_y^+\Sigma_{xy}' & I \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_x - \Sigma_{xy}\Sigma_y^+\Sigma_{xy}' & 0 \\ 0 & \Sigma_y \end{bmatrix}\end{aligned}$$

Since z and y are uncorrelated and Gaussian, they are independent. Then for any $a \in \mathcal{R}(\Sigma_y)$,

$$x|\{\, y = a \,\} = z + \Sigma_{xy}\Sigma_y^+ a \sim \mathrm{N}(\mu_x + \Sigma_{xy}\Sigma_y^+(y - \mu_y), \Sigma_x - \Sigma_{xy}\Sigma_y^+\Sigma_{xy}')$$

□

E.8 Optimization

Below, we prove lemma 10 using the method outlined in [18, pp. 141–142].

Proof of lemma 10. By lemma 3, the feasible set $\mathcal{S} = \{\, x \in \mathbb{R}^p \mid Ax = b \,\}$ is nonempty and $\mathcal{S} = A^+b + \mathcal{N}(A)$. We have that $x^0 \in \mathbb{R}^p$ solves (11) if and only if $x^0 \in \mathcal{S}$ and

$$(x - x^0)'\frac{df}{dx}(x^0) \geq 0 \qquad \forall x \in \mathcal{S} \tag{120}$$

Moreover, if $x^0 \in \mathcal{S}$, we have that $Ax^0 = b$, $A^+b - x^0 \in \mathcal{N}(A)$, and $\mathcal{S} = x^0 + \mathcal{N}(A)$. Therefore we can rewrite the condition (120) as

$$v'\frac{df}{dx}(x^0) \geq 0 \qquad \forall v \in \mathcal{N}(A) \tag{121}$$

But $v'(df/dx)(x^0)$ is linear in v, so for it to be nonnegative for all $v \in \mathcal{N}(A)$, it must be zero for all $v \in \mathcal{N}(A)$. Therefore (121) is equivalent to

$$v'\frac{df}{dx}(x^0) = 0 \qquad \forall v \in \mathcal{N}(A) \tag{122}$$

As a range space condition, (122) can be written as $(df/dx)(x^0) \in \mathcal{R}(A')$, which is true if and only if there exists $\lambda^0 \in \mathbb{R}^p$ such that

$$\frac{df}{dx}(x^0) + A'\lambda^0 = 0 \tag{123}$$

Taking derivatives of the Lagrangian (12), we get that $x^0 \in \mathcal{S}$ and (123) are collectively equivalent to (13). □

F Block matrix pseudoinversion proof

In this appendix, we prove lemmas 12 and 13. In this appendix, we prove lemmas 12 and 13 and corollary 14

Proof of lemma 12. (1.) $\mathcal{R}(V) \subseteq \mathcal{R}(V + XEX') = \mathcal{R}(V_0)$.

(2.) The hypothesis $\mathcal{R}(X) \subseteq \mathcal{R}(V_0)$ and first statement $\mathcal{R}(V) \subseteq \mathcal{R}(V_0)$ are equivalent to $V_0 V_0^+ X = X$ and $V_0 V_0^+ V = V$ by corollary 4.

(3.) First, let $F := (V_0^+)^{1/2}$ so that $\mathcal{R}(X') \supseteq \mathcal{R}(X'F) = \mathcal{R}(X'V_0^+X) = \mathcal{R}(W_0)$. Next we show $\mathcal{R}(X') \subseteq \mathcal{R}(W_0)$. Let $G := X'F$ so that

$$W_0 = X'V_0^+X = X'F^2X = GG'$$

By the second statement and symmetry of V_0 and V_0^+, we have

$$X' = (V_0V_0^+X)' = X'V_0^+V_0$$

Finally, using properties of the psuedoinverse, we have

$$\begin{aligned} W_0W_0^+X' &= W_0W_0^+X'V_0^+V_0 = (GG')(GG')^+GFV_0 \\ &= GFV_0 = X'F^2V_0 = X'V_0^+V_0 = X' \end{aligned}$$

which by corollary 4 is equivalently stated $\mathcal{R}(X') = \mathcal{R}(W_0)$. □

Proof of lemma 13. Let N be defined as

$$N = \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} = \begin{bmatrix} V_0^+ - V_0^+XW_0^+X'V_0^+ & V_0^+XW_0^+ \\ W_0^+X'V_0^+ & W_0W_0^+EW_0W_0^+ - W_0^+ \end{bmatrix}.$$

Then we can write MN as

$$MN = \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} \begin{bmatrix} N_{11} & N_{12} \\ N_{21} & N_{22} \end{bmatrix} = \begin{bmatrix} VN_{11} + XN_{21} & VN_{12} + XN_{22} \\ X'N_{11} & X'N_{12} \end{bmatrix}$$

Using lemma 12 we can rewrite each block in MN as

$$\begin{aligned}
VN_{11} + XN_{21} \\
&= (V_0 - XEX')(V_0^+ - V_0^+XW_0^+X'V_0^+) + XW_0^+X'V_0^+ \\
&= V_0V_0^+ - XEX'V_0^+ - V_0V_0^+XW_0^+X'V_0^+ + XEW_0W_0^+X'V_0^+ + XW_0^+X'V_0^+ \\
&= V_0V_0^+ - XEX'V_0^+ - XW_0^+X'V_0^+ + XEX'V_0^+ + XW_0^+X'V_0^+ \\
&= V_0V_0^+ \\
VN_{12} + XN_{22} \\
&= (V_0 - XEX')V_0^+XW_0^+ + X(W_0W_0^+EW_0W_0^+ - W_0^+) \\
&= V_0V_0^+XW_0^+ - XEW_0W_0^+ + XW_0W_0^+EW_0W_0^+ - XW_0^+ \\
&= XW_0^+ - XEW_0W_0^+ + XEW_0W_0^+ - XW_0^+ \\
&= 0 \\
X'N_{11} &= X'V_0^+ - X'V_0^+XW_0^+X'V_0^+ \\
&= X'V_0^+ - W_0W_0^+X'V_0^+ \\
&= X'V_0^+ - X'V_0^+ \\
&= 0 \\
X'N_{12} &= X'V_0^+XW_0^+ \\
&= W_0W_0^+
\end{aligned}$$

which gives

$$MN = \begin{bmatrix} V_0V_0^+ & 0 \\ 0 & W_0W_0^+ \end{bmatrix} = (MN)' = N'M' = NM$$

since M and N are symmetric. Using lemma 12 we can write MNM as

$$MNM = \begin{bmatrix} V_0V_0^+ & 0 \\ 0 & W_0W_0^+ \end{bmatrix} \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} = \begin{bmatrix} V_0V_0^+V & V_0V_0^+X \\ W_0W_0^+X' & 0 \end{bmatrix} = \begin{bmatrix} V & X \\ X' & 0 \end{bmatrix} = M$$

and NMN as

$$\begin{aligned}
NMN &= \begin{bmatrix} V_0^+ - V_0^+XW_0^+X'V_0^+ & V_0^+XW_0^+ \\ W_0^+X'V_0^+ & W_0W_0^+EW_0W_0^+ - W_0^+ \end{bmatrix} \begin{bmatrix} V_0V_0^+ & 0 \\ 0 & W_0W_0^+ \end{bmatrix} \\
&= \begin{bmatrix} V_0^+V_0V_0^+ - V_0^+XW_0^+X'V_0^+V_0V_0^+ & V_0^+XW_0^+W_0W_0^+ \\ W_0^+X'V_0^+V_0V_0^+ & (W_0W_0^+EW_0W_0^+ - W_0^+)W_0W_0^+ \end{bmatrix} \\
&= \begin{bmatrix} V_0^+ - V_0^+XW_0^+X'V_0^+ & V_0^+XW_0^+ \\ W_0^+X'V_0^+ & W_0W_0^+EW_0W_0^+ - W_0^+ \end{bmatrix} = N
\end{aligned}$$

and therefore $N = M^+$ is the pseudoinverse of M, and $MN = NM = MM^+ = M^+M$ are the orthogonal projectors. □

Proof of corollary 14. By lemma 13, we have the pseudoinverse

$$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}^+ = \begin{bmatrix} V_0^+ - V_0^+XW_0^+X'V_0^+ & V_0^+XW_0^+ \\ W_0^+X'V_0^+ & W_0W_0^+EW_0W_0^+ - W_0^+ \end{bmatrix}$$

where $W_0 = X'V_0^+X$. Moreover, since $\mathcal{R}(X) = \mathcal{R}(XX') \subseteq \mathcal{R}(V + XX') = \mathcal{R}(V_1)$, there is an alternate expression for the pseudoinverse,

$$\begin{bmatrix} V & X \\ X' & 0 \end{bmatrix}^+ = \begin{bmatrix} V_1^+ - V_1^+XW_1^+X'V_1^+ & V_1^+XW_1^+ \\ W_1^+X'V_1^+ & W_1W_1^+ - W_1^+ \end{bmatrix}$$

where $W_1 = X'V_1^+X$ and the simplification $W_1W_1^+W_1W_1^+ = W_1W_1^+$ has been applied. By uniqueness of the pseudoinverse, the $(2,1)$ block of each expression must be equal. □

G Pseudoinverse of sum of positive semidefinite matrices

In this appendix we prove lemmas 15 and 16. First however, consider the following definitions

$$T = I - VV^+ \qquad Z = TX \qquad w = Ty \tag{124a}$$

$$B = I - Z^+Z \qquad C = I - XZ^+ \qquad D = I + BX'V^+XB \tag{124b}$$

and the following lemma, which collects identities for the matrices (124).

Lemma 62. *For any $V \in \mathbb{S}_+^n$ and $X \in \mathbb{R}^{n\times p}$,*

$$\begin{aligned} V^+Z &= 0 & Z^+X &= Z^+Z \\ Z^+V &= 0 & CV &= V \\ CX &= XB & CVV^+ &= VV^+ \\ VV^+C &= VV^+ - XZ^+ + ZZ^+ & VV^+XB &= XB \\ CZZ^+ &= ZZ^+ - XZ^+ & D^{-1}BX'V^+XB &= I - D^{-1} \end{aligned}$$

given the definitions (124).

Proof. The first fact is $V^+Z = V^+TX = 0$. Using lemma 5 the next two facts are shown

$$\begin{aligned} Z^+X &= (TX)^+X = (TX)^+TX = Z^+Z \\ Z^+V &= Z^+TV = 0 \end{aligned}$$

The next two facts are corollaries to the previous facts,

$$\begin{aligned} CV &= (I - XZ^+)V = V \\ CX &= (I - XZ^+)X = X - XZ^+Z = XB \\ CVV^+ &= (I - XZ^+)VV^+ = VV^+ \end{aligned}$$

Finally,

$$\begin{aligned} VV^+C &= VV^+(I - XZ^+) = VV^+ - (I - T)XZ^+ \\ &= VV^+ - XZ^+ + ZZ^+ \\ VV^+XB &= (I - T)XB = XB - ZB = XB \\ CZZ^+ &= (I - XZ^+)ZZ^+ = ZZ^+ - XZ^+ \\ D^{-1}BX'V^+XB &= D^{-1}(I + BX'V^+XB) - D^{-1} = I - D^{-1} \end{aligned}$$

□

Before proving lemmas 15 and 16, we show in the following lemma a sufficient condition for which the pseudoinverse of a sum is equal to the sum of the pseudoinverses.

Lemma 63. *For any matrices $A, B \in \mathbb{R}^{n\times m}$ such that $A'B = 0$ and $BA' = 0$, $(A+B)^+ = A^+ + B^+$.*

Proof. Noting that $A'B = 0 \Leftrightarrow \mathcal{R}(B) \subseteq \mathcal{N}(A')$ and $BA' = 0 \Leftrightarrow \mathcal{R}(A') \subseteq \mathcal{N}(B)$, we also have

$$\begin{aligned}
& \mathcal{R}(B) \subseteq \mathcal{N}(A') = \mathcal{N}(A^+) && \Leftrightarrow && A^+B = 0 \\
\Leftrightarrow \quad & \mathcal{R}(A) \subseteq \mathcal{N}(B') = \mathcal{N}(B^+) && \Leftrightarrow && B^+A = 0 \\
& \mathcal{R}(A') = \mathcal{R}(A^+) \subseteq \mathcal{N}(B) && \Leftrightarrow && BA^+ = 0 \\
\Leftrightarrow \quad & \mathcal{R}(B') = \mathcal{R}(B^+) \subseteq \mathcal{N}(A) && \Leftrightarrow && AB^+ = 0
\end{aligned}$$

The result follows by substitution into (6),

$$\begin{aligned}
(A+B)(A^+ + B^+) &= AA^+ + BB^+ \\
(A^+ + B^+)(A+B) &= A^+A + B^+B \\
(A+B)(A^+ + B^+)(A+B) &= (A+B)(A^+A + B^+B) \\
&= AA^+A + BB^+B = A + B \\
(A^+ + B^+)(A+B)(A^+ + B^+) &= (A^+A + B^+B)(A^+ + B^+) \\
&= A^+AA^+ + B^+BB^+ = A^+ + B^+
\end{aligned}$$

□

Finally, we prove lemmas 15 and 16 below.

Proof of lemma 15. Let $H = C'V^+C + Z'^+Z^+ - C'V^+XBD^{-1}BX'V^+C$. We show that this is the unique pseudoinverse of $V_0 = V + XX'$ by directly checking (6). We use the results of lemma 62 throughout and without reference. First consider the following results,

$$\begin{aligned}
C'V^+CV_0 &= C'V^+CV + C'V^+CXX' \\
&= C'V^+V + C'V^+XBX' \\
&= VV^+ - Z'^+X' + ZZ^+ + C'V^+XBX', \\
Z'^+Z^+V_0 &= Z'^+Z^+XX' = Z'^+Z^+ZX' = Z'^+X', \\
C'V^+XBD^{-1}BX'V^+CV_0 &= C'V^+XB(D^{-1}BX'V^+CV + D^{-1}BX'V^+CXX') \\
&= C'V^+XB(D^{-1}BX'V^+V + D^{-1}BX'V^+XBX') \\
&= C'V^+XB(D^{-1}BX' + (I - D^{-1})BX') \\
&= C'V^+XBX'
\end{aligned}$$

From these results we can rewrite the product HV_0 as

$$\begin{aligned}
HV_0 &= (C'V^+C + Z'^+Z^+ - C'V^+XBD^{-1}BX'V^+C)V_0 \\
&= VV^+ - Z'^+X' + ZZ^+ + C'V^+XBX' + Z'^+X' - C'V^+XBX' \\
&= VV^+ + ZZ^+
\end{aligned}$$

which confirms the third Moore-Penrose condition. The fourth condition is verified by noting that V_0 and H are symmetric,

$$V_0 H = (HV_0)' = (VV^+ + ZZ^+)' = VV^+ + ZZ^+$$

The first condition follows from substitution of the above formula,

$$\begin{aligned} V_0 H V_0 &= (VV^+ + ZZ^+)(V + XX') \\ &= VV^+V + VV^+XX' + ZZ^+XX' \\ &= V + XX' - TXX' + ZX' \\ &= V + XX' \end{aligned}$$

Notice that $V_0 H = VV^+ + ZZ^+$ is a projection, and we can show each term of H is in the range space of that projection,

$$\begin{aligned} C'V^+CV_0H &= C'V^+CVV^+ + C'V^+CZZ^+ \\ &= C'V^+ + C'V^+(ZZ^+ - XZ^+) \\ &= C'V^+C, \\ Z'^+Z^+V_0H &= Z'^+Z^+ZZ^+ = Z'^+Z^+, \\ C'V^+XBD^{-1}BX'V^+CV_0H &= C'V^+XBD^{-1}BX'(V^+CVV^+ + V^+CZZ^+) \\ &= C'V^+XBD^{-1}BX'(V^+ + V^+(ZZ^+ - XZ^+)) \\ &= C'V^+XBD^{-1}BX'V^+C \end{aligned}$$

from which the second condition follows,

$$\begin{aligned} HV_0H &= (C'V^+C + Z'^+Z^+ - C'V^+XBD^{-1}BX'V^+C)(VV^+ + ZZ^+) \\ &= C'V^+C + Z'^+Z^+ - C'V^+XBD^{-1}BX'V^+C \end{aligned}$$

Therefore H satisfies (6) and is the unique pseudoinverse of V_0. □

Proof of lemma 16. The first statement is shown directly. Due to lemma 62 and lemma 15,

$$\begin{aligned} V_0^+X &= C'V^+CX + Z'^+Z^+X - C'V^+XBD^{-1}BX'V^+CX \\ &= C'V^+XB + Z'^+Z^+Z - C'V^+XBD^{-1}BX'V^+XB \\ &= C'V^+XB + Z'^+Z^+Z - C'V^+XB(I - D^{-1}) \\ &= Z'^+ + C'V^+XBD^{-1} \end{aligned}$$

which implies

$$\begin{aligned} X'V_0^+X &= X'Z'^+ + X'C'V^+XBD^{-1} \\ &= Z'Z'^+ + BX'V^+XBD^{-1} \\ &= I - B + I - D^{-1} \end{aligned}$$

Using theorem 60, we have

$$I - D^{-1} = I - (I + BX'V^+XB)^{-1} = BAB$$

where $A = X'W(I + W'XBX'W)^{-1}W'X$ and $W = (V^+)^{1/2}$. which implies $(I - B)(I - D^{-1}) = 0$ and $(I - D^{-1})(I - B) = 0$. Moreover, using lemma 5 we have

$$\begin{aligned}
(I - D^{-1})^+ &= B(I - D^{-1})^+ = (I - D^{-1})^+B \\
D^{-1}B &= (I - BAB)B = B(I - BAB) = BD^{-1} \\
(I - D^{-1})^+D^{-1} &= (BAB)^+(I - BAB) \\
&= (I - BAB)(BAB)^+ = D^{-1}(I - D^{-1})^+
\end{aligned}$$

Using lemma 63, we can write

$$(X'V_0^+X)^+ = (I - B + I - D^{-1})^+ = I - B + (I - D^{-1})^+$$

We show that $(I - D^{-1})^+D^{-1} = (BX'V^+XB)^+$ by checking (6),

$$\begin{aligned}
(I - D^{-1})^+D^{-1}BX'V^+XB &= (I - D^{-1})^+(I - D^{-1}) \\
BX'V^+XB(I - D^{-1})^+D^{-1} &= BX'V^+XBD^{-1}(I - D^{-1})^+ \\
&= (I - D^{-1})(I - D^{-1})^+ \\
BX'V^+XB(I - D^{-1})^+D^{-1}&BX'V^+XB \\
&= D^{-1}(I - D^{-1})(I - D^{-1})^+(I - D^{-1}) \\
&= D^{-1}(I - D^{-1}) \\
&= BX'V^+XB \\
(I - D^{-1})^+D^{-1}BX'V^+XB&(I - D^{-1})^+D^{-1} \\
&= (I - D^{-1})^+(I - D^{-1})(I - D^{-1})^+D^{-1} \\
&= (I - D^{-1})^+D^{-1}
\end{aligned}$$

Finally, we have that

$$\begin{aligned}
(X'V_0^+X)^+X'V_0^+ &= (I - B + (I - D^{-1})^+)(Z^+ + D^{-1}BXV^+C) \\
&= (I - B)Z^+ + (I - D^{-1})^+Z \\
&\quad + (I - B)D^{-1}BXV^+C + (I - D^{-1})^+D^{-1}BXV^+C \\
&= Z^+ + (BXV^+XB)^+BXV^+C
\end{aligned}$$

since $(I - D^{-1})^+Z = (I - D^{-1})^+BZ = 0$ and $(I - B)D^{-1}B = (I - B)BD^{-1} = 0$.

To show the second statement, first notice that

$$\begin{aligned}
B\mathcal{N}(XB) &= B\,\{\,(I - (XB)^+XB)q \mid q \in \mathbb{R}^p\,\} \\
&= \{\,B(I - (XB)^+XB)q \mid q \in \mathbb{R}^p\,\} \\
&= \{\,(B - B(XB)^+XB)q \mid q \in \mathbb{R}^p\,\} \\
&= \{\,(I - Z^+Z - (XB)^+XB)q \mid q \in \mathbb{R}^p\,\} \\
\mathcal{N}(X) &= \{\,(I - X^+X)q \mid q \in \mathbb{R}^p\,\}
\end{aligned}$$

Therefore it suffices to show $Z^+Z + (XB)^+XB = X^+X$ to prove $B\mathcal{N}(XB) = \mathcal{N}(X)$. We can show that $Z^+Z + (XB)^+XB = (X^+X)^+$ by checking (6),

$$\begin{aligned}
(Z^+Z + (XB)^+XB)X^+X &= Z^+Z + (XB)^+XB \\
X^+X(Z^+Z + (XB)^+XB) &= Z^+Z + X^+XB(XB)^+XB \\
&= Z^+Z + X^+XB \\
&= Z^+Z + X^+X(I - Z^+Z) \\
&= X^+X \\
X^+X(Z^+Z + (XB)^+XB)X^+X &= X^+XX^+X = X^+X \\
(Z^+Z + (XB)^+XB)X^+X(Z^+Z + (XB)^+XB) &= (Z^+Z + (XB)^+XB)X^+X \\
&= Z^+Z + (XB)^+XB
\end{aligned}$$

Finally, since X^+X is an orthogonal projector, it is its own pseudoinverse, and $Z^+Z + (XB)^+XB = (X^+X)^+ = X^+X$. □

H Global bounds on the perturbed problem

Proof of theorem 17 *([103, 111]).* First, we show that (15b) implies (15a). It follows by substitution into (6) that $(A'DA)^+ = V_1\Sigma_1^{-1}(U_1'DU_1)^{-1}\Sigma_1^{-1}V_1'$ (theorem 1). Moreover,

$$\begin{aligned}
A(A'DA)^+A'D &= U_1(U_1'DU_1)^{-1}U_1'D \\
(A'DA)^+A'D &= V_1\Sigma_1^{-1}(U_1'DU_1)^{-1}U_1'D = A^+A(A'DA)^+A'D
\end{aligned}$$

Using the last equality, we have that, if (15b) holds, then

$$\|(A'DA)^+A'D\| = \|A^+A(A'DA)^+A'D\| \leq \|A^+\| \cdot \|A(A'DA)^+A'D\| \leq \frac{1}{\underline{\sigma}(A)\chi(A)}$$

Therefore it suffices to show (15b).

Next, we show that $\mathbb{X}(A)$ and $\overline{\mathbb{Y}}(A)$ are disjoint, where $\overline{\mathbb{Y}}(A)$ denotes the closure of $\mathbb{Y}(A)$. Suppose that $z \in \mathbb{X}(A) \cap \overline{\mathbb{Y}}(A)$. Then $\|z\| = 1$ and $z = Aw$ for some $w \in \mathbb{R}^p$. Since $z \in \overline{\mathbb{Y}}(A)$, there exists a sequence $\{ z_k \} \subset \mathbb{R}^m$ such that $z_k \to z$, and a sequence of matrices $\{ D_k \} \subset \mathbb{D}^m_{>0}$ such that $A'D_kz_k = 0$ for all $k \in \mathbb{I}_{>0}$. Therefore, $0 = w'A'D_ky_k = yD_ky_k$ for all $k \in \mathbb{I}_{>0}$. But since $z_k \to z$, there must be some $\ell \in \mathbb{I}_{>0}$ sufficiently large such that, for each nonzero entry of z, the corresponding entry of z_ℓ has the same sign. Since $\|z\| = 1$, there is at least one nonzero entry of z (and z_ℓ). Then $z'D_\ell z_\ell > 0$ which contradicts that $z'D_kz_k = 0$ for all $k \in \mathbb{I}_{>0}$.

Next, we show that $\mathbb{X}(A) \cap \overline{\mathbb{Y}}(A) = \emptyset$ implies $\|A(A'DA)^+A'D\| \leq \frac{1}{\chi(A)}$ for all $D \in \mathbb{D}^m_{>0}$. Since $\mathbb{X}(A)$ is compact, $\overline{\mathbb{Y}}(A)$ is closed (by construction), and they are disjoint, there exists $\rho > 0$ such that $\rho \leq \|x - y\|$ for all $x \in \mathbb{X}(A)$ and $y \in \mathbb{Y}(A)$. In other words, $\chi(A) > 0$.

Let $D \in \mathbb{D}^m_{>0}$ and $z \in \mathbb{R}^m$ such that $\|z\| = 1$. Define $x = A(A'DA)^+A'Dz$ and $y = z - x$ so that $A'Dy = A'D(z - x) = A'Dz - A'DA(A'DA)^+A'Dz = 0$ by lemma 12. Then with $\alpha = 1/\|x\|$, we have

$$\alpha x + \alpha y = \alpha z$$

Noting that $\alpha x \in \mathbb{X}(A)$ and $-\alpha y \in \mathbb{Y}(A)$, it is clear that

$$\chi(A) \leq \|\alpha x + \alpha y\| = \|\alpha z\| = \frac{1}{\|x\|}$$

Taking the reciprocal of both sides, we have

$$\frac{1}{\chi(A)} \geq \|x\| = \|A(A'DA)^+A'Dz\|$$

and taking the maximum over $\|z\| = 1$ gives $\frac{1}{\chi(A)} \geq \|A(A'DA)^+A'D\|$.

Finally, we show that $\|A(A'DA)^+A'D\| \geq \frac{1}{\chi(A)}$ for some $D \in \mathbb{D}^m_{>0}$. Let $x \in \mathbb{X}(A)$ and $y \in \mathbb{Y}(A)$. Then there exists $D \in \mathbb{D}^m_{>0}$ and $w \in \mathbb{R}^p$ such that $A'Dy = 0$ and $x = Aw$. Moreover,

$$A'D(x - y) = A'Dx = A'DAw$$

and therefore we can write $w = (A'DA)^+A'D(x - y)$ by lemma 12. Moreover, we have $x = A(A'DA)^+A'D(x - y)$. Taking the norm of x and using submultiplicativity gives

$$1 \leq \|A(A'DA)^+A'D\| \cdot \|x - y\|$$

since $\|x\| = 1$ by $x \in \mathbb{X}(A)$. Taking the infimum of both sides over $x \in \mathbb{X}(A)$ and $y \in \mathbb{Y}(A)$ produces the desired result. □

Proof of corollary 18. The proof follows straightforwardly by taking the SVDs of X and V and rewriting (17) in the form (15). Using the SVDs gives

$$(X'V_D^{-1}X)^+X'V_D^{-1} = V_1(\tilde{X}'\tilde{D}\tilde{X})^{-1}\tilde{X}'\tilde{D}Q'$$
$$X(X'V_D^{-1}X)^+X'V_D^{-1} = Q'\tilde{X}(\tilde{X}'\tilde{D}\tilde{X})^{-1}\tilde{X}'\tilde{D}Q'$$

where $\tilde{D} = (S+D)^{-1}$, $\tilde{X} = Q'U_1\Sigma_1$, and $\tilde{X}'\tilde{D}\tilde{X}$ is clearly positive definite (and invertible). Taking the norm of both sides of both of the above equations and noting that the norm is invariant to orthogonal transformations, we get

$$\|(X'V_D^{-1}X)^+X'V_D^{-1}\| = \|(\tilde{X}'\tilde{D}\tilde{X})^{-1}\tilde{X}'\tilde{D}\|$$
$$\|X(X'V_D^{-1}X)^+X'V_D^{-1}\| = \|\tilde{X}(\tilde{X}'\tilde{D}\tilde{X})^{-1}\tilde{X}'\tilde{D}\|$$

Since the image of $\mathbb{D}^m_{>0}$ through $(S + D)^{-1}$ is a subset of $\mathbb{D}^m_{>0}$ itself, taking the supremum over the former yields a smaller result than taking the supremum over the latter. In other words,

$$\sup_{D\in\mathbb{D}^m_{>0}} \|(X'V_D^{-1}X)^+X'V_D^{-1}\| \leq \sup_{\tilde{D}\in\mathbb{D}^m_{>0}} \|(\tilde{X}'\tilde{D}\tilde{X})^{-1}\tilde{X}'\tilde{D}\|$$
$$\sup_{D\in\mathbb{D}^m_{>0}} \|X(X'V_D^{-1}X)^+X'V_D^{-1}\| \leq \sup_{\tilde{D}\in\mathbb{D}^m_{>0}} \|\tilde{X}(\tilde{X}'\tilde{D}\tilde{X})^{-1}\tilde{X}'\tilde{D}\|$$

Finally, lemma 19 imples (17). □

Proof of lemma 19 *([80, 103]).* Noting that

$$\mathbb{X}(A) = \mathbb{X}(U_1), \qquad \mathbb{Y}(A) = \mathbb{Y}(U_1), \qquad \mathbb{U}(A) = \mathbb{U}(U_1)$$

it is clear that (18) equivalent to

$$\chi(A) = \chi(U_1) = \min_{U \in \mathbb{U}(U_1)} \underline{\sigma}(U) = \min_{U \in \mathbb{U}(A)} \underline{\sigma}(U)$$

and therefore we can assume $A = U_1$ without loss of generality.

($\leq$) First, we show that

$$\chi(U_1) \leq \min_{U \in \mathbb{U}(U_1)} \underline{\sigma}(U) \tag{126}$$

using the method in [103]. Let $U_{1,1} \in \mathbb{R}^{p \times r}$ be the submatrix of U_1 that solves the right hand side of (126). Then there exists a permutation matrix P such that $PU_1 = \begin{bmatrix} U_{1,1}' & U_{2,1}' \end{bmatrix}'$ where $U_{2,1} \in \mathbb{R}^{m-p \times r}$ contains the remaining rows of U_1. Denote the SVDs of $U_{1,1}$ as

$$U_{1,1} = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix} V'$$

where $S_1 = \operatorname{diag}(s_1, \ldots, s_k)$ and $k = \operatorname{rank}(U_{1,1})$. Moreover, the smallest singular value s_k is equal to the right hand side of (126). Noting that U_1 and U_1V' have the same left singular vectors, and that both sides of (126) are invariant to row permutations, we see that (126) is equivalent to

$$\chi(PU_1V') \leq s_k = \min_{U \in \mathbb{U}(PU_1V')} \underline{\sigma}(U) \tag{127}$$

Rewriting PU_1V' in terms of the columns of $U_{1,1}V'$ and $U_{2,1}V'$, we have

$$PU_1V' = \begin{bmatrix} U_{1,1}V' \\ U_{2,1}V' \end{bmatrix} = \begin{bmatrix} s_1w_1 & \ldots & s_kw_k & 0 & \ldots & 0 \\ z_1 & \ldots & z_k & z_{k+1} & \ldots & z_r \end{bmatrix}$$

It is clear that $\{ s_1w_1, \ldots, s_kw_k \}$ and $\{ z_1, \ldots, z_r \}$ are sets of orthogonal vectors since PU_1V' is an orthogonal matrix.

Since $\mathbb{X}(PU_1V')$ is nonempty and $0 \in \mathbb{Y}(PU_1V')$, we have

$$\chi(PU_1V') = \inf_{x \in \mathbb{X}(PU_1V'), y \in \mathbb{Y}(PU_1V')} \|x - y\| \leq 1$$

This implies $s_k \leq 1$, but we can assume $s_k < 1$ without loss of generality. Under this assumption, $\|z_k\| > 0$ because PU_1V' is an orthogonal matrix and therefore $\| \begin{bmatrix} s_kw_k' & z_k' \end{bmatrix}' \| = s_k + \|z_k\| = 1$. Choose $\varepsilon > 0$ and define

$$y = \begin{bmatrix} -\varepsilon s_k w_k \\ z_k \end{bmatrix}, \qquad D = \operatorname{diag}\left(I, \frac{\varepsilon s_k^2}{\|z_k\|} I \right)$$

so that

$$(PU_1V')'Dy = -\varepsilon s_k V U_{1,1}' w_k + \frac{\varepsilon s_k^2}{\|z_k\|} V U_{2,1}' z_k = -\varepsilon s_k^2 e_k + \varepsilon s_k^2 e_k = 0$$

where e_k is the k-th elementary vector in $\mathbb{R}^r$, and clearly $y \in \mathbb{Y}(PU_1V')$. Let $x = \begin{bmatrix} s_k w_k' & z_k' \end{bmatrix}'$ which is clearly in $\mathbb{X}(PU_1V')$. Then

$$\|x - y\| = \left\| \begin{bmatrix} (1+\varepsilon)s_k w_k \\ 0 \end{bmatrix} \right\| = (1+\varepsilon)s_k$$

Taking the limit as $\varepsilon \to 0^+$, we recover (127), and equivalently (126).

($\geq$) Next, we show that $\chi(U_1) \geq \min_{U \in \mathbb{U}(U_1)} \underline{\sigma}(U)$ using the method in [80]. Define the scalar sign function as

$$\operatorname{sign}(\alpha) = \begin{cases} \alpha/|\alpha| & \text{if } \alpha \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

And define the vector sign function component-wise. Let $y \in \mathbb{Y}(U_1)$ and $\tilde{y} \in \mathbb{R}^m$ such that $\operatorname{sign}(y) = \operatorname{sign}(\tilde{y})$. Then define the scaling matrix $S \in \mathbb{D}^m_{>0}$ as

$$S_{ii} = \begin{cases} y_i/\tilde{y}_i & \text{if } y_i \neq 0 \\ 1 & \text{otherwise} \end{cases}$$

which gives $U_1'DS\tilde{y} = U_1'Dy = 0$ and $\tilde{y} \in \mathbb{Y}(U_1)$ for some $D \in \mathbb{D}^m_{>0}$. Note also that since $\|U_1 w\| = \|w\|$ for all $w \in \mathbb{R}^r$, we have $\mathcal{R}(U_1) = \{\, U_1 w \mid \|w\| = 1 \,\}$. Applying the preceeding results to (16a),

$$\begin{aligned} \chi(U_1) &= \inf_{\substack{x \in \mathbb{X}(U_1) \\ y \in \mathbb{Y}(U_1)}} \|x - y\| \\ &= \inf_{y \in \mathbb{Y}(U_1)} \inf_{\substack{x \in \mathbb{X}(U_1) \\ \operatorname{sign}(y) = \operatorname{sign}(\tilde{y})}} \|x - \tilde{y}\| \\ &= \inf_{y \in \mathbb{Y}(U_1)} \inf_{\substack{\|w\| = 1 \\ \operatorname{sign}(y) = \operatorname{sign}(\tilde{y})}} \|U_1 w - \tilde{y}\| \end{aligned}$$

With the signs of each $\tilde{y}$ fixed by the choice of $y \in \mathbb{Y}(U_1)$ in the outer infimum, we can now select, by scaling, the components of $\tilde{y}$ to create a lower bound on the solution to the inner infimum. First, note that for every $\tilde{y} \in \mathbb{Y}(U_1)$ and w such that $\|w\| = 1$, we must have $\operatorname{sign}(U_1 w) \neq \operatorname{sign}(\tilde{y})$. To see this, suppose we had $\operatorname{sign}(U_1 w) = \operatorname{sign}(\tilde{y})$. Then we could always find a scaling matrix $S \in \mathbb{D}^m_{>0}$ so that $Sy = U_1 w \in \mathbb{Y}(U_1)$. But that implies $\chi(U_1) = 0$ which contradicts theorem 17.

Let $\|w\| = 1$. Denote the set of indices i such that $\operatorname{sign}((U_1 w)_i) \neq \operatorname{sign}(\tilde{y}_i)$ as $\mathcal{I}$. For any matrix (or vector) B, let $B_{\mathcal{I}}$ denote the submatrix formed by the rows of B corresponding to the index set $\mathcal{I}$. For this w, define $\tilde{y}$ as

$$\tilde{y}_i = \begin{cases} (U_1 w)_i & \text{if } i \notin \mathcal{I} \\ \varepsilon_i \operatorname{sign}(y_i) & \text{if } i \in \mathcal{I} \end{cases}$$

where $\varepsilon_i > 0$ is arbitrarily small. The resulting value of $\|\tilde{y} - U\alpha\|$ is no less than $\|(U_1 w)_{\mathcal{I}}\| = \|(U_1)_{\mathcal{I}} w\|$ and therefore $\|\tilde{y} - U\alpha\|$ is bounded below by the smallest singular

value of $(U_1)_{\mathcal{I}}$. In other words,

$$\chi(U_1) = \inf_{y \in \mathbb{Y}(U_1)} \inf_{\substack{\|w\|=1 \\ \operatorname{sign}(y)=\operatorname{sign}(\tilde{y})}} \|U_1 w - \tilde{y}\| \geq \min_{U \in \mathbb{U}(U_1)} \underline{\sigma}(U)$$

□

I Proof of the limit of the perturbed problem solution

In this appendix we prove lemma 20. First, we state a few preliminary definitions that are used in the proof. Consider the following SVDs,

$$X = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1' \\ V_2' \end{bmatrix} = U_1 \Sigma_1 V_1' \tag{128a}$$

$$V = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Q_1' \\ Q_2' \end{bmatrix} = Q_1 S_1 Q_1' \tag{128b}$$

$$A = U_2' Q_1 S_1^{1/2} = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \begin{bmatrix} Y_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} Z_1' \\ Z_2' \end{bmatrix} = W_1 Y_1 Z_1' \tag{128c}$$

and use the following definitions,

$$V_\rho := V + \rho I, \qquad Q := \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \qquad B := U_1' Q_1, \tag{129a}$$

$$C := \rho S_1^{-1} - I, \qquad S := I - XX^+ = U_2 U_2' \tag{129b}$$

where $\rho > 0$, $r = \operatorname{rank}(V)$, and $q = \operatorname{rank}(X)$.

To prove lemma 20, we use a series of three approximations to the perturbed solution, which facilitated by the following lemma.

Lemma 64. *Let $X \in \mathbb{R}^{n \times p}$, $V \in \mathbb{S}_+^n$, $\rho > 0$, and consider* (128) *and* (129)*. Denoting the residuals,*

$$R_1(\rho) := (U_1' V_\rho^{-1} U_1)^{-1} U_1' V_\rho^{-1} - (I + BCB')^{-1}(U_1' + BCQ_1') \tag{130a}$$

$$R_2(\rho) := (I + BCB')^{-1}(U_1' + BCQ_1') - U_1' + B(Q_1' S Q_1 + \rho S_1^{-1})^{-1} Q_1' S \tag{130b}$$

$$R_3(\rho) := B(Q_1' S Q_1 + \rho S_1^{-1})^{-1} Q_1' S - U_1' V (VSV)^+ \tag{130c}$$

we have the upper bounds,

$$\|R_i(\rho)\| \leq \frac{\alpha_i \rho}{\beta_i + \rho}, \qquad i = 1, 2, 3 \tag{131}$$

where $\alpha_1 := \frac{\overline{\sigma}(V)}{\chi^2(Q'U_1)\underline{\sigma}(V)}$, $\alpha_2 := \frac{\overline{\sigma}(SV^{1/2})}{\chi(Q'U_1)\underline{\sigma}^{1/2}(V)}$, $\alpha_3 := \frac{\overline{\sigma}^{1/2}(V)}{\underline{\sigma}(SV^{1/2})}$, $\beta_1 := \underline{\sigma}(V)$, $\beta_2 := \beta_3 := \underline{\sigma}^2(SV^{1/2})$*, and* $\chi(\cdot)$ *is defined by* (16a)*.*

Proof. Throughout the proof, we use without reference the submultiplicativity of the matrix 2-norm (i.e., $\|AB\| \leq \|A\|\|B\|$ for all A, B of suitable dimensions) and the equivalence between the maximum singular value and the 2-norm (i.e., $\|A\| = \|A'\| = \overline{\sigma}(A)$ for all A). Note that the second fact implies $\|U\| = \|U'\| = 1$ for all orthogonal matrices U, and $\|D\| = \max_{i=1,\ldots,\max\{m,n\}} |D_{ii}|$ for all $D \in \mathbb{R}^{m\times n}$ such that $D_{ij} = 0$ for all $i \neq j$.

It is also worth pointing out that the singular values of $SV^{1/2}$ and A are equivalent. To see this, we rewrite $SV^{1/2}$ in terms of the SVD of A,

$$SV^{1/2} = U_2U_2'Q_1S_1^{1/2}Q_1' = U_2AQ_1' = (U_2W_1)Y_1(Q_1Z_1)'$$

which is the SVD of $SV^{1/2}$ with left and right singular vectors U_2W_1 and Q_1Z_1.

(R_1) First we write the SVD of V_ρ,

$$V_\rho := V + \rho I = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} S_1 + \rho I & \\ & \rho I \end{bmatrix} \begin{bmatrix} Q_1' \\ Q_2' \end{bmatrix}$$

and therefore V_ρ^{-1} can be written,

$$\begin{aligned} V_\rho^{-1} &= \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} (S_1 + \rho I)^{-1} & \\ & \rho^{-1} I \end{bmatrix} \begin{bmatrix} Q_1' \\ Q_2' \end{bmatrix} \\ &= Q_1(S_1 + \rho I)^{-1}Q_1' + \rho^{-1}Q_2Q_2' \end{aligned}$$

Using theorem 60, we can expand $(S_1 + \rho I)^{-1}$ and rewrite V_ρ^{-1} as

$$\begin{aligned} V_\rho^{-1} &= Q_1[S_1^{-1} - S_1^{-1}(\rho^{-1}I + S_1^{-1})^{-1}S_1^{-1}]Q_1' + \rho^{-1}Q_2Q_2' \\ &= Q_1S_1^{-1}Q_1' - \rho Q_1S_1^{-2}(I + \rho S_1^{-1})^{-1}Q_1' + \rho^{-1}(I - Q_1Q_1') \\ V_\rho^{-1} &= \rho^{-1}(I + Q_1CQ_1') + R_{1,1}(\rho) \end{aligned} \tag{132}$$

where $R_{1,1}(\rho) := -\rho Q_1S_1^{-2}(I + \rho S_1^{-1})^{-1}Q_1'$. Rewriting $U_1'V_\rho^{-1}U_1$,

$$U_1'V_\rho^{-1}U_1 = \rho^{-1}(I + BCB') + U_1'R_{1,1}(\rho)U_1$$

Note that $U_1'V_\rho^{-1}U_1$ is invertible because it is positive definite. Rewriting the sum $I+BCB'$ as a product,

$$I + BCB' = I + U_1'Q_1(\rho S_1^{-1} - I)Q_1'U_1 = U_1'Q \begin{bmatrix} \rho S_1^{-1} & \\ & I \end{bmatrix} Q'U_1 \tag{133}$$

it is clear that $I + BCB'$ is positive definite (and invertible). Using theorem 60, we have

$$(U_1'V_\rho^{-1}U_1)^{-1} = \rho(I + BCB')^{-1} + R_{1,2}(\rho) \tag{134}$$

where $R_{1,2}(\rho) := -\rho(I+BCB)^{-1}U_1'R_{1,1}U_1(U_1'V_\rho^{-1}U_1)^{-1}$. Combining the results (132) and (134) we can write $R_1(\rho)$ as

$$\begin{aligned}
R_1(\rho) &= [\rho(I+BCB')^{-1} + R_{1,2}(\rho)]U_1'[\rho^{-1}(I+Q_1CQ_1') + R_{1,1}(\rho)] \\
&\qquad - (I+BCB')^{-1}(U_1' + BCQ_1') \\
&= \rho(I+BCB')^{-1}U_1'R_{1,1}(\rho) + R_{1,2}(\rho)U_1'[\rho^{-1}(I+Q_1CQ_1') + R_{1,1}(\rho)] \\
&= \rho(I+BCB')^{-1}U_1'R_{1,1}(\rho) + R_{1,2}(\rho)U_1'V_\rho^{-1} \\
&= \rho^2(I+BCB')^{-1}U_1'R_{1,1}(\rho)(I - U_1(U_1'V_\rho^{-1}U_1)^{-1}U_1'V_\rho^{-1}) \\
R_1(\rho) &= \rho R_{1,3}(\rho)R_{1,4}(\rho)R_{1,5}(\rho)
\end{aligned} \tag{135}$$

where $R_{1,3}(\rho) := \rho(I+BCB')^{-1}B$, $R_{1,4}(\rho) := S_1^{-2}(I+\rho S_1^{-1})^{-1}$, and $R_{1,5}(\rho) := Q_1'(I - U_1(U_1'V_\rho^{-1}U_1)^{-1}U_1'V_\rho^{-1})$.

To bound the norm of the residual $R_1(\rho)$, we find bounds on the norms of $R_{1,3}(\rho)$, $R_{1,4}(\rho)$, and $R_{1,5}(\rho)$. First, we use (133) to rewrite $R_{1,3}(\rho)$ and $R_{1,5}(\rho)$,

$$\begin{aligned}
R_{1,3}(\rho) &= \left(U_1'Q\begin{bmatrix}\rho S_1^{-1} & \\ & I\end{bmatrix}Q'U_1\right)^{-1}U_1'Q\begin{bmatrix}\rho S_1^{-1} & \\ & I\end{bmatrix}\begin{bmatrix}S_1\\0\end{bmatrix} \\
R_{1,5}(\rho) &= Q_1'Q\left(I - Q'U_1\left(U_1'Q\begin{bmatrix}\rho S_1^{-1} & \\ & I\end{bmatrix}Q'U_1\right)^{-1}U_1'Q\begin{bmatrix}\rho S_1^{-1} & \\ & I\end{bmatrix}\right)Q'
\end{aligned}$$

Bounds on $R_{1,3}(\rho)$ and $R_{1,5}(\rho)$ follow from theorem 17,

$$\|R_{1,3}(\rho)\| \le \frac{\overline{\sigma}(V)}{\chi(Q'U_1)}, \qquad \|R_{1,5}(\rho)\| \le \frac{1}{\chi(Q'U_1)} \tag{136}$$

The bound on $R_{1,4}(\rho)$ is directly computed,

$$\|R_{1,4}(\rho)\| = \frac{1}{\underline{\sigma}^2(V)(\underline{\sigma}(V)+\rho)} \tag{137}$$

Finally, the desired bound on the norm of $R_1(\rho)$ follows from (135)–(137),

$$\|R_1(\rho)\| \le \rho\|R_{1,3}(\rho)\|\|R_{1,4}(\rho)\|\|R_{1,5}(\rho)\| \le \frac{\overline{\sigma}(V)}{\underline{\sigma}^2(V)\chi^2(Q'U_1)}\frac{\rho}{\underline{\sigma}(V)+\rho} = \frac{\alpha_1\rho}{\beta_1+\rho}$$

(R_2) For this residual, it suffices to derive a bound on the norm for all $\rho > 0$ such that $I + CB'B$ is invertible. This is because $I + CB'B$ is invertible for almost every $\rho > 0$, and since the residual is continuous for all $\rho > 0$, we can use the limit to ensure the bound holds for any $\rho > 0$ such that $I + CB'B$ is singular.

Using theorem 60, we have

$$\begin{aligned}
&(I+BCB')^{-1}(U_1' + BCQ_1') \\
&\qquad = (I - B(I+CB'B)^{-1}CB')U_1' + (I+BCB')^{-1}BCQ_1' \\
&\qquad = U_1' - B(I+CB'B)^{-1}CQ_1'U_1U_1' + (I+BCB')^{-1}BCQ_1' \\
&\qquad = U_1' + B(I+CB'B)^{-1}CQ_1'S \\
&\qquad = U_1' - B(I+CB'B)^{-1}Q_1'S + R_{2,1}(\rho)
\end{aligned} \tag{138}$$

where $R_{2,1}(\rho) := \rho B(I + CB'B)^{-1}S_1^{-1}Q_1'S$. Using theorem 60 and the fact that $B'B = Q_1'U_1U_1'Q_1 = I - Q_1'SQ_1$, we have

$$\begin{aligned}(I + CB'B)^{-1} &= (I - B'B + \rho S_1^{-1}B'B)^{-1} \\ &= (Q_1'SQ_1 + \rho S_1^{-1} - \rho S_1^{-1}Q_1'SQ_1)^{-1}Q_1'S \\ &= (Q_1'SQ_1 + \rho S_1^{-1})^{-1} + R_{2,2}(\rho) \end{aligned} \tag{139}$$

where $R_{2,2}(\rho) := \rho(I + CB'B)^{-1}S_1^{-1}Q_1'SQ_1(Q_1'SQ_1 + \rho S_1^{-1})^{-1}$, and $Q_1'SQ_1 + \rho S_1^{-1}$ is positive definite (and invertible). Combining the results (138) and (139), we have

$$\begin{aligned}&(I + BCB')^{-1}(U_1' + BCQ_1') \\ &\qquad = U_1' - B[(Q_1'SQ_1 + \rho S_1^{-1})^{-1} + R_{2,2}(\rho)]Q_1'S + R_{2,1}(\rho) \\ &\qquad = U_1' - B(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S + R_{2,1}(\rho) - BR_{2,2}(\rho)Q_1'S\end{aligned}$$

which implies

$$\begin{aligned}R_2(\rho) &= R_{2,1}(\rho) - BR_{2,2}(\rho)Q_1'S \\ &= \rho B(I + CB'B)^{-1}S_1^{-1}Q_1'S(I - Q_1(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S) \\ &= \rho(I + BCB')^{-1}BS_1^{-1}Q_1'S(I - Q_1(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S) \\ &= \rho^2(I + BCB')^{-1}BS_1^{-2}(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S\end{aligned}$$

where the third equality follows from theorem 60 and the fourth is shown below,

$$\begin{aligned}&Q_1'S(I - Q_1(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S) \\ &\qquad = Q_1'S - Q_1'SQ_1(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S \\ &\qquad = Q_1'S - (Q_1'SQ_1 + \rho S_1^{-1} - \rho S_1^{-1})(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S \\ &\qquad = \rho S_1^{-1}(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S\end{aligned}$$

Therefore, we can rewrite $R_2(\rho)$ as follows,

$$R_2(\rho) = \rho R_{2,3}(\rho)R_{2,4}(\rho) \tag{140}$$

where $R_{2,3}(\rho) := \rho(I + BCB')^{-1}BS_1^{-3/2}$ and $R_{2,4}(\rho) := S_1^{-1/2}(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S$.

To bound the norm of $R_2(\rho)$, we again find bounds on the norms of $R_{2,3}(\rho)$ and $R_{2,4}(\rho)$. Since $R_{2,3}(\rho) = R_{1,3}(\rho)S_1^{-3/2}$, we can use (136) to rewrite it as

$$R_{2,3}(\rho) = \left(U_1'Q\begin{bmatrix}\rho S_1^{-1} & \\ & I\end{bmatrix}Q'U_1\right)^{-1}U_1'Q\begin{bmatrix}\rho S_1^{-1} & \\ & I\end{bmatrix}\begin{bmatrix}S_1^{-1/2} \\ 0\end{bmatrix}$$

and by theorem 17 we have

$$\|R_{2,3}(\rho)\| \le \frac{1}{\chi(Q'U_1)\underline{\sigma}^{1/2}(V)} \tag{141}$$

by (136). Second, we rewrite $R_{2,4}(\rho)$ in terms of $A := U_2'Q_1S_1^{1/2} = W_1Y_1Z_1'$,

$$\begin{aligned}
R_{2,4}(\rho) &= (S_1^{1/2}Q_1'SQ_1S_1^{1/2} + \rho I)^{-1}S_1^{1/2}Q_1'S \\
&= (A'A + \rho I)^{-1}A'U_2' \\
&= \left(\begin{bmatrix} Z_1 & Z_2 \end{bmatrix} \begin{bmatrix} Y_1^2 + \rho I & \\ & \rho I \end{bmatrix} \begin{bmatrix} Z_1' \\ Z_2' \end{bmatrix}\right)^{-1} Z_1Y_1W_1'U_2' \\
&= Z_1(Y_1^2 + \rho I)^{-1}Y_1W_1'U_2'
\end{aligned}$$

and therefore

$$\|R_{2,4}(\rho)\| \leq \frac{\overline{\sigma}(A)}{\underline{\sigma}^2(A) + \rho} \tag{142}$$

Combining the results (140)–(142), we have

$$\|R_2(\rho)\| \leq \rho\|R_{2,3}(\rho)\|\|R_{2,4}(\rho)\| \leq \frac{\overline{\sigma}(A)}{\chi(Q'U_1)\underline{\sigma}^{1/2}(V)}\frac{\rho}{\underline{\sigma}^2(A) + \rho} = \frac{\alpha_2\rho}{\beta_2 + \rho}$$

(R_3) By lemma 2, we have

$$(A'A + \rho I)^{-1}A' = A^+ + R_{3,1}(\rho) \tag{143}$$

where $R_{3,1}(\rho) = \rho Z_1(Y_1^2 + \rho I)^{-1}Y_1^{-1}W_1'$. Moreover,

$$\begin{aligned}
Q_1S_1^{1/2}A^+U_2' &= Q_1S_1^{1/2}A'(AA')^+U_2' \\
&= Q_1S_1Q_1'U_2(U_2'Q_1SQ_1'U_2)^+U_2' \\
&= Q_1S_1Q_1'U_2U_2'(U_2U_2'Q_1SQ_1'U_2U_2')^+U_2U_2' \\
&= VS(SVS)^+S
\end{aligned} \tag{144}$$

Combining results (143) and (144) gives

$$\begin{aligned}
&B(Q_1'SQ_1 + \rho S_1^{-1})^{-1}Q_1'S \\
&\quad = U_1'Q_1S_1^{1/2}(A'A + \rho I)^{-1}A'U_2' \\
&\quad = U_1'Q_1S_1^{1/2}A^+U_2' + U_1'Q_1S_1^{1/2}R_{3,1}(\rho)U_2' \\
&\quad = U_1'VS(SVS)^+S + U_1'Q_1S_1^{1/2}R_{3,1}(\rho)U_2'
\end{aligned}$$

and therefore

$$R_3(\rho) = BS_1^{1/2}R_{3,1}(\rho)U_2' \tag{145}$$

Taking the norm of (145) gives

$$\|R_3(\rho)\| \leq \|S_1^{1/2}\|\|R_{3,1}(\rho)\| \leq \frac{\overline{\sigma}^{1/2}(V)}{\underline{\sigma}(A)}\frac{\rho}{\underline{\sigma}^2(A) + \rho} = \frac{\alpha_3\rho}{\beta_3 + \rho}$$

□

The proof of lemma 20 follows directly from lemma 64.

Proof of lemma 20. Using the SVD (128a), we can rewrite the residual matrix in terms of the intermediate residuals of lemma 64 as follows,

$$\begin{aligned} R(\rho) &:= (X'V_\rho^{-1}X)^+X'V_\rho^{-1} - X^+ + X^+VS(SVS)^+S \\ &= V_1\Sigma_1^{-1}[(U_1'V_\rho^{-1}U_1)^{-1}U_1'V_\rho^{-1} - U_1' + U_1'VS(SVS)^+S] \\ &= V_1\Sigma_1^{-1}[R_1(\rho) + R_2(\rho) - R_3(\rho)] \end{aligned}$$

and therefore we have the following bounds on $\|R(\rho)\|$,

$$0 \le \|R(\rho)\| \le \|\Sigma_1^{-1}\|(\|R_1(\rho)\| + \|R_2(\rho)\| + \|R_3(\rho)\|) \le \frac{1}{\underline{\sigma}(X)}\sum_{i=1}^{3}\frac{\alpha_i\rho}{\beta_i+\rho}$$

Absorbing the factor $1/\underline{\sigma}(X)$ into the constants α_i, we get (19). Taking the limit on the inequalities gives $\lim_{\rho\to 0^+}\|R(\rho)\| = 0$ and therefore $\lim_{\rho\to 0^+} R(\rho) = 0$, which is equivalent to (20). □

J Miscellaneous results

In this appendix, we prove lemma 21 and corollary 22.

Proof of lemma 21. Let $r = \operatorname{rank}(X)$ and denote the SVDs of X and V as

$$X = \begin{bmatrix} U_1 & U_2 \end{bmatrix}\begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} V_1' \\ V_2' \end{bmatrix}, \qquad V = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}\begin{bmatrix} S_1 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} Q_1' \\ Q_2' \end{bmatrix},$$

Since $\mathcal{R}(U_1) = \mathcal{R}(X) \subseteq \mathcal{R}(V_0)$, we have $\mathcal{R}(U_1') = \mathcal{R}(U_1'V_0^+U_1)$ by lemma 12. Moreover, U_1' is full row rank so $\mathbb{R}^r = \mathcal{R}(U_1') = \mathcal{R}(U_1'V_0^+U_1)$ and $U_1'V_0^+U_1$ is nonsingular.

We show $Z = V_1\Sigma_1^{-1}(U_1'V_0^+U_1)^{-1}\Sigma_1^{-1}V_1'$ is the pseudoinverse of $X'V_0^+X$ by checking (6),

$$\begin{aligned} ZX'V_0^+X &= V_1\Sigma_1^{-1}(U_1'V_0^+U_1)^{-1}\Sigma_1^{-1}V_1'V_1\Sigma_1U_1'V_0^+U_1\Sigma_1V_1' = V_1V_1' \\ X'V_0^+XZ &= (X'V_0X)'Z' = (ZX'V_0^+X)' = V_1V_1' \\ X'V_0^+XZX'V_0^+X &= X'V_0^+U_1\Sigma_1V_1'V_1V_1' = X'V_0^+U_1\Sigma_1V_1' = X'V_0^+X \\ ZX'V_0^+XZ &= V_1\Sigma_1^{-1}(U_1'V_0^+U_1)^{-1}\Sigma_1^{-1}V_1'V_1V_1' \\ &= V_1\Sigma_1^{-1}(U_1'V_0^+U_1)^{-1}\Sigma_1^{-1}V_1' = Z \end{aligned}$$

and therefore $Z = (X'V_0^+X)^+$.

Let $Y = F + U_1'(V - VU_2(U_2'VU_2)^+U_2V)U_1$ where $F = \Sigma_1V_1'EV_1\Sigma_1$, which we propose is the inverse of $U_1'V_0^+U_1$. By corollary 4 and the fact that $\mathcal{R}(U_1) \subseteq \mathcal{R}(V_0)$,

$$\begin{aligned} V_0^+V_0U_1 &= V_0V_0^+U_1 = U_1 \\ U_1'V_0^+V_0U_1 &= U_1'U_1 = I \end{aligned}$$

Since $U_2'U_1 = 0$ and $V_0 = V + XEX' = V + U_1FU_1'$, we have

$$\begin{aligned} VU_2 &= VU_2 + U_1FU_1'U_2 = V_0U_2 \\ U_1'V_0^+VU_2 &= U_1'V_0^+V_0U_2 = U_1'U_2 = 0 \end{aligned}$$

By corollary 4 and the fact that $\mathcal{R}(U_2'V) \subseteq \mathcal{R}(U_2'VU_2)$,

$$U_2'VU_2(U_2'VU_2)^+U_2'V = U_2'V$$

Using the above identities,

$$\begin{aligned} U_1'V_0^+U_1Y &= U_1'V_0^+U_1[F + U_1'(V - VU_2(U_2'VU_2)^+U_2V)U_1] \\ &= U_1'V_0^+[(V + U_1FU_1')U_1 - (I - U_1U_1')VU_1 \\ &\qquad - U_1U_1'VU_2(U_2'VU_2)^+U_2VU_1] \\ &= U_1'V_0^+[V_0U_1 - U_2U_2'VU_1 - VU_2(U_2'VU_2)^+U_2VU_1 \\ &\qquad + U_2U_2'VU_2(U_2'VU_2)^+U_2VU_1] \\ &= U_1'V_0^+[V_0U_1 - U_2U_2'VU_1 - V_0U_2(U_2'VU_2)^+U_2VU_1 + U_2U_2VU_1] \\ &= I \\ YU_1'V_0^+U_1 &= Y'(U_1'V_0^+U_1)' = (U_1'V_0^+U_1Y)' = I \end{aligned}$$

and therefore $Y = (U_1'V_0^+U_1)^{-1}$. Combining these results,

$$\begin{aligned} (X'V_0^+X)^+X' &= V_1\Sigma_1^{-1}(U_1'V_0^+U_1)^{-1}\Sigma_1^{-1}V_1'V_1\Sigma_1U_1' \\ &= V_1\Sigma_1^{-1}(F + U_1'(V - VU_2(U_2'VU_2)^+U_2V)U_1)U_1' \\ &= V_1\Sigma_1^{-1}(\Sigma_1V_1'EV_1\Sigma_1 + U_1'(V - VU_2(U_2'VU_2)^+U_2V)U_1)U_1' \\ &= X^+XEX' + X^+(V - VU_2(U_2'VU_2)^+U_2V)U_1U_1' \end{aligned}$$

Before deriving the final result, note that

$$\begin{aligned} VU_2(U_2'VU_2)^+U_2V_0V_0^+ &= V(SVS)^+V_0V_0^+ \\ &= V(V^{1/2}S)^+(SV^{1/2})^+V_0V_0^+ \\ &= V(V^{1/2}S)^+(SV^{1/2})^+ \\ &= V(SVS)^+ \end{aligned}$$

where we have used properties of the pseudoinverse, lemma 5, and

$$\mathcal{R}((V^{1/2}S)^+) = \mathcal{R}(V^{1/2}S) \subseteq \mathcal{R}(V^{1/2}) = \mathcal{R}(V) \subseteq \mathcal{R}(V_0)$$

Combining the above identities gives

$$\begin{aligned} (X'V_0^+X)^+X'V_0^+ &= [X^+XEX' + X^+(V - VU_2(U_2'VU_2)^+U_2'V)U_1U_1']V_0^+ \\ &= [X^+V_0 - X^+V(I - U_1U_1') - X^+VU_2(U_2'VU_2)^+U_2'VU_1U_1']V_0^+ \\ &= X^+ - [X^+VU_2U_2' + X^+VU_2(U_2'VU_2)^+U_2'V \\ &\qquad - X^+VU_2(U_2'VU_2)^+U_2'VU_2U_2']V_0^+ \\ &= X^+ - [X^+VU_2U_2' + X^+VU_2(U_2'VU_2)^+U_2'V_0 - X^+VU_2U_2']V_0^+ \\ &= X^+ - X^+VU_2(U_2'VU_2)^+U_2'V_0V_0^+ \\ &= X^+ - X^+V(SVS)^+ \end{aligned}$$

□

Proof of corollary 22. By lemma 21,

$$(X'V_0^+X)^+X'V_0^+ = X^+ - X^+V(SVS)^+ = (X'V_1^+X)^+X'V_1^+$$

□

Proof of lemma 23. Suppose $\beta \in \mathbb{R}^p$ such that (LGM) with nonzero probability. Noting that $e \in \mathcal{R}(V) \subseteq \mathcal{R}(V_0)$ (almost surely) because $e \sim \mathrm{N}(0, V)$, we have $VV^+e = e$ (almost surely) by lemma 3. Then

$$y = X\beta + e = X\beta + VV^+e = \begin{bmatrix} V & X \end{bmatrix} \begin{bmatrix} V^+e \\ \beta \end{bmatrix} \in \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$$

almost surely. Suppose that $y \in \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$ (almost surely). Then there exists $\theta \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ such that

$$y = \begin{bmatrix} V & X \end{bmatrix} \begin{bmatrix} \theta \\ \beta \end{bmatrix} = X\beta + V\theta$$

where $e = V\theta$ with nonzero probability.

Noting that $\mathcal{R}(V) \subseteq \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$ and $\mathcal{R}(X) \subseteq \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})$, we have

$$\begin{aligned}
& V_0V_0^+ \begin{bmatrix} V & X \end{bmatrix} = \begin{bmatrix} V_0V_0^+V & V_0V_0^+X \end{bmatrix} = \begin{bmatrix} V & X \end{bmatrix} \\
\Leftrightarrow \quad & \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) \subseteq \mathcal{R}(V_0) \\
& \begin{bmatrix} V & X \end{bmatrix} \begin{bmatrix} V & X \end{bmatrix}^+ V_0 = \begin{bmatrix} V & X \end{bmatrix} \begin{bmatrix} V & X \end{bmatrix}^+ V + \begin{bmatrix} V & X \end{bmatrix} \begin{bmatrix} V & X \end{bmatrix}^+ XEX' \\
& \qquad = V + XEX' = V_0 \\
\Leftrightarrow \quad & \mathcal{R}(V_0) \subseteq \mathcal{R}(\begin{bmatrix} V & X \end{bmatrix})
\end{aligned}$$

by corollary 4. Therefore $\mathcal{R}(\begin{bmatrix} V & X \end{bmatrix}) = \mathcal{R}(V_0)$ and the proof is complete. □

References

[1] A. C. Aitken, *On least squares and linear combinations of observations*, Proc. R. Soc. Edingburgh, 55 (1935), pp. 42–48.

[2] A. Albert, *Regression and the Moore-Penrose pseudoinverse*, vol. 94 of Mathematics in Science and Engineering, Academic Press, New York and London, 1972.

[3] A. Albert, *The Gauss–Markov Theorem for Regression Models with Possibly Singular Covariances*, SIAM J. Appl. Math., 24 (1973), pp. 182–187, `https://doi.org/10.1137/0124019`.

[4] A. Ali, J. Z. Kolter, and R. J. Tibshirani, *A Continuous-Time View of Early Stopping for Least Squares Regression*, in Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics, PMLR, Apr 2019, pp. 1370–1378.

[5] A. Alouani and W. Blair, *Use of a kinematic constraint in tracking constant speed, maneuvering targets*, IEEE Trans. Auto. Cont., 38 (1993), pp. 1107–1111.

[6] B. D. O. Anderson and J. B. Moore, *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, N. J., 1979.

[7] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1990.

[8] M. Arashi and S. Nadarajah, *On singular elliptical models*, Comm. in Statist. Theory Methods, 46 (2017), pp. 247–258, `https://doi.org/10.1080/03610926.2014.990103`.

[9] R. Bellman, *Introduction to Matrix Analysis, Second Edition*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, Jan 1997, `https://doi.org/10.1137/1.9781611971170`.

[10] M. Benning and M. Burger, *Modern regularization methods for inverse problems*, Acta Num., 27 (2018), pp. 1–111.

[11] M. Benzi and G. H. Golub, *A Preconditioner for Generalized Saddle Point Problems*, SIAM J. Matrix Anal. and Appl., 26 (2004), pp. 20–41.

[12] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Num., 14 (2005), pp. 1–137.

[13] D. P. Bertsekas, *Incremental Least Squares Methods and the Extended Kalman Filter*, SIAM J. Optim., 6 (1996), pp. 807–822.

[14] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, second ed., 2008.

[15] D. Bertsimas, A. King, and R. Mazumder, *Best subset selection via a modern optimization lens*, Ann. Stat., 44 (2016), pp. 813–852.

[16] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*, vol. I–II, Chapman and Hall/CRC, New York, 2015.

[17] S. Bourguignon, J. Ninin, H. Carfantan, and M. Mongeau, *Exact Sparse Approximation Problems via Mixed-Integer Programming: Formulations and Computational Performance*, IEEE Trans. Signal Process., 64 (2016), pp. 1405–1419.

[18] S. P. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[19] A. M. Bruckstein, D. L. Donoho, and M. Elad, *From Sparse Solutions of Systems of Equations to Sparse Modeling of Signals and Images*, SIAM Rev., 51 (2009), pp. 34–81.

[20] F. M. Callier and J. Winkin, *Convergence of the time-invariant Riccati differential equation towards its strong solution for stabilizable systems*, J. Math. Anal. Appl., 192 (1995), pp. 230–257.

[21] D. Calvetti, F. Pitolli, E. Somersalo, and B. Vantaggi, *Bayes Meets Krylov: Statistically Inspired Preconditioners for CGLS*, SIAM Rev., 60 (2018), pp. 429–461, `https://doi.org/10.1137/15M1055061`.

[22] D. Calvetti and E. Somersalo, *Inverse problems: From regularization to Bayesian inference*, Wiley Interdiscip. Rev. Comput. Stat., 10 (2018), p. e1427, `https://doi.org/10.1002/wics.1427`.

[23] E. J. Candès, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Comm. Pure Appl. Math., 59 (2006), pp. 1207–1223.

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

[25] G. De Nicolao and M. Gevers, *Difference and differential Riccati equations: A note on the convergence to the strong solution*, IEEE Trans. Auto. Cont., 37 (1992), pp. 1055–1057.

[26] C. E. de Souza, M. R. Gevers, and G. C. Goodwin, *Riccati equation in optimal filtering of nonstabilizable systems having singular state transition matrices*, IEEE Trans. Auto. Cont., 31 (1986), pp. 831–838.

[27] J. A. Díaz-García, V. Leiva-Sánchez, and M. Galea, *Singular Elliptical Distribution: Density and Applications*, Comm. in Statist. Theory Methods, 31 (2002), pp. 665–681, `https://doi.org/10.1081/STA-120003646`.

[28] H. S. Dollar, N. I. M. Gould, M. Stoll, and A. J. Wathen, *Preconditioning Saddle-Point Systems with Applications in Optimization*, SIAM J. Sci. Comp., 32 (2010), pp. 249–270.

[29] D. Donoho, M. Elad, and V. Temlyakov, *Stable recovery of sparse overcomplete representations in the presence of noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 6–18.

[30] H. Drygas, *Sufficiency and Completeness in the General Gauss-Markov Model*, Sankhyā A, 45 (1983), pp. 88–98, `https://www.jstor.org/stable/25050416`.

[31] H. C. Elman and G. H. Golub, *Inexact and Preconditioned Uzawa Algorithms for Saddle Point Problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1645–1661.

[32] R. Estrin and C. Greif, *On Nonsingular Saddle-Point Systems with a Maximally Rank Deficient Leading Block*, SIAM J. Matrix Anal. and Appl., 36 (2015), pp. 367–384.

[33] A. Ferrante and L. Ntogramatzidis, *Generalized Finite-Horizon Linear-Quadratic Optimal Control*, in Encyclopedia of Systems and Control, Springer-Verlag, 2014, pp. 1–8.

[34] A. Ferrante and L. Ntogramatzidis, *A note on finite-horizon LQ problems with indefinite cost*, Automatica, 52 (2015), pp. 290–293.

[35] A. Ferrante and L. Ntogramatzidis, *On the generalized algebraic Riccati equations*, IFAC–P. Online, 50 (2017), pp. 9555–9560.

[36] A. Frommer, R. Nabben, and D. B. Szyld, *Convergence of Stationary Iterative Methods for Hermitian Semidefinite Linear Systems and Applications to Schwarz Methods*, SIAM J. Matrix Anal. and Appl., 30 (2008), pp. 925–938.

[37] G. M. Fung and O. L. Mangasarian, *Equivalence of Minimal ℓ_0- and ℓ_p-Norm Solutions of Linear Equalities, Inequalities and Linear Programs for Sufficiently Small p*, J. Optim. Theory Appl., 151 (2011), pp. 1–10.

[38] C. F. Gauss, *Theoria motus corporum coelestium in sectionibus conicis solem ambientium, perthas et besser, werke*, Werke, 7 (1809), pp. 240–254.

[39] S. Gnot, H. Knautz, G. Trenkler, and R. Zmyslony, *Nonlinear unbiased estimation in linear models*, Statistics, 23 (1992), pp. 5–16, `https://doi.org/10.1080/02331889208802348`.

[40] A. Goldman and M. Zelen, *Weak generalized inverses and minimum variance linear unbiased estimation*, J. Res. Nat. Bur. Standards Sect. B, 68 (1964), pp. 151–172, `https://doi.org/10.6028/jres.068B.021`.

[41] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, fourth ed., 2013.

[42] J. Gorman and A. Hero, *Lower bounds for parametric estimation with constraints*, IEEE Trans. Inform. Theory, 36 (1990), pp. 1285–1301, `https://doi.org/10.1109/18.59929`.

[43] N. Gould, D. Orban, and T. Rees, *Projected Krylov Methods for Saddle-Point Systems*, SIAM J. Matrix Anal. and Appl., 35 (2014), pp. 1329–1343.

[44] J. Gross, *The general Gauss-Markov model with possibly singular dispersion matrix*, Stat. Pap., 45 (2004), pp. 311–336.

[45] M. Gruber, *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*, CRC Press, 1998.

[46] E. Haber, U. M. Ascher, and D. Oldenburg, *On optimization techniques for solving nonlinear inverse problems*, Inverse Problems, 16 (2000), pp. 1263–1280.

[47] B. E. Hansen, *A Modern Gauss–Markov Theorem*, Econometrica, 90 (2022), pp. 1283–1294.

[48] P. C. Hansen, *Discrete Inverse Problems*, Fundamentals of Algorithms, SIAM, Jan 2010.

[49] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity: the lasso and generalizations*, Chapman and Hall/CRC, 2019.

[50] J. P. Hespanha, *Linear Systems Theory*, Princeton University Press, Princeton and Oxford, second ed., 2018.

[51] A. E. Hoerl and R. W. Kennard, *Ridge Regression: Applications to Nonorthogonal Problems*, Technometrics, 12 (1970), pp. 69–82, https://doi.org/10.1080/00401706.1970.10488635.

[52] A. E. Hoerl and R. W. Kennard, *Ridge Regression: Biased Estimation for Nonorthogonal Problems*, Technometrics, 12 (1970), pp. 55–67, https://doi.org/10.1080/00401706.1970.10488634.

[53] J. Humpherys, P. Redd, and J. West, *A Fresh Look at the Kalman Filter*, SIAM Rev., 54 (2012), pp. 801–823.

[54] J. L. Jerez, E. C. Kerrigan, and G. A. Constantinides, *A sparse and condensed QP formulation for predictive control of LTI systems*, Automatica, 48 (2012), pp. 999–1002.

[55] A. M. Kagan and O. Salaevskii, *The admissibility of least-squares estimates is an exclusive property of the normal law*, Mat. Zametki, 6 (1969), pp. 81–89.

[56] R. E. Kalman, *A new approach to linear filtering and prediction problems*, Trans. ASME, J. Basic Engineering, (1960), pp. 35–45.

[57] H. B. Keller, *On the Solution of Singular and Semidefinite Linear Systems by Iteration*, SIAM J. Numer. Anal., 2 (1965), pp. 281–290.

[58] H. Knautz, *Nonlinear unbiased estimation in the linear regression model with nonnormal disturbances*, Journal of Statistical Planning and Inference, 81 (1999), pp. 293–309.

[59] K. Kuratowski, *Topology*, vol. I & II, Academic Press, New York, 1966.

[60] H. Kwakernaak and R. Sivan, *Linear Optimal Control Systems*, John Wiley and Sons, New York, 1972.

[61] P. S. Laplace, *Théorie analytique des probabilités*, vol. 7, Paris: Courcier, 1820.

[62] S. Le Borne and M. Wende, *Iterative Solution of Saddle-Point Systems from Radial Basis Function (RBF) Interpolation*, SIAM J. Sci. Comp., 41 (2019), pp. A1706–A1732.

[63] H. A. Le Thi, H. M. Le, and T. Pham Dinh, *Feature selection in machine learning: an exact penalty approach using a Difference of Convex function Algorithm*, Machine Learning, 101 (2015), pp. 163–186.

[64] A. M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comète*, Paris: Courcier, 2nd ed., 1806.

[65] E. L. Lehmann and H. Scheffé, *Completeness, Similar Regions, and Unbiased Estimation: Part I*, Sankhyā, 10 (1950), pp. 305–340, `https://www.jstor.org/stable/25048038`.

[66] E. L. Lehmann and H. Scheffé, *Completeness, Similar Regions, and Unbiased Estimation: Part II*, Sankhyā, 15 (1955), pp. 219–236, `https://www.jstor.org/stable/25048243`.

[67] J. R. Magnus and H. Neudecker, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley & Sons Ltd., Hoboken, NJ, third ed., 2019.

[68] A. A. Markov, *Wahrscheinlichkeits-rechnung*, Liepzig and Berlin, 2nd ed., 1912. Trans. H. Liebmann.

[69] G. Marsaglia, *Conditional Means and Covariances of Normal Variables with Singular Covariance Matrix*, J. Am. Stat. Assoc., 59 (1964), pp. 1203–1204, `https://doi.org/10.1080/01621459.1964.10480761`.

[70] T. Marzetta, *A simple derivation of the constrained multiple parameter Cramer-Rao bound*, IEEE Trans. Signal Process., 41 (1993), pp. 2247–2249, `https://doi.org/10.1109/78.218151`.

[71] N. Minamide, *An Extension of the Matrix Inversion Lemma*, SIAM J. Alg. Disc. Meth., 6 (1985), pp. 371–377, `https://doi.org/10.1137/0606038`.

[72] S. K. Mitra, *Unified Least Squares Approach to Linear Estimation in a General Gauss–Markov Model*, SIAM J. Appl. Math., 25 (1973), pp. 671–680, `https://doi.org/10.1137/0125065`.

[73] E. H. Moore, *On the reciprocal of the general algebraic matrix*, Bull. Amer. Math. Soc., 26 (1920), pp. 394–395.

[74] J. L. Mueller and S. Siltanen, *Linear and Nonlinear Inverse Problems with Practical Applications*, Computational Science & Engineering, SIAM, Oct 2012.

[75] R. Nabben and D. B. Szyld, *Schwarz Iterations for Symmetric Positive Semidefinite Problems*, SIAM J. Matrix Anal. and Appl., 29 (2007), pp. 98–116.

[76] S. Nelson and M. Neumann, *Generalizations of the projection method with applications to SOR theory for hermitian positive semidefinite linear systems*, Numer. Math., 51 (1987), pp. 123–141.

[77] A. Neumaier, *Solving Ill-Conditioned and Singular Linear Systems: A Tutorial on Regularization*, SIAM Rev., 40 (1998), pp. 636–666.

[78] A. Y. Ng, *Feature selection, L_1 vs. L_2 regularization, and rotational invariance*, in Twenty-first international conference on Machine learning - ICML '04, Banff, Alberta, Canada, 2004, ACM Press, p. 78.

[79] T. T. Nguyen, C. Soussen, J. Idier, and E.-H. Djermoune, *NP-hardness of ℓ_0 minimization problems: revision and extension to the non-negative setting*, in 2019 13th International conference on Sampling Theory and Applications (SampTA), 2019, pp. 1–4.

[80] D. P. O'Leary, *On bounds for scaled projections and pseudoinverses*, Linear Algebra Appl., 132 (1990), pp. 115–117, https://doi.org/10.1016/0024-3795(90)90056-I.

[81] R. Penrose, *A generalized inverse for matrices*, in Mathematical Proceedings of the Cambridge Philosophical Society, vol. 51, Cambridge University Press, Jul 1955, pp. 406–413, https://doi.org/10.1017/S0305004100030401.

[82] R. L. Plackett, *A Historical Note on the Method of Least Squares*, Biometrika, 36 (1949), pp. 458–460, https://doi.org/10.2307/2332682.

[83] J. Porrill, *Optimal Combination and Constraints for Geometrical Sensor Data*, Int. J. Robot. Res., 7 (1988), pp. 66–77.

[84] S. Portnoy, *Linearity of Unbiased Linear Model Estimators*, Amer. Statist., (2022), pp. 1–4.

[85] R. M. Pringle and A. A. Rayner, *Expressions for Generalized Inverses of a Bordered Matrix with Application to the Theory of Constrained Linear Models*, SIAM Rev., 12 (1970), pp. 107–115, https://doi.org/10.1137/1012007.

[86] B. M. Pötscher and D. Preinerstorfer, *A Modern Gauss-Markov Theorem? Really?*, 2022, https://doi.org/10.48550/arXiv.2203.01425.

[87] C. Radhakrishna Rao, *Information and the accuracy attainable in the estimation of statistical parameters*, Bull. Calcutta Math. Soc., 37 (1945), pp. 81–91.

[88] C. R. Rao, *Unified theory of linear estimation*, Sankhyā A, 33 (1971), pp. 371–394.

[89] C. R. Rao, *A Note on the IPM Method in the Unified Theory of Linear Estimation*, Sankhyā A, 34 (1972), pp. 285–288.

[90] C. R. Rao, *Linear Statistical Inference and Its Applications*, John Wiley and Sons, New York, second ed., 1973.

[91] C. R. Rao, *Unified theory of least squares*, Commun. Stat., 1 (1973), pp. 1–8.

[92] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*, John Wiley & Sons, Inc., New York, 1971.

[93] D. Rappaport and L. Silverman, *Structure and stability of discrete-time optimal systems*, IEEE Trans. Auto. Cont., 16 (1971), pp. 227–233, https://doi.org/10.1109/TAC.1971.1099702.

[94] J. B. Rawlings, D. Q. Mayne, and M. M. Diehl, *Model Predictive Control: Theory, Design, and Computation*, Nob Hill Publishing, Santa Barbara, CA, 2nd, paperback ed., 2020. 770 pages, ISBN 978-0-9759377-5-4.

[95] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Springer-Verlag, 1998.

[96] J. Seely, *A complete sufficient statistic for the linear model under normality and a singular covariance matrix*, Comm. in Statist. Theory Methods, 7 (1978), pp. 1465–1473.

[97] J. Sherman and W. J. Morrison, *Adjustment of an inverse matrix corresponding to a change in one element of a given matrix*, Ann. Math. Stat., 21 (1950), pp. 124–127.

[98] L. M. Silverman, *Discrete Riccati equations: Alternative algorithms, asymptotic properties, and system theory interpretations*, in Control and Dynamic Systems, C. T. Leondes, ed., vol. 12 of Control and Dynamic Systems, Academic Press, 1976, pp. 313–386.

[99] D. Simon, *Kalman filtering with state constraints: a survey of linear and nonlinear algorithms*, IET Control Theor. and Appl., 4 (2010), pp. 1303–1318.

[100] H. W. Sorenson, *Least–squares estimation: from Gauss to Kalman*, IEEE Spectrum, (1970), pp. 63–68.

[101] E. Soubies, L. Blanc-Féraud, and G. Aubert, *A Continuous Exact ℓ_0 Penalty (CEL0) for Least Squares Regularized Problem*, SIAM J. Imag. Sci., 8 (2015), pp. 1607–1639.

[102] E. Soubies, L. Blanc-Féraud, and G. Aubert, *A Unified View of Exact Continuous Penalties for ℓ_2-ℓ_0 Minimization*, SIAM J. Optim., 27 (2017), pp. 2034–2060.

[103] G. W. Stewart, *On scaled projections and pseudoinverses*, Linear Algebra Appl., 112 (1989), pp. 189–193, https://doi.org/10.1016/0024-3795(89)90594-6.

[104] S. M. Stigler, *Gauss and the Invention of Least Squares*, Ann. Stat., 9 (1981), pp. 465–474, https://www.jstor.org/stable/2240811.

[105] A. M. Stuart, *Inverse problems: A Bayesian perspective*, Acta Num., 19 (2010), pp. 451–559.

[106] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*, Other Titles in Applied Mathematics, SIAM, Jan 2005.

[107] H. Theil, *Principles of econometrics*, Wiley, New York, 1971.

[108] R. Tibshirani, *Regression Shrinkage and Selection Via the Lasso*, J. Roy. Stat. Soc. Ser. B, 58 (1996), pp. 267–288, https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

[109] A. N. Tikhonov, *On the solution of ill-posed problems and the method of regularization*, in Doklady Akademii Nauk, vol. 151, Russian Academy of Sciences, 1963, pp. 501–504.

[110] A. N. Tikhonov, A. V. Goncharsky, V. V. Stepanov, and A. G. Yagola, *Numerical methods for the solution of ill-posed problems*, vol. 328 of Mathematics and Its Applications, Springer Science & Business Media, 1995.

[111] S. A. Vavasis, *Stable Numerical Algorithms for Equilibrium Systems*, SIAM J. Matrix Anal. and Appl., 15 (1994), pp. 1108–1131, https://doi.org/10.1137/S0895479892230948.

[112] L.-S. Wang, Y.-T. Chiang, and F.-R. Chang, *Filtering method for nonlinear systems with constraints*, IEE Proc. Control Theory Appl., 149 (2002), pp. 525–531.

[113] Y. Wang and S. Boyd, *Fast model predictive control using online optimization*, IEEE Ctl. Sys. Tech., 18 (2010), pp. 267–278.

[114] M. A. Woodbury, *Inverting modified matrices*, 1950.

[115] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.

[116] X. Wu, B. Silva, and J. Yuan, *Conjugate Gradient Method for Rank Deficient Saddle Point Problems*, Numer. Algorithms, 35 (2004), pp. 139–154.

[117] H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, J. Roy. Stat. Soc. Ser. B, 67 (2005), pp. 301–320, https://doi.org/10.1111/j.1467-9868.2005.00503.x.

[118] G. Zyskind and F. B. Martin, *On Best Linear Estimation and General Gauss-Markov Theorem in Linear Models with Arbitrary Nonnegative Covariance Structure*, SIAM J. Appl. Math., 17 (1969), pp. 1190–1202, https://doi.org/10.1137/0117110.