Learning (simple) regularizers for inverse problems

Giovanni S. Alberti
MaLGa Center, Department of Mathematics, University of Genoa

Joint work with:
E. De Vito, M. Santacesaria, S. Sciutto (U. Genoa), T. Helin (LUT), M. Lassas (U. Helsinki), L. Ratti (U. Bologna)

Data-Enabled Science Seminar
University of Houston
December 6, 2024

Co-funded by the European Union (ERC, SAMPDE, 101041040). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the EU nor the granting authority can be held responsible for them.

Inverse problems

Linear inverse problems

Recover $x \in X$ from the noisy measurement $y \in Y$:

$$y = Ax + \varepsilon$$

- X, Y: separable Hilbert spaces
- $A\colon X \to Y$: bounded linear **injective** operator, A^{-1} possibly **unbounded**

UniGe | MaLGa

Inverse problems

Linear inverse problems

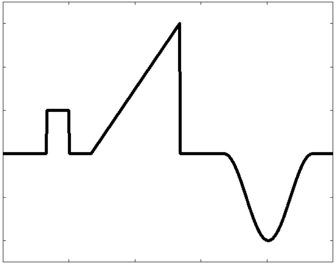
Recover $x \in X$ from the noisy measurement $y \in Y$:

$$y = Ax + \varepsilon$$

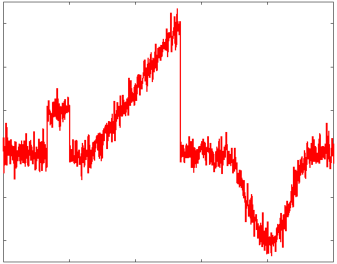
- X, Y: separable Hilbert spaces
- $A\colon X \to Y$: bounded linear **injective** operator, A^{-1} possibly **unbounded**

Denoising - $A = \mathrm{Id}$: identity operator

Unknown to be recovered, x



Observed quantity, y



UniGe | MaLGa

Linear inverse problems

Recover $x \in X$ from the noisy measurement $y \in Y$:

$$y = Ax + \varepsilon$$

- X, Y: separable Hilbert spaces
- $A\colon X \to Y$: bounded linear **injective** operator, A^{-1} possibly **unbounded**

Image deblurring - A: convolution with a smooth kernel

Unknown to be recovered, x

Observed quantity, y

Inverse problems

Linear inverse problems

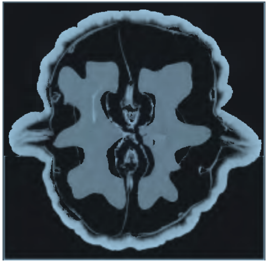
Recover $x \in X$ from the noisy measurement $y \in Y$:

$$y = Ax + \varepsilon$$

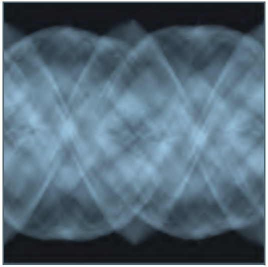
- X, Y: separable Hilbert spaces
- $A\colon X \to Y$: bounded linear **injective** operator, A^{-1} possibly **unbounded**

Computed Tomography - A: Radon transform

Unknown to be recovered, x



Observed quantity, y



Regularization - optimization problem

Given $\quad y = Ax + \varepsilon, \qquad$ solve $\quad \min_{x \in X} \{d_Y(Ax, y) + J(x)\}$

- $d_Y(Ax, y)$ data fidelity term, e.g. $\frac{1}{2}\|Ax - y\|_Y^2$

Regularization - optimization problem

Given $y = Ax + \varepsilon,$ solve $\min_{x \in X} \{d_Y(Ax, y) + J(x)\}$

- $d_Y(Ax, y)$ data fidelity term, e.g. $\frac{1}{2}\|Ax - y\|_Y^2$
- $J\colon X \to \mathbb{R}$ regularization term

Regularization

Regularization - optimization problem

Given $y = Ax + \varepsilon,$ solve $\min_{x \in X} \{d_Y(Ax, y) + J(x)\}$

- $d_Y(Ax, y)$ data fidelity term, e.g. $\frac{1}{2}\|Ax - y\|_Y^2$
- $J\colon X \to \mathbb{R}$ regularization term

How to choose the regularization functional? [1]

J should encode and promote prior information available on the solution

[1] Classical theory: [Engl, Hanke, Neubauer, 1996], Data-driven methods: [JC De los Reyes et al, 2017], [Calatroni et al, 2017], [Lunz et al, 2018], [Arridge et al., 2019], [Li et al., 2020], [Aspri et al. 2021], [De Hoop et al., 2021], [Kabri et al., 2024]

Regularization

Regularization - optimization problem

Given $y = Ax + \varepsilon,$ solve $\min_{x \in X} \{d_Y(Ax, y) + J(x)\}$

- $d_Y(Ax, y)$ data fidelity term, e.g. $\frac{1}{2}\|Ax - y\|_Y^2$
- $J\colon X \to \mathbb{R}$ regularization term

How to choose the regularization functional? [1]

J should encode and promote prior information available on the solution

Ex.1) Tikhonov regularization: $J(x) = \lambda\|x\|_X^2$

Ex.2) Sparsity-promoting regularization: $J(x) = \lambda\|x\|_1 = \lambda\|\{\langle x, \varphi_i\rangle_X\}_i\|_{\ell^1}$

Ex.3) Total Variation: $J(x) = \lambda\|\nabla x\|_1$

Ex.4) A neural network (e.g. unrolling, plug-and-play, adversarial regularizers, etc.)

[1] Classical theory: [Engl, Hanke, Neubauer, 1996],
Data-driven methods: [JC De los Reyes et al, 2017], [Calatroni et al, 2017], [Lunz et al, 2018], [Arridge et al., 2019], [Li et al., 2020], [Aspri et al. 2021], [De Hoop et al., 2021], [Kabri et al., 2024]

Regularization - optimization problem

Given $\quad y = Ax + \varepsilon, \qquad$ solve $\quad \min_{x \in X} \{d_Y(Ax, y) + J(x)\}$

- $d_Y(Ax, y)$ data fidelity term, e.g. $\frac{1}{2}\|Ax - y\|_Y^2$
- $J\colon X \to \mathbb{R}$ regularization term

How to choose the regularization functional? [1]

J should encode and promote prior information available on the solution

Ex.1) Tikhonov regularization: $J(x) = \lambda\|x\|_X^2$

Ex.2) Sparsity-promoting regularization: $J(x) = \lambda\|x\|_1 = \lambda\|\{\langle x, \varphi_i\rangle_X\}_i\|_{\ell^1}$

Ex.3) Total Variation: $J(x) = \lambda\|\nabla x\|_1$

Ex.4) A neural network (e.g. unrolling, plug-and-play, adversarial regularizers, etc.)

[1] Classical theory: [Engl, Hanke, Neubauer, 1996],
Data-driven methods: [JC De los Reyes et al, 2017], [Calatroni et al, 2017], [Lunz et al, 2018], [Arridge et al., 2019], [Li et al., 2020], [Aspri et al. 2021], [De Hoop et al., 2021], [Kabri et al., 2024]

Disclaimer[2]

This talk

State of the art

[2] Joke stolen from Ernesto De Vito, talking about kernel methods in machine learning

Disclaimer[2]

This talk

State of the art

[2]Joke stolen from Ernesto De Vito, talking about kernel methods in machine learning

Disclaimer[2]

This talk

State of the art

[2]Joke stolen from Ernesto De Vito, talking about kernel methods in machine learning

Disclaimer[2]

This talk

State of the art

[2]Joke stolen from Ernesto De Vito, talking about kernel methods in machine learning

Outline

Learning the optimal generalized Tikhonov regularizer

Learning the optimal ℓ^1 regularizer

Sparse regularization via Gaussian mixtures

Generalized Tikhonov regularization

$$R_{h,B}(y) = \arg\min_{x \in X} \left\{ d_Y(Ax, y) + \|B^{-1}(x - h)\|_X^2 \right\}$$

where $h \in X$ and $B\colon X \to X$ is positive and bounded

Generalized Tikhonov regularization

Generalized Tikhonov regularization

$$R_{h,B}(y) = \operatorname*{arg\,min}_{x \in X} \left\{ d_Y(Ax, y) + \|B^{-1}(x-h)\|_X^2 \right\}$$

where $h \in X$ and $B\colon X \to X$ is positive and bounded

Examples:

- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h = 0$: Tikhonov regularization $\rightsquigarrow$ x has small norm

Generalized Tikhonov regularization

$$R_{h,B}(y) = \arg\min_{x \in X} \left\{ d_Y(Ax, y) + \|B^{-1}(x-h)\|_X^2 \right\}$$

where $h \in X$ and $B\colon X \to X$ is positive and bounded

Examples:

- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h = 0$: Tikhonov regularization $\rightsquigarrow$ x has small norm
- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h \in X$ $\rightsquigarrow$ x is a small variation of a reference object h

Generalized Tikhonov regularization

$$R_{h,B}(y) = \arg\min_{x \in X} \left\{ d_Y(Ax, y) + \|B^{-1}(x - h)\|_X^2 \right\}$$

where $h \in X$ and $B \colon X \to X$ is positive and bounded

Examples:

- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h = 0$: Tikhonov regularization $\rightsquigarrow$ x has small norm
- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h \in X$ $\rightsquigarrow$ x is a small variation of a reference object h
- $B^{-1} = \Delta^s$: Sobolev regularization $\rightsquigarrow$ enforces smoothness of x

Generalized Tikhonov regularization

Generalized Tikhonov regularization

$$R_{h,B}(y) = \arg\min_{x \in X} \left\{ d_Y(Ax, y) + \|B^{-1}(x-h)\|_X^2 \right\}$$

where $h \in X$ and $B \colon X \to X$ is positive and bounded

Examples:

- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h = 0$: Tikhonov regularization $\rightsquigarrow$ x has small norm
- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h \in X$ $\rightsquigarrow$ x is a small variation of a reference object h
- $B^{-1} = \Delta^s$: Sobolev regularization $\rightsquigarrow$ enforces smoothness of x
- B^{-1} = arbitrary differential operator $\rightsquigarrow$ enforces arbitrary smoothness of x

Generalized Tikhonov regularization

Generalized Tikhonov regularization

$$R_{h,B}(y) = \arg\min_{x \in X} \left\{ d_Y(Ax, y) + \|B^{-1}(x-h)\|_X^2 \right\}$$

where $h \in X$ and $B\colon X \to X$ is positive and bounded

Examples:

- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h = 0$: Tikhonov regularization $\rightsquigarrow$ x has small norm
- $B^{-1} = \lambda^{1/2}\,\mathrm{Id},\ h \in X$ $\rightsquigarrow$ x is a small variation of a reference object h
- $B^{-1} = \Delta^s$: Sobolev regularization $\rightsquigarrow$ enforces smoothness of x
- $B^{-1} =$ arbitrary differential operator $\rightsquigarrow$ enforces arbitrary smoothness of x

Learning the regularizer: key questions

1. What are the optimal B and h?
2. How can we learn them? How large should the training set be?

Statistical setting: finite dimension

Model for x: square-integrable random vector in $\mathbb{R}^N$;
mean: $\mu_x \in \mathbb{R}^N$; covariance: $\Sigma_x \in \mathbb{R}^{N \times N}$ invertible.

Statistical setting: finite dimension

Model for x: square-integrable random vector in $\mathbb{R}^N$;
mean: $\mu_x \in \mathbb{R}^N$; covariance: $\Sigma_x \in \mathbb{R}^{N \times N}$ invertible.

Model for ε: square-integrable random vector in $\mathbb{R}^N$, $\varepsilon \perp x$;
mean: $0 \in \mathbb{R}^N$; covariance: $\Sigma_\varepsilon \in \mathbb{R}^{N \times N}$ invertible.

Statistical setting: finite dimension

Model for x: square-integrable random vector in $\mathbb{R}^N$;
mean: $\mu_x \in \mathbb{R}^N$; covariance: $\Sigma_x \in \mathbb{R}^{N\times N}$ invertible.

Model for ε: square-integrable random vector in $\mathbb{R}^N$, $\varepsilon \perp x$;
mean: $0 \in \mathbb{R}^N$; covariance: $\Sigma_\varepsilon \in \mathbb{R}^{N\times N}$ invertible.

Regularizer:

$$R_{h,B}(y) = \operatorname*{arg\,min}_{x\in X} \left\{ \|\Sigma_\varepsilon^{-1/2}(Ax - y)\|_Y^2 + \|B^{-1}(x-h)\|_X^2 \right\}$$

Statistical setting: finite dimension

Model for x: square-integrable random vector in $\mathbb{R}^N$;
mean: $\mu_x \in \mathbb{R}^N$; covariance: $\Sigma_x \in \mathbb{R}^{N\times N}$ invertible.

Model for ε: square-integrable random vector in $\mathbb{R}^N$, $\varepsilon \perp x$;
mean: $0 \in \mathbb{R}^N$; covariance: $\Sigma_\varepsilon \in \mathbb{R}^{N\times N}$ invertible.

Regularizer:

$$R_{h,B}(y) = \arg\min_{x\in X} \left\{ \|\Sigma_\varepsilon^{-1/2}(Ax-y)\|_Y^2 + \|B^{-1}(x-h)\|_X^2 \right\}$$

$$\Downarrow$$

Regularizer – explicit formula:

$$\begin{aligned} R_{h,B}(y) &= (A^*\Sigma_\varepsilon^{-1}A + B^{-*}B^{-1})^{-1}(A^*\Sigma_\varepsilon^{-1}y + B^{-*}B^{-1}h) \\ &= h + B^*BA^*(AB^*BA^* + \Sigma_\varepsilon)^{-1}(y - Ah) \end{aligned}$$

Statistical setting: Infinite-dimension

Model for x: square-integrable random variable in X;
mean: $\mu_x \in X$; covariance: $\Sigma_x : X \to X$ trace-class, injective operator.

Model for ε: square-integrable random variable in Y, $\varepsilon \perp x$;
mean: $0 \in Y$; covariance: $\Sigma_\varepsilon : Y \to Y$ trace-class, injective operator.

Statistical setting: Infinite-dimension

Model for x: square-integrable random variable in X;
mean: $\mu_x \in X$; covariance: $\Sigma_x : X \to X$ trace-class, injective operator.

Model for ε: square-integrable random variable in Y, $\varepsilon \perp x$;
mean: $0 \in Y$; covariance: $\Sigma_\varepsilon : Y \to Y$ trace-class, injective operator.
$\Rightarrow$ **Problem:** white noise not included! ($\Sigma_\varepsilon = \mathrm{Id}$ is not trace-class)

Statistical setting: Infinite-dimension

Model for x: square-integrable random variable in X;
mean: $\mu_x \in X$; covariance: $\Sigma_x : X \to X$ trace-class, injective operator.

Model for ε: zero-mean random process on Y, $\varepsilon \perp x$;
mean: $0 \in K^*$; covariance: $\iota^* \circ \Sigma_\varepsilon \circ \iota : K^* \to K$ trace class, injective.

Gelfand triple:

$$K \overset{\iota}{\hookrightarrow} Y \overset{\iota^*}{\hookrightarrow} K^*$$

Statistical setting: Infinite-dimension

Model for x: square-integrable random variable in X;
mean: $\mu_x \in X$; covariance: $\Sigma_x : X \to X$ trace-class, injective operator.

Model for ε: zero-mean random process on Y, $\varepsilon \perp x$;
mean: $0 \in K^*$; covariance: $\iota^* \circ \Sigma_\varepsilon \circ \iota : K^* \to K$ trace class, injective.

Inverse problem:

$$y = Ax + \varepsilon$$

Statistical setting: Infinite-dimension

Model for x: square-integrable random variable in X;
mean: $\mu_x \in X$; covariance: $\Sigma_x : X \to X$ trace-class, injective operator.

Model for ε: zero-mean random process on Y, $\varepsilon \perp x$;
mean: $0 \in K^*$; covariance: $\iota^* \circ \Sigma_\varepsilon \circ \iota : K^* \to K$ trace class, injective.

Inverse problem:
$y = Ax + \varepsilon$

$\rightsquigarrow$

$$y = \iota^* Ax + \varepsilon \quad \text{in } K^*$$
$$\langle y, v \rangle_{K^* \times K} = \langle Ax, v \rangle_Y + \langle \varepsilon, v \rangle_{K^* \times K} \quad \forall v \in K$$

UniGe | MaLGa

Statistical setting: Infinite-dimension

Model for x: square-integrable random variable in X;
mean: $\mu_x \in X$; covariance: $\Sigma_x : X \to X$ trace-class, injective operator.

Model for ε: zero-mean random process on Y, $\varepsilon \perp x$;
mean: $0 \in K^*$; covariance: $\iota^* \circ \Sigma_\varepsilon \circ \iota : K^* \to K$ trace class, injective.

Inverse problem:
$y = Ax + \varepsilon$

$\rightsquigarrow$

$$y = \iota^* Ax + \varepsilon \quad \text{in } K^*$$
$$\langle y, v \rangle_{K^* \times K} = \langle Ax, v \rangle_Y + \langle \varepsilon, v \rangle_{K^* \times K} \quad \forall v \in K$$

Regularizer: desired form

$$R_{h,B}(y) = \arg\min_{x \in X} \left\{ \|\Sigma_\varepsilon^{-1/2}(Ax - y)\|_Y^2 + \|B^{-1}(x - h)\|_X^2 \right\}$$

UniGe | MaLGa

Statistical setting: Infinite-dimension

Model for x: square-integrable random variable in X;
mean: $\mu_x \in X$; covariance: $\Sigma_x : X \to X$ trace-class, injective operator.

Model for ε: zero-mean random process on Y, $\varepsilon \perp x$;
mean: $0 \in K^*$; covariance: $\iota^* \circ \Sigma_\varepsilon \circ \iota : K^* \to K$ trace class, injective.

Inverse problem:
$y = Ax + \varepsilon$

$\rightsquigarrow$

$$y = \iota^* Ax + \varepsilon \quad \text{in } K^*$$
$$\langle y, v\rangle_{K^* \times K} = \langle Ax, v\rangle_Y + \langle \varepsilon, v\rangle_{K^* \times K} \quad \forall v \in K$$

Regularizer: desired form

$$R_{h,B}(y) = \arg\min_{x \in X} \left\{ \|\Sigma_\varepsilon^{-1/2}(Ax - y)\|_Y^2 + \|B^{-1}(x - h)\|_X^2 \right\}$$

$\Rightarrow$ **Problem:** $\Sigma_\varepsilon^{-1/2} y \notin Y$

Statistical setting: Infinite-dimension

Model for x: square-integrable random variable in X;
mean: $\mu_x \in X$; covariance: $\Sigma_x : X \to X$ trace-class, injective operator.

Model for ε: zero-mean random process on Y, $\varepsilon \perp x$;
mean: $0 \in K^*$; covariance: $\iota^* \circ \Sigma_\varepsilon \circ \iota : K^* \to K$ trace class, injective.

Inverse problem:
$y = Ax + \varepsilon$

$\rightsquigarrow$

$$y = \iota^* Ax + \varepsilon \quad \text{in } K^*$$
$$\langle y, v\rangle_{K^* \times K} = \langle Ax, v\rangle_Y + \langle \varepsilon, v\rangle_{K^* \times K} \quad \forall v \in K$$

Regularizer: desired form

$$R_{h,B}(y) = \arg\min_{x \in X} \left\{ \|\Sigma_\varepsilon^{-1/2}(Ax - y)\|_Y^2 + \| \underbrace{B^{-1}(x-h)}_{x'} \|_X^2 \right\}$$

Regularizer: well-defined form - assume compatibility condition $\mathrm{Im}(AB) \subset \mathrm{Im}(\Sigma_\varepsilon \iota)$

$$R_{h,B}(y) = h + B\widehat{x}'$$

$$\widehat{x}' = \arg\min_{x' \in X} \left\{ \|\Sigma_\varepsilon^{-1/2} ABx'\|_Y^2 - 2\langle y - \iota^* Ah, (\Sigma_\varepsilon \iota)^{-1} ABx' \rangle_{K^* \times K} + \|x'\|_X^2 \right\}$$

The optimal regularizer

Mean squared error/expected loss:

$$L(h, B) = \mathbb{E}_{(x,\varepsilon)} \left[\| R_{h,B}(Ax + \varepsilon) - x \|_X^2 \right]$$

[3] Learning the optimal Tikhonov regularizer for inverse problems, NeurIPS 2021

The optimal regularizer

Mean squared error/expected loss:

$$L(h, B) = \mathbb{E}_{(x,\varepsilon)} \left[\| R_{h,B}(Ax + \varepsilon) - x \|_X^2 \right]$$

Theorem [A, De Vito, Lassas, Ratti, Santacesaria][3]

Let Σ_x satisfy $\mathrm{Im}(A\Sigma_x^{1/2}) \subseteq \mathrm{Im}(\Sigma_\varepsilon \iota)$ (compatibility). Then $(h^\star, B^\star)$ is a global minimizer of

$$\min_{h,B} L(h, B)$$

if and only if

$$h^\star = \mu_x \quad \text{and} \quad (B^\star)^2 = \Sigma_x.$$

[3] Learning the optimal Tikhonov regularizer for inverse problems, NeurIPS 2021

The optimal regularizer

Mean squared error/expected loss:

$$L(h, B) = \mathbb{E}_{(x,\varepsilon)} \left[\| R_{h,B}(Ax + \varepsilon) - x \|_X^2 \right]$$

Theorem [A, De Vito, Lassas, Ratti, Santacesaria][3]

Let Σ_x satisfy $\operatorname{Im}(A\Sigma_x^{1/2}) \subseteq \operatorname{Im}(\Sigma_\varepsilon \iota)$ (compatibility). Then $(h^\star, B^\star)$ is a global minimizer of

$$\min_{h,B} L(h, B)$$

if and only if

$$h^\star = \mu_x \quad \text{and} \quad (B^\star)^2 = \Sigma_x.$$

Remarks

- The optimal regularization parameters $B^\star = \Sigma_x^{1/2}$ and $h^\star = \mu_x$ are independent of A and ϵ

[3] Learning the optimal Tikhonov regularizer for inverse problems, NeurIPS 2021

The optimal regularizer

Mean squared error/expected loss:

$$L(h, B) = \mathbb{E}_{(x,\varepsilon)} \left[\| R_{h,B}(Ax + \varepsilon) - x \|_X^2 \right]$$

Theorem [A, De Vito, Lassas, Ratti, Santacesaria][3]

Let Σ_x satisfy $\mathrm{Im}(A\Sigma_x^{1/2}) \subseteq \mathrm{Im}(\Sigma_\varepsilon \iota)$ (compatibility). Then $(h^\star, B^\star)$ is a global minimizer of

$$\min_{h,B} L(h, B)$$

if and only if

$$h^\star = \mu_x \quad \text{and} \quad (B^\star)^2 = \Sigma_x.$$

Remarks

- The optimal regularization parameters $B^\star = \Sigma_x^{1/2}$ and $h^\star = \mu_x$ are independent of A and ϵ
- Expression of the optimal regularizer $R^\star = R_{h^\star, B^\star}$ (LMMSE estimator):

$$R^\star(y) = \mu_x + \Sigma_x A^* (\iota^* (A \Sigma_x A^* + \Sigma_\varepsilon))^{-1} (y - \iota^* A \mu_x)$$

[3] Learning the optimal Tikhonov regularizer for inverse problems, NeurIPS 2021

Learning the regularizer: two approaches

Goal: given a sample $z=\{(x_j,y_j)\}_{j=1}^m \in (X \times K^*)^m$, approximate $(h^\star, B^\star)$

UniGe | MaLGa

Learning the regularizer: two approaches

Goal: given a sample $z = \{(x_j, y_j)\}_{j=1}^m \in (X \times K^*)^m$, approximate $(h^\star, B^\star)$

Supervised learning: find $(\widehat{h}_S, \widehat{B}_S)$ minimizing the empirical risk $\widehat{L}$,

$$(\widehat{h}_S, \widehat{B}_S) = \underset{(h,B)\in\Theta}{\operatorname{argmin}} \ \widehat{L}(h, B), \qquad \widehat{L}(h, B) = \frac{1}{m} \sum_{j=1}^{m} \|R_{h,B}(y_j) - x_j\|_X^2,$$

where Θ is a suitable subset of $X \times \mathcal{L}(X, X)$.

UniGe | MaLGa

Learning the regularizer: two approaches

Goal: given a sample $z=\{(x_j,y_j)\}_{j=1}^m \in (X \times K^\star)^m$, approximate $(h^\star, B^\star)$

Supervised learning: find $(\widehat{h}_S, \widehat{B}_S)$ minimizing the empirical risk $\widehat{L}$,

$$(\widehat{h}_S, \widehat{B}_S) = \operatorname*{argmin}_{(h,B)\in\Theta} \widehat{L}(h,B), \qquad \widehat{L}(h,B) = \frac{1}{m}\sum_{j=1}^m \|R_{h,B}(y_j) - x_j\|_X^2,$$

where Θ is a suitable subset of $X \times \mathcal{L}(X,X)$.

Unsupervised learning: since $h^\star = \mu_x$ and $B^\star = \Sigma_x^{1/2}$, use only the sample $\{x_j\}_{j=1}^m$ to estimate

$$\widehat{h}_U = \widehat{\mu_x} = \frac{1}{m}\sum_{j=1}^m x_j, \qquad \widehat{B}_U = \widehat{\Sigma_x}^{1/2}, \quad \widehat{\Sigma}_x = \frac{1}{m}\sum_{j=1}^m (x_j - \hat{\mu_x}) \otimes (x_j - \hat{\mu_x}).$$

UniGe | MaLGa

Learning the regularizer: two approaches

Goal: given a sample $z=\{(x_j,y_j)\}_{j=1}^m \in (X \times K^\star)^m$, approximate $(h^\star, B^\star)$

Supervised learning: find $(\widehat{h}_S, \widehat{B}_S)$ minimizing the empirical risk $\widehat{L}$,

$$(\widehat{h}_S, \widehat{B}_S) = \operatorname*{argmin}_{(h,B)\in\Theta} \widehat{L}(h,B), \qquad \widehat{L}(h,B) = \frac{1}{m}\sum_{j=1}^{m} \|R_{h,B}(y_j) - x_j\|_X^2,$$

where Θ is a suitable subset of $X \times \mathcal{L}(X,X)$.

Unsupervised learning: since $h^\star = \mu_x$ and $B^\star = \Sigma_x^{1/2}$, use only the sample $\{x_j\}_{j=1}^m$ to estimate

$$\widehat{h}_U = \widehat{\mu_x} = \frac{1}{m}\sum_{j=1}^{m} x_j, \qquad \widehat{B}_U = \widehat{\Sigma_x}^{1/2}, \quad \widehat{\Sigma_x} = \frac{1}{m}\sum_{j=1}^{m} (x_j - \hat{\mu_x}) \otimes (x_j - \hat{\mu_x}).$$

How to evaluate the quality of $(\widehat{h}, \widehat{B})$?

Bounds on the excess error: $L(\widehat{h}, \widehat{B}) - L(h^\star, B^\star)$

Supervised learning - assumptions and main result

$$(h^*, B^*) = \underset{(h,B)\in\Theta}{\arg\min} \underbrace{\mathbb{E}_{x,y}[\|R_{h,B}(y) - x\|_X^2]}_{L(h,B)}, \qquad (\widehat{h}_S, \widehat{B}_S) = \underset{(h,B)\in\Theta}{\arg\min} \sum_{j=1}^{m} \|R_{h,B}(y_j) - x_j\|_X^2$$

1. $\Theta \subset H \times \mathrm{HS}(H^*, H) \subset X \times \mathcal{L}(X, X)$ is compact.

Example: $X = L^2(\mathbb{T}^d)$, $H = H^\sigma(\mathbb{T}^d)$ Sobolev space, smoothness σ

2. *quantify* compactness via s (Sobolev example: $s = \sigma/d$)
3. $(h^\star, B^\star) = (\mu_x, \Sigma_x^{1/2}) \in \Theta$

UniGe | MaLGa

Supervised learning - assumptions and main result

$$(h^*, B^*) = \arg\min_{(h,B)\in\Theta} \underbrace{\mathbb{E}_{x,y}[\|R_{h,B}(y) - x\|_X^2]}_{L(h,B)}, \qquad (\widehat{h}_S, \widehat{B}_S) = \arg\min_{(h,B)\in\Theta} \sum_{j=1}^{m} \|R_{h,B}(y_j) - x_j\|_X^2$$

1. $\Theta \subset H \times \mathrm{HS}(H^*, H) \subset X \times \mathcal{L}(X, X)$ is compact.

Example: $X = L^2(\mathbb{T}^d)$, $H = H^\sigma(\mathbb{T}^d)$ Sobolev space, smoothness σ

2. *quantify* compactness via s (Sobolev example: $s = \sigma/d$)
3. $(h^\star, B^\star) = (\mu_x, \Sigma_x^{1/2}) \in \Theta$

Theorem [A, De Vito, Lassas, Ratti, Santacesaria][4]

Take $\tau > 0$, $s' \in (0, s)$. Then, with probability exceeding $1 - e^{-\tau}$,

$$|L(\widehat{h}_S, \widehat{B}_S) - L(h^\star, B^\star)| \leq \left(\frac{c_1 + c_2\sqrt{\tau}}{\sqrt{m}}\right)^{1-\frac{1}{2s'+1}}.$$

[4] Learning the optimal Tikhonov regularizer for inverse problems, NeurIPS 2021

$$\widehat{h}_U = \widehat{\mu_x} = \frac{1}{m}\sum_{j=1}^{m} x_j, \qquad \widehat{B}_U = \widehat{\Sigma_x}^{1/2}, \quad \widehat{\Sigma_x} = \frac{1}{m}\sum_{j=1}^{m}(x_j - \hat{\mu_x}) \otimes (x_j - \hat{\mu_x}).$$

1. x is a κ-sub-Gaussian random variable

   Example: Gaussian r.v., ~~bounded~~ r.v.

2. technical assumptions

$$\widehat{h}_U = \widehat{\mu_x} = \frac{1}{m}\sum_{j=1}^{m} x_j, \qquad \widehat{B}_U = \widehat{\Sigma_x}^{1/2}, \quad \widehat{\Sigma_x} = \frac{1}{m}\sum_{j=1}^{m}(x_j - \hat{\mu_x}) \otimes (x_j - \hat{\mu_x}).$$

1. x is a κ-sub-Gaussian random variable

   Example: Gaussian r.v., ~~bounded~~ r.v.

2. technical assumptions

Theorem [A, De Vito, Lassas, Ratti, Santacesaria][5]

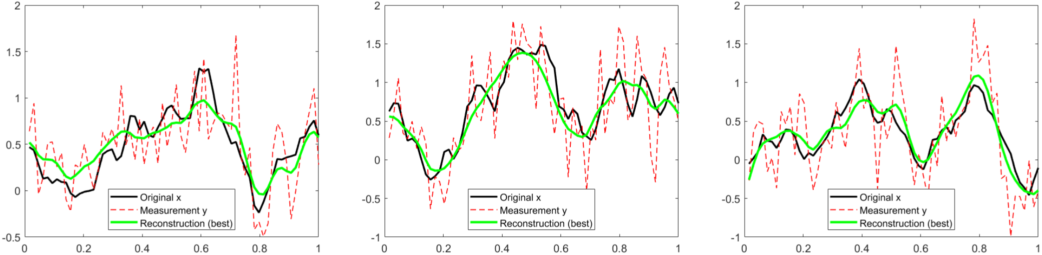
Take $\tau > 0$. Then, with probability exceeding $1 - e^{-\tau}$,

$$|L(\widehat{h}_U, \widehat{B}_U) - L(h^\star, B^\star)| \leq \frac{c_3 + c_4\sqrt{\tau}}{\sqrt{m}}.$$

[5] Learning the optimal Tikhonov regularizer for inverse problems, NeurIPS 2021

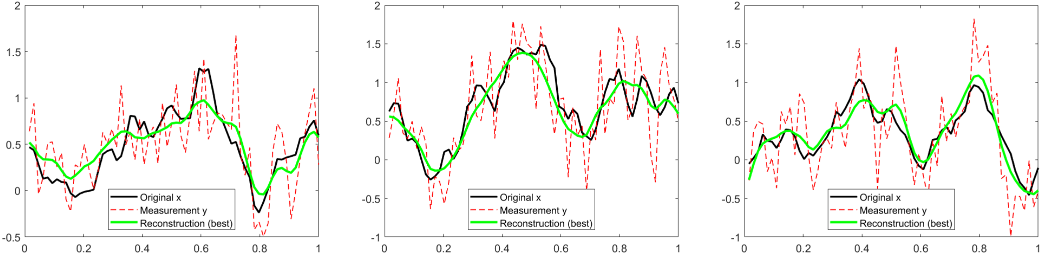
A denoising problem - experimental setup

- $X = Y = L^2(\mathbb{T}^1)$, $\mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ the one-dimensional torus
- $A = \mathrm{Id}$: determine a signal x from $y = x + \varepsilon$



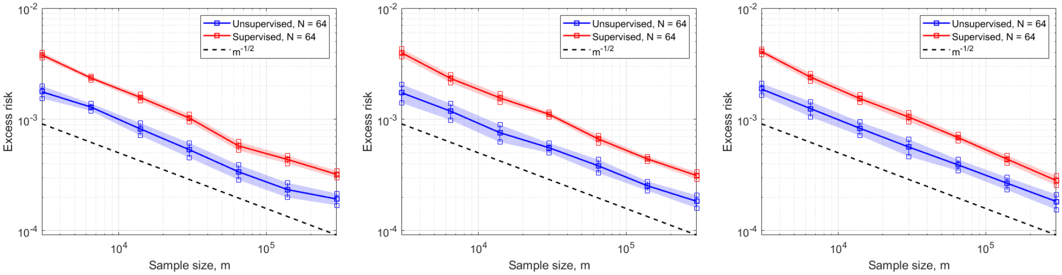
A denoising problem - experimental setup

- $X = Y = L^2(\mathbb{T}^1)$, $\mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ the one-dimensional torus
- $A = \mathrm{Id}$: determine a signal x from $y = x + \varepsilon$



- $x \sim \mathcal{N}(\mu_x, \Sigma_x)$, $\mu_x = 1 - |2x - 1|$, Σ_x: smooth convolution operator
- ε: white noise process, with zero mean and $\Sigma_\varepsilon = \sigma^2 I$
- Discretization: $X = \mathbb{R}^N$ (N dimensional 1D-pixel basis)

Experiment 1: verify the generalization bounds



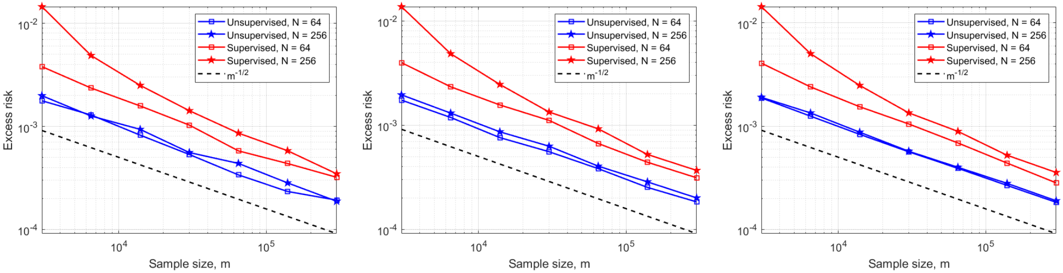
(a) (b) (c)

Decay in m of the excess risks

$$\color{red}{|L(\widehat{\theta}_S) - L(\theta^\star)|} \quad \text{and} \quad \color{blue}{|L(\widehat{\theta}_U) - L(\theta^\star)|}$$

with Gaussian variable x and
(a) Gaussian white noise ε
(b) uniform white noise ε
(c) white noise ε whose wavelet transform has uniform distribution

Experiment 2: dimension-independence



(a) (b) (c)

Outline

Learning the optimal generalized Tikhonov regularizer

Learning the optimal ℓ^1 regularizer

Sparse regularization via Gaussian mixtures

ℓ^1 regularization

Analysis formulation

$$\min_{x \in X} \left\{ \frac{1}{2} \|Ax - y\|_Y^2 + \|\Phi x\|_{\ell^1} \right\}$$

[6] H. Huang, E. Haber, and L. Horesh, Optimal estimation of ℓ^1-regularization prior from a regularized empirical Bayesian risk standpoint, Inverse Probl. Imaging, 2012

ℓ^1 regularization

Analysis formulation

$$\min_{x \in X} \left\{ \frac{1}{2}\|Ax - y\|_Y^2 + \|\Phi x\|_{\ell^1} \right\}$$

$\rightsquigarrow$

Synthesis formulation

$$\min_{u \in U \subset \ell^1} \left\{ \frac{1}{2}\|ABu - y\|_Y^2 + \|u\|_{\ell^1} \right\}$$

where

$$x = Bu, \qquad B \colon \ell^2 \to X \text{ bounded}$$

[6] H. Huang, E. Haber, and L. Horesh, Optimal estimation of ℓ^1-regularization prior from a regularized empirical Bayesian risk standpoint, Inverse Probl. Imaging, 2012

ℓ^1 regularization

Analysis formulation

$$\min_{x \in X} \left\{ \frac{1}{2}\|Ax - y\|_Y^2 + \|\Phi x\|_{\ell^1} \right\}$$

$\rightsquigarrow$

Synthesis formulation

$$\min_{u \in U \subset \ell^1} \left\{ \frac{1}{2}\|ABu - y\|_Y^2 + \|u\|_{\ell^1} \right\}$$

where

$$x = Bu, \qquad B\colon \ell^2 \to X \text{ bounded}$$

Examples

- canonical/pixel-based basis: few activated pixels

[6] H. Huang, E. Haber, and L. Horesh, Optimal estimation of ℓ^1-regularization prior from a regularized empirical Bayesian risk standpoint, Inverse Probl. Imaging, 2012

ℓ^1 regularization

Analysis formulation

$$\min_{x \in X} \left\{ \frac{1}{2}\|Ax - y\|_Y^2 + \|\Phi x\|_{\ell^1} \right\}$$

$\rightsquigarrow$

Synthesis formulation

$$\min_{u \in U \subset \ell^1} \left\{ \frac{1}{2}\|ABu - y\|_Y^2 + \|u\|_{\ell^1} \right\}$$

where

$$x = Bu, \qquad B\colon \ell^2 \to X \text{ bounded}$$

Examples

- canonical/pixel-based basis: few activated pixels
- Fourier basis: band-limited functions, smooth functions

[6] H. Huang, E. Haber, and L. Horesh, Optimal estimation of ℓ^1-regularization prior from a regularized empirical Bayesian risk standpoint, Inverse Probl. Imaging, 2012

ℓ^1 regularization

Analysis formulation

$$\min_{x\in X}\left\{\frac{1}{2}\|Ax-y\|_Y^2+\|\Phi x\|_{\ell^1}\right\}$$

$\rightsquigarrow$

Synthesis formulation

$$\min_{u\in U\subset\ell^1}\left\{\frac{1}{2}\|ABu-y\|_Y^2+\|u\|_{\ell^1}\right\}$$

where

$$x = Bu, \qquad B\colon \ell^2 \to X \text{ bounded}$$

Examples

- canonical/pixel-based basis: few activated pixels
- Fourier basis: band-limited functions, smooth functions
- wavelet bases: isolated discontinuities in some points

[6] H. Huang, E. Haber, and L. Horesh, Optimal estimation of ℓ^1-regularization prior from a regularized empirical Bayesian risk standpoint, Inverse Probl. Imaging, 2012

ℓ^1 regularization

Analysis formulation

$$\min_{x \in X} \left\{ \frac{1}{2} \|Ax - y\|_Y^2 + \|\Phi x\|_{\ell^1} \right\}$$

$\rightsquigarrow$

Synthesis formulation

$$\min_{u \in U \subset \ell^1} \left\{ \frac{1}{2} \|ABu - y\|_Y^2 + \|u\|_{\ell^1} \right\}$$

where

$$x = Bu, \qquad B \colon \ell^2 \to X \text{ bounded}$$

Examples

- canonical/pixel-based basis: few activated pixels
- Fourier basis: band-limited functions, smooth functions
- wavelet bases: isolated discontinuities in some points
- curvelet/shearlet frames: isolated discontinuities along curves

[6] H. Huang, E. Haber, and L. Horesh, Optimal estimation of ℓ^1-regularization prior from a regularized empirical Bayesian risk standpoint, Inverse Probl. Imaging, 2012

ℓ^1 regularization

Analysis formulation

$$\min_{x \in X} \left\{ \frac{1}{2} \|Ax - y\|_Y^2 + \|\Phi x\|_{\ell^1} \right\}$$

$\rightsquigarrow$

Synthesis formulation

$$\min_{u \in U \subset \ell^1} \left\{ \frac{1}{2} \|ABu - y\|_Y^2 + \|u\|_{\ell^1} \right\}$$

where

$$x = Bu, \qquad B \colon \ell^2 \to X \text{ bounded}$$

Examples

- canonical/pixel-based basis: few activated pixels
- Fourier basis: band-limited functions, smooth functions
- wavelet bases: isolated discontinuities in some points
- curvelet/shearlet frames: isolated discontinuities along curves

Goal: learn the optimal choice of B based on sample data[6]

[6] H. Huang, E. Haber, and L. Horesh, Optimal estimation of ℓ^1-regularization prior from a regularized empirical Bayesian risk standpoint, Inverse Probl. Imaging, 2012

Sparsity promotion and ℓ^1 - assumptions

- ⊖ No explicit formula for the solution of the inner problem $\widehat{u}_B$

UniGe | MaLGa

Sparsity promotion and ℓ^1 - assumptions

- ⊖ No explicit formula for the solution of the inner problem $\widehat{u}_B$
- ⊖ No characterization of the optimal choice B

Sparsity promotion and ℓ^1 - assumptions

- ⊖ No explicit formula for the solution of the inner problem $\widehat{u}_B$
- ⊖ No characterization of the optimal choice B
- ⊖ No straightforward unsupervised approach

Sparsity promotion and ℓ^1 - assumptions

- ⊖ No explicit formula for the solution of the inner problem $\widehat{u}_B$
- ⊖ No characterization of the optimal choice B
- ⊖ No straightforward unsupervised approach
- ⊕ Supervised approach: extend the Tikhonov approach, based on **stability** + **covering**

Sparsity promotion and ℓ^1 - assumptions

- ⊖ No explicit formula for the solution of the inner problem $\widehat{u}_B$
- ⊖ No characterization of the optimal choice B
- ⊖ No straightforward unsupervised approach
- ⊕ Supervised approach: extend the Tikhonov approach, based on **stability** + **covering**

Our assumptions

a) $A\colon X \to Y$ is bounded and compact

b) Enriched compatibility: $\mathrm{Im}(A) \subset \mathrm{Im}(\Sigma_\varepsilon)$ and $\Sigma_\varepsilon^{-1} A$ is compact

c) x, ε sub-Gaussian random variables

d) minimize over a compact set

$$\mathcal{B} \subseteq \mathcal{B}_{\mathrm{adm}} := \{B\colon \ell^2 \to X \text{ bdd} : AB \text{ satisfies the finite basis injectivity (FBI)}\}$$

ℓ^1 regularization - theoretical results[7]

What we are able to prove under these assumptions:

- for every $B \in \mathcal{B}$, there exist a minimizer $\widehat{u}_B = R_B(y)$

[7] Learning sparsity-promoting regularizers for linear inverse problems, preprint, 2024

ℓ^1 regularization - theoretical results[7]

What we are able to prove under these assumptions:

- for every $B \in \mathcal{B}$, there exist a minimizer $\widehat{u}_B = R_B(y)$
- Hölder stability with respect to B:

$$\|R_{B_1}(y) - R_{B_2}(y)\|_{\ell^2} \leq c\|B_1 - B_2\|^{1/2}, \qquad B_1, B_2 \in \mathcal{B}$$

[7] Learning sparsity-promoting regularizers for linear inverse problems, preprint, 2024

ℓ^1 regularization - theoretical results[7]

What we are able to prove under these assumptions:

- for every $B \in \mathcal{B}$, there exist a minimizer $\widehat{u}_B = R_B(y)$
- Hölder stability with respect to B:

$$\|R_{B_1}(y) - R_{B_2}(y)\|_{\ell^2} \leq c\|B_1 - B_2\|^{1/2}, \qquad B_1, B_2 \in \mathcal{B}$$

- Generalization estimates:

$$|L(\widehat{B}_S) - L(B^\star)| \leq \left(\frac{c_1 + c_2\sqrt{\tau}}{\sqrt{m}}\right)^{1-\frac{1}{s+1}},$$

where s measures the compactness of $\mathcal{B}$ via covering numbers

$$\log(\mathcal{N}(\mathcal{B}, r)) \lesssim r^{-1/s}$$

UniGe | MaLGa

[7] Learning sparsity-promoting regularizers for linear inverse problems, preprint, 2024

Examples of classes $\mathcal{B}$

- compact perturbation of a reference operator

$$\mathcal{B} = \{B_0(\mathrm{Id} + K) : K \in \mathcal{H}\},$$

being $\mathcal{H}$ a compact set of compact operators

Examples of classes $\mathcal{B}$

- compact perturbation of a reference operator

$$\mathcal{B} = \{B_0(\mathrm{Id} + K) : K \in \mathcal{H}\},$$

  being $\mathcal{H}$ a compact set of compact operators

- learning the mother wavelet:

$$\mathcal{B} = \{B_\phi : \phi \in \Phi\}$$

  where Φ is a compact class of mother wavelets

In both cases, it is possible to quantify compactness via covering numbers

Outline

Learning the optimal generalized Tikhonov regularizer

Learning the optimal ℓ^1 regularizer

Sparse regularization via Gaussian mixtures

Alternative approach to sparsity promotion: Gaussian mixture prior

Motivation

Generalized Tikhonov $\leftrightsquigarrow$ (Linear) MMSE estimator $\leftrightsquigarrow$ x, ε Gaussians

[8] Learning a Gaussian Mixture for Sparsity Regularization in Inverse Problems, arXiv:2401.16612
see also: [Bocchinfuso, Calvetti, Somersalo 2023]

Alternative approach to sparsity promotion: Gaussian mixture prior

Motivation

Generalized Tikhonov $\leftrightsquigarrow$ (Linear) MMSE estimator $\leftrightsquigarrow$ x, ε Gaussians

Goal: statistical model for sparse signals such that the MMSE/Bayes estimator can be computed

[8] Learning a Gaussian Mixture for Sparsity Regularization in Inverse Problems, arXiv:2401.16612
see also: [Bocchinfuso, Calvetti, Somersalo 2023]

Motivation

Generalized Tikhonov $\leftrightsquigarrow$ (Linear) MMSE estimator $\leftrightsquigarrow$ x, ε Gaussians

Goal: statistical model for sparse signals such that the MMSE/Bayes estimator can be computed

Our model for (group) sparsity[8]: degenerate Gaussian mixtures in $\mathbb{R}^n$

$$X = \sum_{i=1}^{L} X_i \mathbb{1}_{\{i\}}(I), \quad X_i \sim \mathcal{N}(\mu_i, \Sigma_i), \quad \operatorname{rank}(\Sigma_i) \leq s \ll n$$

- s sparsity
- I random variable on $\{1, \dots, L\}$
- $w_i := \mathbb{P}(I = i)$ *weights of the mixture*

[8] Learning a Gaussian Mixture for Sparsity Regularization in Inverse Problems, arXiv:2401.16612
see also: [Bocchinfuso, Calvetti, Somersalo 2023]

MMSE/Bayes estimator for Gaussian mixtures and linear observations

$$X = \sum_{i=1}^{L} X_i \mathbb{1}_{\{i\}}(I), \quad X_i \sim \mathcal{N}(\mu_i, \Sigma_i), \quad \mathrm{rank}(\Sigma_i) \leq s \ll n$$

Lemma[9]

Let $E \sim \mathcal{N}(0, \Sigma_E)$ be independent of X_i and I. The Bayes estimator of $Y = AX + E$ is

$$R^\star(y) = \mathbb{E}[X|Y = y] = \sum_{i=1}^{L} \frac{c_i}{\sum_{j=1}^{L} c_j} (\mu_i + \Sigma_i A^T (A\Sigma_i A^T + \Sigma_E)^{-1}(y - A\mu_i)), \tag{1}$$

where

$$c_i = \frac{w_i}{\sqrt{|A\Sigma_i A^T + \Sigma_E|}} \exp\left(-\frac{1}{2} \|(A\Sigma_i A^T + \Sigma_E)^{-\frac{1}{2}}(y - A\mu_i)\|_2^2 \right) \tag{2}$$

[9] Kundu, Chatterjee, Murthy, Sreenivas, 2008

MMSE/Bayes estimator for Gaussian mixtures and linear observations

$$X = \sum_{i=1}^{L} X_i \mathbb{1}_{\{i\}}(I), \quad X_i \sim \mathcal{N}(\mu_i, \Sigma_i), \quad \mathrm{rank}(\Sigma_i) \leq s \ll n$$

Lemma[9]

Let $E \sim \mathcal{N}(0, \Sigma_E)$ be independent of X_i and I. The Bayes estimator of $Y = AX + E$ is

$$R^\star(y) = \mathbb{E}[X|Y=y] = \sum_{i=1}^{L} \frac{c_i}{\sum_{j=1}^{L} c_j}(\mu_i + \Sigma_i A^T (A\Sigma_i A^T + \Sigma_E)^{-1}(y - A\mu_i)), \tag{1}$$

where

$$c_i = \frac{w_i}{\sqrt{|A\Sigma_i A^T + \Sigma_E|}} \exp\Big(-\frac{1}{2}\|(A\Sigma_i A^T + \Sigma_E)^{-\frac{1}{2}}(y - A\mu_i)\|_2^2\Big) \tag{2}$$

Useful parametrization:

$$R^*(y) = R_\theta(y), \qquad \theta = \big(\{w_i\}_{i=1}^L, \{\mu_i\}_{i=1}^L, \{\Sigma_i\}_{i=1}^L\big)$$

[9] Kundu, Chatterjee, Murthy, Sreenivas, 2008

The Bayes estimator is a neural network

Proposition[10]

We have that

$$R_\theta(y) = \sum_{i=1}^{L} \mathrm{softmax}(f(y))_i \, g_i(y), \qquad \theta = \big(\{w_i\}_i, \{\mu_i\}_i, \{\Sigma_i\}_i\big)$$

where

$$f_i(y) = b(w_i, \Sigma_i) - \frac{1}{2}\|(A\Sigma_i A^T + \Sigma_E)^{-\frac{1}{2}}(y - A\mu_i)\|_2^2 \qquad \text{(quadratic)}$$

$$g_i(y) = \mu_i + \Sigma_i A^T (A\Sigma_i A^T + \Sigma_E)^{-1}(y - A\mu_i) \qquad \text{(affine)}$$

UniGe | MaLGa

[10] A, Ratti, Santacesaria, Sciutto, Learning a Gaussian Mixture for Sparsity Regularization in Inverse Problems, 2024

The Bayes estimator is a neural network

Proposition[10]

We have that

$$R_\theta(y) = \sum_{i=1}^{L} \mathrm{softmax}(f(y))_i \, g_i(y), \qquad \theta = \big(\{w_i\}_i, \{\mu_i\}_i, \{\Sigma_i\}_i\big)$$

where

$$f_i(y) = b(w_i, \Sigma_i) - \frac{1}{2}\|(A\Sigma_i A^T + \Sigma_E)^{-\frac{1}{2}}(y - A\mu_i)\|_2^2 \qquad \text{(quadratic)}$$

$$g_i(y) = \mu_i + \Sigma_i A^T (A\Sigma_i A^T + \Sigma_E)^{-1}(y - A\mu_i) \qquad \text{(affine)}$$

$\longrightarrow$ similar to the **attention mechanism** of transformers

UniGe | MaLGa

[10] A, Ratti, Santacesaria, Sciutto, Learning a Gaussian Mixture for Sparsity Regularization in Inverse Problems, 2024

The Bayes estimator is a neural network

Proposition [10]

We have that

$$R_\theta(y) = \sum_{i=1}^{L} \mathrm{softmax}(f(y))_i \, g_i(y), \qquad \theta = \big(\{w_i\}_i, \{\mu_i\}_i, \{\Sigma_i\}_i\big)$$

where

$$f_i(y) = b(w_i, \Sigma_i) - \frac{1}{2}\|(A\Sigma_i A^T + \Sigma_E)^{-\frac{1}{2}}(y - A\mu_i)\|_2^2 \qquad \text{(quadratic)}$$

$$g_i(y) = \mu_i + \Sigma_i A^T (A\Sigma_i A^T + \Sigma_E)^{-1}(y - A\mu_i) \qquad \text{(affine)}$$

$\longrightarrow$ similar to the **attention mechanism** of transformers

Two training approaches:

[10] A, Ratti, Santacesaria, Sciutto, Learning a Gaussian Mixture for Sparsity Regularization in Inverse Problems, 2024

The Bayes estimator is a neural network

Proposition[10]

We have that

$$R_\theta(y) = \sum_{i=1}^{L} \mathrm{softmax}(f(y))_i \, g_i(y), \qquad \theta = \big(\{w_i\}_i, \{\mu_i\}_i, \{\Sigma_i\}_i\big)$$

where

$$f_i(y) = b(w_i, \Sigma_i) - \frac{1}{2}\|(A\Sigma_i A^T + \Sigma_E)^{-\frac{1}{2}}(y - A\mu_i)\|_2^2 \qquad \text{(quadratic)}$$

$$g_i(y) = \mu_i + \Sigma_i A^T (A\Sigma_i A^T + \Sigma_E)^{-1}(y - A\mu_i) \qquad \text{(affine)}$$

$\longrightarrow$ similar to the **attention mechanism** of transformers

Two training approaches:

1. supervised: minimize

$$\widehat{L}(\theta) = \frac{1}{N}\sum_{j=1}^{N} \|x_j - R_\theta(y_j)\|_2^2,$$

[10] A, Ratti, Santacesaria, Sciutto, Learning a Gaussian Mixture for Sparsity Regularization in Inverse Problems, 2024

The Bayes estimator is a neural network

Proposition[10]

We have that

$$R_\theta(y) = \sum_{i=1}^{L} \mathrm{softmax}(f(y))_i \, g_i(y), \qquad \theta = \big(\{w_i\}_i, \{\mu_i\}_i, \{\Sigma_i\}_i\big)$$

where

$$f_i(y) = b(w_i, \Sigma_i) - \frac{1}{2}\|(A\Sigma_i A^T + \Sigma_E)^{-\frac{1}{2}}(y - A\mu_i)\|_2^2 \qquad \text{(quadratic)}$$

$$g_i(y) = \mu_i + \Sigma_i A^T (A\Sigma_i A^T + \Sigma_E)^{-1}(y - A\mu_i) \qquad \text{(affine)}$$

$\longrightarrow$ similar to the **attention mechanism** of transformers

Two training approaches:

1. supervised: minimize

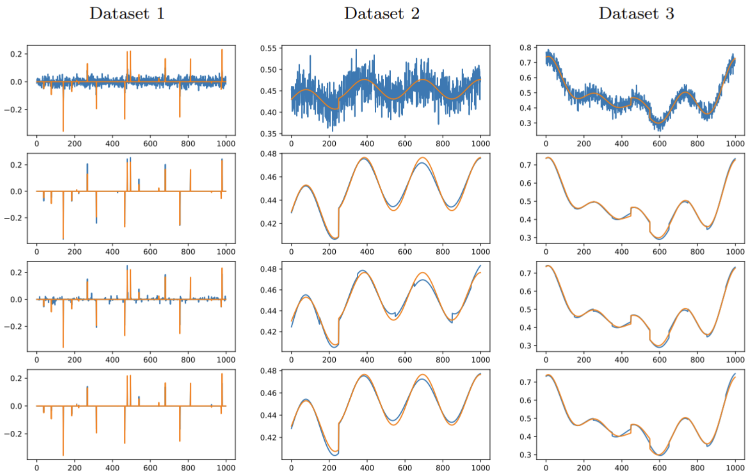
$$\widehat{L}(\theta) = \frac{1}{N}\sum_{j=1}^{N} \|x_j - R_\theta(y_j)\|_2^2,$$

2. unsupervised: approximate w_i, μ_i and Σ_i from $\{x_j\}$

[10] A, Ratti, Santacesaria, Sciutto, Learning a Gaussian Mixture for Sparsity Regularization in Inverse Problems, 2024

Numerical experiments: deblurring with 10% noise

Rows: Data, Unsupervised approach, dictionary learning, group dictionary learning



Numerical experiments: deblurring with 10% noise

Table: Relative MSE values

	Dataset 1	Dataset 2	Dataset 3
Unsupervised	$\mathbf{3.68}\%$	$\mathbf{2.65\ 10^{-3}}\%$	$\mathbf{1.01\ 10^{-2}}\%$
Dictionary learning	14.32%	$6.61\ 10^{-3}\%$	$1.28\ 10^{-2}\%$
Group dictionary learning	13.51%	$4.62\ 10^{-3}\%$	$3.41\ 10^{-2}\%$

Also experiments with denoising and comparisons with Lasso, Group Lasso and iterative hard thresholding

Conclusions

Learning (simple) regularizers for inverse problems:
generalized Tikhonov and sparsity promoting regularization

Infinite-dimensional framework:
discretization-independent results for the learning problem

Gaussian mixtures as model for (group) sparsity:
a non-iterative and learnable approach to sparse optimization

Supervised and **unsupervised** techniques:
comparable theoretical guarantees and numerical effectiveness

Conclusions

Learning (simple) regularizers for inverse problems:
generalized Tikhonov and sparsity promoting regularization

Infinite-dimensional framework:
discretization-independent results for the learning problem

Gaussian mixtures as model for (group) sparsity:
a non-iterative and learnable approach to sparse optimization

Supervised and **unsupervised** techniques:
comparable theoretical guarantees and numerical effectiveness

Further extensions:

1. careful study of the connection between sparsity promotion and the attention mechanism
2. more complex regularization terms & nonlinear inverse problems

Slides