LIKELIHOOD GEOMETRY OF CORRELATION MODELS

C. AMÉNDOLA - P. ZWIERNIK

Correlation matrices are standardized covariance matrices. They form an affine space of symmetric matrices defined by setting the diagonal entries to one. We study the geometry of maximum likelihood estimation for this model and linear submodels that encode additional symmetries. We also consider the problem of minimizing two closely related functions of the covariance matrix: the Stein's loss and the symmetrized Stein's loss. Unlike the Gaussian log-likelihood these two functions are convex and hence admit a unique positive definite optimum. Some of our results hold for general affine covariance models.

1. Introduction

Learning a covariance matrix is arguably one of the most basic problems in statistics. Learning structured covariance matrices has many applications ranging from signal processing to high-dimensional statistics. A type of structure that is most relevant in this article is given by affine constraints. More specifically, we focus on the problem of estimating a correlation, that is, a positive definite matrix Σ such that the diagonal entries are 1. This problem has appeared in statistics in many different contexts [4, 9, 11, 14, 16, 17].

The literature on estimating the correlation matrix focuses on two canonical problems. In the first, the data are represented by a symmetric positive

semidefinite (PSD) matrix S and the goal is to estimate the correlation matrix Σ by finding a suitable projection of S on the space of correlation matrices; see e.g. [4, 20]. In the second problem, each entry of Σ can be estimated from a different sample and, in consequence, the data is represented by a symmetric but not necessarily PSD matrix [9]. In this paper we focus on the first scenario assuming that S is positive definite.

The maximum likelihood estimation problem in this context has been studied recently in [20, Section 4] from an optimization and probabilistic point of view. In this paper we analyze it through the lenses of algebra and geometry.

The likelihood geometry of the bivariate correlation model has been analyzed in some detail; see [11, Example 18.3], [2, Example 2.38], and [17, Section 2.1]. In this bidimensional case, estimating the correlation parameter reduces to solving a cubic equation, so that the likelihood can have either one or three (real) critical points. To distinguish between these two situations it is enough to study the discriminant of the cubic polynomial. In this paper we perform a careful analysis of the discriminant locus to identify not only when there are three critical points, but also when all three points correspond to positive definite solutions (Theorem 3.1). We provide a similar study of the second derivative of the likelihood function to discover a rather surprising fact: for most statistically relevant data points, the likelihood function is concave (Theorem 3.2). Moreover, restricting to the equicorrelation model we are able to generalize the results on the number and nature of the critical points of the likelihood to any dimension. In particular, estimating the equicorrelation parameter reduces to solving a generalized cubic equation (Theorem 5.4).

A similarly detailed analysis is typically not possible for general linear correlation models. For the unrestricted correlation model we show that even in low-dimensional cases, the likelihood function can have many positive definite critical points. There are, however, some unexpected and positive features that are common to all linear correlation models. Although the likelihood function is typically not a concave function, we show that for linear correlation models there are always some data points for which this function is globally concave (Proposition 4.4). This is in sharp contrast to the general covariance model. As an alternative to the non-convex problem of optimizing the likelihood function, we propose minimizing two other functions that are also popular in covariance matrix estimation: the Stein's loss and the symmetrized Stein's loss. We show that, like the former, the latter is strictly convex (Proposition 2.1). All three functions appear in statistics in the context of likelihood-based inference for Gaussian data and they also lead to pseudolikelihood estimators for a wide range of non-Gaussian scenarios. Minimizing the Stein's loss in the correlation model was recently considered in [4], or more generally, as a Bregman projection on

affine models in [6, 12].

The paper is organized as follows. In Section 2 we introduce the three loss functions we consider over the space of correlation matrices. We analyze in detail the bivariate correlation model in Section 3, which serves as the basis for the study that follows. In Section 4 we look at the likelihood geometry of the most general unrestricted correlation model, and provide a simple efficient algorithm to optimize Stein's loss. On the other extreme, we consider the simplest one dimensional model, the equicorrelation model, in Section 5. We show that it exhibits many similarities with the bivariate case. Finally, we conclude in Section 6 with natural follow-up questions and possible future directions.

2. Three loss functions

Denote by $\mathbb{S}^n$ the set of $n \times n$ symmetric matrices and by $\mathbb{S}^n_+$ its subset given by all positive definite matrices. Define the divergence function $I : \mathbb{S}^n_+ \times \mathbb{S}^n_+ \to \mathbb{R}$

$$I(\Sigma_1 \| \Sigma_2) = \langle \Sigma_1, \Sigma_2^{-1} \rangle - \log \det(\Sigma_1 \Sigma_2^{-1}) - n, \tag{1}$$

where $\langle \Sigma_1, \Sigma_2^{-1} \rangle = \mathrm{trace}(\Sigma_1 \Sigma_2^{-1})$ is the standard inner product on $\mathbb{S}^n$. Note that $I(\Sigma_1 \| \Sigma_2)$ is convex both in Σ_1 and $K_2 = \Sigma_2^{-1}$ (but not in Σ_2). Fenchel duality also implies that $I(\Sigma_1 \| \Sigma_2) \geq 0$ with equality if and only $\Sigma_1 = \Sigma_2$.

Based on this divergence function we define two functions. For a fixed $S \in \mathbb{S}^n_+$, the *entropy loss* is $I(S \| \Sigma)$. Similarly, the *Stein's loss* is defined as $I(\Sigma \| S)$. The Stein's loss is a convex function of Σ but the entropy loss is not. Up to constant factors, the entropy loss function is the negative of the Gaussian log-likelihood and the Stein's loss is the negative dual likelihood; see [19]. For this reason the optimum of the entropy loss is called here the maximum likelihood estimator (MLE) and the optimum of the Stein's loss is called the dual MLE.

Since the divergence is not symmetric, it is useful to define the *symmetrized Stein's loss*:

$$L(\Sigma, S) = \frac{1}{2}\Big(I(S \| \Sigma) + I(\Sigma \| S)\Big) = \frac{1}{2}\Big(\langle S, \Sigma^{-1} \rangle + \langle \Sigma, S^{-1} \rangle\Big) - n.$$

The symmetrized Stein's loss has been recently used in the context of high-dimensional matrix estimation, e.g. [13, 18]. Following [15] we list some basic properties of the symmetrized Stein's loss:

(i) $L(\Sigma, S)$ is nonnegative, and zero if and only if $\Sigma = S$.

(ii) $L(\Sigma, S) = L(S, \Sigma)$,

(iii) $L(\Sigma, S) = L(\Sigma^{-1}, S^{-1})$,

(iv) $L(\Sigma, S) = L(P^T \Sigma P, P^T S P)$ for all invertible $P \in \mathbb{R}^{n \times n}$.

These properties made the symmetrized Stein's loss a convenient tool for theoretical analysis in statistics. As we argue in this paper, it can also offer a valuable and numerically tractable alternative to both the maximum likelihood estimator and the dual likelihood estimator. This observation is based on the following, rather surprising, result.

Proposition 2.1. *The symmetrized Stein's loss $L(\Sigma, S)$ is a strictly convex function both in $\Sigma \in \mathbb{S}_+^n$ and in $K = \Sigma^{-1}$.*

Remark 2.2. In the following proof and throughout we use the standard expressions for directional derivatives of functions of Σ. Denoting by ∇_U the directional derivative in the direction $U \in \mathbb{S}^n$ we have $\nabla_U \langle \Sigma, S^{-1} \rangle = \langle U, S^{-1} \rangle$ and

$$\nabla_U \langle S, \Sigma^{-1} \rangle = \langle S, \nabla_U \Sigma^{-1} \rangle = -\langle S, \Sigma^{-1} U \Sigma^{-1} \rangle.$$

Moreover,

$$\nabla_U \log \det \Sigma = \langle \Sigma^{-1}, U \rangle.$$

From this it immediately follows that the gradient of $\langle S, \Sigma^{-1} \rangle$ is $-\Sigma^{-1} S \Sigma^{-1}$ and the gradient of $\log \det \Sigma$ is Σ^{-1}.

Proof of Proposition 2.1. We show that $L(\Sigma, S)$ is a strictly convex function in Σ. The proof that it is also strictly convex in $K = \Sigma^{-1}$ is analogous. Since $\langle \Sigma, S^{-1} \rangle$ is linear in Σ, it is enough to show that $\langle S, \Sigma^{-1} \rangle$ is strictly convex. We check the second order conditions for convexity. The directional derivative in the direction $U \in \mathbb{S}^n$ satisfies

$$\nabla_U \langle S, \Sigma^{-1} \rangle = -\langle S, \Sigma^{-1} U \Sigma^{-1} \rangle.$$

Computing the directional derivative of $\nabla_U L(\Sigma, S)$ in the direction $V \in \mathbb{S}^n$ gives

$$\nabla_V \nabla_U \langle S, \Sigma^{-1} \rangle = \langle S, \Sigma^{-1} V \Sigma^{-1} U \Sigma^{-1} \rangle + \langle S, \Sigma^{-1} U \Sigma^{-1} V \Sigma^{-1} \rangle.$$

The function $\langle S, \Sigma^{-1} \rangle$ is strictly convex if and only if $\nabla_U \nabla_U \langle S, \Sigma^{-1} \rangle > 0$ for all $U \in \mathbb{S}^n$, $U \neq 0$. We have

$$\nabla_U \nabla_U \langle S, \Sigma^{-1} \rangle = 2 \langle S, KUKUK \rangle$$

where S is positive definite and $KUKUK$ is positive semidefinite. It follows that $\nabla_U \nabla_U \langle S, \Sigma^{-1} \rangle \geq 0$, and $\nabla_U \nabla_U \langle S, \Sigma^{-1} \rangle = 0$ if and only if $KUKUK = 0$ (equiv. $UKU = 0$). It remains to show that, for any positive definite K and symmetric U, $UKU = 0$ if and only if $U = 0$. But this is clear: if $U \neq 0$ then there exists x such that $y = Ux \neq 0$. This means $x^T U K U x = y^T K y > 0$ and thus $UKU \neq 0$. $\square$

All three loss functions are continuously differentiable with respect to Σ, making them suitable to optimization by standard techniques. Their gradients take a simple form. We consistently use the notation $K = \Sigma^{-1}$ and $W = S^{-1}$.

Proposition 2.3. *Let* $S \in \mathbb{S}_n^+$ *be fixed. Then the gradient of the entropy loss is*

$$\nabla I(S \| \Sigma) = K - KSK.$$

The gradient of the Stein's loss is

$$\nabla I(\Sigma \| S) = W - K.$$

The gradient of the symmetrized Stein's loss is

$$\nabla L(\Sigma, S) = \frac{1}{2}(W - KSK).$$

Proof. The expression for the gradient of $L(\Sigma, S)$ follows from the proof of Proposition 2.1. The other two cases can be verified in a similar way. $\square$

Denote by $\mathbb{S}_0^n$ the set of all $n \times n$ symmetric matrices with zeros on the diagonal. The set of correlation matrices is precisely $(I_n + \mathbb{S}_0^n) \cap \mathbb{S}_+^n$, where I_n is the identity matrix. Given a vector subspace $\mathcal{L} \subseteq \mathbb{S}_0^n$ we consider the linear correlation model

$$(I_n + \mathcal{L}) \cap \mathbb{S}_+^n. \tag{2}$$

Given a fixed matrix $S \in \mathbb{S}_+^n$, the goal is to optimize the entropy loss, the Stein's loss, and the symmetrized Stein's loss over this model.

The following result gives equations describing the critical points in the problem of minimizing the three loss functions over any spectrahedron $(A + \mathcal{L}) \cap \mathbb{S}_+^n$. The case of correlation models is recovered by taking $A = I_n$ and $\mathcal{L} \subseteq \mathbb{S}_0^n$.

Theorem 2.4. *Consider the model obtained by restricting $\mathbb{S}_+^n$ to an affine subspace $A + \mathcal{L}$ for $A \in \mathbb{S}_+^n$ and a linear subspace $\mathcal{L} \subseteq \mathbb{S}^n$. A necessary condition for $\Sigma \in \mathbb{S}^n$ to be a local minimum of the entropy loss is that*

$$\Sigma - A \in \mathcal{L} \quad and \quad K - KSK \in \mathcal{L}^\perp. \tag{3}$$

A necessary condition for Σ to be a local minimum of the Stein's loss

$$\Sigma - A \in \mathcal{L} \quad and \quad K - W \in \mathcal{L}^\perp. \tag{4}$$

A necessary condition for Σ to be a local minimum of the symmetrized Stein's loss

$$\Sigma - A \in \mathcal{L} \quad and \quad W - KSK \in \mathcal{L}^\perp. \tag{5}$$

If $\Sigma \in \mathbb{S}_+^n$, the last two conditions are also sufficient.

Proof. All three loss functions are continuously differentiable and the model $(A + \mathcal{L}) \cap \mathbb{S}_+^n$ is relatively open in $A + \mathcal{L}$. Therefore, a necessary condition for $\Sigma \in A + \mathcal{L}$ to be a local minimum is that all directional derivatives in the directions $U \in \mathcal{L}$ are zero. Conditions (3), (4), and (5) then follow from the gradient formula in Proposition 2.3. The last two conditions are also sufficient if $\Sigma \in \mathbb{S}_+^n$ because these optimization problems are convex, see Proposition 2.1.   □

Since the Stein's loss and the symmetrized Stein's loss are convex functions of Σ, these two functions admit a unique optimum in $(I_n + \mathcal{L}) \cap \mathbb{S}_+^n$. However, the entropy loss may admit several local optima. In all cases there will typically be many complex critical points. Given the algebraic nature of the critical equations, this number famously does not depend on S if it is generic.

Definition 2.5. Let S be a generic covariance matrix. The *ML degree* is the number of complex solutions to (3), the *dual ML degree* the number of complex solutions to (4), and the *SSL degree* the number of complex solutions to (5).

While the ML degree and dual ML degree[1] for linear covariance models have been studied before in algebraic statistics (see [5, 19]), the SSL degree we propose is novel and it should be interesting to study it further. Note that, unlike the function $I(\Sigma_1 \| \Sigma_2)$ whose critical points give rise to the first two degrees, the symmetrized Stein's loss $L(\Sigma, S)$ is itself an algebraic function.

3.   The bivariate correlation model

In this section we study in detail the bivariate correlation model. Thus, $n = 2$ and the model consists of all matrices of the form

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \qquad \text{for } \rho \in (-1, 1).$$

The problem of optimizing the entropy loss over this model is relatively well studied; see [11, Example 18.3], [2, Example 2.38], and [17, Section 2.1]. Here we largely extend the existing results.

3.1.   Critical points of the entropy loss

The derivative of the entropy loss $I(S \| \Sigma)$ with respect to ρ is $\frac{2}{(1-\rho^2)^2} f(\rho)$, where

$$f(\rho) = \rho(\rho^2 - 1) - S_{12}(1 + \rho^2) + \rho(S_{11} + S_{22}).$$

---

[1]The term *reciprocal ML degree* has been used recently to refer to the dual ML degree [3, 8]. We prefer *dual* from an optimization standpoint, following the original definition from Efron [7].

This polynomial has three roots, two of which may be complex. Note that $f(-1) = -2S_{12} - S_{11} - S_{22} < 0$ and $f(1) = -2S_{12} + S_{11} + S_{22} > 0$ (since S is positive definite) and hence f has at least one root in the open interval $(-1,1)$. As a consequence, if $f(\rho)$ has a unique root in $\mathbb{R}$, this unique root corresponds to a positive definite correlation matrix.

The polynomial $f(\rho)$ depends on S only through $a = \frac{S_{11}+S_{22}}{2}$ and $b = S_{12}$ and with this notation we have

$$f(\rho) = \rho^3 - b\rho^2 - \rho(1-2a) - b. \tag{6}$$

The condition that S is positive definite implies that $a > |b|$. Following [11] we use the discriminant to get the conditions on S that ensure there is a single critical point. The discriminant is

$$\Delta_f(b,a) = -4[b^4 - (a^2 + 8a - 11)b^2 + (2a-1)^3] \tag{7}$$

and $f(\rho)$ has a single root in $\mathbb{R}$ if and only if $\Delta_f(b,a) < 0$. When $\Delta_f(b,a) = 0$ there is an additional double root and there are three real roots when $\Delta_f(b,a) > 0$.
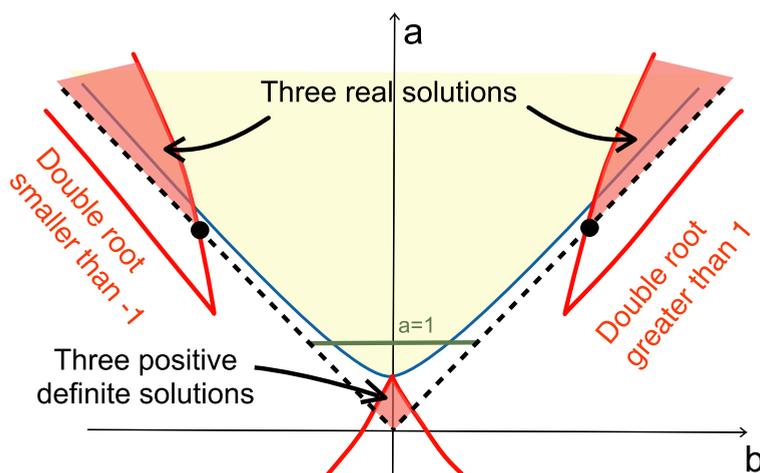

Figure 1: Likelihood geometry for the bivariate correlation model. The boundary of the positive definite cone is given by $a = |b|$ (dashed black line). The zero set of the discriminant $\Delta_f(b,a)$ is given by a red curve. In the red regions, the cubic in (6) has three real solutions. The yellow shaded region corresponds to (b,a) for which the entropy loss is convex; see Figure 2.

To add to the standard discussion of this example we note that even when $f(\rho)$ has three real roots, the other two roots may not lie in $(-1,1)$ in which case the entropy loss still has a unique optimum in the constrained region. To get the

full understanding of the likelihood geometry we depict the situation in Figure 1. The red curves represent the zero locus of the discriminant $\Delta_f(b,a)$. This zero locus divides the cone $a > |b|$ into four chambers: the central region where the discriminant is negative and three red regions where it is positive. All (b,a) that lie in the central chamber lead to a unique real critical point: the MLE. The points on the discriminant locus correspond to an additional double root. Since $f(1) = 2(a-b)$, f has a root at $\rho = 1$ if and only if $a = b$. This shows that the right top part of the discriminant corresponds to a double point in $(1,\infty)$. Indeed, the double point depends continuously on (b,a) and the only point where it could be equal to 1, namely when $a = b = 3 + 2\sqrt{2}$ (the right black dot in Figure 1), it is equal to $\rho = 1 + \sqrt{2} > 1$. Similarly, the fact that $f(-1) = -2(a+b)$ implies that the left component of the discriminant corresponds to double roots in $(-\infty, -1)$. At the point where it could be equal to -1, $a = -b = 3 + 2\sqrt{2}$, it is actually equal to $\rho = -1 - \sqrt{2} < -1$. Finally, the central component of the discriminant corresponds to double points $\rho \in (-1,1)$. The two points where the discriminant locus intersects the boundary point are $a = b = 3 - 2\sqrt{2}$ and $a = -b = 3 - 2\sqrt{2}$. These correspond to the double roots $\rho = 1 - \sqrt{2}$ and $\rho = -1 + \sqrt{2}$ respectively. This completes the picture of the situation.

Theorem 3.1. *For the bivariate correlation model, there is at least one positive definite critical point of the entropy loss function. There are three positive definite critical points if and only if $\Delta_f(b,a) > 0$ and $a < \frac{1}{2}$ (corresponding to the bottom red region in Figure 1).*

Proof. Since S is positive definite we have that $a > |b|$. Since $f(-1) < 0$ and $f(1) > 0$, it is clear that f has always at least one root in $(-1,1)$. As noted above, we get a unique real root (and hence a unique root in (-1,1)) when (b,a) lies in the central chamber of the cone $a > |b|$. Crossing the discriminant into the right red region introduces two extra real roots in $(1,\infty)$. Since the roots of the cubic are continuous in (b,a), we are guaranteed that none of them moves to $(-1,1)$ unless we cross the dashed line. By the same continuity argument, we see that the left red region has exactly one solution in $(-1,1)$ and that the bottom red region corresponds to exactly three solutions in $(-1,1)$. □

3.2. Convexity of the entropy loss

Even when the entropy loss has a unique minimum, it does not need to be convex for $\rho \in (-1,1)$. To understand better the complexity of this optimization procedure we now study precisely when the entropy loss is convex. We do it by studying the second order condition for convexity. The second derivative of the entropy loss with respect to ρ is equal to $\frac{2}{(1-\rho^2)^3} g(\rho)$, where

$$g(\rho) = 2a - 1 - 6b\rho + 6a\rho^2 - 2b\rho^3 + \rho^4.$$

The entropy loss is convex if and only if $g(\rho) \geq 0$ for $\rho \in (-1, 1)$. We have $g(-1) = 8(a + b)$ and $g(1) = 8(a - b)$. Both quantities are strictly positive in the cone $a > |b|$ and they can be zero only if $a = -b$ and $a = b$ respectively.

As for the polynomial f, we get a full picture by careful analysis of the discriminant locus. The discriminant $\Delta_g(b, a)$ of $g(\rho)$ is

$$\Delta_g := -256\Big(27b^6 - 27(2a^2 + 6a - 5)b^4 + 9(3a^4 + 36a^3 - 32a^2 + 8a + 1)b^2$$
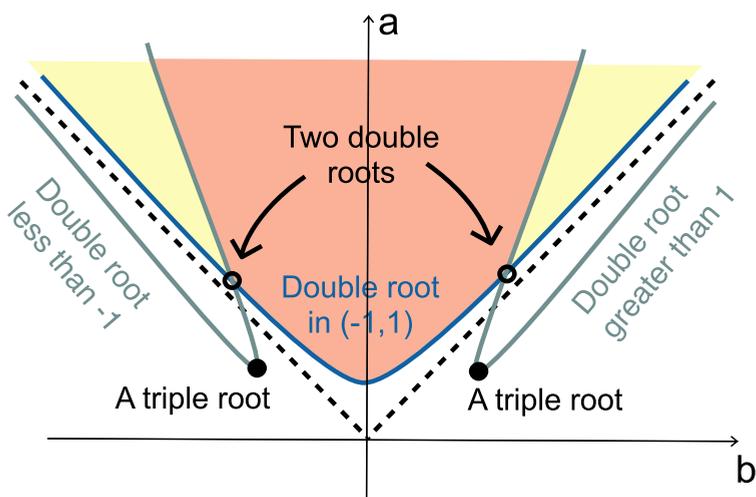$$-(2a - 1)(9a^2 - 2a + 1)^2\Big) \quad (8)$$

Figure 2: Likelihood geometry for the bivariate correlation model. To the plot of the cone in Figure 1 we add in the zero set of the sextic discriminant $\Delta_g(b, a)$. It is given by the blue curve (the central component) and the two green curves.

The somewhat complex boundary where the sextic $\Delta_g(b, a) = 0$ vanishes is sketched in Figure 2. The *central component of the discriminant* is given by the blue line that crosses the a-axis at $(b, a) = (0, \frac{1}{2})$. Two extra components that intersect the graph of $a = |b|$ (dashed line) are given in green. The points (b, a) where the discriminant is non-negative, given by the reddish cone in the center, corresponds to the situation when $g(\rho) \geq 0$ for all $\rho \in \mathbb{R}$. In this case the entropy function is of course convex in $(-1, 1)$ but convexity still holds under weaker conditions. The discriminant locus corresponds to double roots of the equation $g(\rho) = 0$. Since $g(1) = 8(a - b)$, g can have a root at $\rho = 1$ if and only if $a = b$. This implies that the upper right branch of the discriminant locus corresponds to double roots in $(1, \infty)$. We check this in the same way as earlier for the polynomial f using the fact that double roots are continuous in (b, a) and the point (b, a) on the intersection of the right branch of the discriminant

locus on the dashed line corresponds to a double root greater than 1. Similarly, $g(-1) = 8(a+b)$ is zero if and only if $a = -b$, which implies that the left upper branch of the discriminant locus corresponds to a double root in $(-\infty, -1)$. The central component of this locus corresponds to a double root in $(-1, 1)$.

Theorem 3.2. *The entropy function in the bivariate correlation model is convex over $(-1, 1)$ if and only (b, a) lies in the region above the central component of the discriminant $\Delta_g(b, a)$ (corresponding to the union of the reddish and two yellow regions in Figure 2).*

Proof. Since $g(\rho)$ is globally positive for (b, a) inside the central chamber determined by the components (the reddish region in the figure) and because the roots are continuous in (b, a), crossing one of the green lines can introduce roots only in $(-\infty, -1)$ or in $(1, +\infty)$. We conclude that $g(\rho)$ has no root in $(-1, 1)$ unless we are below the central component. In particular, it remains always positive in the interval $(-1, 1)$ and this is enough to guarantee convexity. Crossing the central component of the discriminant locus introduces two roots $-1 < \rho_1 < \rho_2 < 1$. Then g is negative in (ρ_1, ρ_2) and therefore the entropy loss is not convex in this region. $\square$

We conclude this section by noting that the analysis is much simpler for the Stein's loss, where the unique maximizer is given explicitly. We study this and the symmetrized Stein's loss in Section 5 as the special case of the equicorrelation model. There we show that the dual ML degree is 2 and give the formula for the optimum of the Stein's loss. We also show that the SSL degree is 4. We find this last number to be surprising. It indicates that, although the symmetrized Stein's loss is convex (and hence easier to handle from an optimization point of view), from an algebraic standpoint it is more complicated.

4.   The unrestricted correlation model

The unrestricted model corresponds to the case when $\mathcal{L} = \mathbb{S}_0^n$ in (2). Recall that we denote $K = \Sigma^{-1}$ and $W = S^{-1}$. By Theorem 2.4, any optimal correlation matrix must satisfy:

(a) for the entropy loss: $K_{ij} = (KSK)_{ij}$ for $i \neq j$,

(b) for the Stein's loss: $K_{ij} = W_{ij}$ for $i \neq j$,

(c) for the symmetrized Stein's loss: $(KSK)_{ij} = W_{ij}$ for $i \neq j$.

Solving these equations leads to high-degree polynomial systems. Some results on the ML degree, dual ML degree, and SSL degree for small n are

reported in Table 1. The case $n = 2$ was covered in great detail in Section 3. If $n = 3$ the dual ML degree is 5, the ML degree is 15, and the SSL degree is 28. Overall, the table indicates that the SSL degree increases faster than the ML degree, which increases faster than the dual ML degree. We currently have no conjectures on a general formula for these sequences. The numbers for $n > 4$ were computed using `LinearCovarianceModels.jl` as introduced in [19].

n	1	2	3	4	5	6	7	8	9
SSL degree	1	4	28	292	?	?	?	?	?
ML degree	1	3	15	109	1077	13695	?	?	?
dual ML degree	1	2	5	14	43	144	522	2028	8357

Table 1: The different algebraic degrees for the space of correlation matrices.

4.1. Likelihood geometry

Positive definite critical points The entropy loss may lead to many local optima, which can substantially complicate the maximum likelihood estimation. We provide basic analysis moving directly to the case $n = 3$ with the ML degree equal to 15. To obtain a better understanding of the multi-modality of the entropy loss, we study what happens to the MLE and critical points when we observe the ray $S = tI_n$ (multiple of the identity matrix). Denote by (s_{12}, s_{13}, s_{23}) the coordinates of the linear space $I_3 + \mathbb{S}_0^3$. The 15 critical points are exactly:

(a) The origin $(0,0,0)$ (i.e. the identity matrix)

(b) The six points $(\pm\sqrt{1-2t}, 0, 0)$, $(0, \pm\sqrt{1-2t}, 0)$ and $(0, 0, \pm\sqrt{1-2t})$.

(c) The four points $(\alpha, -\alpha, \alpha), (-\alpha, -\alpha, \alpha), (\alpha, \alpha, -\alpha), (-\alpha, -\alpha, -\alpha)$.

(d) The four points $(\beta, -\beta, \beta), (-\beta, -\beta, \beta), (\beta, \beta, -\beta), (-\beta, -\beta, -\beta)$.

where

$$\alpha = \frac{t - 1 + \sqrt{t^2 - 18t + 9}}{4} \qquad \beta = \frac{t - 1 - \sqrt{t^2 - 18t + 9}}{4}.$$

First we see the nature of these critical points. All of them are real when $0 < t < \frac{1}{2}$. The roots of $t^2 - 18t + 9$ are $t = 3(3 \pm 2\sqrt{2})$ and the points in (c) and (d) are real whenever $0 < t < 3(3 - 2\sqrt{2})$ or $t > 3(3 + 2\sqrt{2})$. However, if we want them all to be positive definite, we need to restrict to the first interval. Thus the number of positive definite critical points is 15 when $0 < t < 1/2$. When $t = 1/2$ seven of them collapse to the origin which gives 9 positive definite critical points

which is then true for all $1/2 \leq t < 3(3 - 2\sqrt{2})$. When $t = 3(3 - 2\sqrt{2})$ the eight points in (c) and (d) collapse to four points giving 5 positive definite critical points. For $t > 3(3 - 2\sqrt{2})$ we have a single positive definite solution.

By evaluating each of these critical points in the entropy loss, we obtain expressions that we can compare for every $t > 0$. All the points in the same group (a), (b), (c), (d) or (e) evaluate to the same likelihood value. Thus, for example, when $t < \frac{1}{2}$, we see that the points in (d) give the MLE. On the other hand, when $t > 3(3 - 2\sqrt{2})$, then the origin is the MLE.

Convexity of the entropy function   The entropy function is, in general, not convex. In the bivariate case we have seen however that convexity over the set of correlation matrices $I_n + \mathbb{S}_0^n$ happens for many data points. This motivates the following definition.

Definition 4.1. For an affine model $M = (A + \mathcal{L}) \cap \mathbb{S}_n^+$, its *convexity cone* $C(M)$ is the set of all $S \in \mathbb{S}_+^n$ for which $I(S\|\Sigma)$ is a convex function on M.

In the bivariate correlation model the convexity cone is described in Theorem 3.2. The name of the convexity cone is justified by the following result.

Proposition 4.2. *Fix an affine model $M \subseteq \mathbb{S}_+^n$. Then $C(M)$ is closed under multiplications by a scalar $t \geq 1$.*

Proof. Suppose $S \in C(M)$. If $t > 0$ then

$$I(tS\|\Sigma) = tI(S\|\Sigma) + (t-1)(n - \log\det(S^{-1}\Sigma)) - n\log t.$$

On the right hand side the only two terms that depend on Σ are $tI(S\|\Sigma)$ and $-(t-1)\log\det(\Sigma)$. The first is convex over M by assumption whenever $t > 0$. The second is convex over $\mathbb{S}_+^n$ whenever $t \geq 1$. $\square$

Remark 4.3. For the unrestricted covariance model $M = \mathbb{S}_+^n$ the convexity cone $C(M)$ is empty. By [20, Proposition 3.1], $I(S\|\Sigma)$ is convex in and only in the region where $2S - \Sigma$ is positive definite and so there is no S for which $I(S\|\Sigma)$ is convex over $\mathbb{S}_+^n$. So a first interesting question is to characterize models M for which $C(M)$ is nonempty.

Regarding the question in Remark 4.3, the following result shows that the convexity cone is always non-empty for any linear correlation model.

Proposition 4.4. *If $S = tI_n$ for $t \geq n/2$ then $I(S\|\Sigma)$ is convex over the set of all correlation matrices. In particular, for every linear correlation model M, its convexity cone is non-empty.*

Proof. In light of Proposition 4.2 it is enough to show that the statement holds for $S = \frac{n}{2} I_n$. We will verify the second order conditions for convexity. By Proposition 2.3, $\nabla_U I(S\|\Sigma) = \langle K - KSK, U \rangle$ for any $U \in \mathbb{S}^n$. Similarly, if $V \in \mathbb{S}^n$ then

$$\nabla_V \nabla_U I(S\|\Sigma) = \langle \nabla_V (K - KSK), U \rangle = \langle -KVK + KVKSK + KSKVK, U \rangle.$$

Thus, if $U = V$ and $S = \frac{n}{2} I_n$

$$\nabla_U \nabla_U I(S\|\Sigma) = \mathrm{tr}(KUKUK(2S - \Sigma)) = \mathrm{tr}(KUKUK(nI_n - \Sigma)).$$

To show that this is always nonnegative for every $U \in \mathbb{S}^n$, we use the fact that $KUKUK$ and $nI_n - \Sigma$ are always positive semidefinite. For $nI_n - \Sigma$, this follows from the fact that the maximum eigenvalue of any correlation matrix is bounded above by $\mathrm{tr}(\Sigma) = n$. $\qquad\square$

4.2. Numerical optimization of the Stein's loss

Let $K(x)$ be the matrix equal to $W = S^{-1}$ outside the diagonal and whose diagonal is $x = (x_1, \ldots, x_n)$. By item (b) in the beginning of this section we get that the unique correlation matrix $\check{\Sigma}$ that optimizes the Stein's loss satisfies $\check{K}_{ij} = W_{ij}$ for $i \neq j$. In other words, $\check{K} = \check{\Sigma}^{-1}$ satisfies $\check{K} = K(x^*)$ for some $x^* \in \mathbb{R}^n$. We now exploit this dimensionality reduction (from matrices to vectors) to construct a simple numerical scheme to optimize the Stein's loss, which provides a computationally efficient alternative to the algorithm developed in [4].

Proposition 4.5. *The unique point $\check{\Sigma}$ optimizing the Stein's loss over the set of correlation functions satisfies $\check{K} = K(x^*)$, where $x^* \in \mathbb{R}^n$ is the maximum of the concave function*

$$f(x) = \log \det K(x) - \mathrm{tr}(K(x)), \qquad x \in \mathbb{R}^n. \tag{9}$$

Proof. The Lagrangian of the problem of minimizing $I(\Sigma\|S)$ over the set of correlation matrices is

$$\mathcal{L}(\Sigma, \Lambda) = \langle \Sigma, W \rangle - \log \det(\Sigma W) - n + \langle \Lambda, \Sigma - I_n \rangle,$$

where Λ is a diagonal $n \times n$ matrix of Lagrange multipliers. The dual function is $h(\Lambda) = \inf_{\Sigma \in \mathbb{S}_+^n} \mathcal{L}(\Sigma, \Lambda)$ and the infimum is obtained for $\Sigma = (W + \Lambda)^{-1}$. After plugging this in we obtain

$$
\begin{aligned}
h(\Lambda) &= \log \det(W + \Lambda) - \log \det W - \mathrm{tr}(\Lambda) \\
&= \log \det(W + \Lambda) - \mathrm{tr}(W + \Lambda) - (\log \det W - \mathrm{tr}(W)).
\end{aligned}
$$

The convex dual problem is given by maximizing $h(\Lambda)$ over all diagonal matrices Λ. A simple reparametrization $K(x) = W + \Lambda$ concludes the proof. $\qquad\square$

The optimization problem of maximizing (9) can be very reliably done using gradient descent, which can be complemented with Newton's method after some burn-in steps. Denoting $\Sigma(x) = (K(x))^{-1}$ we get

$$\nabla f(x) = \operatorname{diag}(\Sigma(x)) - \mathbf{1} \quad \text{and} \quad \nabla^2 f(x) = -\Sigma(x) \circ \Sigma(x), \qquad (10)$$

where $\circ$ denotes the Hadamard product. Our basic implementation of this can easily handle n in the hundreds.

Finally we note that, alternatively, a simple coordinate descent scheme is possible. For any $A \in \mathbb{S}^n$ denote by $A_{\searrow i, i}$ the vector $\mathbb{R}^{n-1}$ whose elements are A_{ij} for $j \neq i$. Similarly, $A_{\searrow i}$ denotes the matrix in $\mathbb{S}^{n-1}$ obtained from A by removing its i-th row and i-th column.

Lemma 4.6. *The maximum of $f(x)$ with respect to x_i with other entries of x fixed is*

$$x_i^* = 1 + W_{i, \searrow i}(K_{\searrow i}(x))^{-1} W_{\searrow i, i}.$$

Proof. By (10), $\frac{\partial f}{\partial x_i}(x) = 0$ if and only if $\Sigma_{ii}(x) = 1$. Since $\Sigma_{ii}(x) = \frac{\det K_{\searrow i}(x)}{\det K(x)}$, equivalently $\det K_{\searrow i}(x) = \det K(x)$. By a standard Schur complement argument,

$$\det K(x) = (K_{ii}(x) - K_{i, \searrow i}(x)(K_{\searrow i})^{-1} K_{\searrow i, i}) \det K_{\searrow i}(x).$$

Since $K_{ii}(x) = x_i$ and $K_{i, \searrow i}(x) = W_{i, \searrow i}$ the corresponding partial derivative vanishes if $x_i - W_{i, \searrow i}(K_{\searrow i}(x))^{-1} W_{\searrow i, i} = 1$. $\qquad \square$

Note that since f is a smooth and strictly concave, it is clear that this coordinate descent procedure converges.

5.   Equicorrelation model

An interesting example of a tractable linear correlation models is the equicorrelation model, which consists of positive definite matrices for which $\Sigma_{ij} = \rho$ for all $i \neq j$. In other words $\Sigma = (1 - \rho)I_n + \rho \mathbf{11}^T$, where $\mathbf{1} \in \mathbb{R}^n$ is the vector of ones and so $\mathcal{L} = \operatorname{span}\{\mathbf{11}^T - I_n\}$ in (2). We denote this model by E_n. Note that E_2 is precisely the bivariate correlation model studied in Section 3.

The common off-diagonal entry ρ for every matrix in E_n necessarily satisfies $-\frac{1}{n-1} < \rho < 1$ for $\Sigma = (1 - \rho)I_n + \rho \mathbf{11}^T$ to be positive definite. This is a consequence of the following well-known result.

Lemma 5.1. *If $A = (x - y)I_n + y\mathbf{11}^T$ then $\det A = (x - y)^{n-1}(x + y(n - 1))$ and $A \in \mathbb{S}_+^n$ if and only if*

$$x > 0 \qquad \text{and} \qquad -\frac{x}{n-1} < y < x.$$

Furthermore, if $\det(A) \neq 0$ *then*

$$A^{-1} = \frac{1}{x-y}\left(I_n - \frac{y}{x+(n-1)y}\mathbf{1}\mathbf{1}^T\right).$$

In this way, if $\Sigma = (1-\rho)I_n + \rho\mathbf{1}\mathbf{1}^T$ then from Lemma 5.1 we obtain that $K = (c-d)I_n + d\mathbf{1}\mathbf{1}^T$, with

$$c = \frac{1}{1-\rho}\frac{1+(n-2)\rho}{1+(n-1)\rho}, \qquad d = -\frac{1}{1-\rho}\frac{\rho}{1+(n-1)\rho}. \tag{11}$$

5.1. Algebraic degrees

In this section we consider the problem of maximizing the entropy loss, the Stein's loss, and the symmetrized Stein's loss for the equicorrelation model E_n. We start with a useful result that exploits the symmetry of the model to simplify our calculations. Define the symmetrized versions of S and $W = S^{-1}$ as

$$\overline{S} = \frac{1}{n!}\sum_{P\in\mathcal{S}_n}PSP^T, \qquad \overline{W} = \frac{1}{n!}\sum_{P\in\mathcal{S}_n}PWP^T,$$

where $\mathcal{S}_n$ is the set of all $n\times n$ permutation matrices. The crucial observation is that the equicorrelation model is invariant under the permutation group action.

Lemma 5.2. *For every* $S \in \mathbb{S}^n$ *and* $\Sigma \in E_n$ *it holds that*

$$\langle S, \Sigma^{-1}\rangle = \langle \overline{S}, \Sigma^{-1}\rangle \qquad and \qquad \langle \Sigma, W\rangle = \langle \Sigma, \overline{W}\rangle.$$

Proof. If $\Sigma \in E_n$ then $P\Sigma P^T = \Sigma$ and $P^T\Sigma^{-1}P = \Sigma^{-1}$ for every $P \in \mathcal{S}_n$, so that

$$\langle S, \Sigma^{-1}\rangle = \frac{1}{n!}\sum_{P\in\mathcal{S}_n}\langle S, P^T\Sigma^{-1}P\rangle = \frac{1}{n!}\sum_{P\in\mathcal{S}_n}\langle PSP^T, \Sigma^{-1}\rangle = \langle \overline{S}, \Sigma^{-1}\rangle.$$

The equality $\langle \Sigma, W\rangle = \langle \Sigma, \overline{W}\rangle$ is argued in the same way. □

Lemma 5.2 immediately implies that optimizing the entropy loss $I(S\|\Sigma)$ is equivalent to optimizing the entropy loss $I(\overline{S}\|\Sigma)$ (the difference between the two does not depend on Σ). Similarly, optimizing $I(\Sigma\|S)$ is equivalent to optimizing $I(\Sigma\|\overline{W}^{-1})$. Moreover,

$$L(\Sigma, S) = \tfrac{1}{2}(\langle \overline{S}, \Sigma^{-1}\rangle + \langle \Sigma, \overline{W}\rangle) - n \qquad \text{for all } S \in \mathbb{S}^n_+ \text{ and } \Sigma \in E_n.$$

Let $\bar{c}, \bar{d}$ denote the diagonal and the off-diagonal entry of $\overline{W}$ so that $\overline{W} = (\bar{c}-\bar{d})I_n + \bar{d}\mathbf{1}\mathbf{1}^T$. Recall that c, d are the corresponding entries of K.

Theorem 5.3. *The dual ML degree in the equicorrelation model E_n is 2 for any $n \geq 2$. If $\bar{d} = 0$ then the unique critical point is $\check{\rho} = 0$. If $\bar{d} \neq 0$, there are two critical points given by*

$$\check{\rho} = \frac{1 + (n-2)\bar{d} \pm \sqrt{(n\bar{d}+1)^2 - 4\bar{d}}}{2(n-1)\bar{d}}. \tag{12}$$

The unique positive definite solution $\check{\Sigma}$ corresponds to the solution that takes the negative sign in (12). Moreover, $\check{K} = \check{\Sigma}^{-1}$ is equal to $\bar{d}$ outside of the diagonal and all its diagonal entries are equal and given by taking the positive sign in

$$\frac{1}{2}\left(1 - (n-2)\bar{d} \pm \sqrt{(n\bar{d}+1)^2 - 4\bar{d}}\right). \tag{13}$$

Proof. For the Stein's loss the condition that the critical point $\check{K}$ must satisfy is $\langle \check{K} - W, \mathbf{11}^T - I_n \rangle = 0$, or in other words, the sums of all the off-diagonal elements of both $\check{K}$ and W are equal. Using (11) the critical equation becomes

$$-\frac{1}{1-\rho}\frac{\rho}{1+(n-1)\rho} = \bar{d}. \tag{14}$$

If $\bar{d} = 0$ there is a unique solution given by $\rho = 0$. If $\bar{d} \neq 0$, this leads to a quadratic equation in ρ:

$$(n-1)\bar{d}\rho^2 - ((n-2)\bar{d}+1)\rho - \bar{d} = 0$$

with roots satisfying (12). We can rewrite (14) as $-\rho = \bar{d}(1-\rho)(1+(n-1)\rho)$. Geometrically, this corresponds to intersecting the line $-\rho$ with the parabola $\bar{d}(1-\rho)(1+(n-1)\rho)$. If $\bar{d} > 0$, the intersection point with larger ρ coordinate has $\rho > 1$ and thus does not lead to a positive definite solution. The other solution clearly satisfies $-\frac{1}{n-1} < \rho < 0$. This point is obtained by choosing the negative sign in (12). Similarly, if $\bar{d} < 0$ the intersection point with the larger ρ coordinate satisfies $0 < \rho < 1$ and the other intersection point satisfies $\rho < -\frac{1}{n-1}$. The former is given by choosing the negative sign in (12) and the latter by the positive sign. We conclude that in both cases we should take the negative sign in (12) to guarantee $-\frac{1}{n-1} < \rho < 1$. The last part of the statement follows from the formula for the off-diagonal entry of Σ^{-1} by taking $\rho = \check{\rho}$. $\qquad \square$

For the entropy loss, we have no guarantee that there is a unique positive definite solution and the situation becomes slightly more complicated, but also more interesting.

Let a, b be the diagonal and the off-diagonal entry of $\bar{S}$, so that $\bar{S} = (a-b)I_n + b\mathbf{11}^T$. Using Lemma 5.2, to optimize $I(S\|\Sigma)$ over the equicorrelation model E_n, we can equivalently optimize

$$I(\bar{S}\|\Sigma) = (n-1)\log\left(\frac{1-\rho}{a-b}\right) + \log\left(\frac{1+\rho(n-1)}{a+b(n-1)}\right) + n(ac + bd(n-1) - 1), \tag{15}$$

where c, d are the entries of $K = \Sigma^{-1}$ as given in (11). Recall that $\bar{c}, \bar{d}$ denote the entries of $\overline{W} = (\bar{c} - \bar{d})I_n + \bar{d}\mathbf{1}\mathbf{1}^T$. The next result provides both the ML degree and the SSL degree we had promised at the end of Section 3.

Theorem 5.4. *The ML degree of the equicorrelation model E_n is 3 for any $n \geq 2$. The critical points are given by solutions of the cubic equation $f_n(\rho) = 0$, where*

$$f_n(\rho) := (n-1)\rho^3 + ((n-2)(a-1) - (n-1)b)\rho^2 + (2a-1)\rho - b. \qquad (16)$$

The SSL degree of E_n is 4 for any $n \geq 2$. The critical points satisfy the quartic

$$(n-1)^2 \bar{d}\rho^4 - 2(n^2 - 3n + 2)\bar{d}\rho^3 + ((n^2 - 6n + 6)\bar{d} + (n-2)a - (n-1)b)\rho^2$$
$$+ 2(a + (n-2)\bar{d})\rho + \bar{d} - b = 0.$$

Proof. If $K = (c-d)I_n + d\mathbf{1}\mathbf{1}^T$ and $\overline{S} = (a-b)I_n + b\mathbf{1}\mathbf{1}^T$ then the matrix $K\overline{S}K$ has the form $(x-y)I_n + y\mathbf{1}\mathbf{1}^T$ with

$$y = ad(2c + (n-2)d) + b((c + (n-2)d)^2 + (n-1)d^2).$$

The critical point of $I(\overline{S}\|\Sigma)$ must satisfy that $y = d$. Using (11), this translates to a cubic equation (16). To compute the SSL degree, note that, by Lemma 5.2, $L(\Sigma, S) = \frac{1}{2}(\langle \overline{S}, \Sigma^{-1}\rangle + \langle \Sigma, \overline{W}\rangle) - n$. Since $\bar{c}, \bar{d}$ are the entries of $\overline{W}$, we have

$$L(\Sigma, S) = \frac{n}{2}\left(\bar{c} + ac + (n-1)(\bar{d}\rho + bd)\right) - n.$$

The equation $\frac{\partial}{\partial\rho}L(\Sigma, S) = 0$ is equivalent to the vanishing of the given quartic polynomial in ρ. $\qquad \square$

5.2. Positive definite critical points

Recall that the ML degree is three and thus the cubic $f_n(\rho) = 0$ in (16) always has a real root. Note that $f_n(-1) = -\frac{n}{(n-1)^2}(a + (n-1)b)$ and $f_n(1) = n(a-b)$. Since $\overline{S} \in \mathbb{S}_+^n$, by Lemma 5.1 we have that $f_n(1) > 0$ and $f_n(-1) < 0$. This means that in fact f_n always has a root in $(-\frac{1}{n-1}, 1)$ which corresponds to a positive definite solution. It has precisely one real root if its discriminant

$$\begin{aligned}
\Delta_{f,n}(b,a) \;=\; & -4(n-1)^3 b^4 + 12(n-2)(a-1)b^3 \\
& -4(n-1)((3n^2 - 13n + 13)a^2 - 2a(3n^2 - 8n + 8) + 3n^2 - n + 1)b^2 \\
& +4(n-2)(a-1)((n^2 - 6n + 6)a^2 - (2n^2 - n + 1)a + n^2)b \\
& +(2a-1)^2((n-2)^2 a^2 - 2n^2 a + n^2) \qquad\qquad\qquad (17)
\end{aligned}$$

is negative. Conversely, if $\Delta_{f,n} > 0$ there are three real roots and if $\Delta_{f,n} = 0$ one of them is a double root. Note that if $n = 2$ then $\Delta_{f,n}$ is precisely the discriminant

Δ_f in (1) and the underlying geometry was carefully explained in Section 3. To get a complete picture for how multiple roots may arise in this optimization problem, we follow the same strategy as for the bivariate model. The main difference is that the region is no longer symmetric about the a-axis since the left boundary $a = -(n-1)b$ moves toward it; see Figure 3.
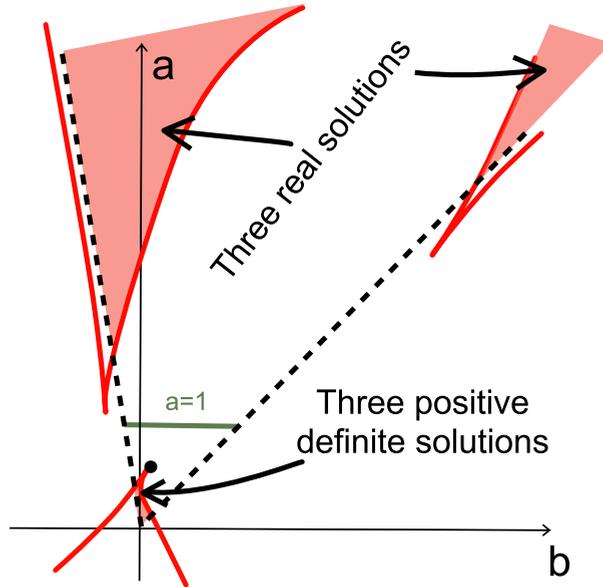

Figure 3: Likelihood geometry for $n > 2$. The set of all points in the positive definite cone such that $a = 1$ is marked in green.

One stark difference from the $n = 2$ case that can be appreciated in Figure 3 is that for $n > 2$ the discriminant $\Delta_{f,n}(b,a)$ crosses the a-axis three times, not only at the point $(0, \frac{1}{2})$ but also when $b = \frac{n(n \pm 2\sqrt{n-1})}{(n-2)^2}$. Note however that both of these extra points converge to 1 as $n \to \infty$.

The discriminant intersects the boundary of the positive definite cone in four points. On the $a = b$ side these are given by $b = \pm 2\sqrt{n(n-1)} + 2n - 1$ (note that as $n \to \infty$, one goes to 0 and the other goes to ∞) while on the $a = -(n-1)b$ side these are: $b = \frac{1}{(\sqrt{n} \pm 1)^2}$. This means that as $n \to \infty$ these points converge to $(b,a) = (0,1)$. One of the consequences is that the red chamber in the bottom shrinks in area towards zero.

In summary, we observe the following pattern for every $n \geq 2$. Inside the positive definite cone there are four chambers induced by the discriminant locus (the non-central chambers are marked in red in Figure 3). The large central chamber leads to only one real critical point of the entropy loss. On the other

hand, if (b, a) lies in one of the other chambers, then there are 3 real critical points in the entropy loss. However, in the two side red chambers there are two extra critical points outside the positive definite cone (that is, there is a unique positive definite solution to the likelihood equations) and only the bottom chamber (which shrinks as n grows) induces all three positive definite solutions.

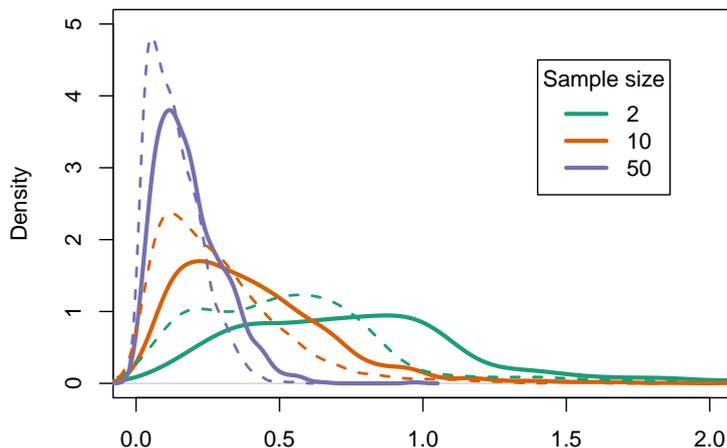**The density of the distance from the truth**

Figure 4: Density plots for the Euclidean distance of (b, a) from $(\rho^*, 1)$ where $\rho^* = 0.5$ for $n = 2$ (solid lines) and $n = 10$ (dashed lines). In both cases we consider sample sizes, $2, 10, 50$.

Remark 5.5. From the mathematical perspective we would like to understand the geometry of the entropy loss for every pair b, a for which the entropy loss is defined. However, it is important to note that statisticians will typically care only about a small neighborhood of $(b^*, a^*) = (\rho^*, 1)$, where ρ^* is the true data generating correlation. To illustrate this point, we show in Figure 4 the distribution of the distance of (b, a) from $(\rho^*, 1)$ assuming that the data follow the Gaussian distribution with covariance in E_n and that $\rho^* = 0.5$. Even for small n and relatively small sample sizes this distance is small with a very high probability. A similar behavior will be observed for any sub-Gaussian distribution.

Remark 5.5 has an interesting consequence for our analysis. Although the entropy loss can have three positive definite critical points, in statistical practice this will happen rarely even for small sample sizes; see Figure 5.
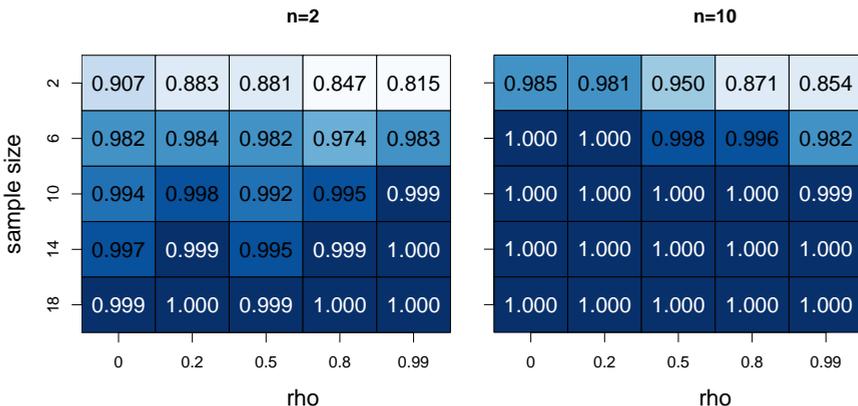
Figure 5: The probability that the entropy loss in the equicorrelation model E_n has one critical point or equivalently, the probability that $\Delta(b, a) < 0$, as a function of the true correlation and the sample size.

5.3. Convexity

In the equicorrelation model E_n, the entropy loss given in (15) depends on S only through the entries a, b of the symmetrized matrix $\overline{S}$ (up to a term that does not depend on ρ). To describe the convexity cone $C(E_n)$ (c.f. Definition 4.1) it is then enough to find for which matrices $\overline{S}$ the entropy loss is convex. We do it by studying the second order condition for convexity. The second derivative of the entropy loss with respect to ρ is equal to

$$\frac{n(n-1)}{(1-\rho)^3(1+(n-1)\rho)^3} g_n(\rho),$$

where

$$
\begin{aligned}
g_n(\rho) \;=\;& 2a - 1 + 2b(n-2) - (6b(n-1) + n - 2)\rho + 6a(n-1)\rho^2 \quad (18) \\
+\;& ((2a-1)(n-2) - 2b(n-1))(n-1)\rho^3 + (n-1)^2\rho^4.
\end{aligned}
$$

The entropy loss is convex in the positive definite cone, which by Lemma 5.1 consists of equicorrelation matrices with $\rho \in \left(-\frac{1}{n-1}, 1\right)$, if and only if $g_n(\rho) \geq 0$ in this interval.

Since $g_n\left(-\frac{1}{n-1}\right) = \frac{2n^2(a+b(n-1))}{(n-1)^2}$ and $g_n(1) = 2(a-b)n^2$ are positive in $\mathbb{S}^n_+$, we believe that a similar analysis to the one for the bivariate case in Section 3 is possible. Alternatively, a simple lower bound for $C(E_n)$ can be easily obtained without studying the discriminant $\Delta_{g,n}$ of g_n.

Lemma 5.6. *In the equicorrelation model E_n, if the entries a and b of $\overline{S}$ satisfy*

$$-\frac{a}{n-1} + \frac{n}{2(n-1)} \leq b \leq a - \frac{n}{2(n-1)} \qquad (19)$$

then the entropy loss function is strictly convex in the positive definite cone.

Proof. We use the result from [20, Proposition 3.1] which states that whenever $2\overline{S} - \Sigma > 0$, the conclusion follows. In our case, the diagonal entries of $2\overline{S} - \Sigma$ are $2a - 1$ and the off-diagonal entries equal $2b - \rho$. Hence, by Lemma 5.1, $2\overline{S} - \Sigma > 0$ if and only if $2a > 1$ and $-\frac{2a-1}{n-1} < 2b - \rho < 2a - 1$, or equivalently,

$$2a > 1 \qquad \text{and} \qquad 1 + 2(b - a) < \rho < 2b + \frac{2a - 1}{n - 1}.$$

Since $-\frac{1}{n-1} < \rho < 1$ always holds for all ρ in the correlation model, to conclude that the function is globally convex, we need that $2b + \frac{2a-1}{n-1} \geq 1$ and $1 + 2(b - a) \leq -\frac{1}{n-1}$. We easily show that this is equivalent to $-\frac{a}{n-1} + \frac{n}{2(n-1)} \leq b \leq a - \frac{n}{2(n-1)}$. $\square$

6. Further examples and open questions

Our study of the likelihood geometry of linear correlation models leads to several new insights both for linear correlation models and general affine covariance models. We believe that some of them can be further explored from an statistical as well as an algebraic perspective. In what follows we offer some guidance on potential open questions that we found particularly interesting.

The convexity cone   One of the new objects to study is the convexity cone, which we showed to be always non-empty for every linear correlation model. It would be useful to describe the convexity cone for the unconstrained correlation model. Another question is whether we can describe the convexity cone of one-dimensional linear correlation models directly in terms of their generator. It follows from the proof of Proposition 4.4 that the convexity cone of an affine covariance model M is always non-empty when we can uniformly bound the maximal eigenvalue of the matrices $\Sigma \in M$. It would be interesting to see if this eigenvalue condition is necessary for non-emptiness of the convexity cone.

The symmetrized Stein's loss and other losses   In this paper we proposed the study of the SSL degree, the number of complex solutions to the critical equations of the symmetrized Stein's loss. In all our calculations we observed that the dual ML degree is smaller than the ML degree, which is then smaller than the SSL degree. If these inequalities hold for general models, this would indicate that, despite its algebraicity, the symmetrized Stein's loss may be relatively harder to handle from an algebraic point of view. Currently we do not know how to prove this.

Data leading to the same critical point   One interesting aspect of the likeli-
hood geometry, which we did not study in detail here, concerns the geometry
of the data matrices S that lead to the same critical point Σ^*. Directly from the
critical equations in Theorem 2.4 we get the following remark.

Remark 6.1. If Σ^* is a fixed point in an affine covariance model $M = A + \mathcal{L}$,
then the set of all $S \in \mathbb{S}^n_+$ for which $I(S\|\Sigma)$ has Σ^* as one of its critical points is
a linear subspace of $\mathbb{S}^n_+$. Indeed, if $U_1, \ldots, U_k$ are the generators of $\mathcal{L}$ then Σ^*
is a critical point of $I(S\|\Sigma)$ for data S if and only if $\langle K^* - K^* S K^*, U_i \rangle = 0$ for
$i = 1, \ldots, k$. This is a linear condition in S.

   The closely related study of all S that lead to the same global minimum of
the entropy loss is more subtle. To illustrate, consider the bivariate correlation
model in Section 3. Fix Σ^* in the model with an underlying correlation ρ. Then
the linear space from Remark 6.1 is given in (b, a) by the equation $f(\rho) = 0$
(linear in a, b), that is, $\rho^3 - b\rho^2 - \rho(1 - 2a) - b = 0$. Consider any point (b, a) on
this line that does not lie in the bottom red region in Figure 1. By Theorem 3.1
every such point leads to a unique critical point of the entropy loss, and thus
the fixed ρ is the optimal point for the data (b, a). However, if (b, a) lies in the
bottom red region then ρ may not be the optimum, so careful analysis is needed
to understand the full picture. The set of all data S for which Σ^* is a global
maximum is a generalization of a *Voronoi cell*. The analogue sets for discrete
statistical models have been studied recently in [1].

   We note that Remark 6.1 extends to the Stein's loss: the set of all $W = S^{-1} \in$
$\mathcal{S}^n_+$ for which $I(\Sigma\|S)$ has Σ^* as a critical point forms a linear subspace of $\mathcal{S}^n_+$.

Other interesting correlation models   There are certainly other tractable and
interesting linear correlation models that one can consider. A one-dimensional
example is given by the tridiagonal equicorrelation model, where Σ is a corre-
lation matrix satisfying $\Sigma_{ij} = \rho$ if $|i - j| = 1$, and $\Sigma_{ij} = 0$ for all other pairs $i \neq j$.
In this case the linear subspace $\mathcal{L}$ defining this model is generated by a single
element U such that $U_{ij} = 1$ if $|i - j| = 1$ and $U_{ij} = 0$ otherwise.

   If $n = 3$, the results are as in the equicorrelation case: the three algebraic
degrees are 2,3,4. It is a natural question whether there is a deeper reason for
this. If $n = 4$, the dual ML degree is 4, the ML degree is 7, and the SSL-degree
is 8. The dual ML degree sequence is actually $2\lfloor \frac{n}{2} \rfloor$. To prove this fact we use
the following auxiliary result.

Lemma 6.2. *The determinant δ_n of the symmetric tridiagonal matrix of dimen-
sion $n \geq 2$, with 1's on the diagonal and ρ in the upper/lower diagonal, satisfies*

$$\delta_n = \sum_{i=0}^{\lfloor n/2 \rfloor} (-1)^i \binom{n-i}{i} \rho^{2i},$$

where in addition we set $\delta_0 = \delta_1 = 1$. Moreover,

$$\det(\Sigma)K_{ij} = (-1)^{i+j}\rho^{j-i}\delta_{i-1}\delta_{n-j} \qquad for\ i < j.$$

Proof. This can be inferred from the formulas (5) and (9) in [10] noting that their matrix M_n links to ours via $\Sigma = \rho M_n$ with their D equal to $\frac{1}{\rho}$. $\square$

Theorem 6.3. *The dual ML degree in the tridiagonal equicorrelation model is $2\lfloor\frac{n}{2}\rfloor$ for any $n \geq 1$. The critical points are the solutions to the polynomial:*

$$s\delta_n + \rho \sum_{i=1}^{n-1} \delta_{i-1}\delta_{n-i-1} = 0, \tag{20}$$

where $s = \sum_{i=1}^{n-1} W_{i,i+1}$ and δ_n is as in Lemma 6.2.

Proof. The critical equation is that s is equal to $\sum_{i=1}^{n-1} K_{i,i+1}$. Using Lemma 6.2, this is equivalent to (20). It remains to show that the polynomial in (20) has degree $2\lfloor\frac{n}{2}\rfloor$ as long as $s \neq 0$. The degree of $\rho\delta_{i-1}\delta_{n-i-1}$ is $2\left(\lfloor\frac{i-1}{2}\rfloor + \lfloor\frac{n-i-1}{2}\rfloor\right) + 1$, which is *strictly* less than the degree of δ_n. To show this, we consider four cases depending on the parity of n and i. If n is odd then the degree of $\rho\delta_{i-1}\delta_{n-i-1}$ is $n-2$, which is less than $\deg(\delta_n) = n-1$. If n is even then $\deg(\rho\delta_{i-1}\delta_{n-i-1}) \in \{n-1, n-3\}$, which again is less than $\deg(\delta_n) = n$. This means that the polynomial (20) in ρ has degree $2\lfloor n/2\rfloor$, with leading coefficient

$$s(-1)^{\lfloor n/2\rfloor}\binom{n-\lfloor n/2\rfloor}{\lfloor n/2\rfloor}. \qquad \square$$

REFERENCES

[1] Yulia Alexandr - Alexander Heaton, *Logarithmic Voronoi cells*, arXiv:2006.09431 (2020).

[2] Ole E Barndorff-Nielsen - David Roxbee Cox, *Inference and asymptotics*, Chapman\& Hall, 1994.

[3] Tobias Boege - Jane Ivy Coons - Christopher Eur - Aida Maraj - Frank Röttger, *Reciprocal Maximum Likelihood Degrees of Brownian Motion Tree Models*, arXiv:2009.11849 (2020).

[4] Christian T Brownlees - Jordi Llorens-Terrazas, *Projected Dynamic Conditional Correlations*, Available at SSRN 3576985 (2020).

[5] Jane Ivy Coons - Orlando Marigliano - Michael Ruddy, *Maximum likelihood degree of the two-dimensional linear Gaussian covariance model*, Algebraic Statistics 11 no. 2 (2020), 107–123.

[6] Inderjit S Dhillon - Joel A Tropp, *Matrix nearness problems with Bregman divergences*, SIAM Journal on Matrix Analysis and Applications 29 no. 4 (2008), 1120–1146.

[7] Bradley Efron, *The geometry of exponential families*, The Annals of Statistics 6 no. 2 (1978), 362–376.

[8] Claudia Fevola - Yelena Mandelshtam - Bernd Sturmfels, *Pencils of Quadrics: Old and New*, arXiv:2009.04334.

[9] Nicholas J Higham, *Computing the nearest correlation matrix—a problem from finance*, IMA journal of Numerical Analysis 22 no. 3 (2002), 329–343.

[10] GY Hu - Robert F O'Connell, *Analytical inversion of symmetric tridiagonal matrices*, Journal of Physics A: Mathematical and General 29 no. 7 (1996), 1511.

[11] MG Kendall - Alan Stuart, *The advanced theory of statistics: Inference and relationship, vol. 2*, London: Charles Griffin (1961).

[12] Brian Kulis - Mátyás A Sustik - Inderjit S Dhillon, *Low-Rank Kernel Learning with Bregman Matrix Divergences.*, Journal of Machine Learning Research 10 no. 2 (2009).

[13] Olivier Ledoit - Michael Wolf, *Optimal estimation of a large-dimensional covariance matrix under Stein's loss*, Bernoulli 24 no. 4B (2018), 3791–3832.

[14] Kung-Yee Liang - Scott L Zeger, *Longitudinal data analysis using generalized linear models*, Biometrika 73 no. 1 (1986), 13–22.

[15] Maher Moakher - Philipp G Batchelor, *Symmetric positive-definite matrices: From geometry to applications and visualization*, Visualization and Processing of Tensor Fields, Springer, 2006, pp. 285–298.

[16] Peter J Rousseeuw - Geert Molenberghs, *The shape of correlation matrices*, The American Statistician 48 no. 4 (1994), 276–279.

[17] Christopher G Small - Jinfang Wang - Zejiang Yang, *Eliminating multiple root problems in estimation*, Statistical Science 15 no. 4 (2000), 313–341.

[18] Jake A Soloff - Adityanand Guntuboyina - Michael I Jordan, *Covariance estimation with nonnegative partial correlations*, arXiv:2007.15252 (2020).

[19] Bernd Sturmfels - Sascha Timme - Piotr Zwiernik, *Estimating linear covariance models with numerical nonlinear algebra*, Algebraic Statistics 11 no. 1 (2020), 31–52.

[20] Piotr Zwiernik - Caroline Uhler - Donald Richards, *Maximum likelihood estimation for linear Gaussian covariance models*, J. R. Stat. Soc. Ser. B. Stat. Methodol. 79 no. 4 (2017), 1269–1292.

C. AMÉNDOLA
Technical University of Munich, Germany
e-mail: `carlos.amendola@tum.de`

P. ZWIERNIK
Pompeu Fabra University and
Barcelona Graduate School of Economics, Spain
e-mail: `piotr.zwiernik@upf.edu`