Learning theory for matrix inverse problems with gaussian priors

Jonas Adler
KTH, Elekta

Olivier Verdier
KTH

1 Maximum a posteriori estimate

Suppose $A : \mathbb{R}^n \to \mathbb{R}^m$ is a linear forward operator (a matrix) and we have data of the form

$$x \in \mathcal{N}(0, \Sigma)$$
$$y \in \mathcal{N}(Ax, \Gamma)$$

eq:inv_prob

prop:mapest

Theorem 1.1 ([**Mu2012**])**.** *The maximum a posteriori (MAP) estimate of x given y*

$$A^{-1}_{MAP}(y) = \arg\max_{x'} P(x'|y)$$

is

$$A^{-1}_{MAP}(y) = (A^T\Gamma^{-1}A + \Sigma^{-1})^{-1}A^T\Gamma^{-1}y$$

Proof. We have

$$P(x'|y) \propto \underbrace{P(y|x')}_{\mathcal{N}(Ax',I)}\ \underbrace{P(x')}_{\mathcal{N}(0,I)}$$

Instead of maximizing the posterior, we minimize the log of the posterior

$$\arg\max_{x'} P(x'|y) = \arg\min \log P(x'|y) = \arg\min \left(\log P(y|x') + \log P(x')\right)$$

where we let

$$\begin{aligned} f(x') &= \log P(y|x') + \log P(x') \\ &= \frac{1}{2}\|Ax - y\|^2_{\Gamma^{-1}} + \frac{1}{2}\|x\|^2_{\Sigma^{-1}} \end{aligned}$$

we compute

$$\nabla f = A^T\Gamma^{-1}(Ax - y) + \Sigma^{-1}x$$

Setting this to zero gives the conditions for a maximum

$$(A^T\Gamma^{-1}A + \Sigma^{-1})x = A^T\Gamma^{-1}y$$

which gives a maximum likelihood solution as

$$x = (A^T\Gamma^{-1}A + \Sigma^{-1})^{-1}A^T\Gamma^{-1}y$$

□

Connection to Tikhonov regularization: By identification, this is equivalent to solving the Tikhonov regularized problem:

$$\min_x \|\Gamma^{-1/2}\big(Ax - y\big)\| + \|\Sigma^{-1/2}x\|$$

2 Best function estimate

prop:bestfunest

Theorem 2.1 ([**Bi2006**])**.** *We have*

$$\arg\min_B \mathbb{E}_{(x,y)}\|B(y) - x\|_2^2 = \mathbb{E}_x\big(x \mid y\big)$$

where the minimization is taken over all functions $B\colon \mathbb{R}^m \to \mathbb{R}^n$*.*

Proof. By the law of total probability

$$\mathbb{E}_{(x,y)}\|B(y) - x\|_2^2 = \mathbb{E}_y\big(\mathbb{E}_x\big(\|B(y) - x\|_2^2 \mid y\big)\big)$$

By the monotonicity of the expectation we have

$$B(y) = \arg\min_z \mathbb{E}_x\big(\|z - x\|_2^2 \mid y\big)$$

we find this by differentiating and setting to zero

$$0 = \mathbb{E}_x\big(2(z - x) \mid y\big) = 2\mathbb{E}_x\big(z \mid y\big) - 2\mathbb{E}_x\big(x \mid y\big)$$

which gives

$$B(y) = \mathbb{E}_x\big(x \mid y\big)$$

□

Theorem 2.2. *A neural network, which we denote by* A_{NN}^{-1}*, will in the limit of infinite capacity and trained to solve the inverse problem equation 1 according to*

$$\arg\min_{A_{NN}^{-1}} \mathbb{E}_{(x,y)}\|A_{NN}^{-1}(y) - x\|_2^2$$

is given by the MAP estimate:

$$A_{NN}^{-1}(y) = (A^T\Gamma^{-1}A + \Sigma^{-1})^{-1}A^T\Gamma^{-1}y$$

Proof. By Theorem 2.1, we have that the optimum solution of the training problem is given by the conditional expectation. Since neural networks are universal function approximators, this will be attainable in the limit of infinite capacity.

Further, for Gaussian distributions the mean and the mode coincide. Thus the conditional expectation, being a gaussian, is the MAP estimate. □

Denoising Consider the case of denoising where instead of observing y as in 1, we first apply the pseudo-inverse

$$z = A^\dagger y$$

By standard arguments this is also a gaussian and we find

$$z \in \mathcal{N}(A^\dagger A x, A^\dagger \Gamma (A^\dagger)^T)$$

Hence the MAP estimate is (with $A^\dagger A = P_{A^*}$, $AA^\dagger = P_A$ being projection onto the range of A^* and A respectively) via straightforward application of theorems 1.1 and 2.1:

$$\begin{aligned} A^{-1}_{\text{denoising}}(A^\dagger y) &= \left((A^\dagger A)^T (A^\dagger \Gamma (A^\dagger)^T)^\dagger (A^\dagger A) + \Sigma^{-1}\right)^\dagger (A^\dagger A)^T (A^\dagger \Gamma (A^\dagger)^T)^\dagger A^\dagger y \\ &= \left(P^T_{A^*} (A^T \Gamma^\dagger A) P_A + \Sigma^{-1}\right)^\dagger P^T_{A^*} (A^T \Gamma^\dagger A) A^\dagger y \\ &= \left((AP_{A^*})^T \Gamma^\dagger (AP_{A^*}) + \Sigma^{-1}\right)^\dagger (AP_{A^*})^T \Gamma^\dagger P_A y \end{aligned}$$

Now via identification, $Az = AA^\dagger y$, so This shows that:

$$A^{-1}_{\text{denoising}}(A^\dagger y) = (AP_{A^*})^{-1}_{NN}(P_A y)$$