Sparse regression

Tuesday, February 27, 2024 9:05 AM

Today sparse regression and dictionary learning

Problem 1: Given time series data, can we fit dynamics to it in a Bayesian context?

Given $\vec{y}_d = \left(y^d_{t_1} \cdots y^d_{t_N} \right)$

Hypothesize $\begin{cases} \dot{y} = F(t, y \mid \theta) \\ y(t=0) = y_0 \end{cases}$

Minimize a norm $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \mathcal{L}\left(y - \left(y_0 + \int_0^t F dt \right) \right)$

Problem 2 Assume direct measurements of relevant DOFs z are unavailable

$$z = G(y \mid \theta_1)$$

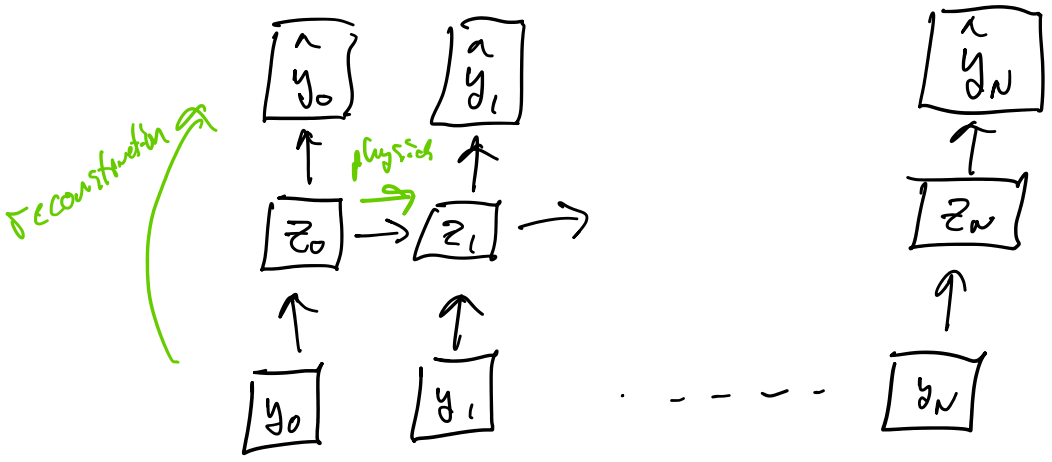
$$\begin{cases} \dot{z} = F(z, t \mid \theta_2) \\ z(t=0) = G(y(t=0)) \end{cases}$$

examples Image-based simulation

Candidate architectures

G → auto encoders

F → neural ODEs, universal differential eqns, dictionary learning



We'll start w/ one simple strategy called SINDy

"Sparse Identification of nonlinear dynamics" 2016 Brunton Proctor Kutz

IDEA $\dot{x} = \Theta(x)\,\xi$

$\Theta(x)$: dictionary; ξ: unknown coefficients

SPARSITY Want to only choose a few terms, so data reveals governing ODE

few terms, so data reveals governing ODE

"Occams Razor", Parsimony

Sparse Reg vs Polynomial Fitting

Not combining basis functions to approximate more complicated ones

ex images of pendulum swinging

Know $\ddot{z} = -kz$ is model

Can we infer θ coordinate from images?

Necessary Background

Revisiting overfitting (Murphy §4.5, § 11.3)

$$\mathcal{L}(\theta;\lambda) = \left[\frac{1}{N}\sum_{i=1}^{N} l(y_i, \theta; x_i)\right] + \lambda C(\theta)$$

λ: reg. param; $C(\theta)$: complexity

Typical Bayesian setting

$l = \log p(y_i|x_i,\theta)$

$C(\theta) = \log p(\theta)$ ← prior

def Maximum a posteriori (MAP) estimation

$$\hat{\theta} = \arg\max_{\theta} \log p(\theta|\mathcal{D})$$

$$\hat{\theta} = \arg\max_{\theta} \log \left(\frac{p(\mathcal{D}|\theta)\, p(\theta)}{p(\mathcal{D})} \right)$$

$$\hat{\theta} = \arg\max_{\theta} \left[\log p(\mathcal{D}|\theta) + \log p(\theta) \right]$$

($\log p(\mathcal{D}|\theta)$: log-likelihood; $\log p(\theta)$: prior)

note the unknown marginal likelihood $p(\mathcal{D})$ drops out

Bayesian regression revisited

Approximate $y(x) = \sum_i \phi_i(x)\, c_i$

$= \Phi(x)^T c$

Model $p(y|x,c) = N(y;\ \mu = \Phi(x)^T c,\ \sigma^2)$

$$\mu_{MLE} = \Phi(x)\, (\Phi(x)\Phi(x)^T)^{-1}\, \Phi(x)\, y$$

Ridge Regression

Adopt the ~~prior~~

$$p(c) = N(c;\ 0,\ \lambda^{-1} I)$$

$$\mathcal{L} = \log p(y|x,c)\, p(c) =$$

$$\log N(y;\ \Phi(x)^T c,\ \sigma^2)\, N(c;\ 0,\ \lambda^{-1} I)$$

$$= \sum_d -\frac{1}{2}\log\sigma^2 - \frac{1}{2}\left(\frac{\Phi(x_d)^T c - y_d}{\sigma}\right)^2 - \frac{1}{2}\,\frac{c^2}{\lambda^2}$$

$$0 = \nabla_c \mathcal{L} = \nabla_c \sum_d \left(\frac{\| P(x_d)^T c - y_d \|^2}{\sigma^2} + \frac{1}{\lambda^2} \|c\|^2 \right)$$

$$0 = \sum_d 2 \left(P(x_d)^T c - y_d \right) P(x_d) + 2 \left(\frac{\sigma}{\lambda} \right)^2 c \Big)$$

$$\sum_d \left(P(x_d) P(x_d)^T + \frac{\sigma^2}{\lambda^2} I \right) c_{MAP} = \sum_d P(x_d) y_d$$

↑ ridge stabilization

- penalizing magnitude of coefficients
- aka - ridge regression
  - l_2 regularization
  - weight decay
- Gives invertible matrix in small data limits

For further reading see Murphy §11.7
- Conjugate priors
- Bayesian Linear regression

Lasso / L^1 regression for sparsity

- many applications want coefficients to not just be small but as few nonzero entries as possible

Goal

L_0 norm $\|w\|_0 = \sum_{d=1}^{D} \mathbb{1}_{|w_d| > 0}$ minimization

- Important for feature selection & dictionary

$\alpha = 1$

- Important for feature selection & dictionary learning

Least Absolute shrinkage + selection operator (Tibshirani 96)

$$\mathcal{L}(w) = -\log p(\mathcal{D}|w) - \log(w|\lambda)$$

If $w|\lambda \sim \prod_d \text{lap}(w_d \mid \mu = 0, b = \lambda^{-1})$

$$\text{lap}(w|\mu, b) = \frac{1}{2b} \exp\left(-\frac{|w-\mu|}{b}\right)$$

Then $\mathcal{L}(w) = \|Xw - y\|^2 + \lambda \|w\|_1$

Note that for other norms $\|w\|_p = \left(\sum |w|^p\right)^{1/p}$ we get sparser solutions, but $p < 1$ is nonconvex

- ℓ_1-norm is tightest convex relaxation of ℓ_0-norm

Why sparse?

$$\min_w -\log p(y|w) + \lambda \|w\|_1$$

is Lagrangian associated w/

$$\min_w -\log p(y|w)$$

$$\text{s.t. } \|w\|_1 \leq B$$

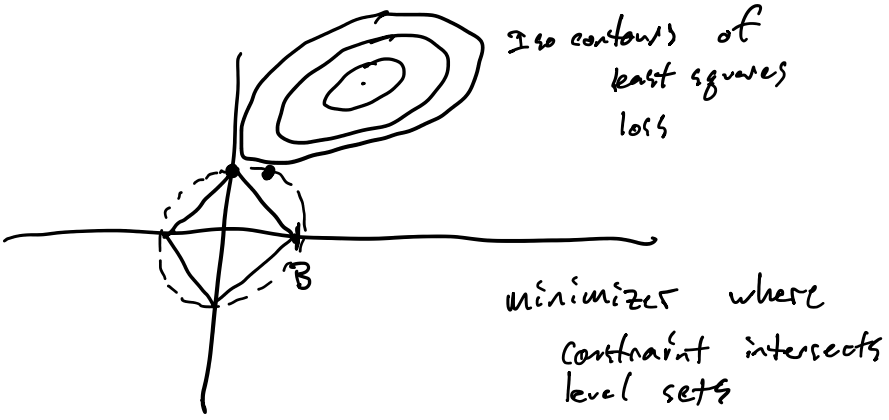
B ← parameter imposing tightness of penalty

Compare to ridge regression

$$\min -\log p(y|w)$$

$\min -\log p(y|w)$

s.t. $\|w\|_2 \le B$



Iso contours of least squares loss

minimizer where constraint intersects level sets

Issue Gradient undefined precisely where entries get sparse

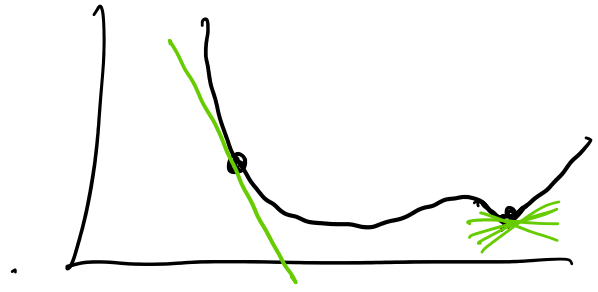
Def Subgradient

$f: \mathbb{R}^n \to \mathbb{R}$, $g \in \mathbb{R}^n$ is a subgradient if for all $z \in \text{dom}(f)$

$$f(z) \ge f(x) + g^T(z-x)$$

← think of like first order Taylor series

Denote $g(x) = \partial f(x)$



At cont. pt, one unique subdifferential

At a kink, many

ex $f(x) = abs(x)$

$$\partial f(x) = \begin{cases} -1 & x < 0 \\ [-1, 1] & x = 0 \\ 1 & x > 0 \end{cases}$$

any number between $[-1, 1]$

Back to lasso (summary of results, see 11.4.3 for details)

$$\mathcal{L}(w) = \underbrace{\|Xw - y\|_2^2}_{①} + \lambda \underbrace{\|w\|_1}_{②}$$

One can show

$$\frac{\partial}{\partial w_d} ① = a_d w_d - c_d$$

$$a_d = \sum_{n=1}^{N} x_{nd}^2$$

intercept as residual → $c_d = \sum_{n=1}^{N} x_{nd} (y_n - \vec{w}_{-d}^T \vec{x}_{n,-d})$

where $\vec{z}_{-d} = \langle z_1, \ldots z_{d-1}, z_{d+1}, \ldots z_m \rangle$

Taking subderivative to acct for discont.

$$\partial_{w_d} (① + ②) = (a_d w_d - c_d) + \lambda \partial_{w_d} \|w\|_1$$

$$= \begin{cases} a_d w_d - c_d - \lambda & \text{if } w_d < 0 \\ [-c_d - \lambda, -c_d + \lambda] & \text{if } w_d = 0 \\ a_d w_d - c_d + \lambda & \text{if } w_d > 0 \end{cases}$$

At $0 = \partial_{w_d} (① + ②)$ we have 3 solutions

$$\hat{w}_d(c_d) = \begin{cases} (c_d + \lambda)/a_d & c_d < -\lambda \\ 0 & c_d \in [-\lambda, \lambda] \end{cases}$$

$$\hat{w}_d(c_d) = \begin{cases} (c_d + \lambda)/a_d & c_d < -\lambda \\ 0 & c_d \in [-\lambda, \lambda] \\ (c_d - \lambda)/a_d & c_d > \lambda \end{cases}$$

Compactly

$$\hat{w}_d = \text{Soft Threshold}\left(\frac{c_d}{a_d}, \lambda/a_d\right)$$

$$\text{SoftThreshold}(x, \delta) = \text{sign}(x)\left(|x| - \delta\right)_+$$

First introduction to Markov Chains

Consider time series data

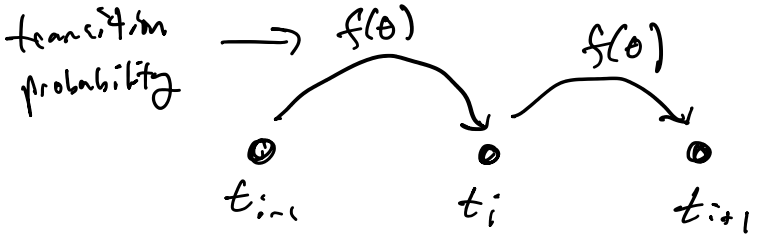
$$\vec{y}_d = \left(y_{t_1}^d, \ldots y_{t_N}^d\right)$$

By chain rule could write

$$p(\vec{y}) = p(y_{t_1})\, p(y_{t_2} | y_{t_1})\, p(y_{t_3} | y_{t_2}, y_{t_1}) \cdots$$
$$= \prod_{i=1}^{N} p(y_{t_i} | y_{t_{1:i-1}})$$

For first-order Markov process, assume

$$p(y_{t_i} | y_{t_{1:i-1}}) = p(y_{t_i} | y_{t_{i-1}})$$
$$= f(\theta)$$



So $$p(\vec{y}_d) = p(y_1) \prod_{i=2}^{N} p(y_i | y_{i-1})$$

To $p(\vec{y}_d) = p(y_1) \prod_{i=2}^{n} p(y_i | y_{i-1})$

↑ initial condition

Note as a special case a stochastic differential equation

$$dX = f(x)\,dt + g(x)\,dB_t$$

Euler-Maruyama $$X^{n+1} - X^n = f(X^n)\Delta t + g(X^n)\xi$$

$$\xi \sim N(0, \Delta t)$$

then $$p(X^{n+1} | X^n) = N\left(X^n + \Delta t\, f(X^n),\ \Delta t\, g(X^n)^2\right)$$

Exercise

Given data for harmonic oscillator

$$\theta(t) = \theta_0 + \sin \omega t$$

We aim to fit

Dynamics of the form

$$\ddot{z} = -\lambda z$$

← single parameter to be fit

Discretize in time

$$\frac{z^{n+1} - 2z^n + z^{n-1}}{\Delta t^2} = -\lambda z^n$$

$$z^{n+1} = (2 + \Delta t^2 \lambda) z^n - z^{n-1}$$

Assume $z^n \sim N(\mu_n, \sigma_n^2)$ univariate gaussian

Recall identities

If $A \sim N(\mu_A, \sigma_A^2)$

$B \sim N(\mu_B, \sigma_B^2)$

From Def of exp, var → $aA + \beta \sim N(a\mu_A + \beta, a^2 \sigma_A^2)$

we didn't show yet → $A + B \sim N(\mu_A + \mu_B, \sigma_A^2 + \sigma_B^2)$

Then $p(z^{n+1} \mid z^n, z^{n-1}, \lambda) =$

$$N\Big(\underbrace{(2 + \Delta t^2 \lambda)\mu_n - \mu_{n-1}}_{M_{n,n-1}}, \underbrace{(2 + \Delta t^2 \lambda)^2 \sigma_n^2 + \sigma_{n-1}^2}_{C_{n,n-1}}\Big)$$

Apply Markovian Assumption

$$p(z_0, \ldots z_N \mid \lambda) = p(z_0, z_1) \prod_{n=2}^{N} p(z_n \mid z_{n-1}, z_{n-2}, \lambda)$$

$p(z_0, z_1)$ ✓ indep of λ

$$\mathcal{L}(\lambda) = -\log p(\vec{z} \mid \lambda) = C + \sum_{n=2}^{N} \log p(z_n \mid z_{n-1}, z_{n-2}, \lambda)$$

Exercise 1

Given noisy data

$$z_n = \theta(t_n) + \varepsilon$$

$$\varepsilon = N(0, 0.01)$$

Infer λ

Exercise 2 Given indirect measurements

$$y_n = \begin{pmatrix} \cos z_n \\ \sin z_n \end{pmatrix}$$

Build Autoencoder

$$z_n \sim N(f_{\theta_1}(y_n), g_{\theta_2}(y_n))$$

Recall ELBO

$$\mathcal{L}(\theta, \phi, \lambda | \mathcal{D}) = \mathbb{E}_{q(\vec{z}|\vec{x})}\left[\log \frac{p_\theta(x|z)}{q_\phi(z|x)}\right]$$

$$= \mathbb{E}_{q(z|x)}[p(x|z)] - KL(q(z|x) \| p(z))$$

Choose prior:

$$p(\vec{z}) = p(z_0, z_1) \prod_{i=2}^{N} p(z_i | z_{i-1}, z_{i-2})$$

Assume separable posterior

$$q(\vec{z}|\vec{x}) = \prod_{i=0}^{N} q(z_i | x_i)$$

Then

$$KL\left(q(\vec{z}|\vec{x}) \,||\, p(z)\right) = \mathbb{E}_{q(\vec{z}|\vec{x})}\left[\log \frac{p(\vec{z})}{q(\vec{z}|\vec{x})}\right]$$

$$= \sum_{i=2}^{N} \mathbb{E}_{q(\vec{z}|\vec{x})}\left[\log p(z_i|z_{i-1}, z_{i-2})\right] + C$$

$$- \sum_{i=0}^{N} \mathbb{E}_{q(\vec{z}|\vec{x})}\left[q(z_i|x_i)\right]$$

$$= \sum_{i=2}^{N} \mathbb{E}_{q(z_i|x_i)}\left[\log p(z_i|z_{i-1}, z_{i-2})\right]$$

$$- \sum_{i=2}^{N} \mathbb{E}_{q(z_i|x_i)}\left[q(z_i|x_i)\right]$$

How can we compute these integrals?

<u>Exercise</u> Let $f(z) \sim N(\mu_1, \sigma_1^2)$
$g(z) \sim N(\mu_2, \sigma_2^2)$

then

$$\int f(z) \log(g(z))\, dz = -\frac{1}{2}\left(\log 2\pi\sigma_2^2 + \frac{\sigma_1^2}{\sigma_2^2}\right.$$

$$\int_{\mathbb{R}^N} f(z) \log\left(g(z)\right) dz = -\frac{1}{2} \sum_j \log 2\pi \sigma_{2,j}^2 + \frac{\sigma_{2,j}^2}{\sigma_{1,j}^2} + \frac{\left(\mu_{1,j} - \mu_{2,j}\right)^2}{\sigma_{2,j}^2}$$