

Globally Optimal Learning for Structured Elliptical Losses

Yoav Wald ^{1,2} Nofar Noy ² Gal Elidan ^{1,2} Ami Wiesel ^{1,2}

¹Google Research ²Hebrew University of Jerusalem

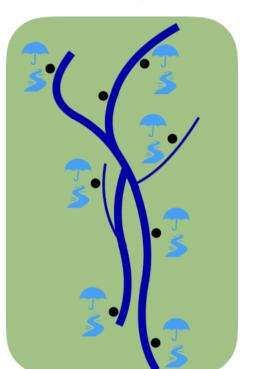
Motivation

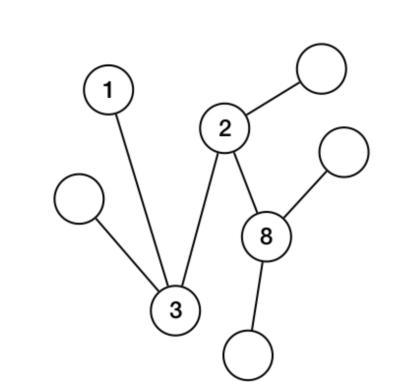
We are interested in regression problems where there are:

- Multiple correlated tasks, calls for Structured Prediction.
- Data that is heavy tailed or contaminated by outliers. Usually handled with **Robust Regression**.

For example, predicting river discharge in multiple locations:

•
$$\mathbf{x} = [\mathbf{f}(t), \mathbf{s}(t)]$$
• $\mathbf{y} = \mathbf{s}(t+1) \in \mathbb{R}^n$





Structured Regression

For correlated regression tasks, it is common to use the **inverse covariance**:

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \ \Gamma^{-1} = \mathbb{E}_{\mathbf{z}} \begin{bmatrix} \mathbf{z} \mathbf{z}^{\top} \end{bmatrix}, \ \Gamma = \begin{bmatrix} \Gamma_{\mathbf{x}\mathbf{x}} & \Gamma_{\mathbf{x}\mathbf{y}} \\ \Gamma_{\mathbf{y}\mathbf{x}} & \Gamma_{\mathbf{y}\mathbf{y}} \end{bmatrix}$$

This matrix satisfies desirable properties:

Optimal linear regressor obtained in closed form.

$$\mathbf{y}(\mathbf{x}) = \Gamma_{\mathbf{y}\mathbf{y}}^{-1} \Gamma_{\mathbf{y}\mathbf{x}} \mathbf{x}$$

• Conditional independence structures are manifested as sparsity patterns in Γ .

$$\Gamma(\mathbf{w}) = \begin{bmatrix} w_{11} & 0 & w_{13} & \cdots & 0 \\ 0 & w_{22} & w_{23} & \cdots & w_{28} \\ \vdots & & & & \\ 0 & w_{28} & \cdots & & w_{88} \end{bmatrix}$$

Standard way to learn such matrices is maximum likelihood of a Gaussian Markov Random Field:

$$(GMRF) \quad \arg\min_{\mathbf{w}:\Gamma(\mathbf{w})\succeq 0} \frac{1}{m} \sum_{i=1}^{m} \mathbf{z}_{i}^{\top} \Gamma(\mathbf{w}) \mathbf{z}_{i} + \log|\Gamma(\mathbf{w})^{-1}|. \tag{1}$$

Objective is **convex**, very efficient solutions. Minimizes **squared loss** if no structure imposed. **Caveat: not robust** to heavy tailed and contaminated data, common in real-world applications.

Robust Elliptical Losses

In regression with single label, robust losses $\rho(\langle \mathbf{w}, \mathbf{x} \rangle - y)$ are suitable. e.g. Huber's loss:

$$\rho(t) = \min\{\frac{1}{2}t^2, \delta(|t| - \frac{1}{2}\delta)\}$$

Suggestion for **structured multitask** setting:

$$(RMRF) \quad \arg\min_{\mathbf{w}:\Gamma(\mathbf{w})\succeq 0} \frac{1}{m} \sum_{i=1}^{m} \rho\left(\sqrt{\mathbf{z}_{i}^{\top}\Gamma(\mathbf{w})\mathbf{z}_{i}}\right) + \log|\Gamma(\mathbf{w})^{-1}|. \tag{2}$$

Based on maximizing likelihood of elliptical distributions. Robust loss, but non-convex.

Globally Optimal Learning

Main result: Formal guarantees on optimization, for most common losses in the literature

Gaussian	$\rho(t) = t^2$
Generalized Gaussian	$\rho(t) = t^{2\beta}, \beta \in (0, 1)$
T distribution	$\rho(t) = \frac{n+\nu}{2} \log(1 + \frac{t^2}{\nu}), \nu > 2$
Angular / Tyler	$\rho(t) = n \log(t^2)$
Huber	$\rho(t) = \min\{\frac{1}{2}t^2, \delta(t - \frac{1}{2}\delta)\}$

Assumptions:

- The loss $\rho(\sqrt{t})$ is twice differentiable and concave in t. Its derivative w.r.t t, denoted by ψ , satisfies $\psi(t) \ge -t\psi'(t)$ for all t > 0.
- Data comes from a Spherically Invariant distribution (close relatives of elliptical distributions), result applies at the population limit.
- Structure is linear. Includes graphical structures, but allows for many other structures. Matrices are defined by weights $\mathbf{w} \in \mathbb{R}^I$:

$$\Gamma(\mathbf{w}) = \sum_{\alpha \in I} w_{\alpha} \mathbf{G}_{\alpha}.$$

Theorem

Let **z** be an SIRV $(g, \Gamma^{-1}(\mathbf{w}^*))$ and ρ a function that satisfies our assumption. Consider the optimization problem:

$$\min_{\mathbf{w}:\Gamma(\mathbf{w})\succeq 0} \mathbb{E}_{\mathbf{z}} \left\{ \rho \left(\sqrt{\mathbf{z}^{\top} \Gamma(\mathbf{w}) \mathbf{z}} \right) \right\} + \log |\Gamma(\mathbf{w})^{-1}|.$$
 (3)

If **w** is a stationary point of the loss, it holds that $\Gamma(\mathbf{w})$ equals $\Gamma(\mathbf{w}^*)$ up to a multiplicative constant.

Proof relies on the following useful result.

Lemma: Let \mathbf{v} be an SIRV (g, Σ) with arbitrary texture g, and ρ a function that satisfies our assumption. Define the matrix:

$$\Sigma^{\rho}(\mathbf{v}) = \mathbb{E}_{\mathbf{v}} \left[\mathbf{v} \mathbf{v}^{\top} \psi(\|\mathbf{v}\|_{2}^{2}) \right].$$

Then $\Sigma^{\rho}(\mathbf{v})$ and Σ commute, and maintain the same order of eigenvalues.

Solution with Minimization-Majorization

Minimization-majorization with minimization steps based on Gaussian MRFs work much better than the alternatives in practice.

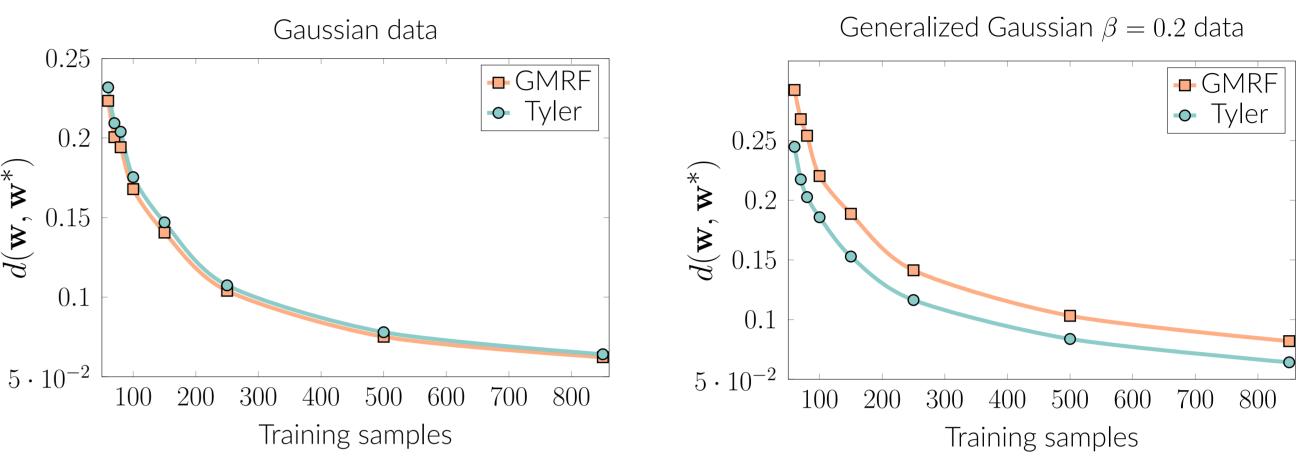
Algorithm 1 Minimization Majorization for Elliptical Markov Random Fields Require: $\rho: \mathbb{R}_{++} \to \mathbb{R}, \{\mathbf{z}_i\}_{i=1}^m$

- 1: Set $\Gamma_0 \leftarrow \boldsymbol{I}$
- 2: for $t = 0 \dots T$ do
- 3: Rescale data $\tilde{\mathbf{z}}_i = \psi(\sqrt{\mathbf{z}_i^{\top} \Gamma_t \mathbf{z}_i})^{\frac{1}{2}} \cdot \mathbf{z}_i \quad \forall i \in [m]$
- Solve convex minimization in (1) with data $\{\tilde{\mathbf{z}}_i\}_{i=1}^m$ and set Γ_{t+1} with the solution
- 5. and for

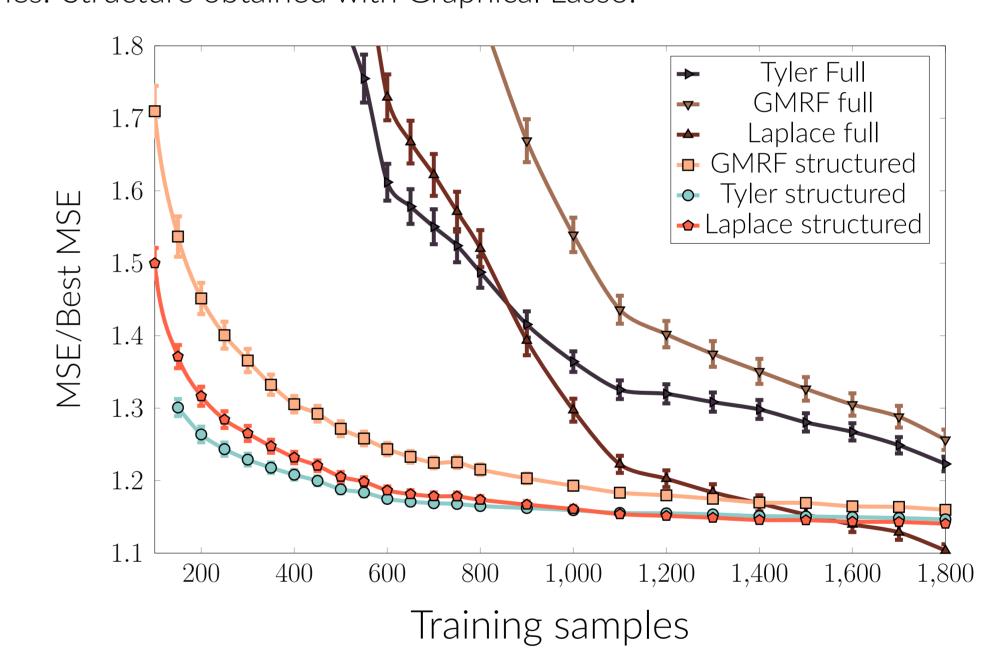
Experiments

We compare the GMRF and elliptical losses of RMRFs on synthetic and real-world data.

Synthetic: random structured covariances, robust losses significantly better for heavy-tailed data.



Kaggle "Huge Stock Market Dataset": Predict intra-day returns of 15 hidden stocks from 105 observed ones. Structure obtained with Graphical Lasso.



River Discharge dataset: Predict river discharge in 34 locations from past precipitation and discharge levels. Structure determined roughly by geographical proximity.

