Warwick Summer School on Complexity and Inference - Lecture I: Inverse Problems

Dalia Chakrabarty[1]

[1]University of Warwick,
Department of Statistics

May 13, 2012

Outline of this lecture

1. The inverse appoach in complex systems.
    - non-parametric learning when: many, coupled degrees of freedom + non-linear physics (defining correlations, evolution).
    - Bayesian mindset.
    - inverse problems and a personal classification - model structure in each case.
    - relevance to quality and quantity of data, sparsity.
    - something to think about.
2. Inference - optimisation, Monte Carlo, MCMC.
3. Case studies - real deprojection problem; inverse learning in dynamical systems.

Why does nonparametric modelling suggest itself readily when the system manifests complexity?

⋆ **Heterogeneous nature of underlying phase or state space - multimodal density function; isolated, sharply declining modes.**

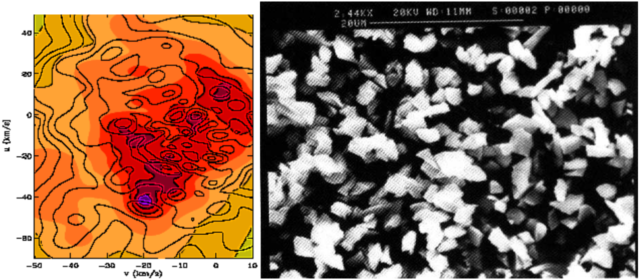


Figure: *Left:* Stellar velocities recorded in the j^{th} **S** cell are used to estimate $f_j(U, V)$ (overlaid in solid black contour lines over) $f_0(U, V)|\mathcal{D}$ (from Chakrabarty, 2007). *Right:* SEM image of real-life material.

Why does nonparametric modelling suggest itself readily when the system manifests complexity?

⋆ **Heterogeneous nature of underlying network structure**

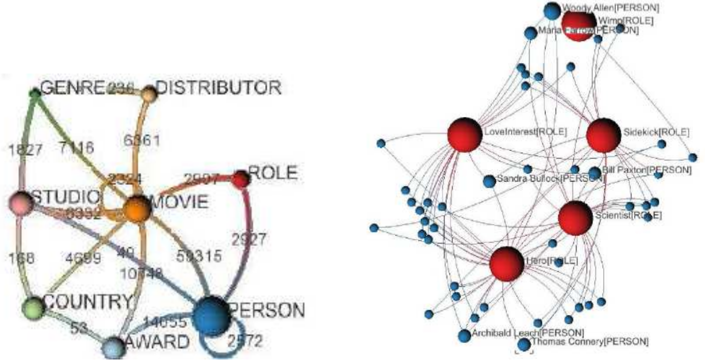


Figure: Heterogeneous complex network using archived movie data bases (Zeqian Shen, Kwan-liu Ma, and Tina Eliassi-Rad, 2006). Network shows eight “node types” - person, movie, role, studio, distributor, genre, award and country. There are 35,312 nodes and 108,212 links. The data is noisy and contains missing relationships.

Parametric models - *aka* analytical approximations - for an average description insufficient?

Mean field approach is a shortcoming when system manifests

- non-uniform structural correlation distribution.
- unequal dynamical correlations (Pair approximation theories-Newman, 2010)
- non-uniform distribution of states over degree-k nodes.

When is an average description sufficient?

Gleeson et. al, Phys Rev E, 2012 - explore pertinence of Mean Field theory in the context of real dynamical systems on networks $\rightsquigarrow$ MF theories are relatively more accurate when most nodes have high-degree neighbours. Barabasi, Albert & Jeong (1999) suggest a MF theory for the scale-free random networks. Lacroix et. al (2011) treat nucleons as independent particles that are playing in a MF potential. They say this model works because nucleons inside the nucleus experience widely different potential from nucleons inside the nucleus. But Cook (2010) contradicts.

- when heterogeneity in dynamical correlation distribution is low.
- when modularity can be assumed.
- when local and global effects are well decoupled.
- ......

Wish list ...

lots of parameters in the model that is free-form . . . parameters to be learnt from the data

- under-abundance of data compared to parameters.
- constricted information - renders large allowed solution set.
- inference in high-dim is hard.

- Introduce correlation among parameters . . . invoke assumptions about topology of phase/state space. Can test for support in data for assumption(s) by designing bespoke statistical tests of hypothesis.
- Improve information content - scrutinise system geometry, extra information from theory, elicit from literature, the *p*-word.
- Improve inference.

Input care into modelling - ease inference if possible

Epistimeologically speaking, inference can be “probabilistic” or “relativistic”

The probability of it raining tomorrow is 0.5

The probability of it raining tomorrow$\in [0,1]$. Out of these, I, as the observer, think that there is more chance it raining rather than staying dry tomorrow $|$ the information I possess.

With respect to the observer.
Helps contract solution space.

The Bayesian mindset and Bayes rule

All are random varibles distributed as a probability distribution - prior. $\Pr(A) = 0.298$ **means that the probability that** $A = a$ **is 0.298.**

- Relative to the practitioner.
- Subjectivity diminishes with data.

The Bayesian mindset and Bayes rule

Probability of A,conditional on B is

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{1}$$

Then if $A :=$ “model” and $B :=$ “data”, we get

$$\Pr(model|data) = \frac{\Pr(data|model)\Pr(model)}{\Pr(data)} \tag{2}$$

Bayes Rule provides for a bridge between inverse problems - that which we actually want to solve in Science - and the forward problem (the kind which find easy to deal with).

Examples ⟶

Inference

Statistical inference, in the context of inverse nonparametric learning of model parameters of a complex system could be

- deterministic if there is “no shortage” of data $\rightsquigarrow$ trivial soln; non-unique/non-existent solutions otherwise.
- In lieu of so much information about the system, practitioner brings in “her own” information - Bayesian. In the presence of “lots of parameters”, inference gets hard.
- If system is itself probabilistic, inference has to account for inherent variation in model parameters.
- All this notwithstanding, variation or noise inherent in measurements acknowledgemed in learning of model parameters.

General inverse problem - statement; finite dimensional sytems

$$\mathbf{I} = \mathcal{P}[\rho] + \varepsilon \tag{3}$$

where data $I \in \mathcal{D}$: $\mathcal{D}$ is the Banach space of functions $\mathbf{I} : \mathcal{U} \subset \mathbb{R}^m \longrightarrow \mathbb{R}$, while the unknown function $\rho \in \mathcal{H}$, $\mathcal{H}$ is the Banach space of functions $\rho : \mathcal{V} \in \mathbb{R}^n \longrightarrow \mathbb{R}$. Here $\mathcal{U}$ and $\mathcal{V}$ are closed intervals in $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, defined by problem at hand.

For the operator $\mathcal{P} : \mathcal{H} \longrightarrow \mathcal{D}$, $n > m$ leads to a fundamentally ill-posed problem (Tarantola 2004, Bereto 1998, Hansen 1998, Cotter et. al 2010, Stuart 2010).

ε is the measurement noise, the distribution of which, may be known, but may or may not be Gaussian!

Ill-posedness

$\mathcal{P}:\mathbb{R}^m \longrightarrow \mathbb{R}^n$, $m \leq n \Longrightarrow$ learning of ρ is fundamentally ill-posed in the Hadamard sense.

- solution $\rho = \mathcal{P}^{-1}(I)$ does not exist.
- solution non-unique.
- solution not continuous with variations in data $\longrightarrow$ ill-conditioned.

In real-life, many interesting inverse problems may not be well-posed in this sense but may be tractable by bringing in *extra information* into the model or if the model is sparse "enough" (Donoho & Tanner, 2005)

, while maintaining **P** as square.

Sparse density functions

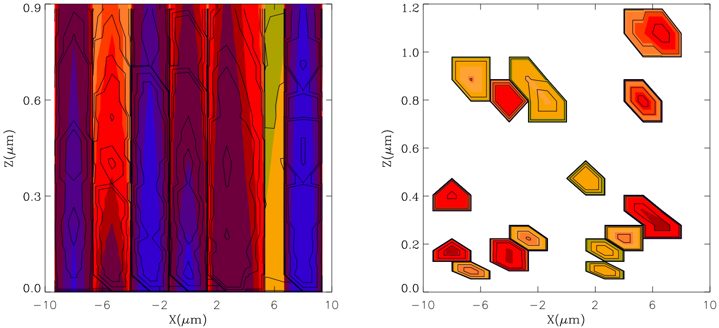


Figure: Simulated images of Copper-Tungsten alloy - sparse (left) and dense (right) density structure.

Two types of inverse problems - a personal classification

$\mathcal{P}$ is known. Typically, then $\mathcal{P}$ is a projection operator. The problems could include learning of density in n-D from inversion of $n-1$-D images, blind deconvolution problems. $\mathbf{I}$ and $\boldsymbol{\rho}$ can be vectors or matrices. Traditionally these are handled using inverse-Radon transforms which fails if data is not Holder continuous or if viewing angle measurements are limited/unavailable. Deprojection is in general ill-posed even when noise is absent. $\mathcal{P}$ is linear or non-linear.

$\mathcal{P}$ is unknown. Data is unknown functional of model parameters. This description lends itself to many problems in which we want to learn vector/matrix of model parameters $\boldsymbol{\rho}$ given data vector/matrix $\mathbf{I}$ which is then an unknown functional of unknown $\boldsymbol{\rho}$. Usually, in these kinds of problems, extra information is needed (eg. the unknown functional is a realisation of a known stochasticprocess); when ρ is a function of model parameters, discretisation has to be invoked.

Linear inverse problems - integral equation formulation

Fredholm equation of the 1^{st} kind

$$i = \int_a^b p(i,x)\rho(x)dx \tag{4}$$

One example of Volterra equation of the 1^{st} kind - in astronomy

$$i(x,y) = p(x,y) * \int_0^{(h(x,y)} \rho(x,y,z)dz \tag{5}$$

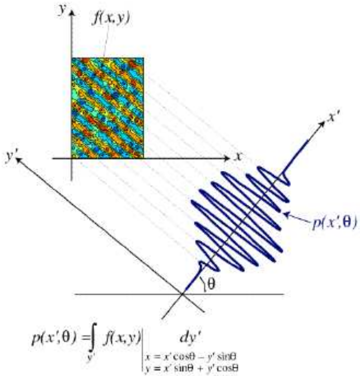
Tricomi book.

$m < n$ - Radon Transform

$f(\mathbf{x}) = f(x, y)$ be a continuous function bounded in $\mathbb{R}^2$. Radon transform of $f(\mathbf{x})$ is the projection of this function on the line segment L:

$$\begin{aligned} \mathcal{R}f(L) &= \int_L f(\mathbf{x})du \qquad (6) \\ u &\text{ is orthogonal to } L \\ p(x', \theta) &= \int_{y'} f(x, y)|_{(x,y)=\mathrm{R}_\theta(x',y')}dy' \end{aligned}$$

2-D RT


y
f(x,y)
y'
x
x'
p(x',θ)
θ
p(x',θ) = ∫_{y'} f(x,y)|_{x = x' cosθ − y' sinθ, y = x' sinθ + y' cosθ} dy'

Inverse RT

- Inverting: RT at a given θ corresponds to the inverse Fourier Transform of a slice taken at the same angle θ in the Fourier space.
- Applications - find f, given $\{\theta_i, p(\cdot, \theta_i)\}_{i=1}^{N}$ ... inverse Fourier methods, (Deans, 1983) - RT in
- Inversion of RT in n-D involves the $n-1^{th}$ x'-derivative - numerically unstable for many real-life samples.

When θ_i unavailable -
$\star$ nature of the problem (astronomy, biology),
$\star$ logistical shortcomings (material science)-
RT is fundamentally inapplicable

Projection and deprojection

If $\mathcal{P}$ is a projection operator in $\mathbf{I} = \mathcal{P}[\rho] + \varepsilon$ where data I is marked by ε and the unknown parameters are in ρ, then in the low noise limit, i.e. $\lim \parallel \varepsilon \parallel \longrightarrow$ **0,**

$$\begin{aligned} \mathbf{P}^{+}\mathbf{I} &= \rho \quad \textit{from} \qquad (7) \\ \mathbf{P}^{\dagger}\mathbf{I} &= \mathbf{P}^{\dagger}\mathbf{P}\rho \\ (\mathbf{P}^{\dagger}\mathbf{P})^{-1}\mathbf{P}^{\dagger}\mathbf{I} &= \rho \end{aligned}$$

where $\mathbf{P}^{+} := (\mathbf{P}^{\dagger}\mathbf{P})^{-1}\mathbf{P}^{\dagger}$ is the Moore-Penrose inverse of the matrix representation of the projection operator $\mathcal{P}$ (C.R.Rao et. al 1971 for generalised inverses, Avner Friedman for projection matrices, Roger & Charles for matrix analysis). It exists if $(\mathbf{P}^{\dagger}\mathbf{P})$ is invertible.

Projection and deprojection

- The projection operator is idempotent: $\mathbf{P}^2 = \mathbf{P}$.
- The operator $\mathcal{P}$ is a projection along its null space $\mathcal{N}(\mathcal{P})$ onto its range $\mathcal{R}(\mathcal{P})$. $\mathcal{R}(\mathcal{P}) \perp \mathcal{N}(\mathcal{P}) \Longleftrightarrow$ **P** is Hermitian. Then projection is orthogonal, else oblique. $\parallel \mathcal{P} \parallel = \parallel \mathcal{I} - \mathcal{P} \parallel \; \forall \mathcal{P}$ (Szyld 2006, refs therein).
- If **P** is an orthogonal projection, real matrix **P**, **P** is self-adjoint, square symmetric matrix.
- Orthogonal projection matrices are then diagnolisable - $\mathbf{P}^+$ then exists.
- The Moore-Penrose inverse when it exists, is unique $\Longrightarrow \rho = \mathbf{P}^+\mathbf{I}$ is unique, if $\mathbf{P}^+$ exists. From an inference point of view,
  $\star$ dimensionality of range of $\mathcal{P}$ = dimensionality of data space $\Longrightarrow$ no **I** outside range of $\mathcal{P}$ and **P** is square - soln. exists and unique.

Projections in a space of functions

If $\mathcal{P}$ is a projection in the Banach space $\mathcal{H}$, we can represent $\mathcal{D}$ as the direct sum of the two orthogonal subspaces, as in $\mathcal{H} = \mathcal{R}(\mathcal{P}) \oplus \mathcal{R}(\mathcal{I} - \mathcal{P})$, where $\mathcal{I} - \mathcal{P}$ is also a projection.

In the case of infinite-dimensional space onto which a continuous projection happens, $\mathcal{R}(\mathcal{P})$ must be a closed subspace.

A subspace $\mathcal{D}$ of $\mathcal{H}$ can be the range of a contractive projection in $\mathcal{H}$, under specific conditions (Lindberg 72, Villanueva 1991), leading to a contractive projection onto $\mathcal{D}$ defined as a real-valued continuous function on the extreme points.

Condition number

What is the fractional uncertainty in the learnt $\hat{\rho}$, $\frac{\| \boldsymbol{\rho} - \hat{\boldsymbol{\rho}} \|}{\| \boldsymbol{\rho} \|}$,

given noise in the data $\frac{\delta \mathbf{I}}{\mathbf{I}}$, (where $\mathbf{I} + \delta\mathbf{I} = \mathbf{P}[\hat{\rho}]$)?

$$\begin{aligned} \| \mathbf{I} \| &= \| \mathbf{P}\boldsymbol{\rho} \| \leq \| \mathbf{P} \| \| \boldsymbol{\rho} \|, \\ \implies \frac{\| \boldsymbol{\rho} - \hat{\boldsymbol{\rho}} \|}{\| \boldsymbol{\rho} \|} &\leq \frac{\| \mathbf{P} \| \| \mathbf{P}^{-1} \| \| \delta\mathbf{I} \|}{\| \mathbf{I} \|} \\ \text{Also} \frac{\| \boldsymbol{\rho} - \hat{\boldsymbol{\rho}} \|}{\| \boldsymbol{\rho} \|} &\geq \frac{\| \delta\mathbf{I} \|}{\| \mathbf{P} \| \| \boldsymbol{\rho} \|} \geq \frac{\| \delta\mathbf{I} \|}{\| \mathbf{P} \| \| \mathbf{P}^{-1} \| \| \mathbf{I} \|}. \end{aligned} \tag{8}$$

Here the condition number is $\kappa = \| \mathbf{P} \| \| \mathbf{P}^{-1} \|$.

How to solve the inverse problem?

- Inversion of a high-dimensional $\mathbf{P}^{(n\times n)}$ matrix is cost-intensive with the computational complexity varying from $O(n^3)$ to approximately $O(n^{2.38})$ with different algorithms - not doable if placed within an iterative scheme.
- So, in the n-th step of the forward model, compute the norm $\| \mathbf{P}\hat{\rho}_n - \mathbf{I} \| \rightsquigarrow$ the likelihood; $\forall\ n$.
- In a Bayesian framework, likelihood $\mathcal{L}(\mathbf{I}|\boldsymbol{\rho})$ is used in conjunction with prior $\pi_0(\boldsymbol{\rho})$, to define the posterior probability of the model parameters, given the data ($\pi(\boldsymbol{\rho}|\mathbf{I})$), as per Bayes rule.

  $\star$ Measurement noise has to be built into all this - maybe in the definition of the likelihood, or by convolving posterior with noise distribution.

  $\star$ Sample from the posterior using MCMC techniques.
- In a frequentist framework, likelihood is optimised, in presence of imposed constraints, *aka* regularisation.

When the mapping to the data space is unknown

This is a bigger challenge but perhaps encountered more frequently in Science, especially when complexity manifests in system behaviour.

Data is considered

- **unknown function of system parameter vector ρ: $\mathbf{I} = \xi(\rho)$, example, (1) observed image data in distant galaxy as an unknown function of the unknown vector of gravitational mass density at different spatial locations in the system, (2) house prices observed at different time points, as an unknown function of the sought vector of pre-identified economic parameter at these times.**
- **unknown functional of $\rho(\mathbf{x})$ where X is a system variable: $\mathbf{I} = \xi[\rho(\mathbf{x})]$, example, partial phase space information - available for a sample of particles in an autonomous dynamical system - is unknown functional of evolution function $g(\mathbf{w})$, where W is phase space vector.**

Inverse problem

Aim is to invert the equation I $= \xi(\rho)$ to learn ρ, given I. Thus, the problem to solve is inverse. Instead, when knowing a new parameter $\rho^{(new)}$, we want to predict $I^{(new)}$, that is a forward problem.

Interested in learning $\rho(\cdot)$ - not necessarily in learning the form of $\xi[\cdot]$. The suggestion is to compute posterior probability $[\rho|\mathbf{I}]$ and sample from this posterior, to learn ρ.

The Bayesian approach

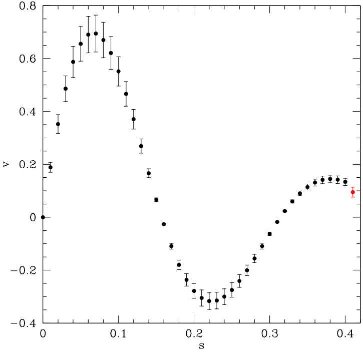
Bayesian approach is a natural choice for such problems because

- when too much is unknown, we can bring in information into the model in the form of priors. Constraints can be built into the prior structure. In lieu of “enough” data, solution is prior-driven (Stuart 2010).
- it readily allows for quantification of uncertainties in learnt parameters.
- non-linear optimisation in high-dimensions is in general much less efficient than MCMC techniques.

Supervised learning

Thus, we are discussing learning of model parameters, as supervised by some training data.

Let $f(\mathrm{s})$ describe the data. Then we want to infer $f(\cdot)$ given the data, i.e. predict the value of measurement v_{n+1} at a new point $\mathrm{s}^{(n+1)}$.



Inference

In the forward model,

- consider $\boldsymbol{\xi}[\cdot]$ to be a realisation from a known stochastic process - popular choices include a Gaussian Process (Rasmussen's book). But we need to be cautious if $\boldsymbol{\rho}(\mathbf{x}) \geq 0$ or if $\boldsymbol{\rho}(\mathbf{x})$ is not continuous. Then compute $\pi[\rho_1, \rho_2, \ldots, \rho_N, \boldsymbol{\Upsilon}|\mathbf{I}]$, where $\boldsymbol{\rho}(x_i) := \rho_i$, $i = 1, \ldots, N$, where we rephrase the aim of wanting to learn the function $\boldsymbol{\rho}(\mathbf{x})$ as wanting to learn vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_N)^T$. In other words, we discretise the relevant closed interval of X.
- Once this posterior is computed, we can marginalise it over $\{\Upsilon_i\}_{i=1}^{M}$. Sample from this marginalised posterior using MCMC.
- Downsides
  1. matrix inversion - less of a problem for smaller data sets
  2. learning smoothness of the chosen process from the data may be difficult - more of a problem for smaller data sets.

Stochastic Processes

A Stochastic Process is a collection of random variables indexed by “time”.

$\star$ discrete time SP $X = \{X_t, t = 0, 1, 2, \ldots\}$; index could include negative values too.

$\star$ continuous time SP $X = \{X(t) | t \in T \subset \mathbb{R}\}$

For $t \in T \subset \mathbb{R}$, $X_t : \Omega \longrightarrow \mathbb{R}$ defined on the probability space $(\Omega, \mathcal{F}, \mathrm{Pr})$, the function $X : T \times \Omega \longrightarrow \mathbb{R}$, is a Stochastic Process. Then $x(t|\omega = \omega_0)$ is a sample function of this SP.

Random variable assigns a real number to outcome $\omega \in \Omega$; SP assigns sample function $x(t, \omega)$ to outcome $\omega \in \Omega$. SP is a distribution of a space of functions. For a given time $t = t_1$, the random function $X_1 = x(t_1, \omega)$ - sample path of X at ω. X_i is a random variable$\longrightarrow$ definition of process. All $\{X_i\}_{i=1}^{N}$ are $i.i.d$.

Poisson Process

Poisson Process is a continuous time, counter process example of an SP, such that number of arrivals in time interval $[a, b]$ is $N(a, b)$ which is distributed as a Poisson distribution.

- Number of arrivals in non-overlapping intervals is independent.
- The inter-arrival times are distributed as an exponential distribution.
- $\Pr[N(t) = n] = \frac{(\lambda t)^n}{n!} \exp(-nt)$, where λ is the rate parameter;
  1. rate parameter is independent of time interval $[a, b]$ for homogeneous Poisson Process - an example of a Levy Process.
  2. rate function $\lambda(t)$ for inhomogeneous Poisson Process.

Gaussian Process

Gaussian Process is an example of a stochastic process (Mehr & McFadden 1965, Shepp 1971, Adler 1981, Abrahamsen 1997, Mackay 1998, Rasmussen & Williams 2006, Santner et. al 2003). $\mathcal{GP}$ can be thought of as a distribution over a function space, such that any finite subset of the data generated by a $\mathcal{GP}$, for any domain, is distributed as a multivariate normal, with specified mean and covariance structures.

$i.i.d\ X_{i}{}_{i=1}^{N}$, in the limit $N \longrightarrow \infty \Longrightarrow$ Gaussian distribution, i.e. $\mathbf{X} := (X_1, X_2 \ldots, X_n)^T$ is dstributed as

$$\frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{\mathbf{detS}}} \exp\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})^2}{2\mathbf{S}}\right] \tag{9}$$

where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{X})$, $\mathbf{S} = [s_{ij}] = \mathbb{E}((x_i - \mu_i)(x_j - \mu_j))$.

Markov Process

Stochastic Process with the property that probability of current event depends only on probability of previous event, but not on other earlier events (the Markov property).

The stochastic process defined as

$$\begin{aligned}\Pr(X_n = x_n | X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, \ldots, X_0 = x_0) &= \\ \Pr(X_n = x_n | X_{n-1} = x_{n-1})\end{aligned} \tag{10}$$

is a Markov Process.

Inference

In the forward model,

- consider $\boldsymbol{\xi}[\cdot]$ to be a realisation from a known stochastic process - popular choices include a Gaussian Process (Rasmussen's book). But we need to be cautious if $\boldsymbol{\rho}(\mathbf{x}) \geq 0$ or if $\boldsymbol{\rho}(\mathbf{x})$ is not continuous. Then compute $\pi[\rho_1, \rho_2, \ldots, \rho_N, \boldsymbol{\Upsilon}|\mathbf{I}]$,
  where $\boldsymbol{\rho}(x_i) := \rho_i$, $i = 1, \ldots, N$, where we rephrase the aim of wanting to learn the function $\boldsymbol{\rho}(\mathbf{x})$ as wanting to learn vector $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_N)^T$. In other words, we discretise the relevant closed interval of X.
- Once this posterior is computed, we can marginalise it over $\{\Upsilon_i\}_{i=1}^{M}$. Sample from this marginalised posterior using MCMC.
- Downsides
  1. matrix inversion - less of a problem for smaller data sets
  2. learning smoothness of the chosen process from the data may be difficult - more of a problem for smaller data sets.

Inference - embed ρ in definition of state space *pdf*

- In the forward model, attempt dimensionality reduction of parameter space by invoking system specific functionals in a discretised model, such that $\boldsymbol{\xi}[\boldsymbol{\rho}(\mathbf{x})] \sim \mathcal{P}(f[\Psi, \Phi_1, \ldots, \Phi_p])$ where Ψ is a function of two vectors $\mathbf{x}$, $\boldsymbol{\rho}$ and $\Phi_j(\cdot)$ is some other function of $\mathbf{x}$, $j = 1, \ldots, p$.
- $f[\cdot]$, is the $p+1$-variate state-space probability density function, defined over the product space given by a direct product of the domains of $\Psi, \Phi_1, \ldots, \Phi_p$. In fact, it is treated in the discretised model as the $(2N) \times (p+1)$-dimensional matrix such that $\mathbf{f_i}^{(2N\times(p+1))} := (\mathbf{f}_{i1}, \ldots, \mathbf{f}_{i(2N)})^T$, $i = 1, \ldots, p$.
- $\mathcal{P}$ is a contractive projection onto a sub-space of $\mathcal{H}$.
- Then likelihood $\mathcal{L} := \mathcal{P}(f[\Psi(\boldsymbol{\rho}, \mathbf{x}), \Phi_1(\mathbf{x}), \ldots, \Phi_p(\mathbf{x})])$.
- Once $\mathcal{L}$ is defined, $\pi(\boldsymbol{\rho}, \mathbf{f_1}, \ldots, \mathbf{f_p} | \mathbf{I})$ is written with the help of the priors, and sampled from using MCMC.

Dimensionality reduction

For our purposes, the choice of nonparametric models implies a very large number of parameters in general. So, dimensionality reduction in inverse nonparametric modelling of complex systems refers to the reduction in number of parameters in the model.

This can be brought about by

- invoking symmetries in topology of state space, thus introducing correlations between certain model parameters.
- identifying details of system geometry.
- identifying sparsity in the model description.
- projection onto an "optimal" subspace of the system (Tokdar et. al 2010), where such a subspace is a characteristic of the system.

Data-driven choice of modelling strategy

The choice of modelling strategy depends primarily on:

- **quantity and quality of available data.**
- **available information (deterministic and/or probabilistic).**
- **degree of non-linearity in the problem, which in turn determines the dimensionality of parameter space.**

Data-driven choice of modelling strategy

For large data sets -

- inverting the **P** matrix is not feasible (unless **P** is sparse)
- writing the likelihood function and optimising or using MCMC methods is a possibility.

Small data sets, under-abundant systems -

- As long as the **P** matrix can be maintained to be a square, direct inversion is feasible, though this might be numerically unstable if data is not Holder continuous (Rullgard 2004, Markoe & Quinto 1985).
- For under-abundant systems, a square **P** matrix is not achievable. Then (and also if mapping to data-space is unknown), dimensionality reduction of parameter space is if vital importance.

What to do with noisy or incomplete data?

For noisy data -

- when the distribution of measurement uncertainties is Gaussian, the likelihood can be defined as Gaussian in $\parallel \boldsymbol{\xi}[\boldsymbol{\rho}] - \mathbf{I} \parallel$, with zero mean and variance defined by the measurement uncertainty.
- when the noise distribution is not Gaussian, but known, this distribution can be convolved with the posterior probability $\pi(\boldsymbol{\rho}|\mathbf{I}_0)$, where data is $\mathbf{I} = \mathbf{I}_0 + \delta\mathbf{I}$.

For incomplete data -

- in the discretised models that are typically of relevance, non-uniform gridding might be called for to accommodate non-uniform sampling from data space.
- risky to interpolate between datum in the presence of non-linearities; learn given whatever quality of data is available.

What to do with “partially sampled” data?

By “partially sampled” is meant data that is sampled from only a subspace of the system state space. This might be due to logistics that affect the experiment or in the nature of the problem. Then,

- projection of $\xi[\rho]$ into this observed subspace is invoked, in order to define likelihood.

Uniqueness considerations

- If problem can be reduced to the form $\mathbf{I} = \mathbf{P}\rho$, where $\mathbf{P}$ is a known squre matrix, then solution for ρ is deterministic ($\mathbf{P}^{+}\mathbf{I}$), given $\mathbf{I}$ - this is equivalent to solving the least squares problem. Then in the presence of noise in data $\mathbf{I}$, solution is no longer deterministic, with condition number given as discussed above. As we saw above, such an uncertainty in the solution is $\| \boldsymbol{\rho} - \boldsymbol{\rho}_0 \| \leq \| \boldsymbol{\rho} \| \kappa \frac{\delta \mathbf{I}}{I}$, where $\boldsymbol{\rho}_0$ is the solution in the zero noise limit and $\kappa \geq 1$ is the condition number for the problem $\mathbf{I} = \mathbf{P}\rho$.
- Otherwise, infinite solutions are feasible. When sampling from posterior is undertaken, uncertainty estimation given multimodal posterior is a challenge.
- Point estimates do not bear information about the range of solutions. Interval estimates are readily available in the Bayesian approach - crucial superiority.

Influence of sparsity

- Ill-posed deconvolution or deprojection problems are solvable (Li & Speed 2000, Ramalau & Teschke 2006) if they admit a sparse representation. In this paradigm, the inverse problem is described by constraints of sparsity with respect to a chosen frame.
- If the problem can be reduced to the form $\mathbf{I} = \mathbf{P}\rho$, where $\mathbf{P}$ is a sparse matrix. Then even if $\mathbf{P}$ is high-dimensional, direct inversion is possible.
- In the Bayesian paradigm, priors on sparsity can be identified.