DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF COPENHAGEN

CENTRE FOR STOCHASTIC GEOMETRY AND ADVANCED BIOIMAGING

Regression on Manifolds

Line Kühnel
kuhnel@di.ku.dk
Supervisors: Stefan Sommer and Mads Nielsen

Teaser

There is a large statistical framework for analyzing data in euclidean space. However, not all data can be assumed to belong to a euclidean space. For anatomical objects or shape data it is not possible to define addition such that the full space of data points are closed under this operation. Instead these kinds of data are assumed to form a non-linear manifold, $\mathcal{M}$. The problem is that a lot of the statistical framework that are already defined, are based on addition of datapoints. Examples are average, variance and several models such as regression. This means that we have to come up with a new theory to be able to perform statistical analysis on manifold valued data.

Our Goal

To define a regression model to describe the relation between multiple covariate variables in $\mathbb{R}^m$ and a response variable in $\mathcal{M}$. An example could be to model how a treatment $x \in \mathbb{R}$ affects the brain structure of patients.

The Regression Model

Let $\mathcal{M}$ denote a d-dimensional manifold embedded in $\mathbb{R}^k$, $k \geq d$ and $X_\alpha \colon \mathbb{R}^d \to T_{y_0}\mathcal{M}$ be a frame for the tangent space at $y_0 \in \mathcal{M}$. We observe

i: n realisations of the response variable $y \in \mathcal{M}$, $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n \in \mathbb{R}^k$

ii: and for each realisation $\boldsymbol{y}_i$, m covariate variables $\boldsymbol{x}_i = (x_i^1, \ldots, x_i^m) \in \mathbb{R}^m$ for $m \geq 1$.

Consider stochastic processes z_t^i solving the stochastic differential equation,

$$dz_t^i = \beta_t' dt + W dX_t^i + d\varepsilon_t^i, \quad i = 1, \ldots, n, \tag{1}$$

in which $\beta_t' dt$ is a fixed drift, W is a $m \times m$-matrix of coefficients, dX_t^i is a brownian bridge with $X_0^i = 0$, $X_1^i = \boldsymbol{x}_i$ and ε_t^i is a brownian motion. Notice that the structure of these processes are similar to a usual regression model, with a general effect ($\beta_t' dt$), a covariate dependency ($W dX_t^i$) and an individual error term ($d\varepsilon_t^i$).
Based on the processes z_t^i, we can define a relation between covariate variables on $\mathbb{R}^m$ and the response variable on $\mathcal{M}$. For each observation i, a sample of the stochastic process z_t^i are transported to $\mathcal{M}$ by stochastic development through the frame bundle $\mathcal{FM}$. Let $Y_i \colon \Omega \to \mathcal{M}$ be a stochastic variable following the distribution of the endpoints of the transported sample paths of z_t^i. Each observation $\boldsymbol{y}_i$ is then modelled as

$$\boldsymbol{y}_i = Y_i + \nu_i \tag{2}$$

where $\nu_i \sim \mathcal{N}(0, \tau^2 I)$ denotes the measurement error in $\mathbb{R}^k$.

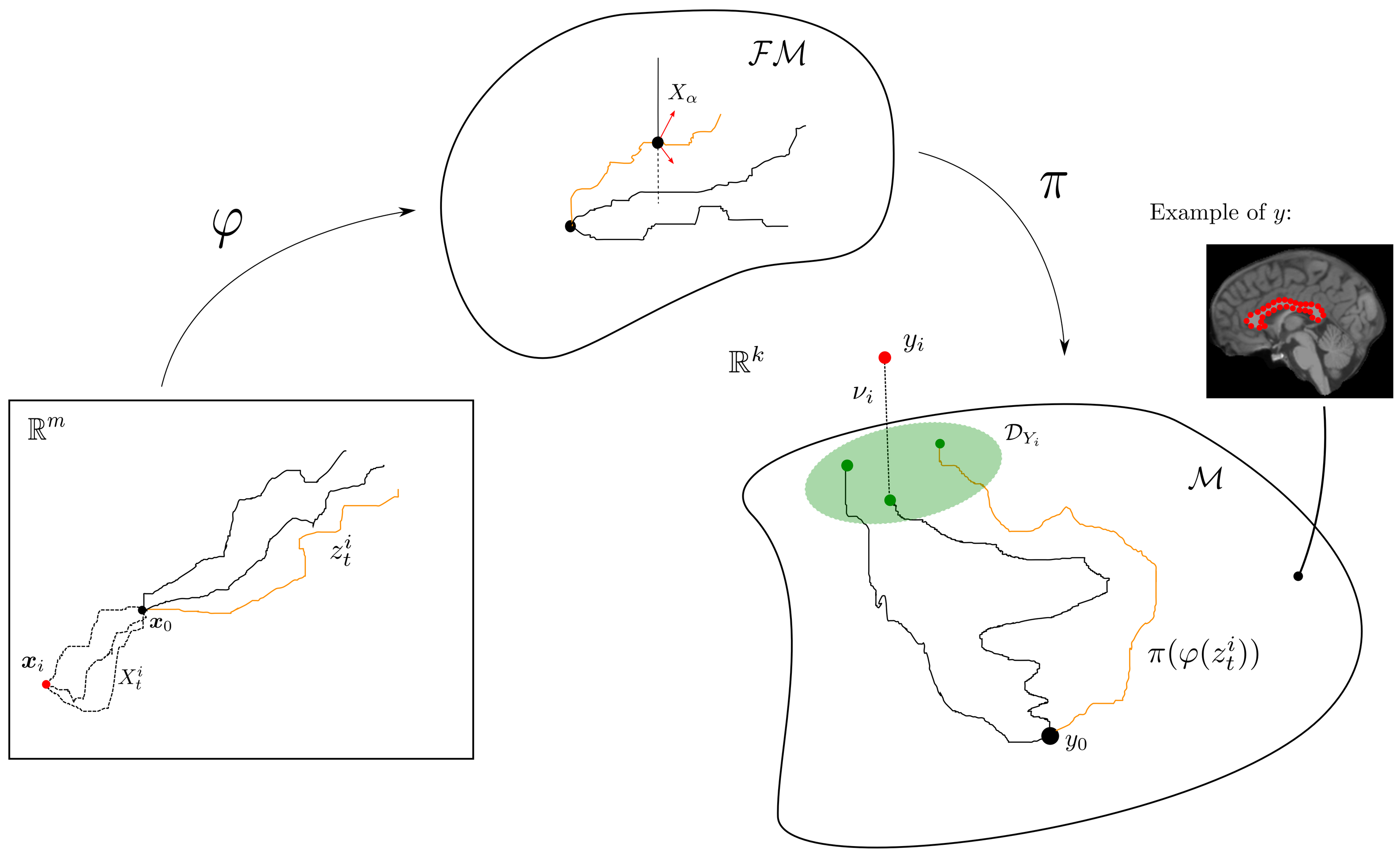
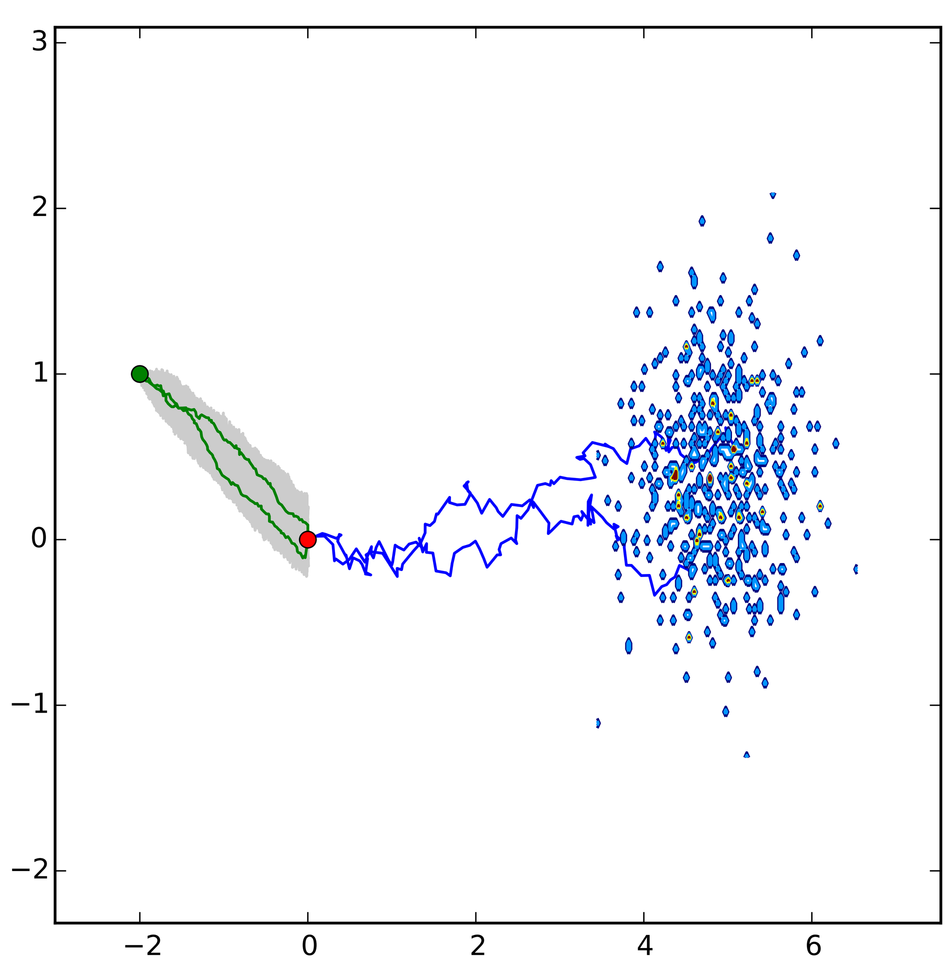


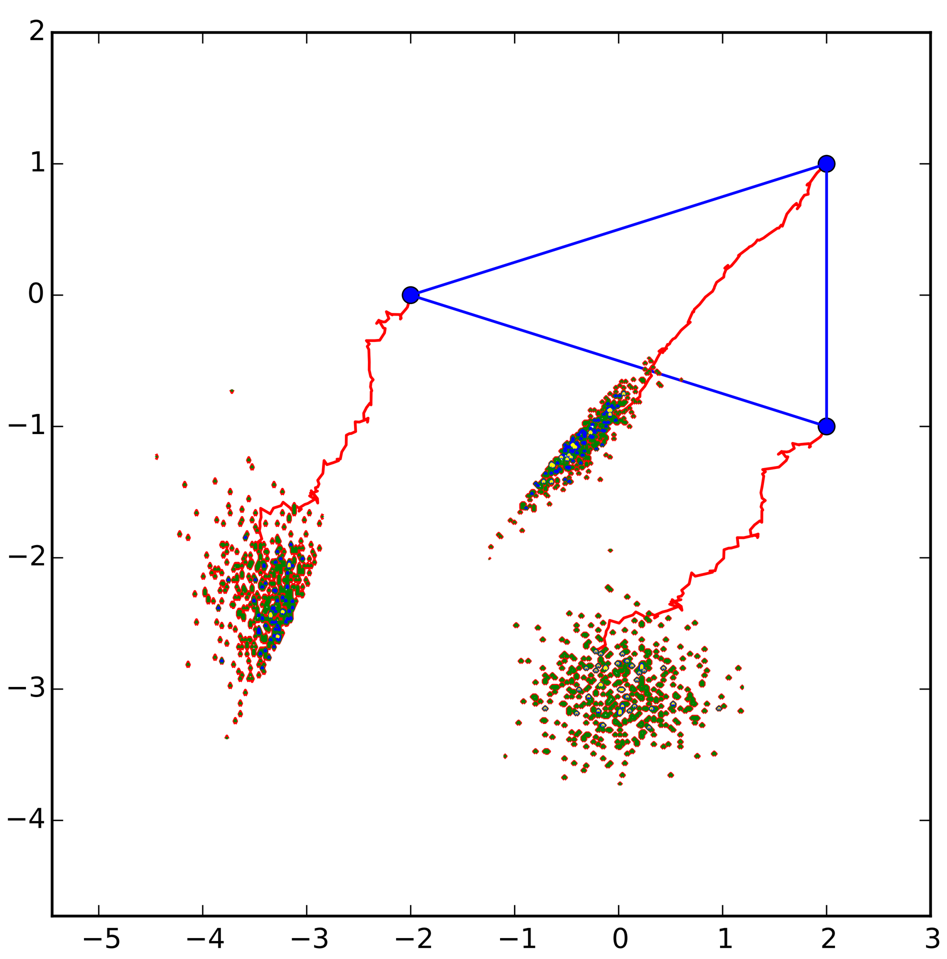
Figure 1: Illustration of the basics of the model. φ denotes the stochastic development to the frame bundle and π is a projection to the manifold.

Example

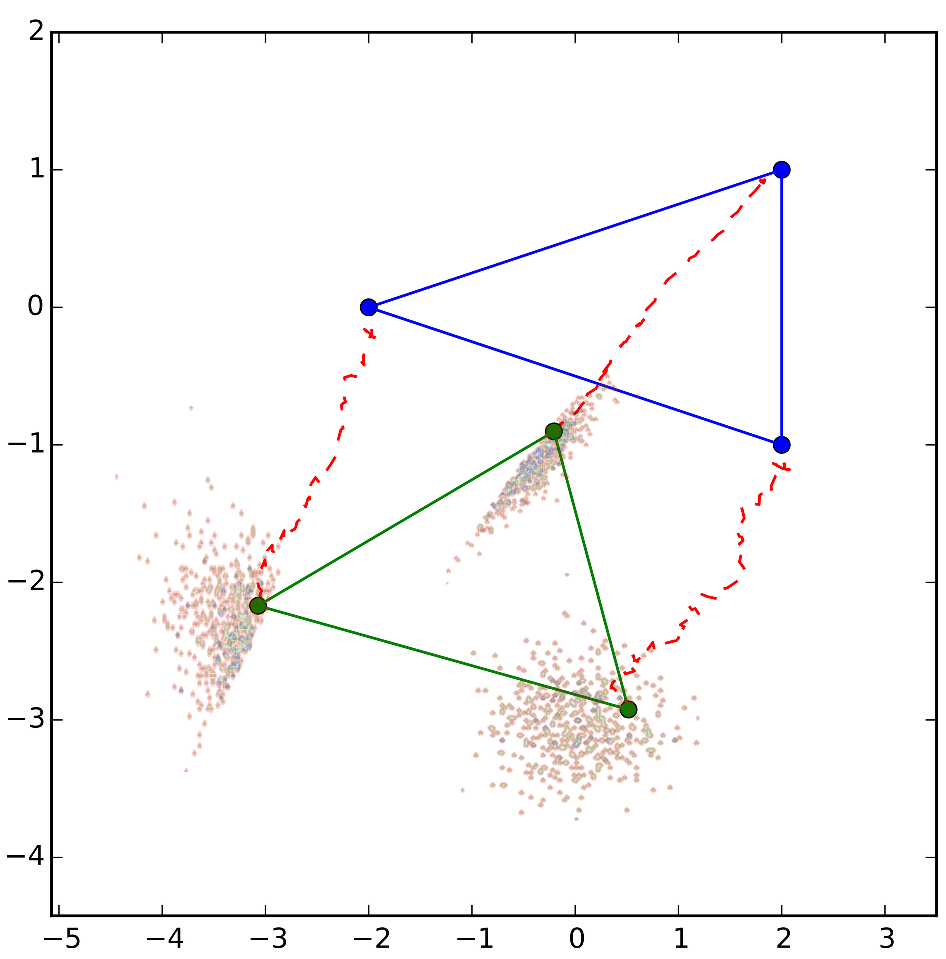
In the figures below we show an example of how the model can be used for prediction. It is a simple example with a response variable y of triangles represented by 3 landmarks. Assume we observe two covariate variables, $x_1 = -2$, $x_2 = 1$ (The green bullet in Figure (a)). We simulate 1000 processes z_t and transport them to $\mathcal{M}$. In Figure (b) are shown the end distribution for the transported processes for each landmark. The blue triangle is the starting point y_0. Based on these end distributions, we are then able to make predictions on y from the model. This means that based on the measured covariates we can predict an observation y. In Figure (c) is shown such a prediction as the green triangle.



(a) Simulated processes on $\mathbb{R}^m$. The green lines represents the X_t brownian bridgess and the blue are the Z_t processes. The green bullet point are the covariate values.



(b) The evolution of each landmark and their end distribution. The blue triangle visualizes the initial value $y_0 \in \mathcal{M}$.



(c) The same picture as in Figure (b), with a prediction of the triangle for the given covariate values.