Heavy-tailed Process Priors for Selective Shrinkage

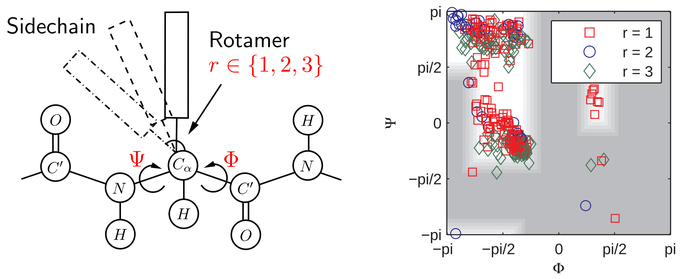
Fabian L. Wauthier and Michael I. Jordan

University of California, Berkeley

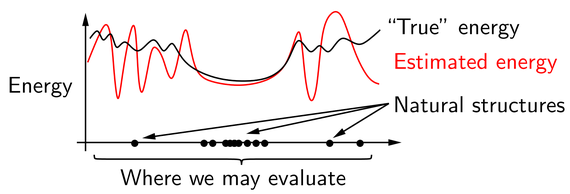
{flw,jordan}@cs.berkeley.edu

Introduction

- Problem: Often sparsity occurs in the **input** rather than the **output** space.
- Working example: Protein structure prediction.
  - Imitate nature by minimizing an estimated energy function.
  - Important part of energy estimate: Conditional probabilities of **rotamers** (discrete sidechain angles) r given continuous **backbone angles** (Φ,Ψ).



  - Have sampling bias. Mostly see structures near the energy minimum. Remaining space is **sparsely sampled**.
  - Result: Estimated energy is most accurate near the energy minimum.



- **Want**: Improved conditional rotamer probabilities in sparse areas.
- **Idea**: **Selectively shrink** probabilities towards conservative values (i.e. 1/3 for 3 rotamers) more strongly in sparse regions than in dense regions.
- **How**: Modify Gaussian process classification (GPC) to use a heavy-tailed process (HP) prior instead of a Gaussian process (GP) prior.

Gaussian Process Classification (GPC)

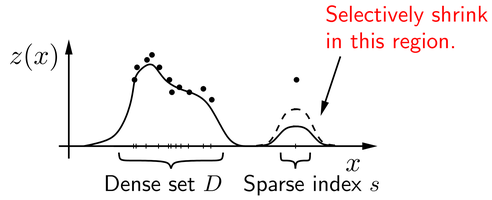
- Write $z(X) \sim p(z(X)) = \mathcal{N}(0, K(X,X))$ for a sample from the GP with kernel matrix $[K(X,X)]_{i,j} = k(x_i, x_j)$.
- Binary GPC: Posterior probability that x_* is labeled as class $y_* = 1$ is

$$p(y_* = 1|X, y, x_*) = \mathbb{E}_{p(z(x_*)|X,y,x_*)}\left(\frac{1}{1+\exp\{-z(x_*)\}}\right)$$

$$p(z(x_*)|X, y, x_*) = \int p(z(x_*)|X, z(X), x_*)p(z(X)|X, y)dz(X).$$

- Intuition: Selectively shrink $p(y_* = 1|X, y, x_*)$ to $1/2$ by selectively shrinking $p(z(x_*)|X, y, x_*)$ towards pointmass at zero.

- Focus on regression case for now. Want:



Heavy-tailed Process Priors via the Copula Trick

Suppose $\text{var}(z(x)) = \sigma^2\ \forall x$. Construct HP $f(X)$ with marginal c.d.f. G_b as

$$f(X) = G_b^{-1}(\Phi_{0,\sigma^2}(z(X))), \tag{1}$$

where $\Phi_{0,\sigma^2}(\cdot)$ is the c.d.f. of a centered Gaussian with variance σ^2.

- Can use $f(X)$ for heavy-tailed process regression (HPR).
- Inference reduces to GPR inference with transformed observations.

Analyzing Selective Shrinkage

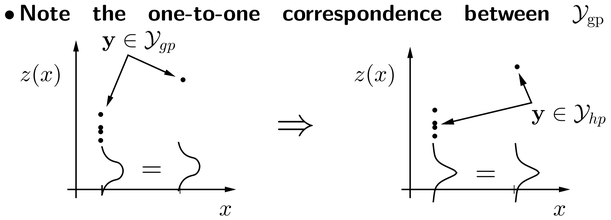
- Assume locations in dense region D are identical and a single sparse location s is far removed (special case of figure above).
- Set up a special GP on these locations so that posterior distributions are identical at the two locations if observations y satisfy

$$y \in \mathcal{Y}_{\text{gp}} \triangleq \left\{ y \,\middle|\, \sum_{d\in D} y_d = y_s \right\}.$$

- Transform this GP to a heavy-tailed process using equation 1.
- Can show that the posterior distributions under this HPR are identical at the two locations if measurements y' satisfy

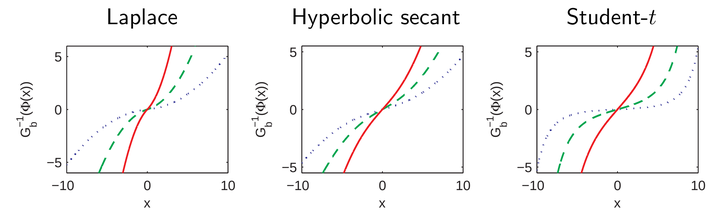
$$y' \in \mathcal{Y}_{\text{hp}} \triangleq \left\{ y' = G_b^{-1}(\Phi_{0,\sigma^2}(y)) | y \in \mathcal{Y}_{\text{gp}} \right\}.$$

- **Note the one-to-one correspondence between $\mathcal{Y}_{\text{gp}}$ and $\mathcal{Y}_{\text{hp}}$**



- Compare sets $\mathcal{Y}_{\text{gp}}$ and $\mathcal{Y}_{\text{hp}}$ for elements $0 \neq y \in \mathcal{Y}_{\text{gp}}$ with $\text{sign}\,(y_d) = \text{sign}\,(y_{d'}), \forall d, d' \in D$.
- For $d^* = \text{argmax}_{d\in D}|y_d|$ we have $|y_s| > |y_{d^*}|$.
- For sufficiently heavy-tailed G_b: $y' = G_b^{-1}(\Phi_{0,\sigma^2}(y)) \in \mathcal{Y}_{\text{hp}}$ satisfies

$$|y'_s| = \left|G_b^{-1}(\Phi_{0,\sigma^2}(y_s))\right| > \left|\frac{G_b^{-1}(\Phi_{0,\sigma^2}(y_{d^*}))}{y_{d^*}} y_s\right| = \left|\frac{y'_{d^*}}{y_{d^*}} y_s\right|.$$



- So y' leads to identical posteriors in HPR even though observation y'_s is disproportionately larger than y_s in GPR $\Rightarrow$ selective shrinkage occurs.

Heavy-tailed Process Classification (HPC)

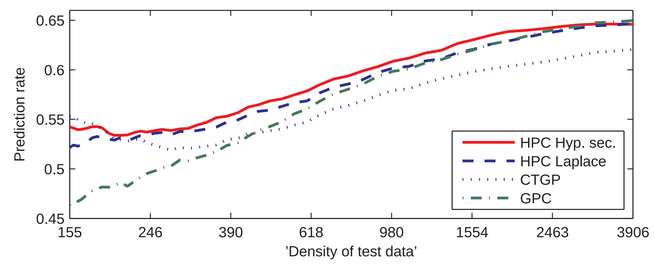
- Embed HP in a multi-class classification model. Posterior inference by Laplace approximation.
- Learn parameters by maximizing an ℓ_2 regularized version of the Laplace approximation to the log marginal likelihood.

Results

- Evaluate rotamer prediction tasks on 17 amino-acids. Input covariates are (Φ, Ψ) angle pairs, output is the rotamer label.
- Set up GPC and HPC models using a von Mises-inspired kernel for d-dimensional angular data

$$k(x_i, x_j) = \sigma^2 \exp\left\{\lambda\left(\left(\sum_{k=1}^{d}\cos(x_{i,k} - x_{j,k})\right) - d\right)\right\}.$$

- Evaluate predictions on sparse regions as a function of region "sparsity".



References

[1] Tamara Broderick and Robert B. Gramacy. Classification and Categorical Inputs with Treed Gaussian Process Models. *Journal of Classification*. To appear.

[2] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[3] Ed Snelson, Carl E. Rasmussen, and Zoubin Ghahramani. Warped Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 16, pages 337–344, 2004.