Geodesic convexity & covariance estimation

Ami Wiesel

School of Engineering and Computer Science
Hebrew University of Jerusalem, Israel

June 28, 2013

Acknowledgments

- Teng Zhang (Princeton).
- Maria Greco (Universita di Pisa).
- Ilya Soloveychik (Hebrew University).
- Alba Sloin (Hebrew University).

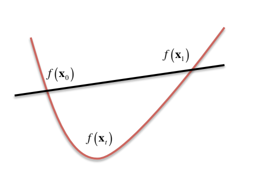
Outline

Convexity

Convex function

$$f\left(\mathbf{x}_t\right) \leq t f\left(\mathbf{x}_1\right) + (1-t) f\left(\mathbf{x}_0\right)$$

$$\mathbf{x}_t = t\mathbf{x}_1 + (1-t)\mathbf{x}_0$$



- Local solutions are easy to find and globally optimal!
- Easy to generalize:
  - Building bricks: linear, quadratic, norms...
  - Rules: convex+convex=convex,...

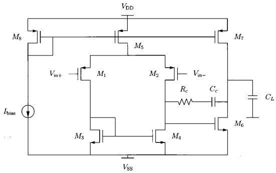
Convex optimization with positive variables

Power control [Chiang:07]

$$
\begin{array}{ll}
\text{minimize} & \prod_{i=1}^{N} \frac{1}{1+\mathsf{SIR}_i} \\
\text{subject to} & (2^{TR_{i,min}} - 1)\frac{1}{\mathsf{SIR}_i} \leq 1, \;\; \forall i, \\
& (\mathsf{SIR}_{th})^{N-1}(1 - P_{o,i,max}) \prod_{j\neq i}^{N} \frac{G_{ij}P_j}{G_{ii}P_i} \leq 1, \;\; \forall i, \\
& P_i(P_{i,max})^{-1} \leq 1, \;\; \forall i.
\end{array}
$$

Variables: powers.

Circuit design [Hershenson:01]



Variables: transistors widths, lengths, currents, capacitors,...

The Geometric Programming (GP) trick

- The above problems are non-convex.
- Can be convexified by a change of variables $q_i = e^{z_i}$.

Convexity with positive variables

- Exp: e^{z_i} are convex in z_i.
- Log-sum-exp: $\log \sum_i e^{z_i}$ is convex in z_i.
- If $f\left(e^{z}\right)$ is convex in z then $f\left(e^{z_1+z_2}\right)$ is convex in z_1, z_2.
- e^{z} transforms sums into products!

The Geometric Programming (GP) trick

- Minimize products of positive numbers $q_i \geq 0$ using e^{z_i}.

Convexity with positive definite matrices $\mathbf{Q}_i \succeq \mathbf{0}$

Today: GP with positive definite matrices

- Can we minimize powers $\mathbf{a}^T\mathbf{Q}^{\pm 1}\mathbf{a}$?
- Can we minimize log determinants $\log|\mathbf{Q}|$?
- Can we minimize products $\mathbf{Q}_1 \otimes \mathbf{Q}_2$?

- The answers are YES!
- But the solution is not a simple change of variables.
- Instead, we turn to geodesic convexity.

Revisiting the GP trick

Convexity

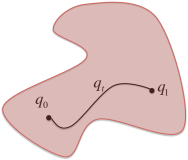
$$f(\overbrace{t\mathbf{z}_1 + (1-t)\mathbf{z}_0}^{\text{line}}) \leq tf(\mathbf{z}_1) + (1-t)f(\mathbf{z}_0)$$

Geodesic convexity $\tilde{f}(\mathbf{q}) = f(\log \mathbf{q})$

$$\tilde{f}(\overbrace{\mathbf{q}_1^t \mathbf{q}_0^{1-t}}^{\text{geodesic}}) \leq t\tilde{f}(\mathbf{q}_1) + (1-t)\tilde{f}(\mathbf{q}_0)$$

Geodesic convexity [Rapcsak 91], [Liberti 04]

- For any $\mathbf{q}_1, \mathbf{q}_0 \in D$ we define a geodesic $\mathbf{q}_t \in D$ parameterized by $t \in [0,1]$.



- A function $f(\mathbf{q})$ is g-convex in $\mathbf{q} \in D$ if

$$f(\mathbf{q}_t) \leq t f(\mathbf{q}_1) + (1-t) f(\mathbf{q}_0) \qquad \forall \quad t \in [0,1].$$

Properties

- Any local minimizer of $f(\mathbf{q})$ over $\mathbf{D}$ is a global minimizer.
- g-convex + g-convex = g-convex.

From scalars to matrices

- We do <u>not</u> know the matrix version of e^x.
- We do know how to generalize the geodesics $q_t = q_1^t q_0^{1-t}$.

Geodesic between $\mathbf{Q}_0 \succ \mathbf{0}$ and $\mathbf{Q}_1 \succ \mathbf{0}$

$$\mathbf{Q}_t = \mathbf{Q}_0^{\frac{1}{2}} \left(\mathbf{Q}_0^{-\frac{1}{2}} \mathbf{Q}_1 \mathbf{Q}_0^{-\frac{1}{2}} \right)^t \mathbf{Q}_0^{\frac{1}{2}}, \qquad t \in [0, 1].$$

Powers (matrix case)

Theorem

The function

$$f(\mathbf{Q}) = \mathbf{a}^T\mathbf{Q}^{\pm 1}\mathbf{a}$$

is g-convex in $\mathbf{Q} \succ \mathbf{0}$.

- Proof: eigenvalue decomposition reduces to scalar case.

Log-sum-exp (matrix case)

Theorem

The function

$$f(\mathbf{Q}) = \log \left| \sum_{i=1}^{n} \mathbf{H}_i \mathbf{Q} \mathbf{H}_i^T \right|$$

is g-convex in $\mathbf{Q} \succ \mathbf{0}$.

- Similarly, $\log |\mathbf{Q}|$ is g-linear.
- Proof: eigenvalue decomposition reduces to scalar case.

Products (matrix case)

Theorem

If $f(\mathbf{W})$ is g-convex in $\mathbf{W} \succ \mathbf{0}$, then

$$g(\mathbf{Q}_1, \cdots, \mathbf{Q}_n) = f(\mathbf{Q}_1 \otimes \mathbf{Q}_2 \otimes \cdots \otimes \mathbf{Q}_n)$$

is g-convex in $\mathbf{Q}_i \succ \mathbf{0}$.

- The operation $\otimes$ is a Kronecker product.

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \cdots & a_{1p}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \cdots & a_{21}\mathbf{B} \\ \vdots & \vdots & & \vdots \\ a_{p1}\mathbf{B} & a_{p2}\mathbf{B} & & a_{pp}\mathbf{B} \end{bmatrix}$$

Invariance to orthogonal operators

A set $\mathcal{S}$ is g-convex if

$$\mathbf{Q}_0, \mathbf{Q}_1 \in \mathcal{S} \qquad \Rightarrow \qquad \mathbf{Q}_t \in \mathcal{S}.$$

Local minimas over g-convex sets are global.

Theorem

For orthonormal $\mathbf{U}$, the set $\{\mathbf{Q} \ : \ \mathbf{Q} = \mathbf{U}\mathbf{Q}\mathbf{U}^T\}$ is g-convex.

- Proof: Matrix commutativity properties $\mathbf{QU} = \mathbf{UQ}$.
- Trivial in scalar case.

Summary

- $\mathbf{a}^T \mathbf{Q}^{\pm 1} \mathbf{a}$ is g-convex.
- $\log \left| \sum_{i=1}^{n} \mathbf{H}_i \mathbf{Q} \mathbf{H}_i^T \right|$ is g-convex.
- $\mathbf{Q}_i \otimes \cdots \otimes \mathbf{Q}_j$ preserves g-convexity.
- $\{\mathbf{Q} \ : \ \mathbf{Q} = \mathbf{U}\mathbf{Q}\mathbf{U}^T\}$ is g-convex.

Outline

Covariance estimation

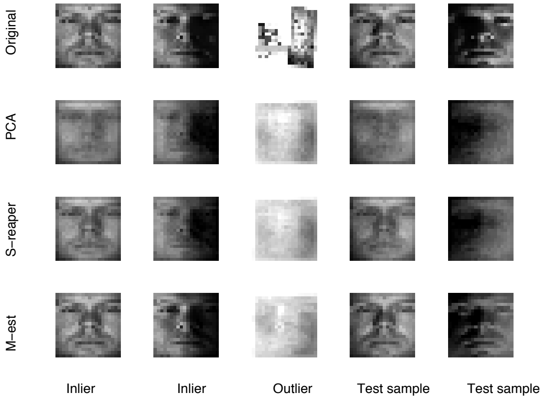
- $\mathbf{x}$: p-dimensional random vector.
- Mean $E\{\mathbf{x}\} = \mathbf{0}$, covariance $\boldsymbol{\Sigma} = E\left[\mathbf{x}\mathbf{x}^T\right]$.
- $\{\mathbf{x}_i\}_{i=1}^n$: n independent & identically distributed realizations.

Goal

- Problem: Derive $\hat{\boldsymbol{\Sigma}}\left(\{\mathbf{x}_i\}_{i=1}^n\right)$ to estimate $\boldsymbol{\Sigma}$.
- Solution: Maximum likelihood.
- Emphasis on the hard non-Gaussian and structured cases.

CIMI on “Optimization and Statistics in Image Processing”

- I work on other stuff: comm, radar, sensor networks...
- I was told this can also be used with images [Zhang:2012].



Outline

A popular robust covariance estimator

- Elliptical distributions, Spherically Invariant Random processes, Compound Gaussian, Multivariate Student, etc..

$$\begin{bmatrix} \vdots \\ \mathbf{x}_i \\ \vdots \end{bmatrix} = \sqrt{q}_i \underbrace{\begin{bmatrix} \vdots \\ \mathbf{u}_i \\ \vdots \end{bmatrix}}_{\mathcal{N}(\mathbf{0},\mathbf{Q})}$$

- Non-convex ML via fixed point iteration:

$$\mathbf{Q}_{k+1} = \frac{p}{n} \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^T}{\mathbf{x}_i^T \mathbf{Q}_k^{-1} \mathbf{x}_i}$$

A bit of background $\mathbf{Q}_{k+1} = \frac{p}{n}\sum_{i=1}^{n} \frac{\mathbf{x}_i\mathbf{x}_i^T}{\mathbf{x}_i^T\mathbf{Q}_k^{-1}\mathbf{x}_i}$

- [Tyler:87] Introduction, fixed point iteration, existence, uniqueness, convergence analysis.
- [Gini:95], [Conte:02] Analysis, array processing.
- [Pascal:08] Analysis and generalizations.
- [Gini:95], [Abramovich:07], [Bandeira:10] Regularization, normalization, diagonal loading, Bayesian priors.
- [Chen:10] Regularization analysis via Perron Frobenius.
- [Bombrun:2011], [Ollila:2012] Generalized Gaussian.

Lots of applications! Lots of difficult theory!
But specific and hard to follow and generalize.

Revisiting Tyler's estimator

The negative log likelihood is

$$L(\mathbf{Q}) = \frac{p}{n}\sum_{i=1}^{n} \log\left(\mathbf{x}_i^T \mathbf{Q}^{-1}\mathbf{x}_i\right) + \log|\mathbf{Q}|$$

- Non-convex optimization problem.
- 25 years of methods that converge to the global solution.

Theorem

[Auderset:05] The negative log likelihood is g-convex.
Actually, jointly g-convex in $\mathbf{q}$ and $\mathbf{Q}$.
Also for other elliptical distributions, e.g., MGGD.

Why is this helpful? Regularization

- Often, we need regularization / prior.
- [Abramovich:07], [Chen:10] difficult design and analysis.
- We propose to use g-convex regularization schemes

Global solution to ML (+ regularization)

$$\min \quad L(\cdot) \quad + \quad \underbrace{\lambda h(\cdot)}_{\text{needs to be g-convex}}$$

Guaranteed to be g-convex, and can be solved efficiently. We can put priors on both the covariance and the scalings.

G-convex scalings penalties

Prior knowledge on the scaling factors via g-convex functions:

- Bounded peak values $L \leq \log q_i \leq U$.
- Bounded second moments $\sum_i \log^2 q_i \leq U$.
- Sparsity (outliers) $\sum_i |\log q_i| \leq U$.
- Smooth time series $|\log q_i - \log q_{i-1}| \leq U$.

Without g-convexity [Bucciareli:96], [Wang:06], [Chitour:08].

We can also change variables and use convex penalties.

G-convex matrix penalties

- Shrinkage to identity ($\mathbf{T} = \mathbf{I}$) or arbitrary target

$$h(\mathbf{Q}) = p \log\left(\mathrm{Tr}\left\{\mathbf{Q}^{-1}\mathbf{T}\right\}\right) + \log|\mathbf{Q}|$$

- Shrinkage to diagonal

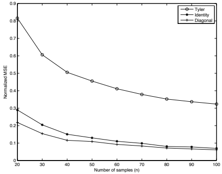
$$h(\mathbf{Q}) = \log \prod_{i=1}^{p} \left[\mathbf{Q}^{-1}\right]_{ii} + \log|\mathbf{Q}|$$

- Regularization of condition number

$$h(\mathbf{Q}) = \frac{\lambda_{\max}(\mathbf{Q})}{\lambda_{\min}(\mathbf{Q})}$$

Non-Gaussian versions of [Stoica:08], [Schafer:05], [Won:09].

Experiments



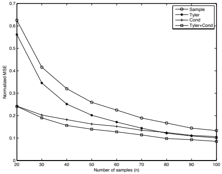
Shrink to diag

Toeplitz $p = 10$

$\mathbf{\Sigma}_{ij} = 0.4^{|i-j|}$

Factor 2 on 1st

cross validation



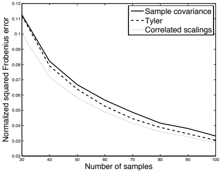
Condition number

Toeplitz $p = 10$

$\mathbf{\Sigma}_{ij} = 0.4^{|i-j|}$

$\kappa = 4.98 \in [1, 10]$

cross validation



Correlated scalings

Toeplitz $p = 10$

$\mathbf{\Sigma}_{ij} = 0.8^{|i-j|}$

MA(2) with

$\|\mathbf{Lz}\|_2 \leq 7.$

Outline

Kronecker (separable, transposable) model $\mathbf{Q}_1 \otimes \mathbf{Q}_2$

- Estimating covariances of random $p_2 \times p_1$ matrices.
- A standard approach is to impose structure

$$\mathbf{X} = \mathbf{Q}_2^{\frac{1}{2}} \mathbf{W} \mathbf{Q}_1^{\frac{1}{2}}$$

  - $\mathbf{W}_{ij}$ are i.i.d. $\mathcal{N}(0,1)$.
  - $\mathbf{Q}_2$ correlates the columns.
  - $\mathbf{Q}_1$ correlates the rows.
- In vector notations, $\mathrm{E}\left[\mathbf{x}\mathbf{x}^T\right] = \mathbf{Q}_1 \otimes \mathbf{Q}_2$
- Examples: Tx $\otimes$ Rx, products $\otimes$ costumers, etc...

A bit of background $\mathbf{Q}_1 \otimes \mathbf{Q}_2$

- [Mardia:93], [Dutilleul:99] Introduction, Flip-Flop.
- [Kermoal:02] Experiments in MIMO radio channels.
- [Lu:05], [Srivastava:08] Testing, uniqueness.
- [Werner:08] Asymptotic analysis and extensions.
- [Allen:10] Regularization and applications in bioinformatics.
- [Zhang:10], [Stegle:11] Sparsity, multitask learning.
- [Tsiligkaridis:12] COMING UP COLLOQUIUM.
- [Akdemir:11] Multiway Kronecker models.

Lots of applications! Lots of difficult theory!
But specific and hard to follow and generalize.

Revisiting the Kronecker model

The Kronecker likelihood function is

$$L(\mathbf{Q}_1, \mathbf{Q}_2) = \sum_{i=1}^{n} \mathbf{x}_i^T (\mathbf{Q}_1 \otimes \mathbf{Q}_2)^{-1} \mathbf{x}_i + \log |\mathbf{Q}_1 \otimes \mathbf{Q}_2|$$

- Non-convex optimization problem.
- 20 years of methods that converge to the global solution.

Theorem

The negative log likelihood is jointly g-convex in $\mathbf{Q}_1$ and $\mathbf{Q}_2$!
Also holds for multiway models with $\mathbf{Q}_1 \otimes \cdots \otimes \mathbf{Q}_n$.

Thus, every local minima is global, and we have lots of extensions.

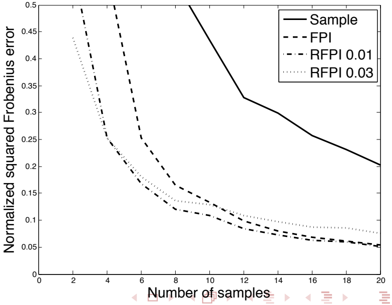
Why is this helpful? Regularized ML

- Kronecker models do not require many samples.
- [Allen:10] one sample + regularization via SVD.
- We propose

$$\min_{\mathbf{Q}_1,\mathbf{Q}_2} \; L(\mathbf{Q}_1, \mathbf{Q}_2) + \alpha \mathrm{Tr}\left\{\mathbf{Q}_1^{-1}\right\} \mathrm{Tr}\left\{\mathbf{Q}_2^{-1}\right\}$$

which is jointly g-convex.

$$p_1 = p_2 = 5$$
$$\mathbf{\Sigma}_{ij} = 0.8^{|i-j|}$$



Why is this helpful? Non-Gaussian & Kronecker ML

- Just for fun: hybrid robust Kronecker model:

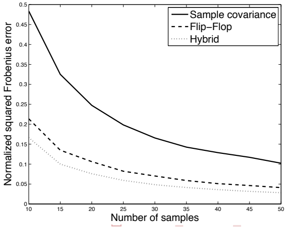
$$q_i \mathbf{Q} \quad + \quad \mathbf{Q}_1 \otimes \mathbf{Q}_2 \quad \Rightarrow \quad q_i \cdot \mathbf{Q}_1 \otimes \mathbf{Q}_2$$

- We propose

$$\min_{\mathbf{q}, \mathbf{Q}_1, \mathbf{Q}_2} \sum_{i=1}^{n} \mathbf{x}_i^T \left(q_i \cdot \mathbf{Q}_1 \otimes \mathbf{Q}_2\right)^{-1} \mathbf{x}_i + \log\left|q_i \cdot \mathbf{Q}_1 \otimes \mathbf{Q}_2\right|$$

which is jointly g-convex.

$$p_1 = 10 \text{ and } p_2 = 2$$
$$\mathbf{\Sigma}_{ij} = 0.8^{|i-j|}$$



Outline

Common symmetry constraints

Symmetry

$$\mathbf{Q} = \mathbf{U}\mathbf{Q}\mathbf{U}^T \quad \forall \quad \mathbf{U} \in \mathcal{K}$$

Applications:

- Circulant, used for approximating Toeplitz = stationary
- Persymmetric, e.g., radar systems using a symmetrically spaced linear array with constant pulse repetition interval

$$\begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-1} \\ c_1 & c_0 & c_1 & \cdots & c_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_1 & c_2 & c_3 & \cdots & c_0 \end{bmatrix} \quad \begin{bmatrix} p_{11} & p_{12} & p_{13} & \cdots & p_{1n} \\ p_{12} & p_{22} & p_{23} & \cdots & p_{1n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{41} & p_{42} & p_{32} & \cdots & p_{12} \\ p_{51} & p_{41} & p_{31} & \cdots & p_{11} \end{bmatrix}$$

More symmetry constraints - properness

Symmetry

$$\mathbf{Q} = \mathbf{U}\mathbf{Q}\mathbf{U}^T \quad \forall \quad \mathbf{U} \in \mathcal{K}$$

Applications:

- Complex normal = double real normal ($\mathcal{CN}_p = \mathcal{N}_{2p}$)
- Plus a symmetry constraint $\mathbf{x} \sim e^{j\theta}\mathbf{x}$.

$$\text{cov}\begin{bmatrix} \text{Re}(\mathbf{x}) \\ \text{Im}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ -\mathbf{B} & \mathbf{A} \end{bmatrix}$$

- Recently, proper Gaussian quaternions $\mathbf{x} = \mathbf{a} + i\mathbf{b} + j\mathbf{c} + k\mathbf{d}$.
- For example, in radar with I/Q phase and polarizations
- Here too: $\mathcal{QN}_p = \mathcal{N}_{4p}$ + special symmetry $\mathbf{x} \sim e^{\nu\theta}\mathbf{x}$.

A bit of background $\mathbf{Q} = \mathbf{U}\mathbf{Q}\mathbf{U}^T$

- Gaussian
  - Genreal symmetry groups [Shah & Chandrasekaran 2012]
  - Everybody knows proper complex (circularly symmetric)
  - Proper quaternion [Miron:06], [Bukhari:11], [Via:11]....
- Non Gaussian
  - Persymmetric [Pailloux:11]
  - Complex elliptical distributions [Bombrun:11], [Ollila:12]

Lots of applications! But specific and hard to follow and generalize.
Easy in the Gaussian case (linear constraint).

Revisiting symmetry constraints

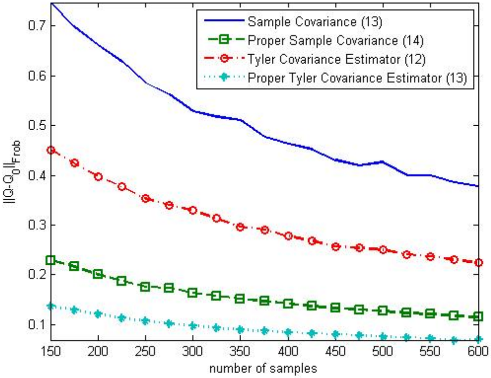
Theorem

The set $\mathbf{Q} = \mathbf{U}\mathbf{Q}\mathbf{U}^T$ is g-convex!

- Can be combined with any g-convex negative-log-likelihood.
- Can be combined with Kronecker models.
- Symmetrically constrained Tyler, MGGD....
- Any descent algorithm should find the global solution.

Experiments

Proper quaternion multivariate T distribution, dimension 10.



Discussion

- Geodesic convexity in positive definite matrices
- Similar to geometric programming in scalars.
- Powers and log determinants are g-convex.
- G-convexity is preserved in Kronecker products.
- Symmetry sets are g-convex.
- Unifies and generalizes many previous results.
- Lots of applications....

Take home message

If you always find the global solution, maybe its (g-)convex!