Bayesian inference with data-driven image priors

Marcelo Pereyra

Heriot-Watt University

& Maxwell Institute

Joint work with

Matthew Holden
(Maxwell Institute)

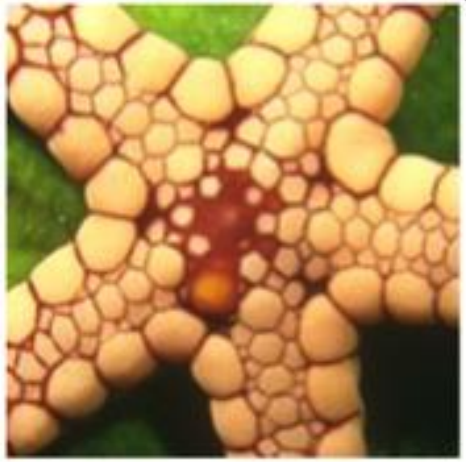
Kostas Zygalakis
(Edinburgh University
& Maxwell Institute)

Andrés Almansa
(CNRS, University of Paris)

Outline

- **Introduction**
- Proposed method
- Experiments

Forward problem

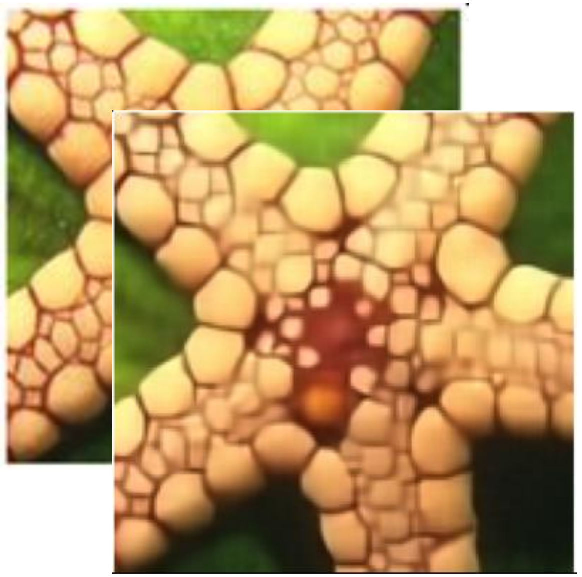


True scene

Imaging device

Observed image

Inverse problem



Estimated scene

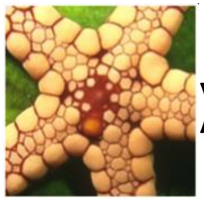
Imaging **method**

Observed image

Problem statement

- We are interested in recovering an unknown image $x \in \mathbb{R}^d$, e.g.,
- We measure y, related to x by some mathematical model.
- For example, many imaging problems involve models of the form

$$y = Ax + w,$$

 =  () + w

for some linear operator A, and some perturbation or “noise” w.

- The recovery of x from y is often ill-posed or ill-conditioned, so we regularize it.

Bayesian statistics

- We formulate the estimation problem in the Bayesian statistical framework, a probabilistic mathematical framework in which we represent x as a random quantity and use probability distributions to model expected properties.

- To derive inferences about x from y we postulate a joint statistical model $\mathrm{p}(x, y)$ typically specified via the decomposition $\mathrm{p}(x, y) = p(y|x)p(x)$.

- The Bayesian framework is equipped with a powerful decision theory to derive solutions and inform decisions and conclusions in a rigorous and defensible way.

Bayesian statistics

- The decomposition $\mathrm{p}(x, y) = p(y|x)p(x)$ has two ingredients:

- The **likelihood**: the conditional distribution $p(y|x)$ that <u>models</u> the data observation process (the forward model).

- The **prior**: the marginal distribution $p(x) = \int p(x, y)\, dy$ that <u>models</u> expected properties of the solutions.

- In imaging, $p(y|x)$ usually has significant identifiability issues and we rely strongly on $p(x)$ to regularize the estimation problem and deliver meaningful solutions.

Bayesian statistics

- We base our inferences on the **posterior** distribution

$$\mathrm{p}(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)}$$

where $p(y) = \int p(x,y)\,dx$ provides an indication of the goodness of fit.

- The conditional distribution $\mathrm{p}(x|y)$ models our knowledge about the solution x after observing the data y, in a manner that is clear, modular and elegant.

- Inferences are then derived by using Bayesian decision theory.

Bayesian statistics

There are three main challenges in deploying Bayesian approaches in imaging sciences:

1. Bayesian computation: calculating probabilities and expectations w.r.t. $\mathrm{p}(x|y)$ is computationally expensive, although algorithms are improving rapidly.

2. Bayesian analysis: we do not usually know what questions to ask $\mathrm{p}(x|y)$, imaging sciences are a field in transition and the concept of *solution* is evolving.

3. Bayesian modelling: while it is true that *all models are wrong, but some are useful,* image models are often too simple to reliably support advanced inferences.

Outline

- Introduction
- **Proposed method**
- Experiments

In this talk

- Instead of specifying an analytic form for $p(x)$, we consider the situation where the prior knowledge about x is available as a set of examples $\{x_i'\}_{i=1}^{M}$ i.i.d. w.r.t x.

- We aim to combine this prior knowledge with a likelihood $p(y|x)$ specified analytically to derive a posterior distribution for $p\left(x\middle|y,\{x_i'\}_{i=1}^{M}\right)$.

- The goal is to construct $p\left(x\middle|y,\{x_i'\}_{i=1}^{M}\right)$ in a way that preserves the modularity and interpretability of analytic Bayesian models, and enables efficient computation.

Bayesian model

- Following the *manifold hypothesis*, we assume that x takes values close to an unknown p–dimensional submanifold of $\mathbb{R}^d$.
- To estimate this submanifold from $\{x_i'\}_{i=1}^M$, we introduce a latent representation $z \in \mathbb{R}^p$ with $p \ll d$, and a mapping $\phi : \mathbb{R}^p \to \mathbb{R}^d$, such that the pushforward measure under ϕ of $z \sim N(0, I_p)$ is close to the empirical distribution of $\{x_i'\}_{i=1}^M$.
- Given ϕ, the likelihood $\mathrm{p}(y|z) = p_{y|x}\,(y|\phi(z))$. We can then easily derive the posterior $\mathrm{p}(z|y) \propto p(y|z)p(z)$ and benefit from greatly reduced dimensionality.
- The posterior $\mathrm{p}(x|y)$ is simply the pushforward measure of $z|y$ under ϕ.

Estimating ϕ

- There are different learning approaches to estimate ϕ, e.g., variational auto-encoders (VAE)s and generative adversarial networks (GAN)s.

- We use a VAE, i.e., we assume x is generated from the latent variable z as follows:

$$z \sim N\left(0, I_p\right), \qquad x \sim p(x|z)$$

- As $p(x|z)$ is unknown, we approximate it by a parameterized distribution $p_\theta(x|z)$ defined by a neural network (the decoder). This typically has form $N(\mu_X(z), \sigma_X^2(z)\ I)$.

- The objective is to set θ to maximize the marginal likelihood $p_\theta(x'_1, \ldots, x'_M)$. This is usually computationally intractable, so we maximize a lower bound instead.

See Kingma P. et at. "Auto-encoding variational Bayes." (2013) *arXiv:1312.6114*.

Variational Auto-Encoders

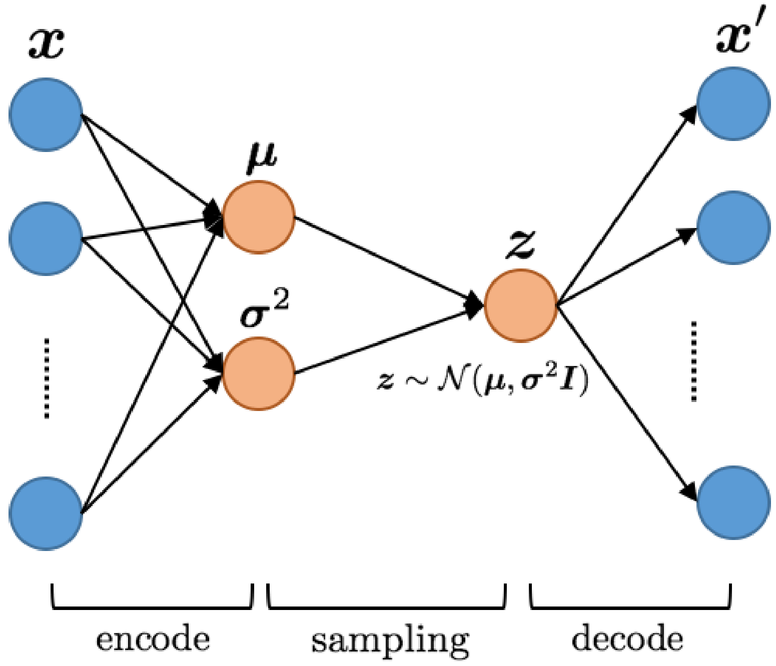
- The **variational lower bound** on the log-likelihood is given by

$$\log p_\theta(x|z) \geq E_{q_\theta}[\log p_\theta(x|z)] - D_{KL}\big(q_\varphi(z|x) \| p_\theta(z)\big)$$

- $q_\varphi(z|x)$ is an approximation of $p_\theta(z|x)$, parameterised by a neural network (the encoder). Typically $N(\mu(x), \sigma^2(x))$.

- In maximising the variational lower bound, the encoder and decoder are trained simultaneously.

- We use the decoder mean to define ϕ, i.e., x = $\mu_X(z)$.



Bayesian computation

- To compute probabilities and expectations for $\mathrm{z}|\mathrm{y}$ we use *a preconditioned Crank Nicolson algorithm,* which is a slow but robust Metropolized MCMC algorithm.
- For additional robustness w.r.t. multimodality, we run M+1 parallel Markov chains targeting $\mathrm{p}(z), \mathrm{p}^{\frac{1}{M}}(z|y), \mathrm{p}^{\frac{2}{M}}(z|y), \ldots, \mathrm{p}(z|y)$, and perform randomized chain swaps.
- Probabilities and expectations for $\mathrm{x}|\mathrm{y}$ are directly available by ϕ-pushing samples.
- We are developing fast gradient-based stochastic algorithms. Naïve off-the-shelf implementations are not robust and have poor theoretical guarantees in this setting.

Cotter, Simon L., et al. "MCMC methods for functions: modifying old algorithms to make them faster." *Statistical Science* (2013): 424-446.

Previous works

- Our work is closely related to the Joint MAP method of M. González et at. (2019) arXiv:1911.06379, which considers a similar setup but seeks to compute the maximiser of $\mathrm{p}(x, z|y)$ by alternating optimization.

- It is also related to works that seek to learn $p\left(x \middle| y, \{x_i'\}_{i=1}^{M}\right)$ by using a GAN, e.g., Adler J et al. (2018) arXiv:1811.05910 and Zhang C et al. (2019) arXiv:1908.01010.

- More generally, part of a literature on data-driven regularization schemes; see Arridge S, Maass P, Oktem O, and Schönlieb CB (2019) Acta Numerica, 28:1-174.

- Underlying vision of Bayesian imaging methodology set in the seminal paper Besag J, Green P, Higdon D, Mengersen K (1995) Statist. Sci., 10 (1), 3--41.

Outline

- Introduction
- Proposed method
- **Experiments**

Experiments

- We illustrate the proposed approach with three imaging problems: denoising, deblurring (Gaussian blur 6x6 pixels), and inpainting (75% of missing pixels).

- For simplicity, we used the MNIST dataset (training set 60,000 images, test set 10,000 images, images of size 28x28 pixels). In our experiments we use approx. 10^5 iterations and 10 parallel chains. Computing times of the order of 5 minutes.

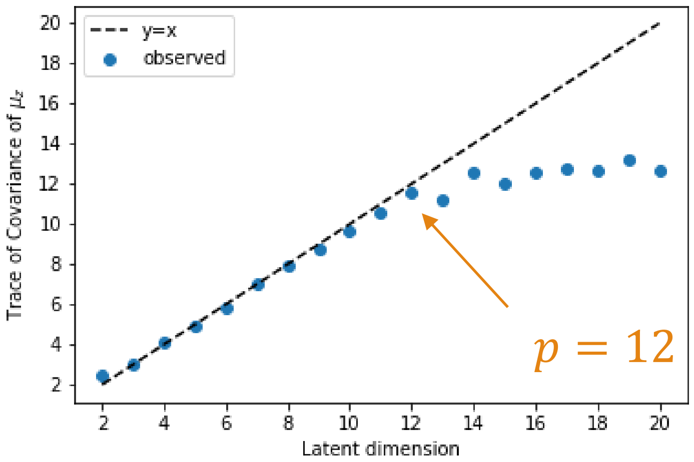
- We report comparisons with J-MAP of Gonzales et al. (2019) and plug-and-play ADMM of Venkatakrishnan (2013) using a deep denoiser specialised for MNIST.

S.V. Venkatakrishnan, C.A. Bouman, and B. Wohlberg, Plug-and-Play Priors for Model Based Reconstruction, GlobalSIP, 2013.

Dimension of the latent space

- The dimension of the latent space plays an important role in the regularization of the inverse problem and strongly impacts the quality of the model.

- We can easily identify suitable dimensions by looking at the empirical marginal $\mathrm{p}(z)$ obtained from encoded training examples, e.g., we look at the trace of $\mathrm{cov}(z)$.



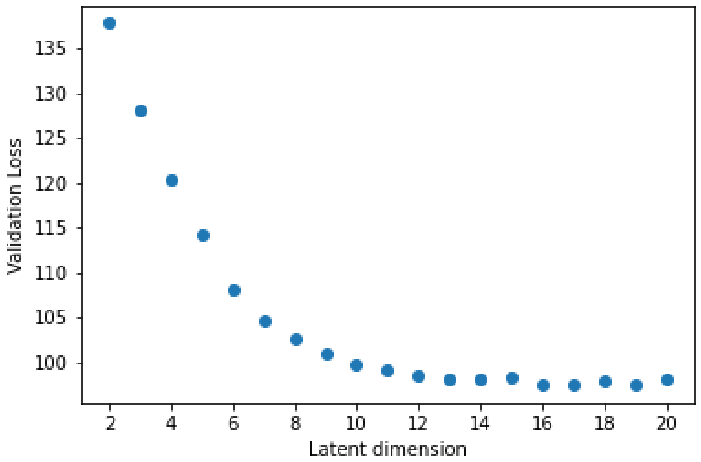


Image denoising

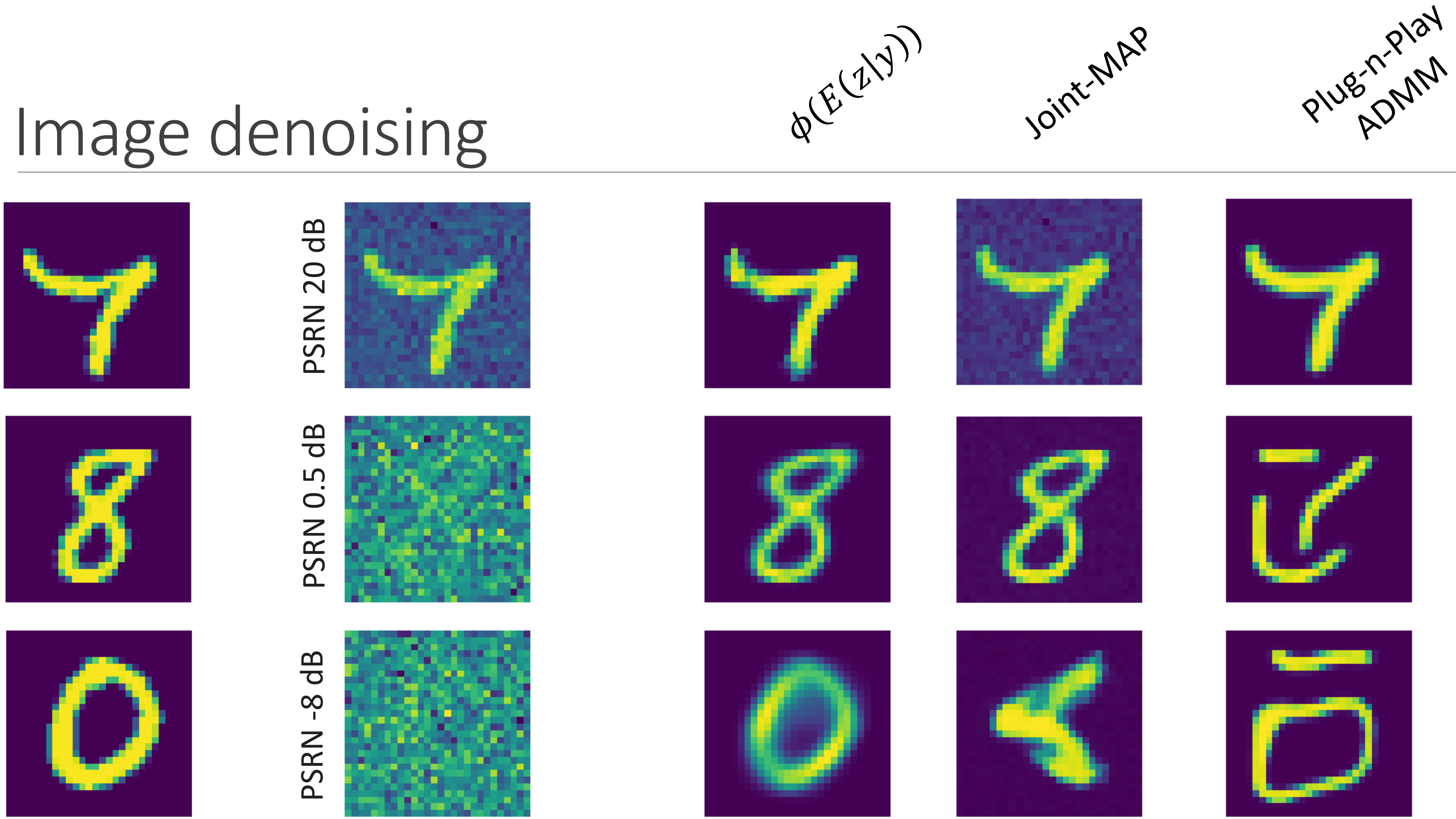


Image deblurring

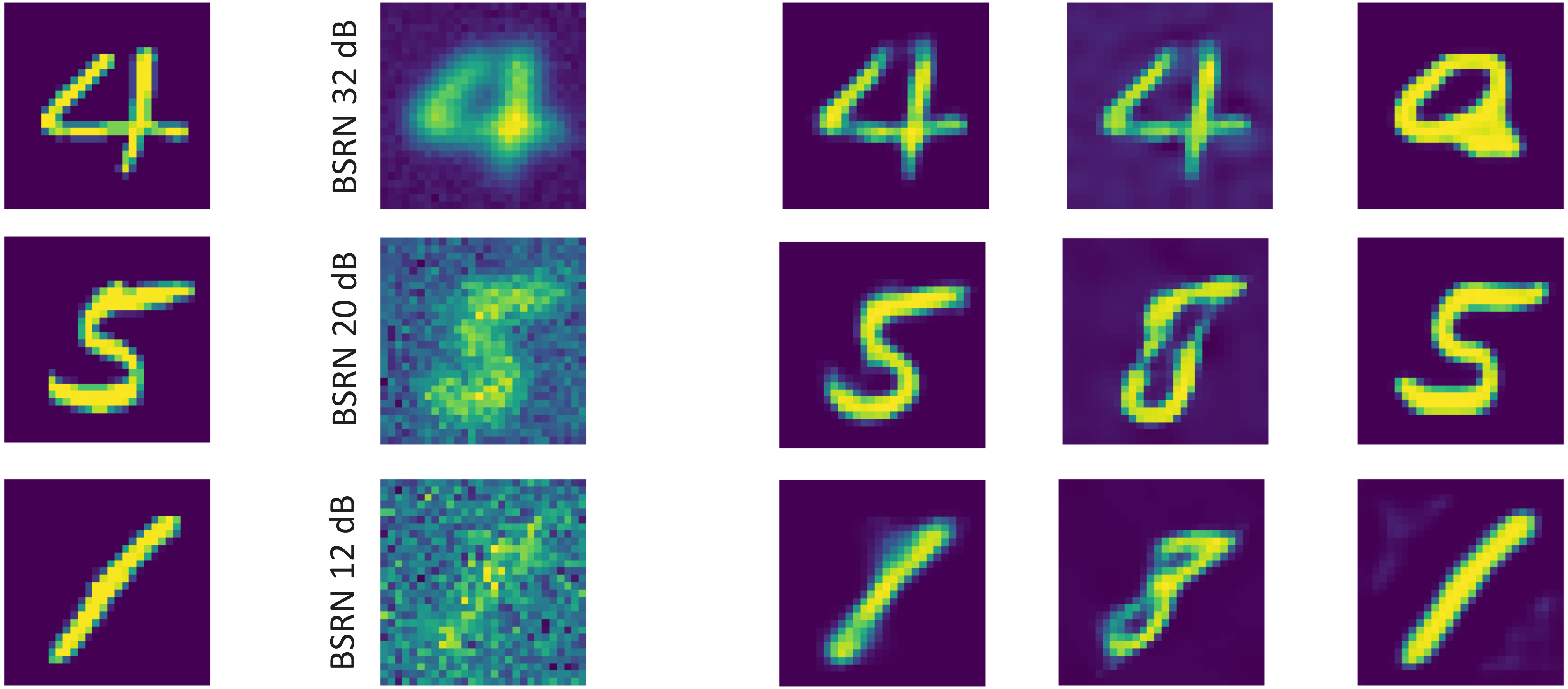
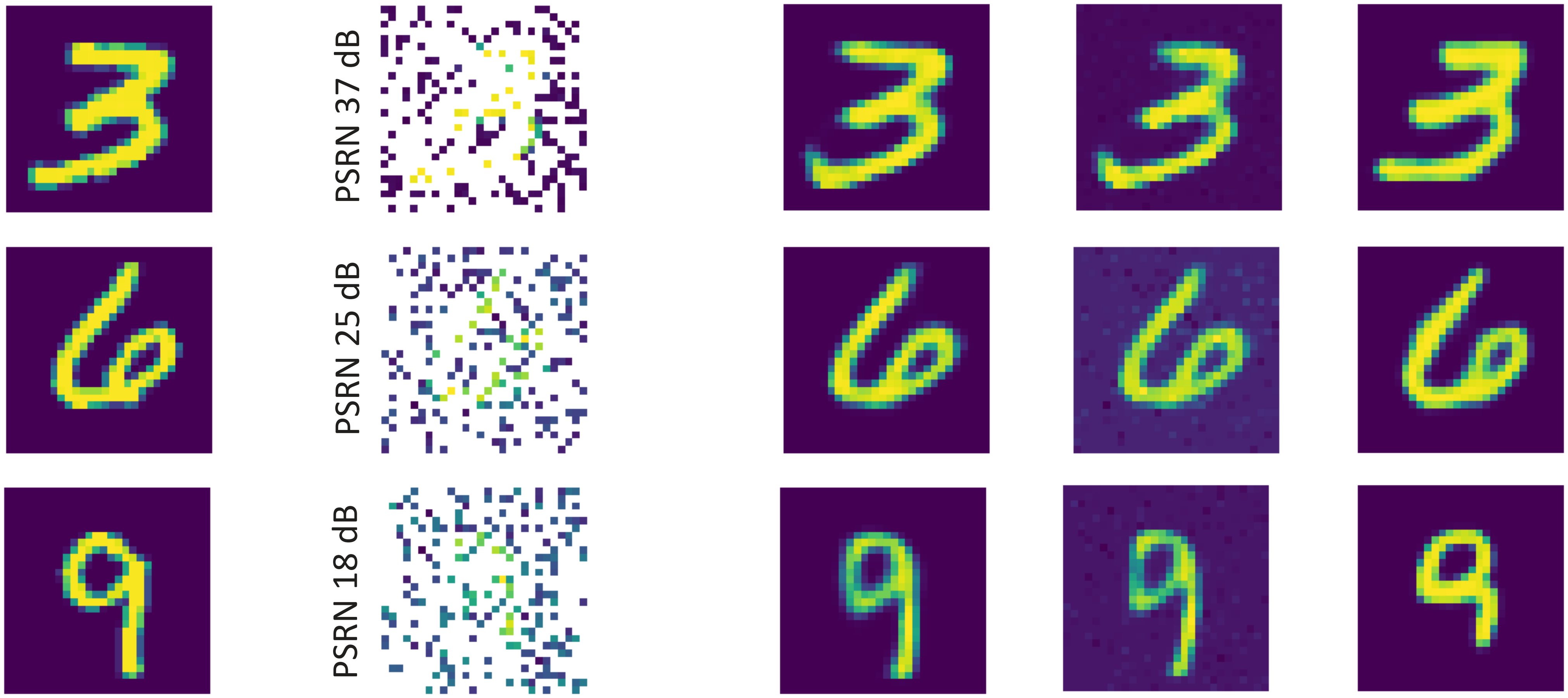


Image inpainting

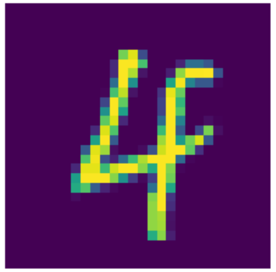


Uncertainty visualization

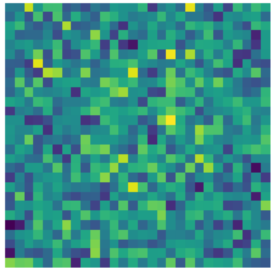
- Inverse problems that are ill-conditioned or ill-posed typically have high levels of intrinsic uncertainty, which are not captured by point estimators.
- As a way of visualizing this uncertainty, we compute an eigenvalue decomposition of the (latent) posterior covariance matrix to identify its two leading eigenvectors.
- We then produce **a grid of solutions** across this two-dimensional subspace.

Visualizing uncertainty

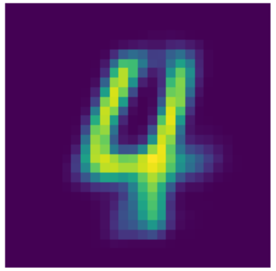
Truth

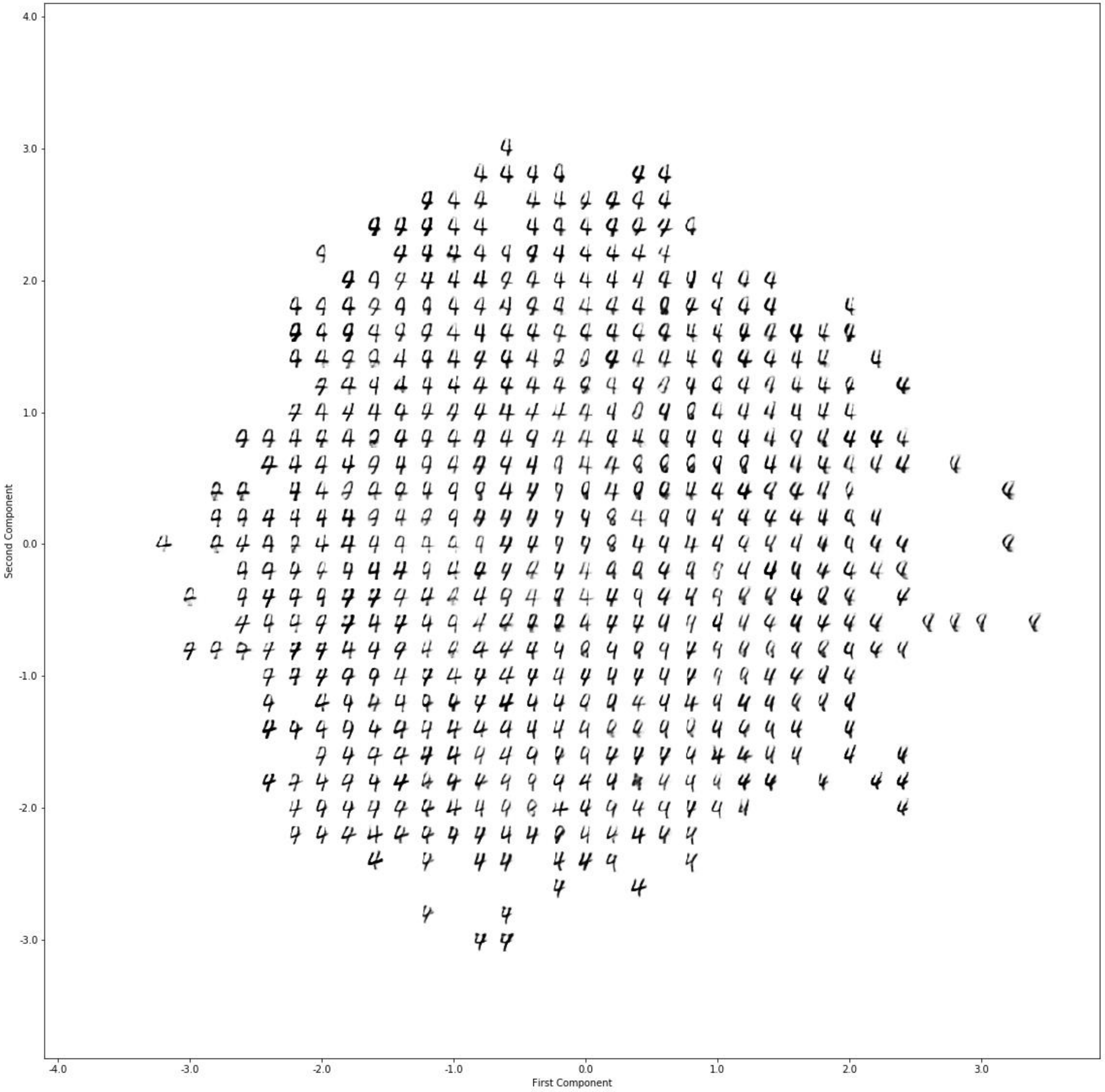


Noisy Observation



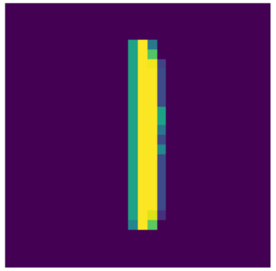
$\phi(E(z|y))$



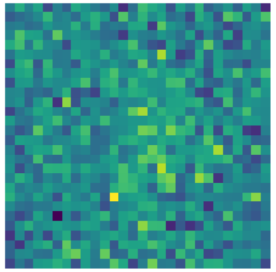


Visualizing uncertainty

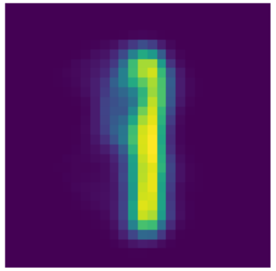
Truth

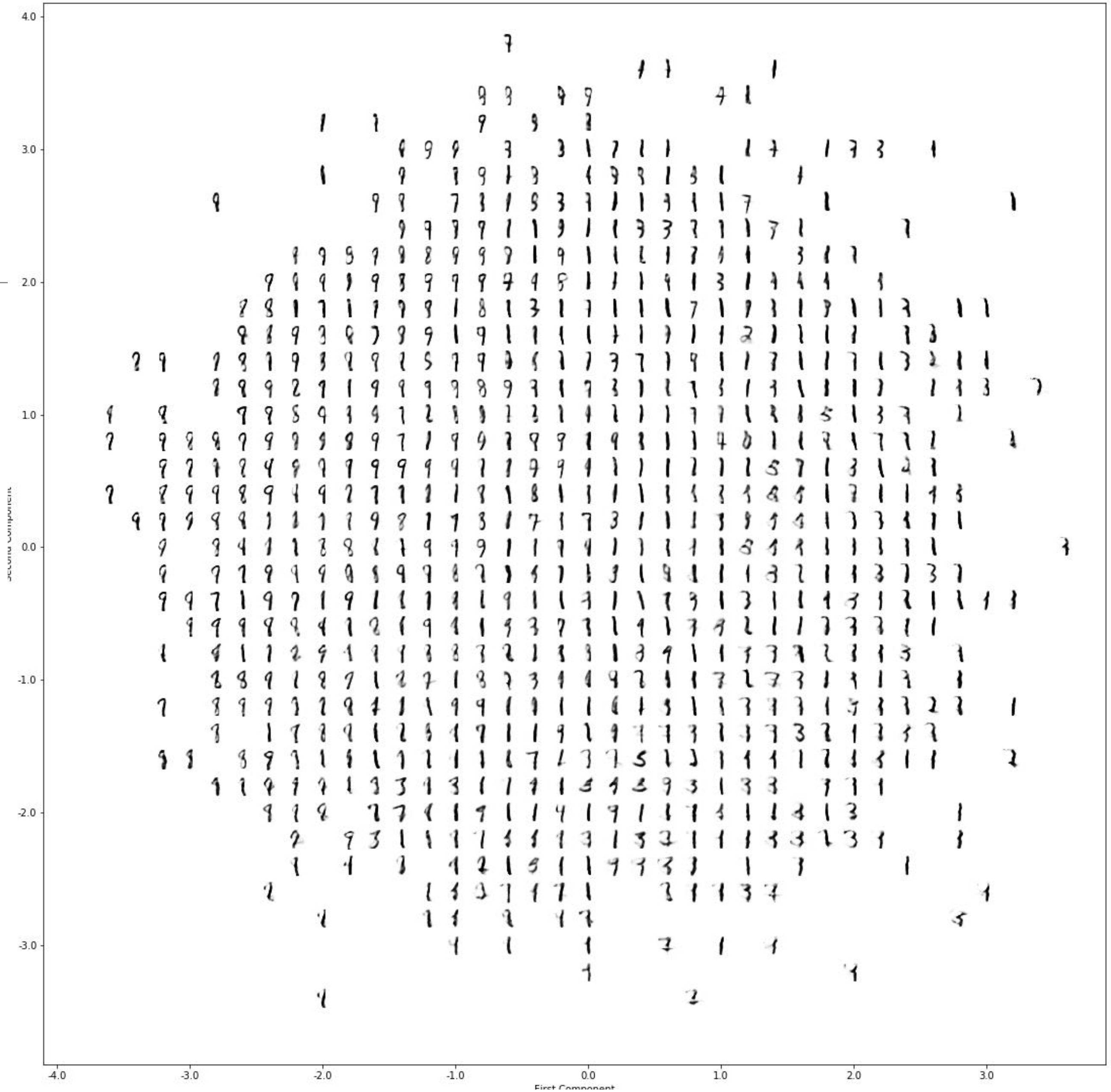


Blurred &
Noisy
Observation



$\phi(E(z|y))$





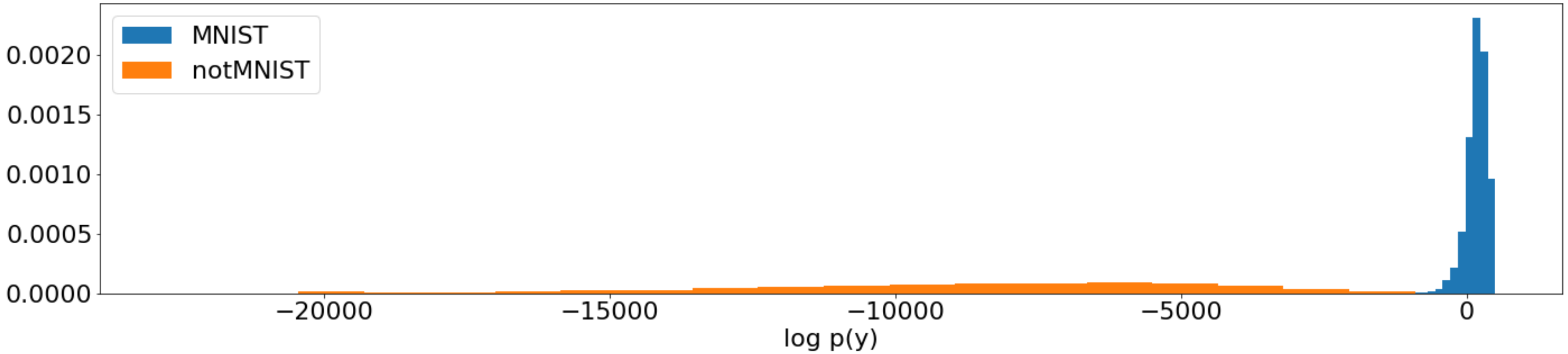
Warning of severe model misspecification

- Data-driven priors strongly concentrate probability mass in specific regions of the solution space.
- When used appropriately, then can deliver impressive results.
- However, **data-driven priors easily override the likelihood** and can lead to severe model misspecification when the truth differs significantly from the training examples.

Model misspecification testing

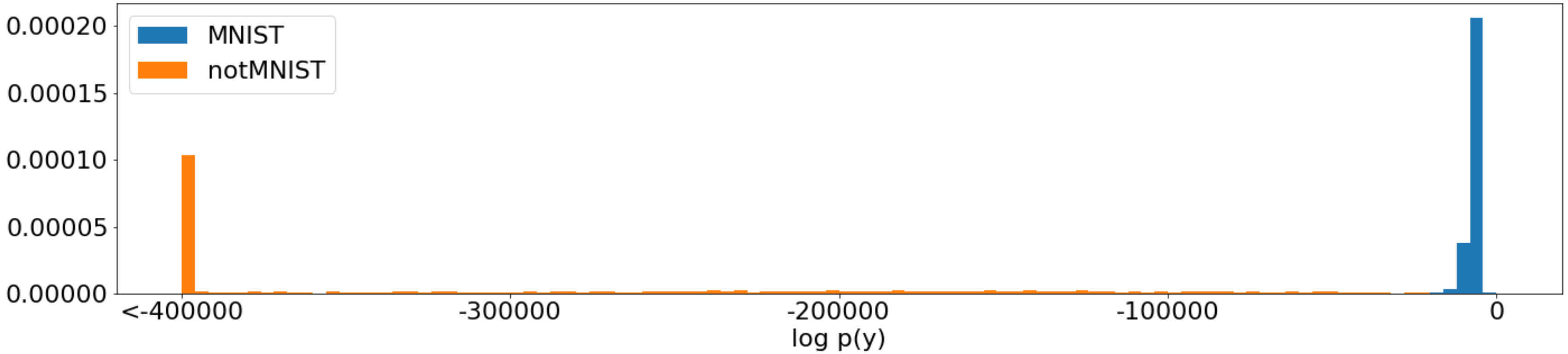
- When using data-driven priors it is important to perform model misspecification diagnosis tests.
- In the spirit of the Neyman-Pearson Lemma, we construct a statistical test based on the marginal likelihood $p(y)$ that we estimate from the chains.
- We compute this statistic for synthetic observations generated from the training dataset to establish the null distribution.
- This then allows misspecification testing and reporting p-values for observed data.

Model misspecification test



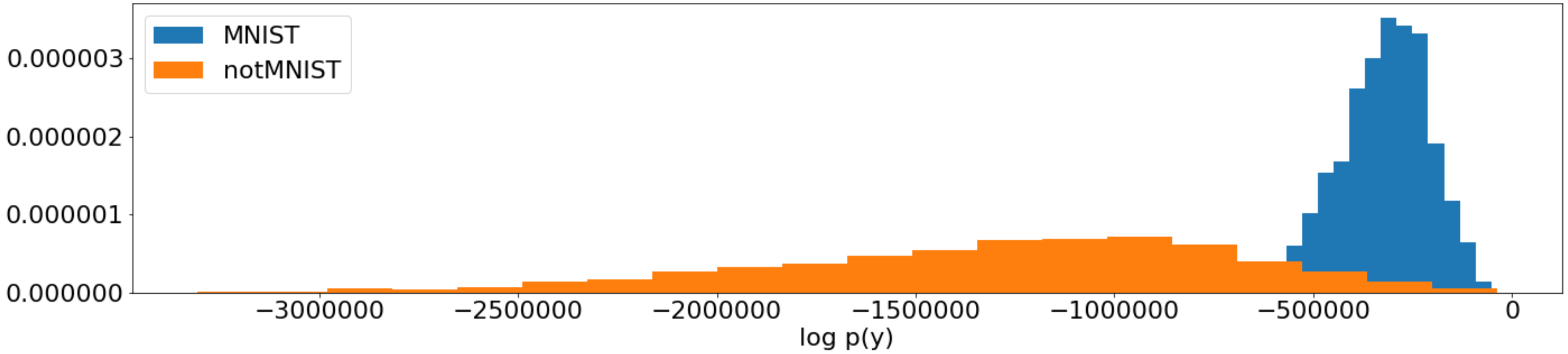
Denoising experiment ($\sigma = 0.1$). Reject null hypothesis (`MNIST`) with 99% confidence, and average power of 99.6% for `NotMNIST` dataset.

Model misspecification test



Deblurring experiment ($\sigma = 0.01$). Reject null hypothesis (`MNIST`) with 99% confidence, and average power of 99.8% for `NotMNIST` dataset.

Model misspecification test



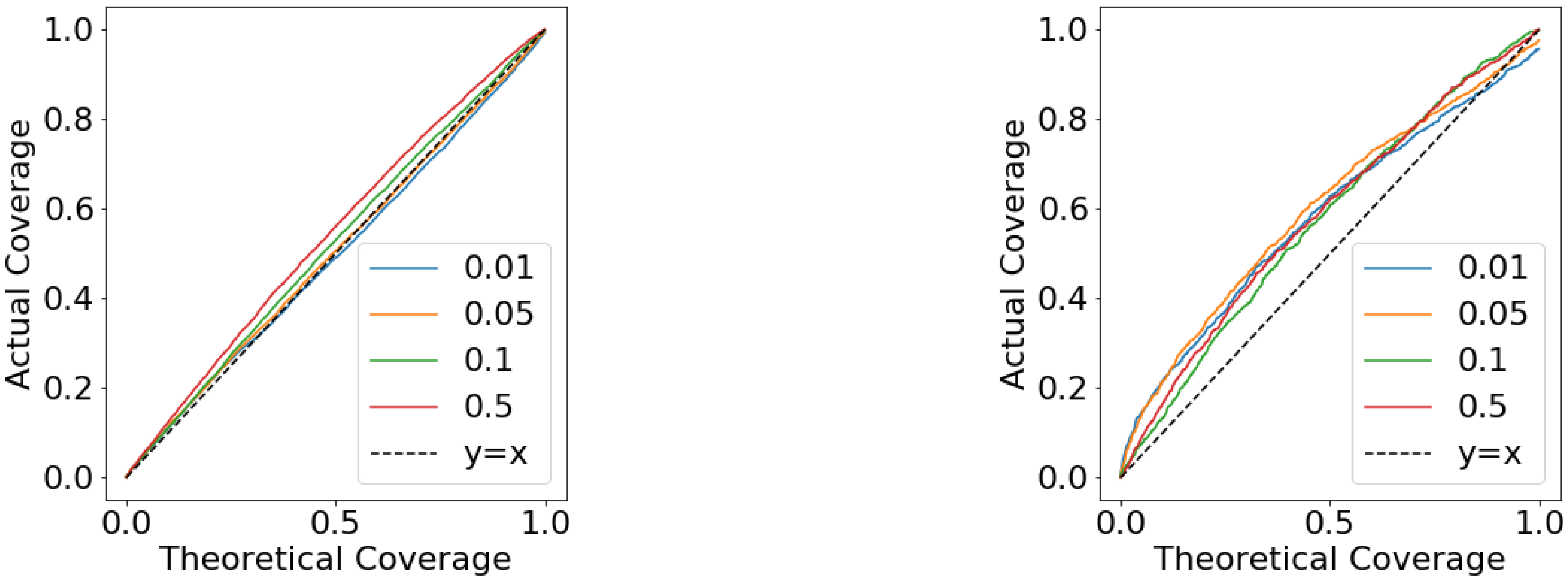
Inpainting experiment ($\sigma = 0.01$). Reject null hypothesis (`MNIST`) with 99% confidence, and average power of 88.5% for `NotMNIST` dataset.

Frequentist coverage of Bayesian probabilities

- Are the Bayesian probabilities reported by our models accurate in a frequentist sense? i.e., are they in agreement with empirical averages from repeated experiments?

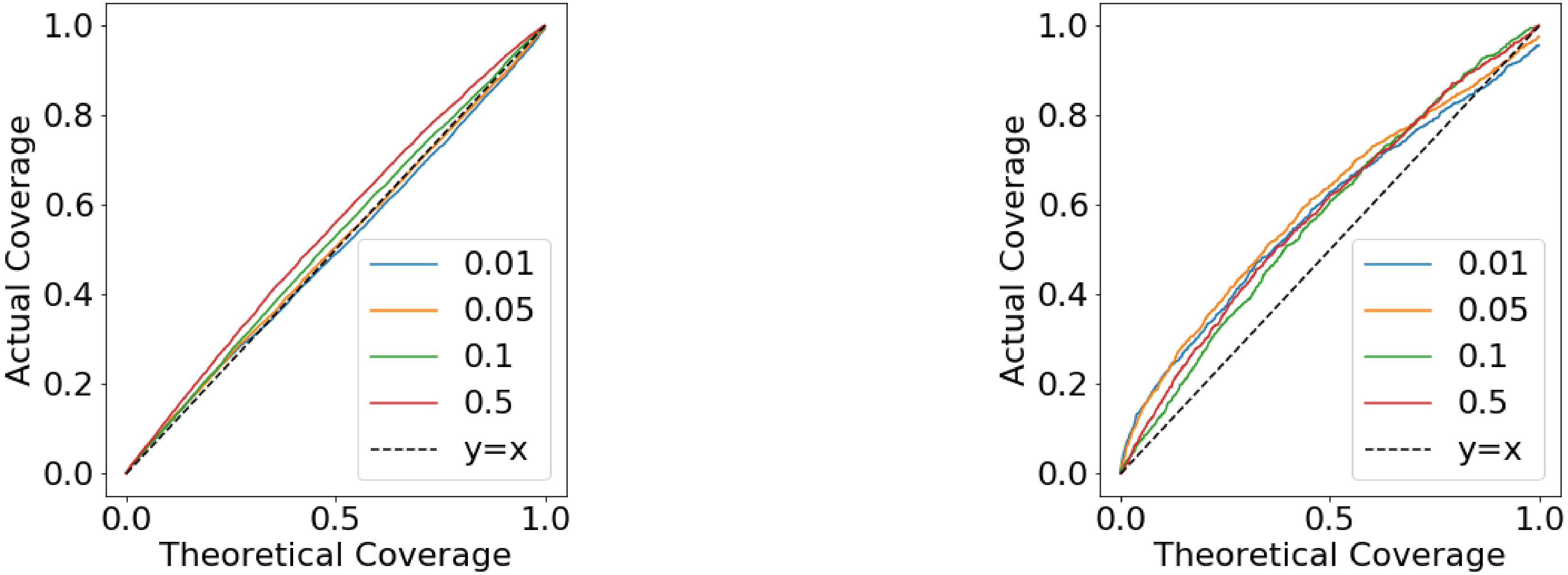
- We explore this question by repeating experiments with 1,000 test images and measuring the empirical probabilities that the truth is within the $(1-\alpha)\%$ highest posterior density credible region.

Frequentist coverage of Bayesian probabilities



Coverage properties for denoising (left) and inpainting (right) for different noise levels (pixel dynamic range [0,1]).

Frequentist coverage of Bayesian probabilities



To the best of our knowledge, this is the first example of a Bayesian model with accurate frequentist coverage properties in an imaging setting, albeit with a very simple image dataset!

Thank you!