

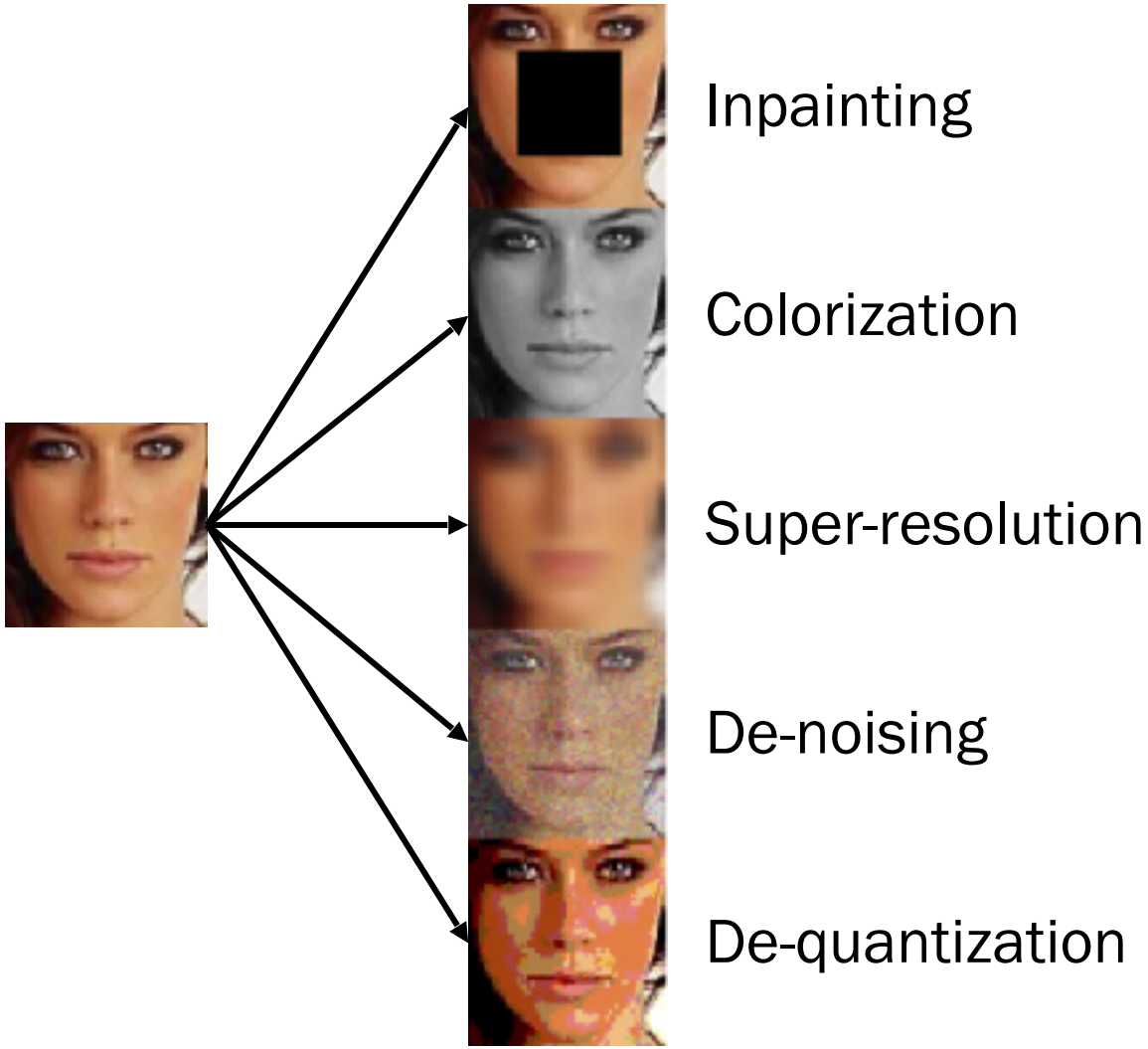
From Statistical Estimation and Generative Modeling to Inverse Problems

LECTURE 3: STATISTICAL ESTIMATION

Prologue

In our three lectures we focus on the following subjects:

- Statistical Estimation methods (Parameter Estimation)
  Measure X but would like to know (estimate) Z
- Random data are described by probability densities
  Not convenient for modern datasets: Usually multi-dimensional but recide on low dimensional surfaces (manifolds)
  Alternative description with Generative Models
- Modification of Statistical Estimation methods to accept generative models in place of probability densities
  Application to inverse problems (mostly from Computer Vision)



Inpainting

Colorization

Super-resolution

De-noising

De-quantization

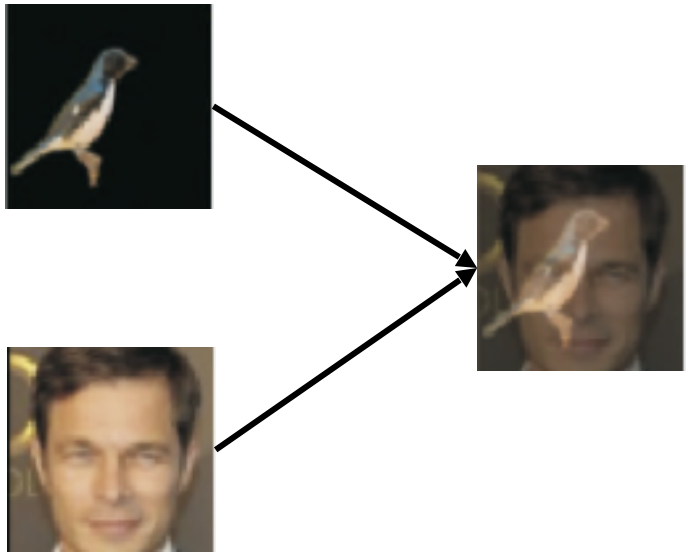


Image Separation

Statistical Estimation

- Problem Definition
- Bayesian Estimation
  - Optimum solution for general cost function
  - Solution for the MSE, MAE, AP
- Non-Bayesian Estimation
  - The ML estimator
  - Asymptotic optimality

Problem Definition

We are given vector X (measurements) but interested in vector Z !
Can we estimate Z from X ?

Z and X must be related otherwise impossible

The relationship can be “soft” (not deterministic) and expressed with a joint probability density f(X,Z)

So Z and X are random and have a random relationship captured by f(X,Z)

The density f(X,Z) also expresses prior knowledge about each individual random quantity Z and X

Prior knowledge about individual X and Z is captured by the marginal densities $\mathsf{g}(X)$ and $\mathsf{h}(Z)$ where

$$\mathsf{g}(X) = \int \mathsf{f}(X, Z)\, dZ, \quad \mathsf{h}(Z) = \int \mathsf{f}(X, Z)\, dX$$

When measurements X become available, knowledge about Z changes to

$$\mathsf{f}(Z|X) = \frac{\mathsf{f}(X, Z)}{\mathsf{g}(X)} = \frac{\mathsf{f}(X, Z)}{\int \mathsf{f}(X, Z)\, dZ}$$

If X and Z independent then $\mathsf{f}(X,Z)=\mathsf{g}(X)\mathsf{h}(Z)$ and

$$\mathsf{f}(Z|X) = \frac{\mathsf{f}(X, Z)}{\mathsf{g}(X)} = \frac{\mathsf{g}(X)\mathsf{h}(Z)}{\mathsf{g}(X)} = \mathsf{h}(Z)$$

X has no information about Z

What is an estimator ?

Given f(X, Z) and measurements X suppose we select "estimate" $\hat{Z}$

If measurements X do not change no reason to change $\hat{Z}$

Estimate $\hat{Z} = \hat{Z}(X)$ is any deterministic function of X

There are "good" estimators and there are "bad" estimators

Statistical Estimation theory identifies the "best"

Bayesian Estimation

There is cost function $\mathsf{C}(\cdot,\cdot)$ such that if
true value is Z and estimate $\hat{Z}$ then cost is $\mathsf{C}(\hat{Z}, Z)$

Interested in average cost

$$\mathscr{C}(\hat{Z}) = \mathbb{E}_{X,Z}\big[\mathsf{C}\big(\hat{Z}(X), Z\big)\big] = \iint \mathsf{C}\big(\hat{Z}(X), Z\big)\mathsf{f}(X, Z)\, dX\, dZ$$

Depends on the estimator function $\hat{Z}(X)$

We like to minimize it

$$\min_{\hat{Z}} \mathscr{C}(\hat{Z}) = \min_{\hat{Z}} \iint \mathsf{C}\big(\hat{Z}(X), Z\big)\mathsf{f}(X, Z)\, dX\, dZ$$

Prior knowledge about X and Z

$$\mathsf{g}(X) = \int \mathsf{f}(X, Z)\, dZ, \quad \mathsf{h}(Z) = \int \mathsf{f}(X, Z)\, dX$$

After we take measurements X our knowledge about Z becomes

$$\mathsf{f}(Z|X) = \frac{\mathsf{f}(X, Z)}{\mathsf{g}(X)} = \frac{\mathsf{f}(X, Z)}{\int \mathsf{f}(X, Z)\, dZ}$$

Posterior probability density

and we can write $\mathsf{f}(X, Z) = \mathsf{f}(Z|X)\mathsf{g}(X)$

$$\mathscr{C}(\hat{Z}) = \int \left(\int \mathsf{C}\big(\hat{Z}(X), Z\big)\mathsf{f}(Z|X)\, dZ \right) \mathsf{g}(X)\, dX$$

call $\mathsf{G}\big(\hat{Z}(X), X\big)$

$$\mathsf{G}(U,X) = \int \mathsf{C}(U,Z)\mathsf{f}(Z|X)\,dZ$$

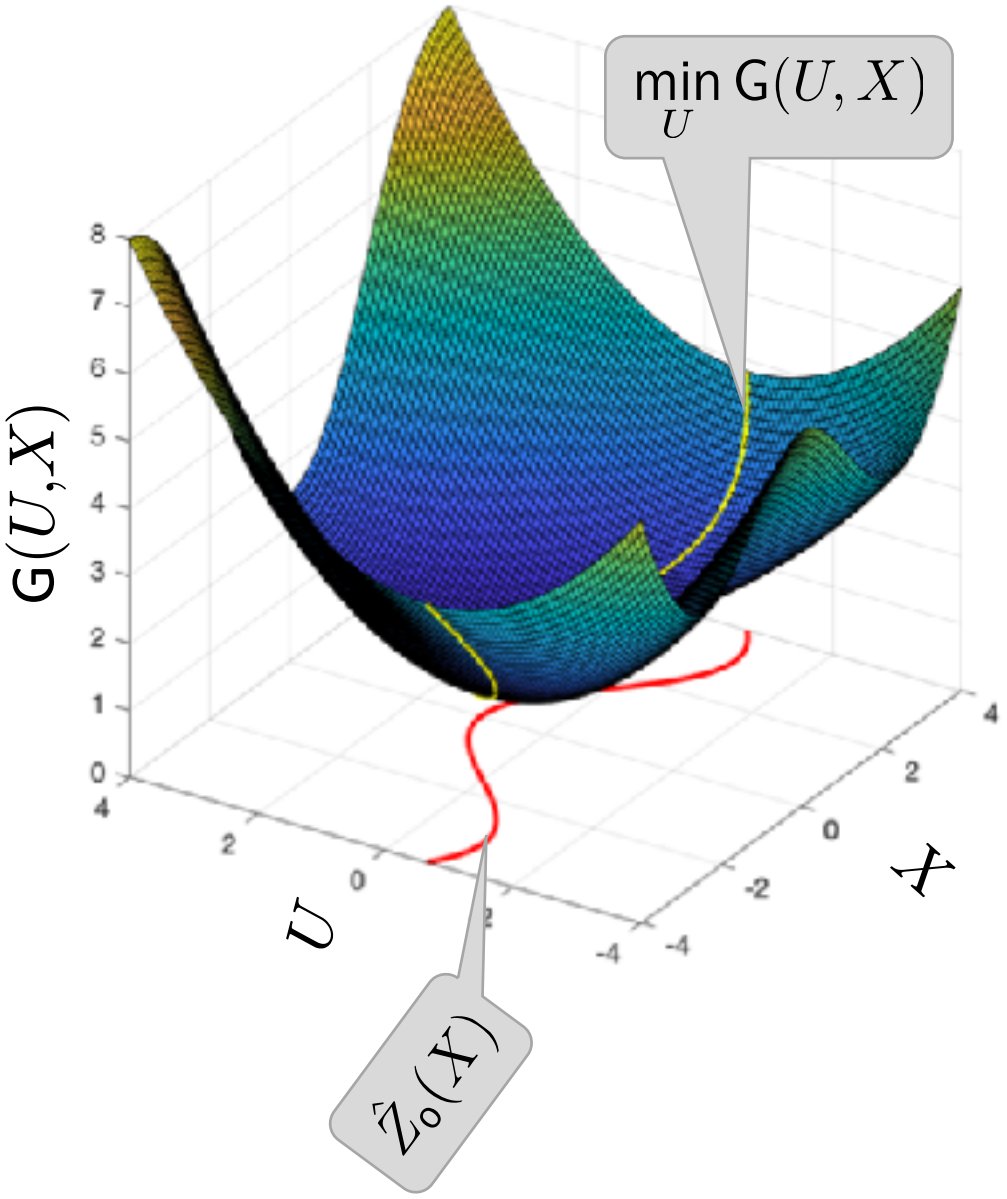
$$\mathscr{C}(\hat{Z}) = \int \mathsf{G}(\hat{Z}(X),X)\mathsf{g}(X)\,dX$$

For each X perform minimization

$$\min_U \mathsf{G}(U,X) = \Phi(X)$$

$$\hat{Z}_{\mathsf{o}}(X) = \arg\min_U \mathsf{G}(U,X)$$

$$\mathsf{G}(\hat{Z}_{\mathsf{o}}(X),X) = \min_U \mathsf{G}(U,X) = \Phi(X)$$



$$\min_U \mathsf{G}(U,X) = \Phi(X) \;\Rightarrow\; \mathsf{G}(??,X) \geq \Phi(X) \;\Rightarrow\; \mathsf{G}\big(\hat{Z}(X),X\big) \geq \Phi(X)$$

$$\mathscr{C}(\hat{Z}) = \int \mathsf{G}\big(\hat{Z}(X),X\big)\mathsf{g}(X)\,dX \;\geq\; \int \Phi(X)\mathsf{g}(X)\,dX$$

Lower bound

If estimator attains lower bound then optimum

Consider $\hat{Z}_\mathsf{o}(X) = \arg\min_U \mathsf{G}(U,X)$

$$\mathscr{C}(\hat{Z}_\mathsf{o}) = \int \mathsf{G}\big(\hat{Z}_\mathsf{o}(X),X\big)\mathsf{g}(X)\,dX \;=\; \int \Phi(X)\mathsf{g}(X)\,dX$$

$$= \Phi(X)$$

$\hat{Z}_\mathsf{o}(X) = \arg\min_U \mathsf{G}(U,X)$ is optimum

Examples

Minimum Mean Square Error (MMSE): $\mathsf{C}(\hat{Z}, Z) = \|\hat{Z} - Z\|^2$

$$\mathsf{G}(U, X) = \int \mathsf{C}(U, Z)\mathsf{f}(Z|X)\, dZ = \int \|U - Z\|^2 \mathsf{f}(Z|X)\, dZ$$

Must compute $\min_U \mathsf{G}(U, X)$

$$\nabla_U \mathsf{G}(U, X) = 0 \;\Rightarrow\; \int \nabla_U \|U - Z\|^2 \mathsf{f}(Z|X)\, dZ = 0$$

$$\int 2(U - Z)\mathsf{f}(Z|X)\, dZ = 0 \;\Rightarrow\; U = \int Z \mathsf{f}(Z|X)\, dZ$$

$$\hat{Z}_{\mathsf{MMSE}}(X) = \mathbb{E}[Z|X] = \int Z\, \mathsf{f}(Z|X)\, dZ = \frac{\int Z\, \mathsf{f}(X, Z)\, dZ}{\int \mathsf{f}(X, Z)\, dZ}$$

Minimum Mean Absolute Error (MMAE): $\mathsf{C}(\hat{Z}, Z) = |\hat{z}_1 - z_1| + \cdots + |\hat{z}_L - z_L|$

$$\hat{Z} = [\hat{z}_1 \cdots \hat{z}_L]^\intercal, \qquad Z = [z_1 \cdots z_L]^\intercal$$

Treat each coordinate separately

Minimize with respect to u : $\mathsf{G}(u, X) = \int |u - z|\, \mathsf{f}(z|X)\, dz$

$$\frac{\partial \mathsf{G}(u, X)}{\partial u} = 0 \;\Rightarrow\; \int \frac{\partial |u - z|}{\partial u}\, \mathsf{f}(z|X)\, dz = \int \mathsf{sign}(u - z)\, \mathsf{f}(z|X)\, dz = 0$$

$$\Rightarrow \int_{-\infty}^{u} \mathsf{f}(z|X)\, dz = \int_{u}^{\infty} \mathsf{f}(z|X)\, dz \;\Rightarrow\; 2 \int_{-\infty}^{u} \mathsf{f}(z|X)\, dz = 1$$

$$\int_{-\infty}^{\hat{z}_{\mathsf{MAE}}} \mathsf{f}(z|X)\, dz = \frac{1}{2},$$ Conditional Median

Maximum Aposteriory Probability (MAP):

$$\mathsf{C}(\hat{Z}, Z) = \begin{cases} 1 & \text{when } \|\hat{Z} - Z\| > \delta \\ 0 & \text{when } \|\hat{Z} - Z\| \leq \delta \end{cases} \quad \delta \to 0$$

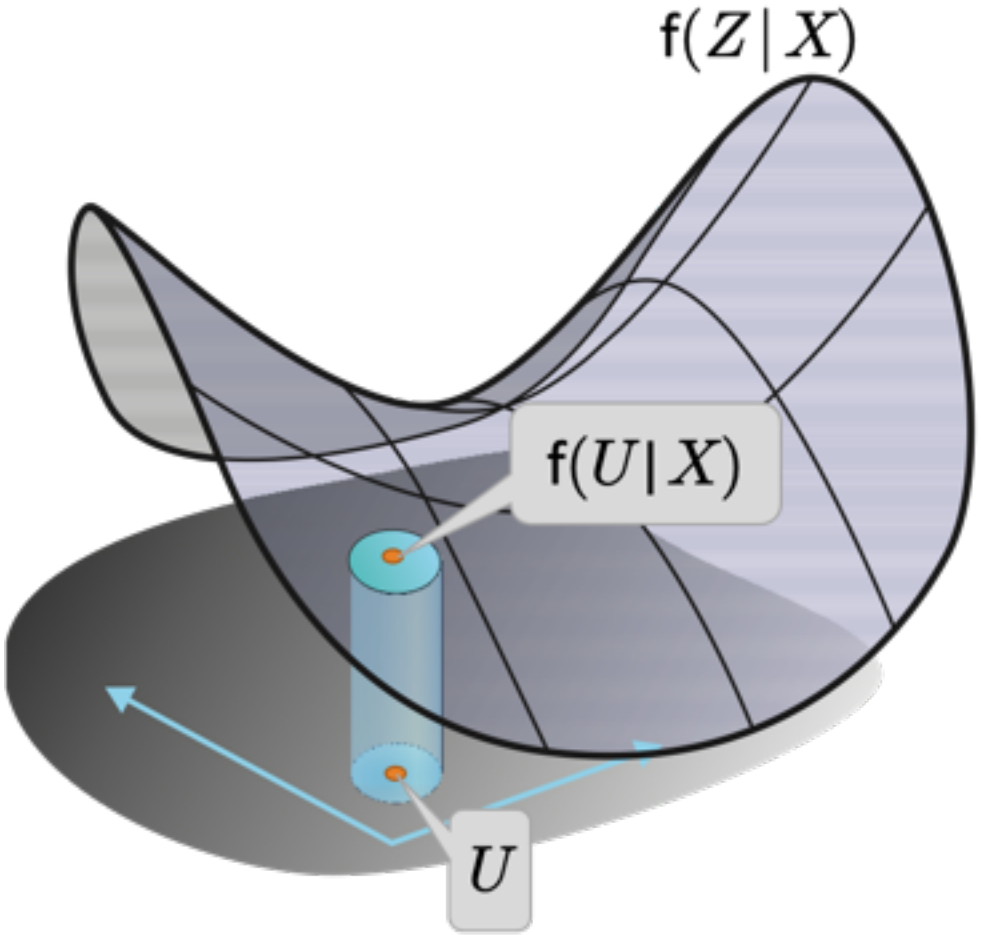
$$\mathsf{G}(U, X) = \int \mathsf{C}(U, Z)\mathsf{f}(Z|X)\, dZ$$

$$= \int_{\|U-Z\|>\delta} \mathsf{f}(Z|X)\, dZ$$

$$= 1 - \int_{\|U-Z\|\leq\delta} \mathsf{f}(Z|X)\, dZ \approx 1 - \mathsf{Ball}(\delta)\mathsf{f}(U|X)$$

$$\min_U \mathsf{G}(U, X) \approx 1 - \mathsf{Ball}(\delta) \max_U \mathsf{f}(U|X)$$

$$\hat{Z}_{\mathsf{MAP}}(X) = \arg\max_Z \mathsf{f}(Z|X)$$



Non-Bayesian Estimation

Measurements X and desired Z assumed to be "weakly" related through joint probability density f(X,Z) considered known

From f(X, Z) computed posterior probability density

$$\mathrm{f}(Z|X) = \frac{\mathrm{f}(X,Z)}{\mathrm{g}(X)} = \frac{\mathrm{f}(X,Z)}{\int \mathrm{f}(X,Z)\, dZ}$$

In many applications no access to Z translates in no availability of f(X,Z) and f$(Z|X)$

It is possible in some applications to know the conditional density f$(X|Z)$



X Z W

$$X = Z + W$$

Conditional density $\mathsf{f}(X|Z)$ requires only $\mathsf{g}_w(W)$

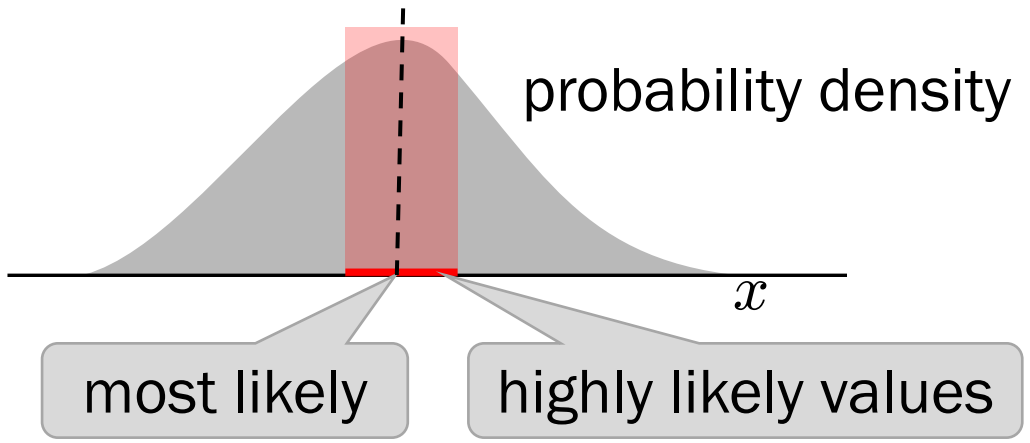
$$\mathsf{f}(X|Z) = \mathsf{g}_w(X - Z)$$

Joint density $\mathsf{f}(X,Z)$ requires $\mathsf{g}_w(W)$ and $\mathsf{h}(Z)$

$$\mathsf{f}(X, Z) = \mathsf{f}(X|Z)\mathsf{h}(Z) = \mathsf{g}_w(X - Z)\mathsf{h}(Z)$$

Knowing noise density $\mathsf{g}_w(W)$ possible. Knowing $\mathsf{h}(Z)$ difficult !

Maximum Likelihood Estimator (MLE):



Density $\mathrm{f}(X \mid Z)$ is likelihood of X (given Z).

For measurements X what is Z that makes X most likely ?

We must solve $\max_Z \mathrm{f}(X|Z)$

$$\hat{Z}_{\mathrm{MLE}} = \arg\max_Z \mathrm{f}(X|Z)$$

$$\hat{Z}_{\text{MAP}} = \arg\max_Z \mathsf{f}(Z|X) = \arg\max_Z \frac{\mathsf{f}(X,Z)}{\mathsf{g}(X)} = \arg\max_Z \mathsf{f}(X,Z)$$

$$= \arg\max_Z \mathsf{f}(X|Z)\mathsf{h}(Z)$$

if $\mathsf{h}(Z)$ constant (degenerate uniform)

$$\hat{Z}_{\text{MLE}} = \arg\max_Z \mathsf{f}(X|Z)$$

Optimality of MLE

Unbiased Estimator $\quad \mathbb{E}_X\left[\hat{Z}(X)|Z\right] = Z$

Interested in Error covariance matrix: $\mathbb{E}_X\left[\left(\hat{Z}(X) - Z\right)\left(\hat{Z}(X) - Z\right)^{\mathsf{T}}|Z\right]$

Cramer-Rao Lower Bound (CRLB)

From unbiased property

$$0 = \mathbb{E}_X\Big[\hat{Z}(X) - Z|Z\Big] = \int \Big(\hat{Z}(X) - Z\Big)\mathsf{f}(X|Z)\, dX$$

Derivative (Jacobian) with respect to Z

$$0 = \int \Big\{ - I\mathsf{f}(X|Z) + \Big(\hat{Z}(X) - Z\Big)\Big(\nabla_Z \mathsf{f}(X|Z)\Big)^\intercal\Big\}\, dX$$

$$I = \int \Big\{\Big(\hat{Z}(X) - Z\Big)\Big(\frac{\nabla_Z \mathsf{f}(X|Z)}{\mathsf{f}(X|Z)}\Big)^\intercal\Big\}\mathsf{f}(X|Z)\, dX$$

$$= \mathbb{E}_X\left[\Big(\hat{Z}(X) - Z\Big)\Big(\frac{\nabla_Z \mathsf{f}(X|Z)}{\mathsf{f}(X|Z)}\Big)^\intercal\Big|Z\right]$$

Define

$$\mathcal{E}(X) = \hat{Z}(X) - Z, \qquad \Delta(X) = \frac{\nabla_Z \mathsf{f}(X|Z)}{\mathsf{f}(X|Z)}$$

Compute covariance matrix

$$\mathbb{E}_X\left[\begin{bmatrix} \mathcal{E}(X) \\ \Delta(X) \end{bmatrix} [\mathcal{E}^\intercal(X)\ \Delta^\intercal(X)] \middle| Z\right] = \begin{bmatrix} \mathbb{E}_X[\mathcal{E}(X)\mathcal{E}^\intercal(X)|Z] & \mathbb{E}_X[\mathcal{E}(X)\Delta^\intercal(X)|Z] \\ \mathbb{E}_X[\Delta(X)\mathcal{E}^\intercal(X)|Z] & \mathbb{E}_X[\Delta(X)\Delta^\intercal(X)|Z] \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{E}_X[\mathcal{E}(X)\mathcal{E}^\intercal(X)|Z] & I \\ I & \mathbb{E}_X[\Delta(X)\Delta^\intercal(X)|Z] \end{bmatrix}$$

$$\mathbb{E}_X[\mathcal{E}(X)\mathcal{E}^\intercal(X)|Z] \geq (\mathsf{FI})^{-1}$$

Fisher Information matrix

$$\mathsf{FI} = \mathbb{E}_X[\Delta(X)\Delta^\intercal(X)|Z] = \mathbb{E}_X\left[\left(\frac{\nabla_Z \mathsf{f}(X|Z)}{\mathsf{f}(X|Z)}\right)\left(\frac{\nabla_Z \mathsf{f}(X|Z)}{\mathsf{f}(X|Z)}\right)^\intercal \middle| Z\right]$$

Estimation error covariance matrix of any unbiased estimator is lower bounded by inverse of Fisher Information matrix

$$\text{CRLB} = (\text{FI})^{-1}$$

Estimation error power of any estimator cannot go below a certain level

Call n the size of measurement data X

Theorem: Under general conditions we have

$$\lim_{n\to\infty} \mathbb{E}_X\left[\left(\hat{Z}_{\text{MLE}}(X) - Z\right)\left(\hat{Z}_{\text{MLE}}(X) - Z\right)^{\intercal} | Z\right] \times \text{FI} = I$$

Asymptotic optimality (for large data size)

Summary

Minimum Mean Square Error: $\hat{Z}_{\text{MMSE}}(X) = \mathbb{E}[Z|X]$

Minimum Mean Absolute Error: $\hat{Z}_{\text{MMAE}}(X) = \arg\left\{ u : \int_{-\infty}^{u} \mathrm{f}(z|X)\, dz = \frac{1}{2} \right\}$

Maximum Aposteriori Probability: $\hat{Z}_{\text{MAP}}(X) = \arg\max_{Z} \mathrm{f}(Z|X)$

Optimum for any measurement datasize.

Maximum Likelihood Estimator: $\hat{Z}_{\text{MLE}}(X) = \arg\max_{Z} \mathrm{f}(X|Z)$

Asymptotically optimum, “large” measurement datasize.

Special Case

Let $Z=\{Z_1,Z_2\}$ and there is prior for Z_1 but not for Z_2

Treat non-existing priors as degenerate uniforms starting with MAP estimator

$$\begin{aligned}
\arg\max_{Z_1,Z_2} \mathsf{f}(Z_1,Z_2|X) &= \arg\max_{Z_1,Z_2} \mathsf{f}(X,Z_1,Z_2) \\
&= \arg\max_{Z_1,Z_2} \mathsf{f}(X|Z_1,Z_2)\mathsf{h}(Z_1,Z_2) \\
&= \arg\max_{Z_1,Z_2} \mathsf{f}(X|Z_1,Z_2)\mathsf{h}_1(Z_1|Z_2)\mathsf{h}_2(Z_2) \\
\mathsf{h}_2(Z_2) \text{ degenerate uniform} \quad &= \arg\max_{Z_1,Z_2} \mathsf{f}(X|Z_1,Z_2)\mathsf{h}_1(Z_1|Z_2)
\end{aligned}$$

If Z_1 and Z_2 independent and interested only in estimating Z_1

$$\hat{Z}_1 = \arg\max_{Z_1} \left\{ \left(\max_{Z_2} \mathsf{f}(X|Z_1,Z_2) \right) \mathsf{h}_1(Z_1) \right\}$$



From Statistical Estimation and Generative Modeling to Inverse Problems

LECTURE 4: GENERATIVE MODELING

Generative Modeling

- Realizations of random variables
- Classical Methods
  - Inverse cdf
  - Acceptance-rejection
- Using transformations
  - Generative models
  - Design with adversarial methods
  - Design with non-adversarial methods
- Probability density vs Generative model

Realizations of Random Variables

Realizations of random variables are needed in Monte-Carlo simulations

Generate synthetic data (images, music, videos)

When densities of interest are usual then classical methods

In case of modern datasets classical methods fail miserably in producing realizations

Need alternative techniques that can handle data that are multi-dimensional and can lie in lower dimensional manifolds

Regarding random variables that lie on lower dimensional manifolds description using probability densities is not efficient

Classical Methods

For given density f(x) we like to generate realizations

Assume we have probability density h(z) for which we can easily generate realizations, for example Uniform or Gaussian

Inverse CDF

If z_1,z_2,... independent realizations, uniform in [0,1], apply transformation

$$x_i = \mathrm{F}^{-1}(z_i)$$

where $\mathrm{F}(\mathrm{x}) = \int_{-\infty}^{\mathrm{x}} \mathrm{f}(w)\, dw$ is the cumulative distribution function (cdf)

and $\mathrm{F}^{-1}(\mathrm{z})$ its inverse function ($\mathrm{F}^{-1}(\mathrm{z}) = \mathrm{x} \Leftrightarrow \mathrm{z} = \mathrm{F}(\mathrm{x})$)

x_i follows f(x)

$$\mathbb{P}(x \leq \mathsf{x}) = \mathbb{P}\big(\mathsf{F}^{-1}(z) \leq \mathsf{x}\big) = \mathbb{P}\big(z \leq \mathsf{F}(\mathsf{x})\big) = \mathsf{F}(\mathsf{x})$$

Can we extend this to multi-dimensional densities?

If we have joint density $\mathsf{f}(x_1,x_2,x_3)$ can we generate triplets?

Bayes Rule: $\mathsf{f}(x_1, x_2, x_3) = \mathsf{f}(x_3|x_2, x_1)\mathsf{f}(x_2|x_1)\mathsf{f}(x_1)$

Conditionally independent

Generate independent uniform realizations $z_1,\ z_2,\ z_3$

$$\begin{aligned}
\mathsf{f}(x_1) \ &\Rightarrow\ \mathsf{F}(x_1) \ \Rightarrow\ x_1 = \mathsf{F}^{-1}(z_1) \\
\mathsf{f}(x_2|x_1) \ &\Rightarrow\ \mathsf{F}(x_2|x_1) \ \Rightarrow\ x_2 = \mathsf{F}^{-1}(z_2|x_1) \\
\mathsf{f}(x_3|x_2, x_1) \ &\Rightarrow\ \mathsf{F}(x_3|x_2, x_1) \ \Rightarrow\ x_3 = \mathsf{F}^{-1}(z_3|x_2, x_1)
\end{aligned}$$

then x_1,x_2,x_3 follow $\mathsf{f}(x_1,x_2,x_3)$

Not practically convenient !!!

Acceptance/Rejection

We are given $\mathsf{f}(x)$, cannot compute $\mathsf{F}(x)$ or $\mathsf{F}^{-1}(z)$, can evaluate $\mathsf{f}(x)$

Assume another density $\mathsf{h}(z)$ for which we can generate realizations (e.g. Gaussian)

Assume we know L such that $\dfrac{\mathsf{f}(x)}{\mathsf{h}(x)} \leq \mathsf{L} < \infty$ for all x

Generate pair (z_i, t_i), independent with $z_i \sim \mathsf{h}(z)$ and $t_i \sim U([0,1])$

If $\dfrac{\mathsf{f}(z_i)}{\mathsf{h}(z_i)} \geq \mathsf{L}\, t_i$ then accept and set $x_i = z_i$

x_i follows $\mathsf{f}(x)$

Otherwise reject pair and try again with a new one

True for vector densities $\mathsf{f}(X)$ when can generate $\mathsf{h}(Z)$

Using Transformations

Generative Models

Want realizations X_i that follow $\mathsf{f}(X)$. Common method using transformations

Start with $Z \sim \mathsf{h}(Z)$, find $\mathsf{G}(Z)$ deterministic so that $X = \mathsf{G}(Z)$ follows $\mathsf{f}(X)$

Does $\mathsf{G}(Z)$ exist ?

THEOREM: Under general conditions

YES IT EXISTS !!!

Pair $\{\mathsf{G}(Z), \mathsf{h}(Z)\}$ called Generative Model

Theorem proves existence

Scalar $z \sim \mathsf{h}(z)$, strictly increasing $\mathsf{G}(z)$, then density of $x = \mathsf{G}(z)$ can be found

$$\mathbb{P}(x \leq \mathsf{x}) = \mathbb{P}(\mathsf{G}(z) \leq \mathsf{x}) = \mathbb{P}\big(z \leq \mathsf{G}^{-1}(\mathsf{x})\big) = \mathsf{H}\big(\mathsf{G}^{-1}(\mathsf{x})\big)$$

$\mathsf{z}(\mathsf{x}) = \mathsf{G}^{-1}(\mathsf{x})$ is the unique solution to $\mathsf{G}(\mathsf{z}) = \mathsf{x}$

$$\Rightarrow \ \mathsf{f}(x) = \mathsf{h}\big(\mathsf{G}^{-1}(x)\big) \times \frac{1}{\mathsf{G}'\big(\mathsf{G}^{-1}(x)\big)} = \mathsf{h}\big(\mathsf{z}(x)\big) \times \frac{1}{\big|\mathsf{G}'\big(\mathsf{z}(x)\big)\big|}$$

If $\mathsf{G}(z)$ not monotone then $\mathsf{G}(\mathsf{z}) = \mathsf{x}$ may have multiple solutions: $\mathsf{z}_1(\mathsf{x}), \ldots, \mathsf{z}_p(\mathsf{x})$

$$\mathsf{f}(x) = \mathsf{h}\big(\mathsf{z}_1(x)\big) \times \frac{1}{\big|\mathsf{G}'\big(\mathsf{z}_1(x)\big)\big|} + \cdots + \mathsf{h}\big(\mathsf{z}_p(x)\big) \times \frac{1}{\big|\mathsf{G}'\big(\mathsf{z}_p(x)\big)\big|}$$

Vector $Z \sim \mathsf{h}(Z)$ and vector transformation $\mathsf{G}(Z)$

Random vector $X = \mathsf{G}(Z)$, how to compute density?

Equation $\mathsf{G}(\mathsf{Z}) = \mathsf{X}$ has multiple solutions $\mathsf{Z}_1(\mathsf{X}), \ldots, \mathsf{Z}_p(\mathsf{X})$

$$\mathsf{f}(X) = \frac{\mathsf{h}\big(\mathsf{Z}_1(X)\big)}{\left|\det\left\{\mathbb{J}_Z \mathsf{G}\big(\mathsf{Z}_1(X)\big)\right\}\right|} + \cdots + \frac{\mathsf{h}\big(\mathsf{Z}_p(X)\big)}{\left|\det\left\{\mathbb{J}_Z \mathsf{G}\big(\mathsf{Z}_p(X)\big)\right\}\right|}$$

Target

$$\mathsf{G}(Z) = \begin{bmatrix} \mathsf{g}_1(z_1, \ldots, z_k) \\ \vdots \\ \mathsf{g}_k(z_1, \ldots, z_k) \end{bmatrix}, \quad \mathbb{J}_Z \mathsf{G}(Z) = \begin{bmatrix} \frac{\partial \mathsf{g}_1}{\partial z_1} & \cdots & \frac{\partial \mathsf{g}_1}{\partial z_k} \\ \vdots & \cdots & \vdots \\ \frac{\partial \mathsf{g}_k}{\partial z_1} & \cdots & \frac{\partial \mathsf{g}_k}{\partial z_k} \end{bmatrix}$$

To design

Why?? To generate realizations of X following $\mathsf{f}(X)$ by generating realizations of Z following $\mathsf{h}(Z)$ and then transforming $X = \mathsf{G}(Z)$

Adversarial Methods (GANs)

Goal: Start with Z following $\mathsf{h}(Z)$. Find generative transformation $\mathsf{G}(Z)$ so that $X{=}\mathsf{G}(Z)$ follows $\mathsf{f}(X)$. Solve problem without knowing $\mathsf{f}(X)$!!!

Let us develop our method following a step-by-step procedure

Suppose we have two probability densities $\mathsf{f}(X)$ and $\mathsf{g}(X)$
How can we test $\mathsf{f}(X) = \mathsf{g}(X)$?

$$\text{Sufficient: } \frac{\mathsf{g}(X)}{\mathsf{f}(X)} = 1 \quad \equiv \quad \omega\left(\frac{\mathsf{g}(X)}{\mathsf{f}(X)}\right) = \omega(1)$$

for ANY $\omega(\cdot)$ strictly increasing

Suppose we do not have $\mathsf{f}(X)$, $\mathsf{g}(X)$ but still desire to test $\mathsf{f}(X) = \mathsf{g}(X)$

For every function $\delta(X)$ mechanism provides $\boxed{\mathbb{E}_{\mathsf{f}}[\delta(X)],\ \ \mathbb{E}_{\mathsf{g}}[\delta(X)]}$

Can we compute $\mathrm{D}(X) = \omega\left(\frac{\mathrm{g}(X)}{\mathrm{f}(X)}\right)$ using averages ? YES !!!

THEOREM:

Specify strictly increasing function $\omega(\mathrm{r})$ and positive function $\rho(z)$

Define: $\phi(z), \psi(z)$: $\quad \psi'(z) = \rho(z), \quad \phi'(z) = -\omega^{-1}(z)\rho(z)$

and for any scalar function $\mathrm{D}(\mathrm{X})$ the cost

$$\mathrm{J}(\mathrm{D}) = \mathbb{E}_{\mathrm{f}}\left[\phi\left(\mathrm{D}(X)\right)\right] + \mathbb{E}_{\mathrm{g}}\left[\psi\left(\mathrm{D}(X)\right)\right]$$

Then the optimum solution to the maximization problem $\max_{\mathrm{D}(X)} \mathrm{J}(\mathrm{D})$ satisfies

$$\mathrm{D}_{\mathrm{o}}(X) = \omega\left(\frac{\mathrm{g}(X)}{\mathrm{f}(X)}\right)$$

Why insist on averages ?

If we do not have $\mathsf{f}(X)$, $\mathsf{g}(Y)$ but two sets of realizations

$$X_1, X_2, \ldots, X_n \sim \mathsf{f}(X) \quad \rightarrow \quad \mathbb{E}_{\mathsf{f}}\big[\delta(X)\big] \approx \frac{1}{n}\sum_{t=1}^{n} \delta(X_t)$$

$$Y_1, Y_2, \ldots, Y_m \sim \mathsf{g}(Y) \quad \rightarrow \quad \mathbb{E}_{\mathsf{g}}\big[\delta(Y)\big] \approx \frac{1}{m}\sum_{t=1}^{m} \delta(Y_t)$$

How can we optimize with respect to unknown function $\mathsf{D}(X)$?

Replace $\mathsf{D}(X)$ with neural network $\mathsf{D}(X,\vartheta)$. Optimize over network parameters ϑ

$$\mathsf{J}(\mathsf{D}) = \mathbb{E}_{\mathsf{f}}\big[\phi\big(\mathsf{D}(X)\big)\big] + \mathbb{E}_{\mathsf{g}}\big[\psi\big(\mathsf{D}(X)\big)\big]$$

$$\hat{\mathsf{J}}(\vartheta) = \frac{1}{n}\sum_{t=1}^{n} \phi\big(\mathsf{D}(X_t,\vartheta)\big) + \frac{1}{m}\sum_{t=1}^{m} \psi\big(\mathsf{D}(Y_t,\vartheta)\big)$$

Solve $\max_{\vartheta} \hat{\mathrm{J}}(\vartheta) \;\Rightarrow\; \vartheta_{\mathrm{o}} \;\Rightarrow\; \mathrm{D}(X, \vartheta_{\mathrm{o}})$

Expect $\mathrm{D}(X, \vartheta_{\mathrm{o}}) \approx \mathrm{D}_{\mathrm{o}}(X) = \omega\left(\frac{\mathrm{g}(X)}{\mathrm{f}(X)}\right)$

Can we use function $\mathrm{D}(X,\vartheta_{\mathrm{o}})$ to examine the two densities $\mathrm{f}(X)$, $\mathrm{g}(Y)$?

Employ $\mathrm{D}(X,\vartheta_{\mathrm{o}})$ to discriminate between $\mathrm{f}(X)$, $\mathrm{g}(Y)$ using only data

$$X_1, X_2, \ldots, X_n \;\sim\; \mathrm{f}(X) \qquad Y_1, Y_2, \ldots, Y_m \;\sim\; \mathrm{g}(Y)$$

If $\mathrm{D}(X, \vartheta_{\mathrm{o}}) \not\approx \omega(1) \;\Rightarrow$ the two datasets have different densities

If $\mathrm{D}(X, \vartheta_{\mathrm{o}}) \approx \omega(1) \;\Rightarrow$ the two datasets have similar densities

$\mathrm{D}(X,\vartheta)$ is known as the Discriminator function

Examples

A: $\omega(\mathsf{r}) = \mathsf{r}$

$$\rho(z) = 1, \quad \Rightarrow \quad \phi(z) = -\frac{z^2}{2}, \;\; \psi(z) = z$$

(Mean Square)

B: $\omega(\mathsf{r}) = \log(\mathsf{r})$

$$\rho(z) = e^{-0.5z} \quad \Rightarrow \quad \phi(z) = -2e^{0.5z}, \;\; \psi(z) = -2e^{-0.5z}$$

(Exponential)

C: $\omega(\mathsf{r}) = \dfrac{\mathsf{r}}{\mathsf{r}+1}$ Goodfellow et al. (2014), NeurIPS

$$\rho(z) = \frac{1}{z}, \quad \Rightarrow \quad \phi(z) = \log(1-z), \;\; \psi(z) = \log(z)$$

(Cross Entropy)

What happened to the Generator ?

Generator $\mathsf{G}(Z)$ transforms Z to $Y = \mathsf{G}(Z)$ with $Z \sim \mathsf{h}(Z)$

We desire $Y \sim \mathsf{f}(.)$ same density as training data $\{X_1,...,X_n\}$

Naïve Method: Select a $\mathsf{G}(Z)$. Test if transformation is the desired
If not, make another selection

How do we test if selection is any good ?

Generate realizations of Z: $\{Z_1,...,Z_m\}$

Apply generator to samples, create $\{Y_1,...,Y_m\}$ where $Y_i = \mathsf{G}(Z_i)$

Test $\{Y_1,...,Y_m\}$ against $\{X_1,...,X_n\}$

$$\hat{\mathsf{J}}(\vartheta) = \frac{1}{n}\sum_{t=1}^{n}\phi\big(\mathsf{D}(X_t,\vartheta)\big) + \frac{1}{m}\sum_{t=1}^{m}\psi\big(\mathsf{D}(Y_t,\vartheta)\big)$$

$$\max_{\vartheta} \hat{\mathsf{J}}(\vartheta) \Rightarrow \vartheta_{\mathsf{o}} \Rightarrow \mathsf{D}(X,\vartheta_{\mathsf{o}})$$

$$\mathsf{D}(X,\vartheta_{\mathsf{o}}) \approx \omega\left(\frac{\mathsf{g}(X)}{\mathsf{f}(X)}\right)$$

$$\mathsf{D}(X,\vartheta_{\mathsf{o}}) \overset{\not\approx}{\approx} \omega(1)$$

$$\hat{\mathsf{J}}(\vartheta) = \frac{1}{n}\sum_{t=1}^{n}\phi\big(\mathsf{D}(X_t,\vartheta)\big) + \frac{1}{m}\sum_{t=1}^{m}\psi\big(\mathsf{D}(Y_t,\vartheta)\big)$$

$$\hat{\mathsf{J}}(\vartheta) = \frac{1}{n}\sum_{t=1}^{n}\phi\big(\mathsf{D}(X_t,\vartheta)\big) + \frac{1}{m}\sum_{t=1}^{m}\psi\Big(\mathsf{D}\big(\mathsf{G}(Z_t),\vartheta\big)\Big)$$

We are looking for function $\mathsf{G}(Z)$. Use neural network $\mathsf{G}(Z,\theta)$

$$\hat{\mathsf{J}}(\theta,\vartheta) = \frac{1}{n}\sum_{t=1}^{n}\phi\big(\mathsf{D}(X_t,\vartheta)\big) + \frac{1}{m}\sum_{t=1}^{m}\psi\Big(\mathsf{D}\big(\mathsf{G}(Z_t,\theta),\vartheta\big)\Big)$$

THEOREM: For fixed θ (generator $\mathsf{G}(Z,\theta)$) we have

$$\max_{\vartheta}\hat{\mathsf{J}}(\theta,\vartheta) \gtrapprox \phi\big(\omega(1)\big) + \psi\big(\omega(1)\big)$$

Equality when $Y = \mathsf{G}(Z,\theta)$ has density $\mathsf{f}(.)$ same as $\{X_1,\ldots,X_n\}$

Since $\max_{\vartheta} \hat{\mathsf{J}}(\theta, \vartheta) \gtrapprox \phi\big(\omega(1)\big) + \psi\big(\omega(1)\big)$

to bring $\max_{\vartheta} \hat{\mathsf{J}}(\theta, \vartheta)$ as close as possible to $\phi\big(\omega(1)\big) + \psi\big(\omega(1)\big)$

we must apply minimization over θ

$$\min_{\theta} \max_{\vartheta} \hat{\mathsf{J}}(\theta, \vartheta)$$

$$= \min_{\theta} \max_{\vartheta} \left\{ \frac{1}{n} \sum_{t=1}^{n} \phi\big(\mathsf{D}(X_t, \vartheta)\big) + \frac{1}{m} \sum_{t=1}^{m} \psi\Big(\mathsf{D}\big(\mathsf{G}(Z_t, \theta), \vartheta\big)\Big) \right\}$$

Adversarial Optimization

G Generative
A Adversarial
N Networks

We have $Z \sim \mathsf{h}(Z)$
Design generator $\mathsf{G}(Z)$ so that $Y = \mathsf{G}(Z)$ has the same density as X for which we have realizations (training set) $\{X_1,\ldots,X_n\}$

Approximate generator with neural network $\mathsf{G}(Z,\theta)$
Define second neural network the discriminator $\mathsf{D}(X,\vartheta)$

For realizations of Z: $\{Z_1,\ldots,Z_m\}$ consider adversarial problem

$$\min_{\theta} \max_{\vartheta} \hat{\mathsf{J}}(\theta,\vartheta)$$
$$= \min_{\theta} \max_{\vartheta} \left\{ \frac{1}{n}\sum_{t=1}^{n} \phi\big(\mathsf{D}(X_t,\vartheta)\big) + \frac{1}{m}\sum_{t=1}^{m} \psi\Big(\mathsf{D}\big(\mathsf{G}(Z_t,\theta),\vartheta\big)\Big) \right\}$$

then generator $\mathsf{G}(Z,\theta_o)$ when applied to realizations of Z yields samples following closely the density of $\{X_1,\ldots,X_n\}$

Example

High definition CelebA (30 000 high definition images 1024 X 1024 of celebrities)



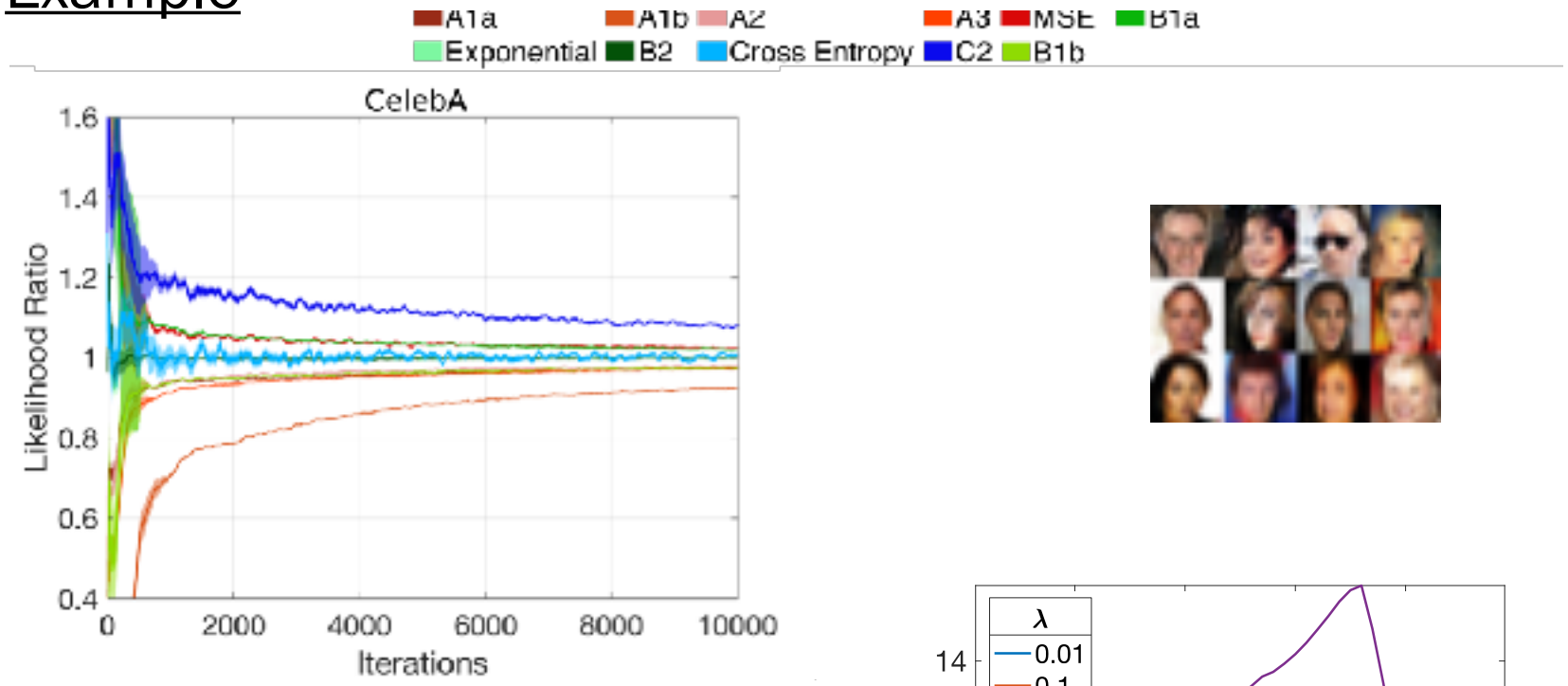


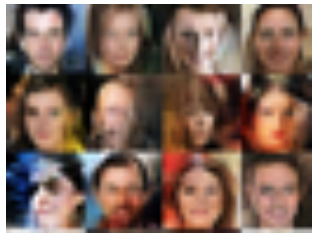

Extremely hard to control convergence of the adversarial problem

NVIDIA used progressive growing of GANs (4X4), (8X8),...,(1024X1024)

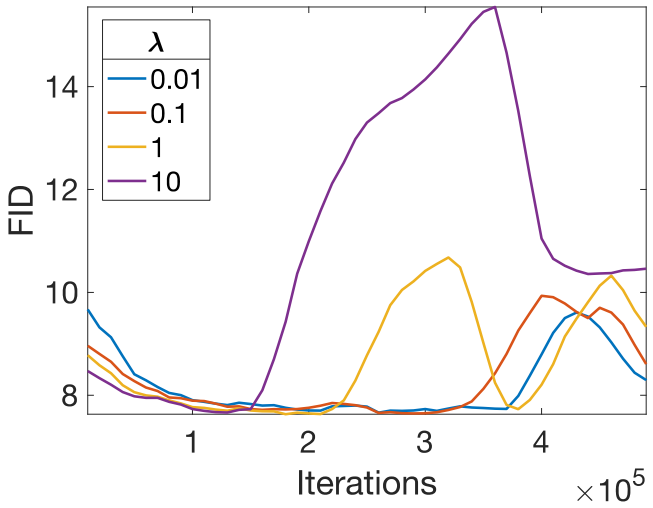


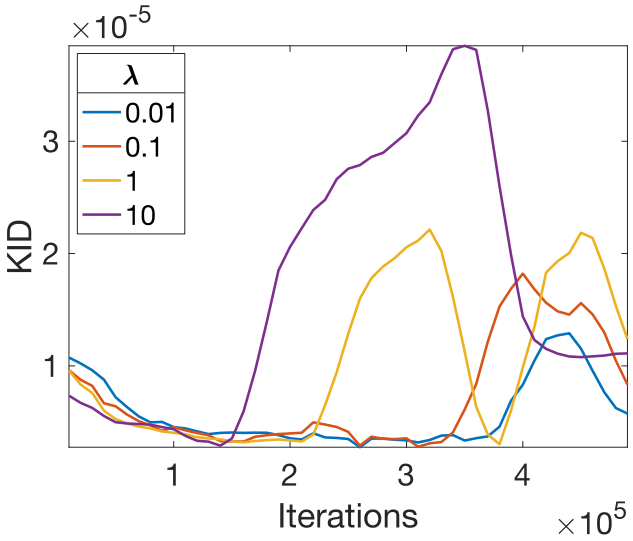
Example





Convergence/Divergence





Non-Adversarial Method

Interested in Generator
Discriminator used during generator design, afterwards useless

Possible (at least in theory) to design Generator without Discriminator by not employing min-max (adversarial) optimization

We can define optimization criteria involving only maximization or minimization with the help of positive definite kernels and design successfully Generators

A symmetric scalar function $\mathrm{K}(X,Y)$ will be positive definite if for every nonzero function $\varphi(X)$ it satisfies

$$\iint \mathrm{K}(X,Y)\varphi(X)\varphi(Y)\,dX\,dY \;>\; 0$$

Gaussian Kernel: $\mathrm{K}(X,Y) = e^{-\frac{1}{h^2}\|X-Y\|^2}$

Validity of the Cauchy-Schwarz inequality

$$\left(\iint \mathsf{K}(X,Y)\phi(X)\psi(Y)\,dX\,dY \right)^2$$
$$\leq \left(\iint \mathsf{K}(X,Y)\phi(X)\phi(Y)\,dX\,dY \right) \left(\iint \mathsf{K}(X,Y)\psi(X)\psi(Y)\,dX\,dY \right)$$

For densities $\mathsf{f}(X)$, $\mathsf{g}(Y)$

$$\frac{\left(\iint \mathsf{K}(X,Y)\mathsf{f}(X)\mathsf{g}(Y)\,dX\,dY \right)^2}{\iint \mathsf{K}(X,Y)\mathsf{f}(X)\mathsf{f}(Y)\,dX\,dY} \leq \iint \mathsf{K}(X,Y)\mathsf{g}(X)\mathsf{g}(Y)\,dX\,dY$$

with equality if and only if $\mathsf{f}(X) = \mathsf{g}(Y)$

$$\frac{\left(\iint \mathsf{K}(X,Y)\mathsf{f}(X)\mathsf{g}(Y)\,dX\,dY\right)^2}{\iint \mathsf{K}(X,Y)\mathsf{f}(X)\mathsf{f}(Y)\,dX\,dY} \leq \iint \mathsf{K}(X,Y)\mathsf{g}(X)\mathsf{g}(Y)\,dX\,dY$$

$$\frac{\left(\mathbb{E}\big[\mathsf{K}(X,Y)\big]\right)^2}{\mathbb{E}\big[\mathsf{K}(Y^1,Y^2)\big]} \leq \mathbb{E}\big[\mathsf{K}(X^1,X^2)\big]$$

where X^1, X^2 independent with the same density $\mathsf{f}(X)$
and Y^1, Y^2 independent with the same density $\mathsf{g}(Y)$

$$\mathsf{J}(\theta) = \frac{\left(\mathbb{E}\big[\mathsf{K}\big(X,\mathsf{G}(Z,\theta)\big)\big]\right)^2}{\mathbb{E}\big[\mathsf{K}\big(\mathsf{G}(Z^1,\theta),\mathsf{G}(Z^2,\theta)\big)\big]} \leq \mathbb{E}\big[\mathsf{K}(X^1,X^2)\big]$$

where Z^1, Z^2 independent and follow $\mathsf{h}(Z)$

$$\max_{\theta}\ \mathsf{J}(\theta)$$

We have $Z \sim \mathsf{h}(Z)$

Design generator $\mathsf{G}(Z)$ so that $Y = \mathsf{G}(Z)$ has the same density as X for which we have realizations (training set) $\{X_1,\ldots,X_n\}$

Generate $\{Z_1,\ldots,Z_m\}$ from $\mathsf{h}(Z)$

$$\hat{\mathsf{J}}(\theta) = \frac{\left(\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\mathsf{K}\big(X_i,\mathsf{G}(Z_j,\theta)\big)\right)^2}{\frac{1}{m(m-1)}\sum_{i=1}^{m}\sum_{j=1,j\neq i}^{m}\mathsf{K}\big(\mathsf{G}(Z_i,\theta),\mathsf{G}(Z_j,\theta)\big)}$$

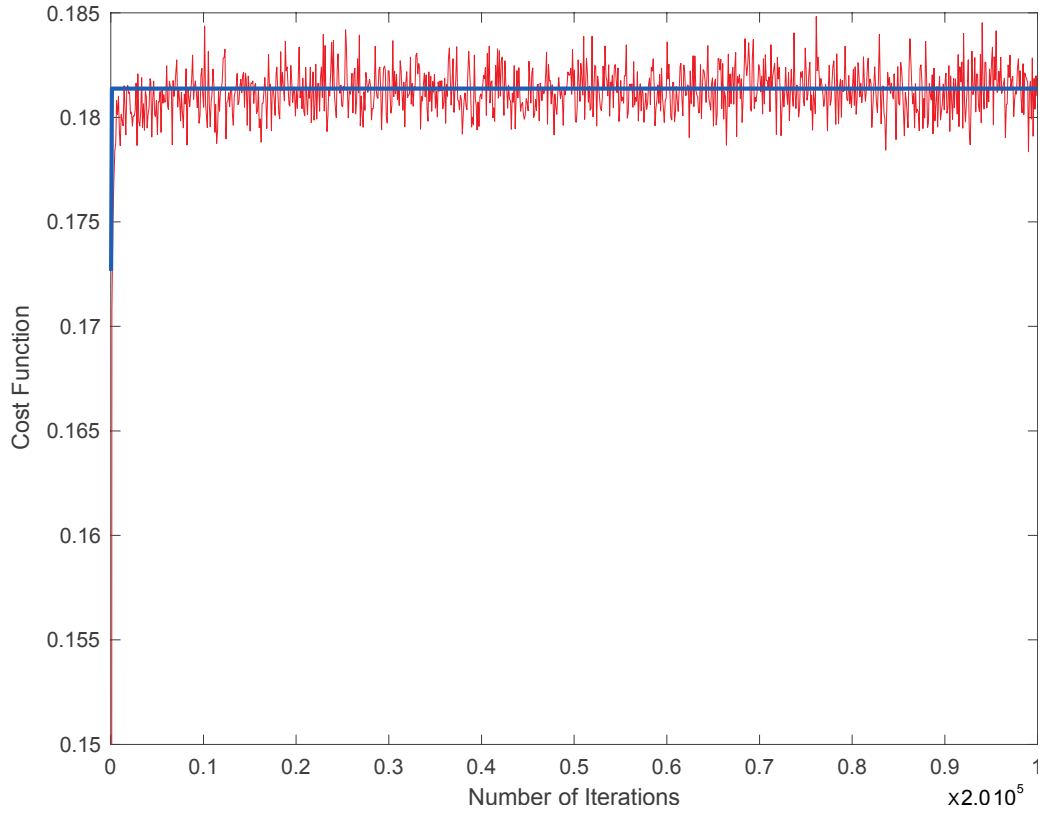
$$\lessapprox \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\mathsf{K}\big(X_i,X_j\big)$$

$$\max_{\theta}\hat{\mathsf{J}}(\theta)$$

No Convergence/Divergence phenomenon!!!!

Example

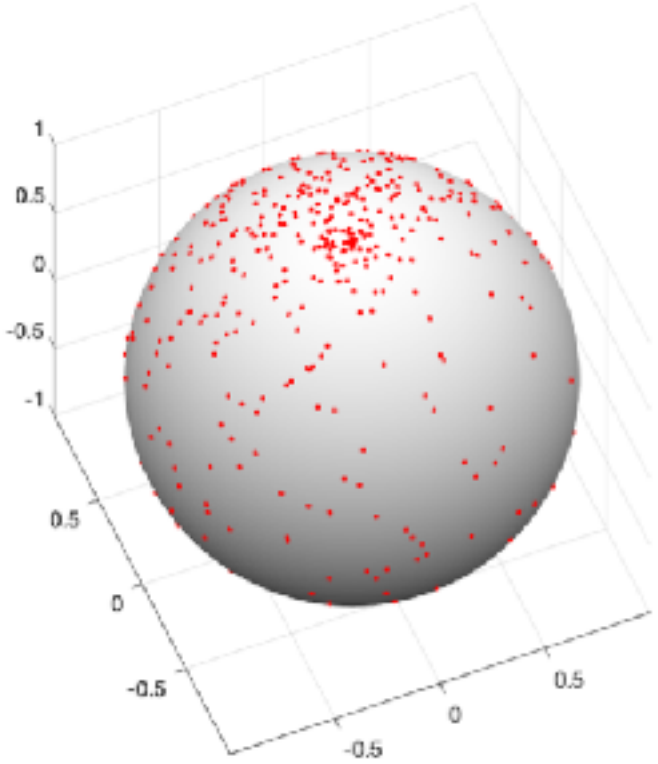
CelebA images 32 X 32





Suitable for small sized and fully connected networks

Probability Density vs Generative Model



Points in N-D space can be random and lie on a lower dimensional surface (manifold)

Example red points on sphere (2-D in 3-D space)

Points are random with coordinates $[x_1,\ x_2,\ x_3]$ related through a deterministic equation

To lie on a sphere of radius r : $\quad x_1^2 + x_2^2 + x_3^2 = r^2$

$$\mathrm{f}(x_1, x_2, x_3) = \delta(x_1^2 + x_2^2 + x_3^2 - r^2)\,\mathrm{h}(x_1, x_2)$$

Dirac $\delta(x)$ generalized function is defined as

$$\delta(x) = \begin{cases} 0 & x \neq 0 \\ \infty & x = 0 \end{cases}, \quad \int_{-\epsilon}^{\epsilon} \delta(x)\,dx = 1$$

Generative model would describe the random data with input density $\mathsf{h}(z_1,z_2)$ and generator vector function $\mathsf{G}(z_1,z_2)$

$$X = \mathsf{G}(z_1, z_2) \Rightarrow \begin{bmatrix} x_1 = \mathsf{G}_1(z_1, z_2) \\ x_2 = \mathsf{G}_2(z_1, z_2) \\ x_3 = \mathsf{G}_3(z_1, z_2) \end{bmatrix} \Rightarrow \begin{bmatrix} x_1 = r\cos(2\pi z_1)\sin(\pi z_2) \\ x_2 = r\sin(2\pi z_1)\sin(\pi z_2) \\ x_3 = r\cos(\pi z_2) \end{bmatrix}$$

$\mathsf{h}(z_1,z_2)$ defined on $[0,1]\times[0,1]$ and $\mathsf{G}(z_1,z_2)$ is an ordinary function

Data are representable as $X = \mathsf{G}(Z)$, Z~$\mathsf{h}(Z)$. Many datasets satisfy

$$\dim(Z) \ll \dim(X)$$

In HD CelebA: $\dim(X) = 3 \times 1024 \times 1024 = 3 \times 10^6$

Input to Generator $\mathsf{G}(Z)$: $\dim(Z) = 500$ (independent Gaussians)

Instead of estimating X, we first estimate Z, then recover X as X=$\mathsf{G}(Z)$



From Statistical Estimation and Generative Modeling to Inverse Problems

LECTURE 5: INVERSE PROBLEMS

Inverse Problems

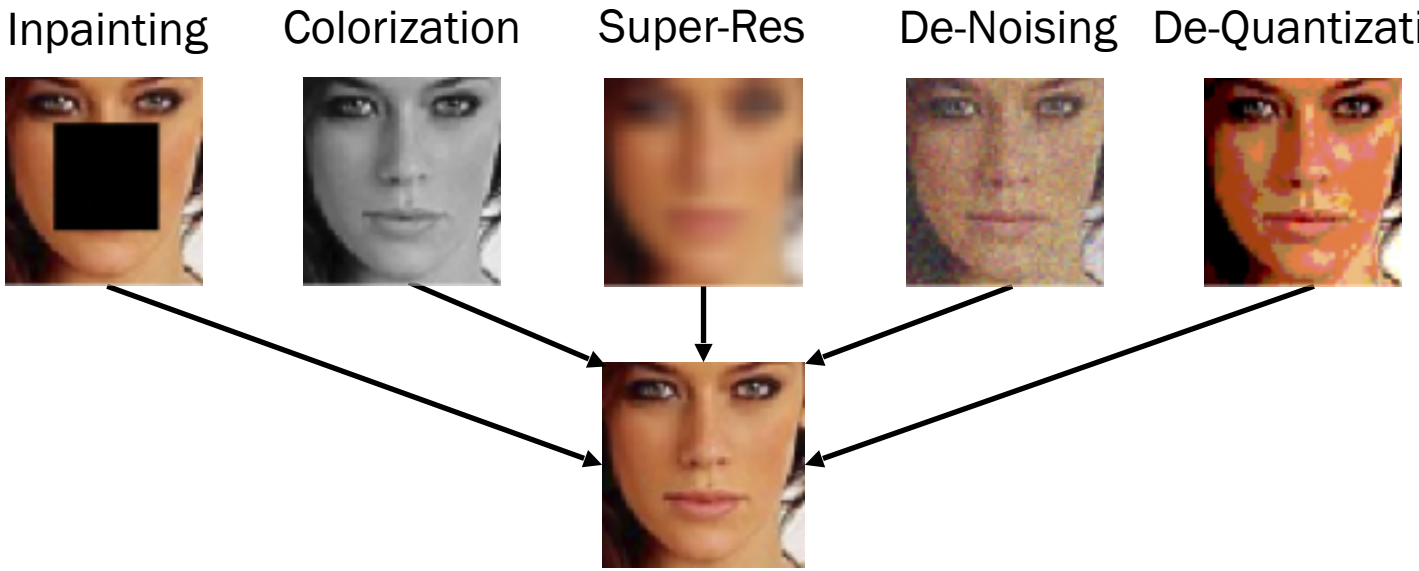
- Problem Definition
- Early Efforts
- Solutions Based on GANs
  - Add-Hoc
  - Applying Statistical Estimation
- Examples

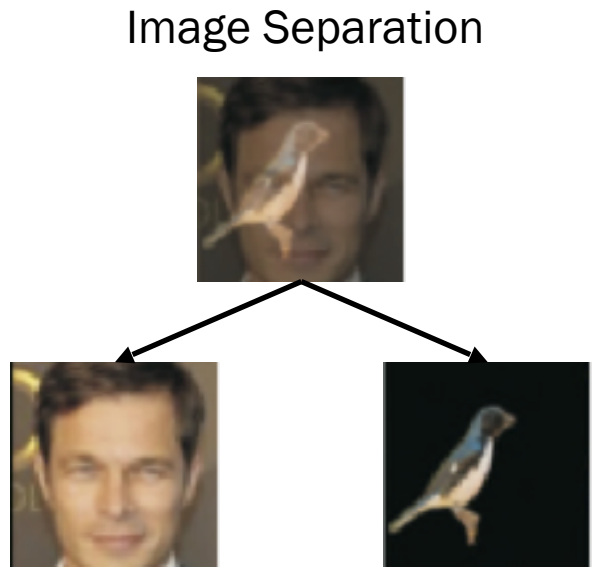
Problem Definition

We are given vector X (measurements) and interested in estimating vector Y

We assume $X = \mathsf{T}(Y) + W$ where $\mathsf{T}(Y)$ general (mostly) known transformation

Basic characteristic: $\dim(X) \leq \dim(Y)$





Early Efforts

Inpainting







Diffusion based inpainting





No prior information, training data not very useful

Solutions Based on GANs

Available training data $\{X_1,...,X_n\}$

Design generator $\mathsf{G}(Z)$ so that when applied to Z with $Z \sim \mathsf{h}(Z)$ then $Y = \mathsf{G}(Z)$ has same density as training data

Generate $\{Z_1,...,Z_m\}$ with density $\mathsf{h}(Z)$

$$\hat{\mathsf{J}}(\theta,\vartheta) = \frac{1}{n}\sum_{t=1}^{n}\phi\big(\mathsf{D}(X_t,\vartheta)\big) + \frac{1}{m}\sum_{t=1}^{m}\psi\Big(\mathsf{D}\big(\mathsf{G}(Z_t,\theta),\vartheta\big)\Big)$$

Discriminator

Generator

$$\min_{\theta}\ \max_{\vartheta}\ \hat{\mathsf{J}}(\theta,\vartheta)$$

Assume known Generative Model $\{\mathsf{G}(Z),\mathsf{h}(Z)\}$, also Discriminator $\mathsf{D}(X)$

General Problem

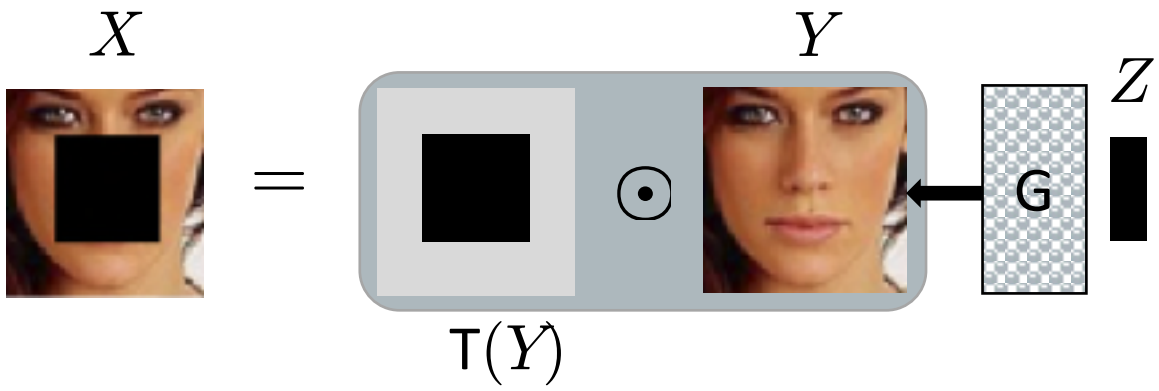
Given vector X (measurements) we are interested in estimating vector Y

with $X = \mathsf{T}(Y) + W$ and $\mathsf{T}(Y)$ known transformation

Y follows Generative Model

There exists Z following density $\mathsf{h}(Z)$ such that $Y = \mathsf{G}(Z)$

Instead of estimating Y from X we estimate input to generative model Z



Measurements $X = \mathsf{T}(Y) + W$

Compute estimate $\hat{Y}$

Measurements $X = \mathsf{T}\big(\mathsf{G}(Z)\big) + W$

Compute estimate $\hat{Z}$

and let $\hat{Y} = \mathsf{G}(\hat{Z})$

Since $\dim(Z) \ll \dim(Y)$

Significant computational gain and more stable processing

Select $\hat{Z}$ so that measurements X and $\mathsf{T}\big(\mathsf{G}(\hat{Z})\big)$ are "close"

$$\min_{Z} \|X - \mathsf{T}\big(\mathsf{G}(Z)\big)\|^2 \;\Rightarrow\; \hat{Z} = \arg\min_{Z} \|X - \mathsf{T}\big(\mathsf{G}(Z)\big)\|^2$$

Well defined optimization, computationally stable

Doesn't Work!!

$\hat{Y} = \mathsf{G}(\hat{Z})$ does not have the correct characteristics

Recall that generative model is the pair $\{\mathsf{G}(Z), \mathsf{h}(Z)\}$

Even if $\mathsf{T}\big(\mathsf{G}(\hat{Z})\big)$ "close" to X, if likelihood $\mathsf{h}(\hat{Z}) \ll 1$

then $\hat{Y} = \mathsf{G}(\hat{Z})$ is "bad" result

Must take into account input density $\mathsf{h}(Z)$

Ad-Hoc Solutions

Yeh et al. (2017), (2018)

Assumes generative model trained with cross entropy method

$\phi(z) = \log(1-z),\ \psi(z) = \log(z),\ z \in (0,1) \ \Rightarrow$

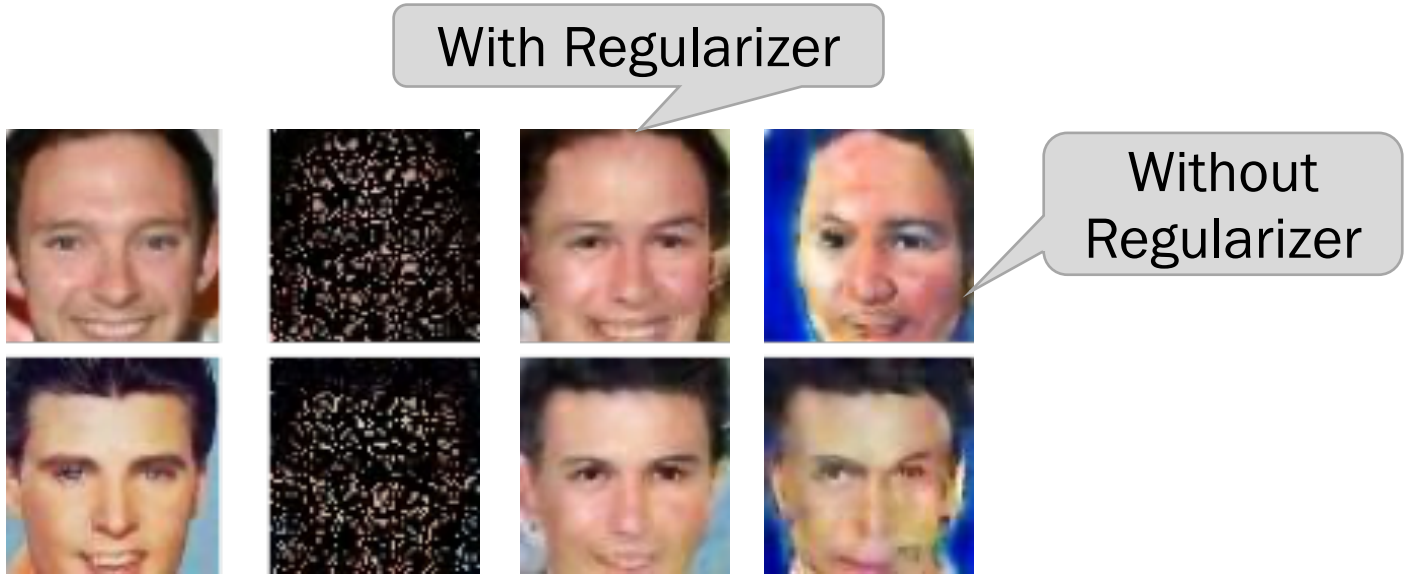
Design $\mathsf{G}(Z)$ generator, $\mathsf{D}(Z)$ discriminator, when $Z \sim \mathsf{h}(Z)$

$$\mathsf{J}(Z) = \|X - \mathsf{T}\big(\mathsf{G}(Z)\big)\|_{L_1} + \lambda \Big\{ \log\Big(1 - \mathsf{D}\big(\mathsf{G}(Z)\big)\Big) - \log\Big(\mathsf{D}\big(\mathsf{G}(Z)\big)\Big) + \log\big(\mathsf{h}(Z)\big) \Big\}$$

Regularizer

Parameter needs tuning

$$\hat{Z} = \arg\min_{Z} \mathsf{J}(Z), \quad \hat{Y} = \mathsf{G}(\hat{Z})$$





To tune parameter λ need ideal images $\{Y_1,...,Y_n\}$ apply transformation $X_i=\mathsf{T}(Y_i)$

For various values of λ for each $\{X_1,...,X_n\}$ compute

$$\hat{Z}_1(\lambda),\ldots,\hat{Z}_n(\lambda) \;\Rightarrow\; \hat{Y}_1(\lambda),\ldots,\hat{Y}_n(\lambda) \qquad \lambda_i = \arg\min_{\lambda} \|Y_i - \hat{Y}_i(\lambda)\|$$

$$\lambda = \frac{1}{n}\{\lambda_1 + \cdots + \lambda_n\}$$

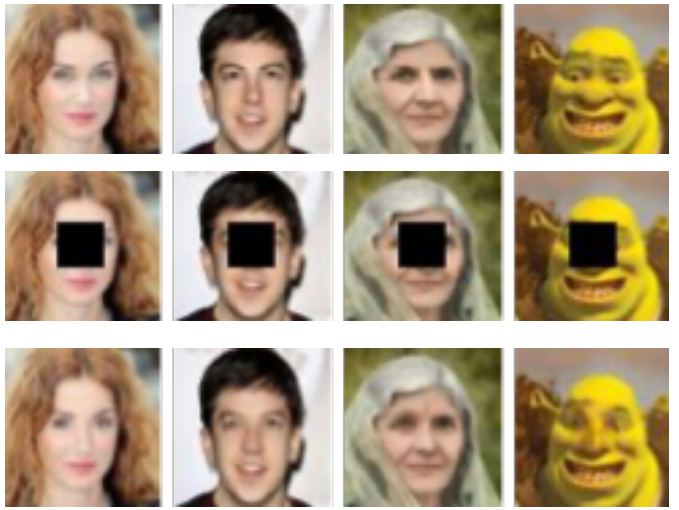
Requires exact knowledge of $\mathsf{T}(Y)$

Asim et al. (2019)

$$\hat{Z} = \arg\min_{Z} \left\{ \|X - \mathsf{T}(\mathsf{G}(Z))\|^2 - \lambda \log\left(\mathsf{h}(Z)\right) \right\}$$

Parameter needs tuning





Requires exact knowledge of $\mathsf{T}(Y)$

Siavelis et al. (2020)

$$\hat{Z} = \arg\min_{Z} \left\{ \|X - \mathsf{T}(\mathsf{G}(Z))\|_{L_1} - \lambda\left(1 - \mathsf{D}(\mathsf{G}(Z))\right) \right\}$$

Parameter needs tuning









Requires exact knowledge of $\mathsf{T}(Y)$

Applying Statistical Estimation

Given densities $\mathsf{f}(X|Z)$ and prior $\mathsf{h}(Z)$, for measurement X estimate Z

MAP estimator is of the form

$$\hat{Z} = \arg\max_Z \mathsf{f}(Z|X) = \arg\max_Z \mathsf{f}(X,Z) = \arg\max_Z \mathsf{f}(X|Z)\mathsf{h}(Z)$$

Let $Z=\{Z_1,Z_2\}$ where there is prior for Z_1 but not for Z_2

Treat non-existing prior as degenerate uniform

$$\begin{aligned}
\arg\max_{Z_1,Z_2} \mathsf{f}(Z_1,Z_2|X) &= \arg\max_{Z_1,Z_2} \mathsf{f}(X,Z_1,Z_2) \\
&= \arg\max_{Z_1,Z_2} \mathsf{f}(X|Z_1,Z_2)\mathsf{h}(Z_1,Z_2) \\
&= \arg\max_{Z_1,Z_2} \mathsf{f}(X|Z_1,Z_2)\mathsf{h}_1(Z_1|Z_2)\mathsf{h}_2(Z_2) \\
\mathsf{h}_2(Z_2) \text{ degenerate uniform} \quad &= \arg\max_{Z_1,Z_2} \mathsf{f}(X|Z_1,Z_2)\mathsf{h}_1(Z_1|Z_2)
\end{aligned}$$

If interested in estimating Z_1 and Z_2 are nuisance parameters then

$$\hat{Z}_1 = \arg\max_{Z_1} \left\{ \max_{Z_2} \mathsf{f}(X|Z_1, Z_2)\mathsf{h}_1(Z_1|Z_2) \right\}$$

where $\mathsf{h}_1(Z_1|Z_2)$ prior of Z_1 given Z_2

If Z_1 does not depend on Z_2 then $\mathsf{h}_1(Z_1|Z_2)=\mathsf{h}_1(Z_1)$ and

$$\hat{Z}_1 = \arg\max_{Z_1} \left\{ \max_{Z_2} \mathsf{f}(X|Z_1, Z_2)\mathsf{h}_1(Z_1) \right\}$$

We are given vector X (measurements) and interested in estimating vector Y

$$\text{We assume } X = \mathsf{T}(Y, \alpha) + W = \mathsf{T}(\mathsf{G}(Z), \alpha) + W$$

$\mathsf{T}(Y, \alpha)$: transformation of known mathematical form possibly containing unknown parameters α. Can be different per measurement X

W: additive noise with density $\mathsf{g}_w(W, \beta)$ possibly containing unknown parameters β. Can be different per measurement X

Z: follows density $\mathsf{h}(Z)$ from generative model $\{\mathsf{G}(Z), \mathsf{h}(Z)\}$

$$Z_1 = Z, \quad Z_2 = \{\alpha, \beta\} \qquad \mathsf{f}(X|Z_1, Z_2) = \mathsf{f}(X|Z, \alpha, \beta) = \mathsf{g}_w\Big(X - \mathsf{T}\big(\mathsf{G}(Z), \alpha\big), \beta\Big)$$

$$\hat{Z} = \arg\max_{Z} \left\{ \max_{\alpha,\beta} \mathsf{f}(X|Z, \alpha, \beta)\mathsf{h}(Z) \right\}$$

$$= \arg\max_{Z} \left\{ \max_{\alpha,\beta} \mathsf{g}_w\Big(X - \mathsf{T}\big(\mathsf{G}(Z), \alpha\big), \beta\Big)\mathsf{h}(Z) \right\}$$

$$\hat{Z} = \arg\max_{Z} \left\{ \max_{\alpha} \max_{\beta} \mathsf{g}_w\Big(X - \mathsf{T}\big(\mathsf{G}(Z), \alpha\big), \beta\Big)\mathsf{h}(Z) \right\}$$

W: additive noise is Gaussian mean 0 and covariance $\beta^2 I$

$$\max_{\beta} \mathsf{g}_w\Big(X - \mathsf{T}\big(\mathsf{G}(Z), \alpha\big), \beta\Big) = \max_{\beta} \frac{e^{-\|X - \mathsf{T}(\mathsf{G}(Z), \alpha)\|^2 / 2\beta^2}}{(\sqrt{2\pi\beta^2})^N}$$

$$= \frac{\mathsf{C}}{\|X - \mathsf{T}(\mathsf{G}(Z), \alpha)\|^N}$$

N: length of measurement vector X

$$\hat{Z} = \arg\max_{Z} \left\{ \max_{\alpha} \frac{\mathsf{C}\,\mathsf{h}(Z)}{\|X - \mathsf{T}(\mathsf{G}(Z), \alpha)\|^N} \right\}$$

$$= \arg\max_{Z} \frac{\mathsf{h}(Z)}{(\min_{\alpha} \|X - \mathsf{T}(\mathsf{G}(Z), \alpha)\|^2)^{N/2}}$$

Z: If input of generative model is Gaussian with mean 0 and covariance identity

$$\hat{Z} = \arg\max_{Z} \frac{\mathsf{C}'\, e^{-\|Z\|^2/2}}{(\min_{\alpha} \|X - \mathsf{T}(\mathsf{G}(Z), \alpha)\|^2)^{N/2}}$$

$$\hat{Z} = \arg\min_{Z} \left\{ \log\left(\min_{\alpha} \|X - \mathsf{T}(\mathsf{G}(Z), \alpha)\|^2 \right) + \frac{1}{N}\|Z\|^2 \right\}$$

$$\Leftrightarrow \quad \min_{Z,\alpha} \left\{ \log\left(\|X - \mathsf{T}(\mathsf{G}(Z), \alpha)\|^2 \right) + \frac{1}{N}\|Z\|^2 \right\}$$

If transformation satisfies

$$\mathsf{T}(Y,\alpha) = \alpha_1 \mathsf{T}_1(Y) + \cdots + \alpha_m \mathsf{T}_m(Y)$$

then $\mathsf{T}\big(\mathsf{G}(Z),\alpha\big) = \alpha_1 \mathsf{T}_1\big(\mathsf{G}(Z)\big) + \cdots + \alpha_m \mathsf{T}_m\big(\mathsf{G}(Z)\big) = \mathcal{G}(Z)A$

where $\mathcal{G}(Z) = \Big[\mathsf{T}_1\big(\mathsf{G}(Z)\big) \cdots \mathsf{T}_m\big(\mathsf{G}(Z)\big)\Big], \;\; A = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix}$

$$\min_A \|X - \mathcal{G}(Z)A\|^2 = \|X\|^2 - X^\intercal \mathcal{G}(Z)\big(\mathcal{G}^\intercal(Z)\mathcal{G}(Z)\big)^{-1}\mathcal{G}^\intercal(Z)X$$

$$\hat{Z} = \arg\min_Z \left\{ \log\left(\|X\|^2 - X^\intercal \mathcal{G}(Z)\big(\mathcal{G}^\intercal(Z)\mathcal{G}(Z)\big)^{-1}\mathcal{G}^\intercal(Z)X \right) + \frac{1}{N}\|Z\|^2 \right\}$$

Examples

Blurring with 3 X 3 kernel

Mth 1 Mth 2 Known Unknown



Colorization (green channel)

Mth 1 Mth 2 Known Unknown



No parameters to tune

De-Quantization

2 levels per RGB channel, 8 colors



De-Quantization

3 levels per RGB channel, 27 colors



De-Quantization

5 levels per RGB channel, 125 colors



De-Quantization and Colorization

RGB → Gray → BW (2 levels)



With threshold estimation

Data Mixtures

We assume two independent data vectors Y_1, Y_2 that are combined as

$$X = \alpha_1 Y_1 + \alpha_2 Y_2 + W$$

From mixture X recover original Y_1, Y_2

Y_1: generative model $\{\mathsf{G}_1(Z_1), \mathsf{h}_1(Z_1)\}$

Y_2: generative model $\{\mathsf{G}_2(Z_2), \mathsf{h}_2(Z_2)\}$

W: additive noise with density $\mathsf{g}_w(W, \beta)$ containing unknown parameters β

If W Gaussian, mean 0, covariance $\beta^2 I$

If inputs of both generative models Gaussian, mean 0, covariance identity

$$\{\hat{Z}_1, \hat{Z}_2\} =$$

known

$$\arg \min_{Z_1, Z_2} \left\{ \log\left(\|X - \alpha_1 \mathsf{G}_1(Z_1) - \alpha_2 \mathsf{G}_2(Z_2)\|^2 \right) + \frac{1}{N}\left(\|Z_1\|^2 + \|Z_2\|^2 \right) \right\}$$

$$\{\hat{Z}_1, \hat{Z}_2\} =$$

unknown

$$\arg \min_{Z_1, Z_2} \left\{ \log \Big(\min_{\alpha_1, \alpha_2} \|X - \alpha_1 \mathsf{G}_1(Z_1) - \alpha_2 \mathsf{G}_2(Z_2)\|^2 \Big) + \frac{1}{N} \big(\|Z_1\|^2 + \|Z_2\|^2 \big) \right\}$$

$$= \arg \min_{Z_1, Z_2} \left\{ \log \left(\|X\|^2 - X^\mathsf{T} \mathcal{G}(Z_1, Z_2) \Big(\mathcal{G}^\mathsf{T}(Z_1, Z_2) \mathcal{G}(Z_1, Z_2) \Big)^{-1} \mathcal{G}^\mathsf{T}(Z_1, Z_2) X \right) \right.$$
$$\left. + \frac{1}{N} \big(\|Z_1\|^2 + \|Z_2\|^2 \big) \right\}$$

where $\mathcal{G}(Z_1, Z_2) = \big[\mathsf{G}_1(Z_1) \; \mathsf{G}_2(Z_2) \big]$

Known coefs Uknown coefs



Known coefs Uknown coefs



Nonlinear Data Mixtures

We assume two independent data vectors Y_1, Y_2 that are combined as

$$X = \mathsf{T}(Y_1, Y_2, \alpha) + W$$

Y_1: generative model $\{\mathsf{G}_1(Z_1), \mathsf{h}_1(Z_1)\}$

Y_2: generative model $\{\mathsf{G}_2(Z_2), \mathsf{h}_2(Z_2)\}$

W: additive noise with density $\mathsf{g}_w(W, \beta)$ containing unknown parameters β

$$\{\hat{Z}_1, \hat{Z}_2\} =$$

$$\arg \min_{Z_1, Z_2} \left\{ \log \left(\min_{\alpha} \|X - \mathsf{T}\big(\mathsf{G}_1(Z_1), \mathsf{G}_2(Z_2), \alpha\big)\|^2 \right) + \frac{1}{N} \left(\|Z_1\|^2 + \|Z_2\|^2 \right) \right\}$$

$$\Rightarrow \min_{Z_1, Z_2, \alpha} \left\{ \log \left(\|X - \mathsf{T}\big(\mathsf{G}_1(Z_1), \mathsf{G}_2(Z_2), \alpha\big)\|^2 \right) + \frac{1}{N} \left(\|Z_1\|^2 + \|Z_2\|^2 \right) \right\}$$