

Algebraic Statistics Final Project:

Sparsity in Covariance Selection for Gaussian Graphical Models

Cynthia Vinzant, 12/11/08

Here we deal with fitting a gaussian multivariate distribution $\mathcal{N}(\mu, \Sigma)$ to data $X^{(1)}, X^{(2)}, \dots, X^{(k)} \in \mathbf{R}^n$. This is generally done by maximizing the probability of seeing the data over the (μ, Σ) in some subset of $\mathbf{R}^n \oplus PD^n$, giving the solution $\mu = \bar{X}$, and

$$(\Sigma^*)^{-1} = X^* = \operatorname{argmax} \log(\det[X]) - \langle S, X \rangle \quad (1)$$

where $\bar{X} = \frac{1}{k} \sum_{i=1}^k X^{(i)}$ and

$$S = \frac{1}{k} \sum_{i=1}^k (X^{(i)} - \bar{X})(X^{(i)} - \bar{X})^T.$$

Unfortunately, because of noisy data, this may not give a sparse $(\Sigma^*)^{-1}$ even if sparsity exists in the underlying unknown distribution. Sparsity can make graphical models easier both for computation and interpretation. Dempster [2] proposes setting a sparsity pattern of Σ^{-1} before data analysis. This problem, called covariance selection, is the same optimization problem as (1) with the constraints $X_{ij} = 0$ for certain $i \neq j$.

Rather than imposing a sparse model prior to data analysis, we explore a method of model selection proposed by Banerjee et. al. [1] which penalizes non-sparsity. The true formulation of the problem is

$$X^* = \operatorname{argmax} \log(\det[X]) - \langle S, X \rangle - \rho \cdot \operatorname{Card}(X) \quad (2)$$

with some penalization parameter ρ and $\operatorname{Card}(X) = \sum_{ij} 1_{\{X_{ij} \neq 0\}}$. To do this convexly, [1] use the heuristic of penalizing the L_1 -norm of the inverse covariance matrix instead, giving

$$X^* = \operatorname{argmax} \log(\det[X]) - \langle S, X \rangle - \rho \|X\|_1. \quad (3)$$

Goal:

I would like to analyze the properties and geometry of this optimization problem. From an optimization standpoint, our goal is to describe a formula for

the optimal point as a piecewise algebraic function in ρ . While heuristics exist for a choice of ρ , it is more desirable to be able to see the progression of solutions as ρ increases. If using this only for model selection, we would like to describe only the progression of sparsity patterns of the solution.

1 A Modified MLE problem:

Here we will examine a modified version of the problem proposed by [1]. The first change I make is to not penalize diagonal elements of X . Since they cannot be zero, they are not penalized in (2) and it seems better to not penalize them in the relaxation. This also makes the relaxation more consistent. For example if S is diagonal, penalizing the diagonal of X will result in a solution with the same sparsity pattern as a solution to (1) but with lower likelihood.

The second alteration made here is to normalize S before data analysis so that it has 1's along it's diagonal. This has the advantage that the solution will be invariant under scaling certain dimensions of $X^{(i)}$. For example, changing the units in which we measure a certain r.v. will not affect the sparsity pattern of X^* . This also (slightly) reduces the problem size and makes analysis easier. So let $\bar{S} = D^{1/2}SD^{1/2}$ where D is a diagonal matrix with $D_{ii} = S_{ii}$. The model selection problem resulting from these modifications is:

$$Y_\rho^* = \operatorname{argmax} \log(\det[Y]) - \langle \bar{S}, Y \rangle - \rho \|Y\|_1 + \rho \cdot \operatorname{tr}(Y). \quad (4)$$

Taking the dual of this problem gives

$$\max \log(\det[\bar{S} + U]) \quad \text{s.t.} \quad \|U\|_\infty \leq \rho, \quad U_{ii} = 0. \quad (5)$$

1.1 Some Properties:

1.1.1 Solving with S on the PD boundary

As noted in [1] for (3), for \bar{S} not of full rank, the problem (4) has a solution for $\rho > 0$ even though (1) does not. Since $\bar{S} \succeq 0$, we can always find U satisfying the constraints of (4) so that $\bar{S} + U \succ 0$.

1.1.2 Finding independence

By a classical result of Fischer, any $n \times n$ positive definite matrix A satisfies

$$\det(A_{[1\dots j]}) \cdot \det(A_{[j+1\dots n]}) \geq \det(A).$$

Thus if for some subset $B \subset [n]$, $\rho \geq s_{ij}$ for all $i \in B$ and $j \in [n] \setminus B$ (where $s_{ij} = \bar{S}_{ij}$), then $(Y_\rho^*)_{ij} = 0$ for all such i, j . In the distribution given by Y^* , $\{X_i : i \in B\}$ is independent of $\{X_j : j \in [n] \setminus B\}$.

1.1.3 The gradient

An equivalent (though not convex) optimization problem to (5) is

$$\max \det(\bar{S} + U) \quad \text{s.t.} \quad \|U\|_\infty \leq \rho, \quad U_{ii} = 0, \quad \bar{S} + U \succ 0.$$

To solve this, we'll need to analyze

$$f(\mathbf{x}) = \det \begin{pmatrix} 1 & x_{12} & \dots & x_{1n} \\ x_{12} & 1 & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1n} & x_{2n} & \dots & 1 \end{pmatrix}.$$

Then

$$\frac{\partial f}{\partial x_{ij}} = 2(-1)^{i+j} m_{[n \setminus \{i\}][n \setminus \{j\}]},$$

where $m_{[a][b]}$ is the minor with rows a and columns b of the above matrix.

1.1.4 Conditional Independence

Because the partial derivatives of f are these almost-principal minors, we have that for \mathbf{x} with

$$\frac{\partial f}{\partial x_{ij}} \Big|_{(\mathbf{x})} = 0,$$

the gaussian distribution given by \mathbf{x} will satisfy $i \perp j$ “rest”.

1.1.5 Piecewise algebraic solution

Because we're optimizing a polynomial over a box, the optimal point \mathbf{x}^* will be on some face defined by some $x_{ij} = s_{ij} \pm \rho$ and some $\partial f / \partial x_{ij} = 0$. Thus the components of \mathbf{x}^* will be algebraic in ρ with algebraic breakpoints. If we are able to identify these break points, we could quickly calculate \mathbf{x}^* for any ρ .

2 Analysis of the 3×3 problem

Here we will completely characterize the behavior of (5) for $n = 3$. Let

$$M(x, y, z) = \begin{bmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{bmatrix},$$

and

$$f(x, y, z) = \det[M(x, y, z)] = 1 - x^2 - y^2 - z^2 + 2xyz$$

Then

$$\frac{\partial f}{\partial x} = 2(yz - x), \quad \frac{\partial f}{\partial y} = 2(xz - y), \quad \text{and} \quad \frac{\partial f}{\partial z} = 2(xy - z).$$

We are maximizing f over the intersection of $M(x, y, z) \succ 0$ with the box

$$B_\rho = [s_{12} - \rho, s_{12} + \rho] \times [s_{13} - \rho, s_{13} + \rho] \times [s_{23} - \rho, s_{23} + \rho].$$

Note that f is invariant under changing the signs of any row and column. In the 3×3 case, this reduces to being invariant under switching the signs of any two of x, y, z . This reduces our analysis to two cases: $|\{s_{12}, s_{13}, s_{23}\} \cap \mathbf{R}_{\geq 0}|$ odd and $|\{s_{12}, s_{13}, s_{23}\} \cap \mathbf{R}_{\geq 0}|$ even. Also, because we also have symmetry among x, y, z , we will do analysis only for $|s_{12}| \leq |s_{13}| \leq |s_{23}|$ and analysis for other cases will follow by symmetry.

2.1 Case 1: $|\{s_{12}, s_{13}, s_{23}\} \cap \mathbf{R}_{\geq 0}|$ odd

Suppose that $s_{12}, s_{13}, s_{23} \geq 0$. The other three cases are symmetric.

Claim 1 For $x, y, z \geq 0$ and $M(x, y, z) \succ 0$, at most one of $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$ can be positive.

Without loss of generality, suppose $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \geq 0$. Then

$$(yz)(xz) \geq xy.$$

Since our $M(x, y, z) \succ 0$, we must have $z < 1$, giving $xy = 0$. Because both our partials are non-negative, we have the $x = 0$ iff $y = 0$.

From this we see that the $\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \geq 0$ implies $0 = x = y = \frac{\partial f}{\partial x} = \frac{\partial f}{\partial y}$.

Claim 2 For ρ less than two of $\{s_{12}, s_{13}, s_{23}\}$, the optimal point will be on one of the three edges connected to $c(\rho) = (s_{12} - \rho, s_{13} - \rho, s_{23} - \rho)$ and we can identify that edge by evaluating $(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z})$ at this point.

Since the gradient $(\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z})$ must be tangent to B_ρ at the optimal point and at most one of these components can be non-negative (for $x, y, z > 0$), we have that optimal point must lie on the corner closest to the origin or one of the three edges connected to it.

If all the partial derivatives are negative at $c(\rho)$, then $c(\rho)$ is a critical point, and thus our optimal point. If one of the partials of f is positive at the corner $c(\rho)$ (say $\frac{\partial f}{\partial x}|_{c(\rho)} > 0$), then by fixing y and z and increasing x , we decrease $\frac{\partial f}{\partial x}$. Either we reach the point $((s_{13} - \rho) \cdot (s_{23} - \rho), s_{13} - \rho, s_{23} - \rho)$ at which $\frac{\partial f}{\partial x}$ is zero, giving us a critical point on this edge, or we will have $\frac{\partial f}{\partial x} \geq 0$ at the point $(s_{12} + \rho, s_{13} - \rho, s_{23} - \rho)$.

Thus we would like to look at $(\frac{\partial f}{\partial x}|_{c(\rho)}, \frac{\partial f}{\partial y}|_{c(\rho)}, \frac{\partial f}{\partial z}|_{c(\rho)})$. Note that each of these are monic quadratics in ρ :

$$\frac{\partial f}{\partial x}|_{c(\rho)} = \rho^2 + (1 - s_{13} - s_{23}) \cdot \rho + s_{13} \cdot s_{23} - s_{12}$$

$$\frac{\partial f}{\partial y}|_{c(\rho)} = \rho^2 + (1 - s_{12} - s_{23}) \cdot \rho + s_{12} \cdot s_{23} - s_{13}$$

$$\frac{\partial f}{\partial z}|_{c(\rho)} = \rho^2 + (1 - s_{12} - s_{13}) \cdot \rho + s_{12} \cdot s_{13} - s_{23}.$$

We're now ready to analyze the behavior of the optimal point and corresponding sparsity progressions as $\rho \rightarrow 1$.

2.1.1 Case 1a) : $\frac{\partial f}{\partial x}|_s, \frac{\partial f}{\partial y}|_s, \frac{\partial f}{\partial z}|_s < 0$ for $s = (s_{12}, s_{13}, s_{23})$

For small ρ , the partial derivatives at $c(\rho)$ will have the same sign patterns as at s , meaning that $c(\rho)$ will be the optimal point. As ρ approaches s_{12} , $\frac{\partial f}{\partial x}|_{c(\rho)}$ will become positive and remain positive for larger ρ . The optimal point will then move to the edge $((s_{13} - \rho) \cdot (s_{23} - \rho), s_{13} - \rho, s_{23} - \rho)$ and remain there until ρ reaches s_{13} . The optimal point will then move to $(0, 0, s_{23} - \rho)$ on the face of B_ρ .

To summarize,

$$(x^*, y^*, z^*) = \begin{cases} (s_{12} - \rho, s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [0, a_2] \\ ((s_{13} - \rho) \cdot (s_{23} - \rho), s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [a_2, s_{13}] \\ (0, 0, s_{23} - \rho) & \text{for } \rho \in [s_{13}, s_{23}] \\ (0, 0, 0) & \text{for } \rho \in [s_{23}, 1] \end{cases}$$

where

$$a_2 = \frac{1}{2} \cdot [y + z - 1 + \sqrt{(y + z - 1)^2 - 4(yz - x)}]$$

is the larger root of $\frac{\partial f}{\partial x}|_{c(\rho)}$ as a polynomial in ρ .

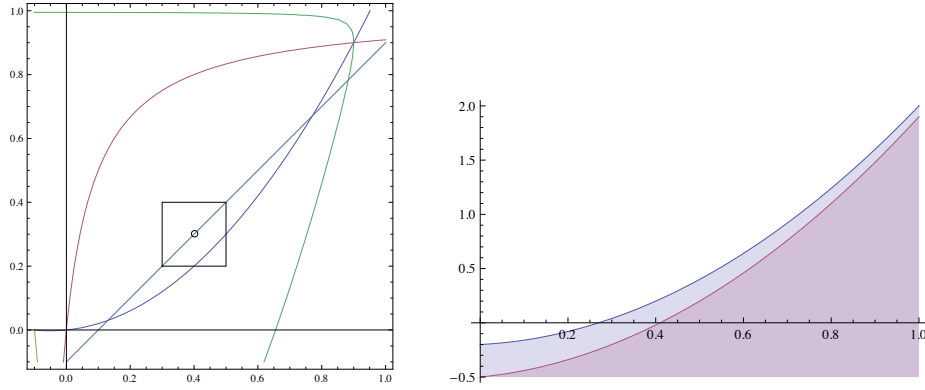


Fig. Above we have (right) a the $z = y + .1$ plane along with the curves $\frac{\partial f}{\partial x} = 0$, $\frac{\partial f}{\partial y} = 0$, and the trajectory of $c(\rho)$ with starting point $(s_{12}, s_{13}, s_{23}) = (.3, .4, .5)$ and a slice of B_ρ . To the left is a plot of $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ evaluated at this point as ρ increases.

2.1.2 Case 1b : $\frac{\partial f}{\partial x}|_s > 0$ and $\frac{\partial f}{\partial y}|_s, \frac{\partial f}{\partial z}|_s < 0$

Note that $\partial f / \partial x|_s > 0$ implies $s_{13} \cdot s_{23} > s_{12}$. Since $S \succ 0$, we have $s_{13}, s_{23} < 1$, giving that $s_{12} < s_{13}, s_{23}$.

Let $a_1 \leq a_2$ be the roots of $\frac{\partial f}{\partial x}|_{c(\rho)}$. If $s_{13} \leq a_1$, then $\frac{\partial f}{\partial x}|_{c(\rho)} \geq 0$ for all $\rho \leq s_{13}$, meaning that the optimal point lives on the edge $(x, s_{13} - \rho, s_{23} - \rho)$.

Then for ρ small enough, the signs of the gradient at $(s_{12} + \rho, s_{13} - \rho, s_{23} - \rho)$ will match those at s , making this our optimal point. As ρ increases, $\frac{\partial f}{\partial x}$ at this

corner will decrease and become negative, and the optimal point will live on the relative interior of this edge. This continues until ρ reaches s_{13} and $1 \perp \{2, 3\}$ is available. Then the optimal point moves to the $x = y = 0$ face and then to the origin.

To summarize,

$$(x^*, y^*, z^*) = \begin{cases} (s_{12} + \rho, s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [0, b_2] \\ ((s_{13} - \rho) \cdot (s_{23} - \rho), s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [b_2, s_{13}] \\ (0, 0, s_{23} - \rho) & \text{for } \rho \in [s_{13}, s_{23}] \\ (0, 0, 0) & \text{for } \rho \in [s_{23}, 1] \end{cases}$$

where

$$b_2 = \frac{1}{2} \cdot [y + z + 1 - \sqrt{(y + z + 1)^2 - 4(yz - x)}]$$

is the smaller root of $\partial f / \partial x$ at $(s_{12} + \rho, s_{13} - \rho, s_{23} - \rho)$.

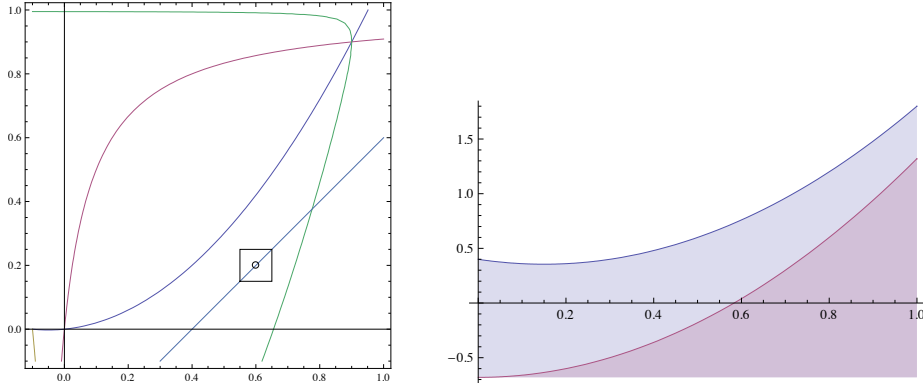


Fig. Above we have (right) a the $z = y + .1$ plane along with the curves $\frac{\partial f}{\partial x} = 0$, $\frac{\partial f}{\partial y} = 0$, and the trajectory of $c(\rho)$ with starting point $(s_{12}, s_{13}, s_{23}) = (.2, .6, .7)$ and a slice of B_ρ . To the left is a plot of $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ evaluated at this point as ρ increases.

On the other hand if $a_1 < s_{13}$, the behavior of the optimal point becomes more complicated. As before it starts at $(s_{12} + \rho, s_{13} - \rho, s_{23} - \rho)$ and moves to the edge $(x, s_{13} - \rho, s_{23} - \rho)$. It then reaches $c(\rho)$ as $\frac{\partial f}{\partial x}|_{c(\rho)}$ dips below zero. As ρ approaches s_{12} , $\frac{\partial f}{\partial x}|_{c(\rho)}$ again becomes positive, and the optimal point comes back to the edge. The behavior then continues as above, giving

$$(x^*, y^*, z^*) = \begin{cases} (s_{12} + \rho, s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [0, b_2] \\ ((s_{13} - \rho) \cdot (s_{23} - \rho), s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [b_2, a_1] \\ (s_{12} - \rho, s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [a_1, a_2] \\ ((s_{13} - \rho) \cdot (s_{23} - \rho), s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [a_2, s_{13}] \\ (0, 0, s_{23} - \rho) & \text{for } \rho \in [s_{13}, s_{23}] \\ (0, 0, 0) & \text{for } \rho \in [s_{23}, 1] \end{cases}$$

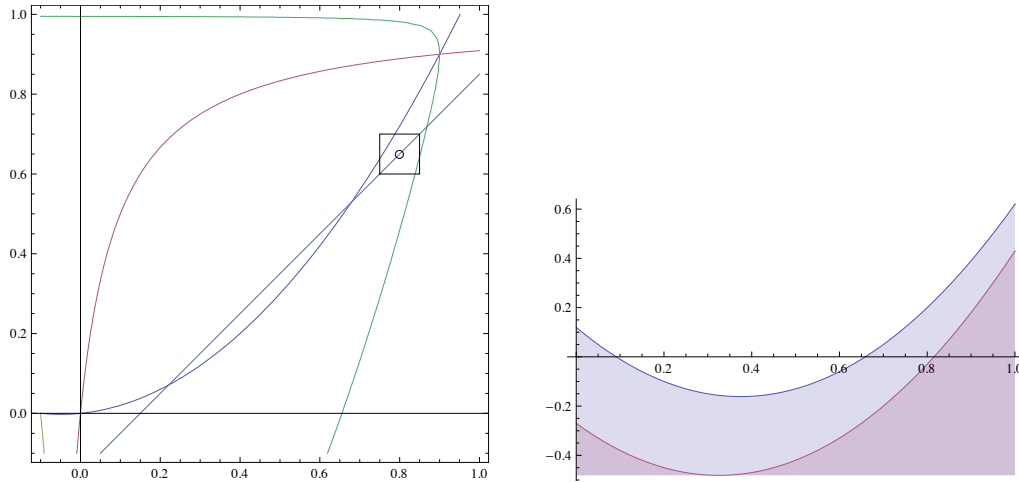


Fig. Above we have (right) a the $z = y + .1$ plane along with the curves $\frac{\partial f}{\partial x} = 0, \frac{\partial f}{\partial y} = 0$, and the trajectory of $c(\rho)$ with starting point $(s_{12}, s_{13}, s_{23}) = (.65, .8, .9)$ and a slice of B_ρ . To the left is a plot of $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ evaluated at this point as ρ increases.

2.2 Case 2 : $|\{s_{12}, s_{13}, s_{23}\} \cap \mathbf{R}_{\geq 0}|$ even

Suppose $s_{12} < 0, s_{13}, s_{23} > 0$. The other cases can be described symmetrically.

Note that if $x \leq 0, y, z \geq 0$, then,

$$\begin{aligned} \frac{\partial f}{\partial x} &= 2(yz - x) \leq 0 \\ \frac{\partial f}{\partial y} &= 2(xz - y) \leq 0 \\ \frac{\partial f}{\partial z} &= 2(xy - z) \geq 0, \end{aligned}$$

Thus while $\rho < |s_{12}|$, B_ρ will be entirely contained in this octant and the optimal point will be the corner of B_ρ closest to the origin, $(s_{12} + \rho, s_{13} - \rho, s_{23} - \rho)$. Since the signs of the partial derivatives don't change as ρ passes $|s_{12}|$, neither does the optimal point.

As ρ approaches s_{13} , we see that for ρ close enough to s_{13} , $\frac{\partial f}{\partial x}$ becomes negative. Then the optimal point resides on the edge $(x^*, s_{13} - \rho, s_{23} - \rho)$, where $x^* = (s_{13} - \rho) \cdot (s_{23} - \rho)$, making $\frac{\partial f}{\partial x} = 0$.

Once $\rho > |s_{13}|$, our optimal point moves to the face $(0, 0, s_{23} - \rho)$. It resides here until $\rho \geq |s_{23}|$, when $(0, 0, 0)$ enters B_ρ and becomes the optimal point.

To summarize,

$$(x^*, y^*, z^*) = \begin{cases} (s_{12} + \rho, s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [0, b_1] \\ ((s_{13} - \rho) \cdot (s_{23} - \rho), s_{13} - \rho, s_{23} - \rho) & \text{for } \rho \in [b_1, s_{13}] \\ (0, 0, s_{23} - \rho) & \text{for } \rho \in [s_{13}, s_{23}] \\ (0, 0, 0) & \text{for } \rho \in [s_{23}, 1) \end{cases}$$

b_1 smaller root of $\partial f / \partial x$ evaluated at $(s_{12} + \rho, s_{13} - \rho, s_{23} - \rho)$.

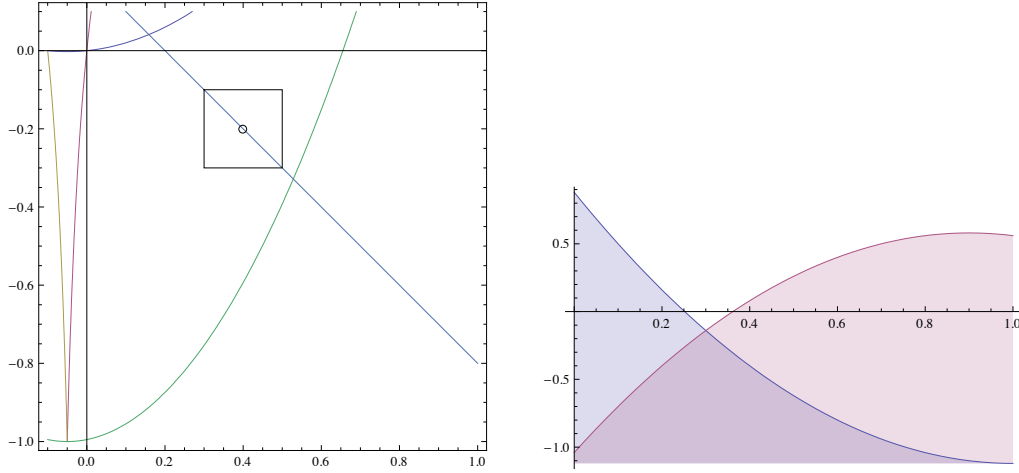

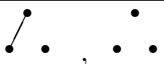
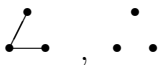
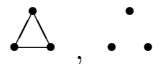



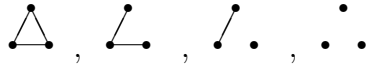
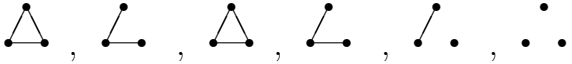


Fig. Above we have (right) a the $z = y + .1$ plane along with the curves $\frac{\partial f}{\partial x} = 0$, $\frac{\partial f}{\partial y} = 0$, and the trajectory of $(s_{12} + \rho, s_{13} - \rho, s_{23} - \rho)$ with starting point $(s_{12}, s_{13}, s_{23}) = (-.2, .4, .5)$ and a slice of B_ρ . To the left is a plot of $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ evaluated at this point as ρ increases.

2.3 Sparsity Patterns

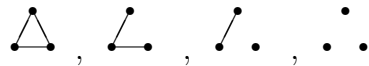
The following are the possible progressions of sparsity patterns for $|s_{12}| \leq |s_{13}| \leq |s_{23}|$:

Sparsity as $\rho \rightarrow 1$	(s_{12}, s_{13}, s_{23})
	$(s_{12}, s_{13}, s_{23}) = (0, 0, 0)$
	$s_{12} = s_{13} = 0$
	$\frac{\partial f}{\partial x} _s = 0$ and $s_{13} = s_{23} \neq 0$
	$s_{12} = s_{13} = s_{23} \neq 0$
	$\frac{\partial f}{\partial x} _s = 0$ and not (*)
	$s_{12} = s_{13} \neq 0$
	$\frac{\partial f}{\partial x} _s = 0$ and (*)
	not (*)
	(*)

where (*) denotes the following constraints:

- $\{s_{12}, s_{13}, s_{23}\} \cap \mathbf{R}_{\geq 0}$ is odd
- $\frac{\partial f}{\partial x}|_s \geq 0$
- $|s_{13}| > \frac{1-q}{2}$ for $q = |s_{23}| - |s_{13}|$,
- $|s_{12}| > |s_{13}| - \frac{1-q^2}{4}$

Note that for $|s_{12}| \leq |s_{13}| \leq |s_{23}|$, the only models that appear are



If we use this technique only for model selection, the relative order $|s_{12}|, |s_{13}|, |s_{23}|$ is all we need to characterize the solution.

3 Conclusion

(Tractability of analysis for larger cases)

We’ve seen that for small cases, it is possible to completely analyze the behavior of the optimal points of (4) and (5). While not done here, it seems possible that with further analysis, these techniques could be implemented on a much larger scale. The number of breakpoints seems to “usually” be n though possibly more, many of which would have to be approximated using root finding techniques. It’s unclear how tractable this analysis would be for many more variables but seems worthy of study.

(Quality of the L1 penalization)

This technique maybe be useful for sparse model selection (finding the right set of conditional independences), but does not seem to work well for finding the actual covariance matrix to be used. While solution to (4) will be sparse for high enough ρ , it will not be the “most likely” solution with this sparsity pattern. A better technique would be to use (4) to find the desired sparse graphical model and optimize likelihood over this model, as proposed by Dempster [2].

(Questions for further research)

- How many break points can there be in high dimensions? What proportion will give have $O(n)$ break points?
- When do the (3) and (4) give different sparsity answers?
- For $n = 3$ we saw that the set of possible sparsity patterns coming from a point (s_{12}, s_{13}, s_{23}) can be determined by the relative sizes of $|s_{12}|$, $|s_{13}|$, and $|s_{23}|$. Is this true for higher n ?

References

- [1] O. Banerjee, A. dAspremont, and L. E. Ghaoui, Sparse Covariance Selection via Robust Maximum Likelihood Estimation, arXiv:cs.CE/0506023 June, 2005
- [2] Dempster, A. (1972), Covariance selection, Biometrics 28, 157175.