Some Concerns About Sparse Approximations for Gaussian Process Regression

Joaquin Quiñonero Candela

Max Planck Institute for Biological Cybernetics

Gaussian Process Round Table
Sheffield, June 9 and 10, 2005

Menu

- Concerns about the quality of the predictive distributions
- Augmentation: a bit more expensive, but gooood ...
- Dude, where's my prior?
- A short tale about sparse greedy support set selection

The Regression Task

- Simplest case, additive independent Gaussian noise of variance σ^2
- Gaussian process prior over functions:

$$p(\boldsymbol{y}|\boldsymbol{f}) \sim \mathcal{N}(\boldsymbol{f}, \sigma^2\,\mathbf{I}) \;\;, \qquad\qquad p(\boldsymbol{f}) \sim \mathcal{N}(0, \boldsymbol{K})$$

- Task: obtain the predictive distribution of f_* at the new input x_*:

$$p(f_*|x_*, \boldsymbol{y}) = \int p(f_*|x_*, \boldsymbol{f})\, p(\boldsymbol{f}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{f}$$

- Need to compute the posterior distribution (expensive):

$$p(\boldsymbol{f}|\boldsymbol{y}) \sim \mathcal{N}\left(\boldsymbol{K}\left(\boldsymbol{K} + \sigma^2\,\mathbf{I}\right)^{-1}\boldsymbol{y}, \sigma^2\boldsymbol{K}\left(\boldsymbol{K} + \sigma^2\,\mathbf{I}\right)^{-1}\right)$$

- ... and integrate $\boldsymbol{f}$ from the conditional distribution of f_*:

$$p(f_*|x_*, \boldsymbol{f}) \sim \mathcal{N}\left(\boldsymbol{K}_{*,\cdot}\,\boldsymbol{K}^{-1}\boldsymbol{y}, \boldsymbol{K}_{*,*} - \boldsymbol{K}_{*,\cdot}\,\boldsymbol{K}^{-1}\boldsymbol{K}_{*,\cdot}^{\top}\right)$$

Usual Reduced Set Approximations

- Consider some very common approximations
  - Naïve process approximation on subset of the data
  - Subset of regressors (Wahba, Smola and Bartlett...)
  - Sparse online GPs (Csató and Opper)
  - Fast Sparse Projected Process Approx (Seeger et al.)
  - Relevance Vector Machines (Tipping)
  - Augmented Reduced Rank GPs (Rasmussen, Quiñonero Candela)
- All based on considering only a subset I of the latent variables

$$p(f_*|x_*, \boldsymbol{y}) = \int p(f_*|x_*, \boldsymbol{f}_I)\, p(\boldsymbol{f}_I|\boldsymbol{y})\, \mathrm{d}\boldsymbol{f}_I$$

- However they differ in:
  - the way the support set I and the hyperparameters are learnt
  - the likelihood and/or predictive distribution approximations
- This has important consequences on the resulting predictive distribution
  - risk of over-fitting
  - degenerate approximations with nonsense predictive uncertainties

Naïve Process Approximation

- Extremely simple idea: throw away all the data outside I!
- The posterior only benefits from the information contained in $\boldsymbol{y}_I$:

$$p(\boldsymbol{f}_I|\boldsymbol{y}_I) \sim \mathcal{N}\left(\boldsymbol{K}_I\,(\boldsymbol{K}_I + \sigma^2\,\mathbf{I})^{-1}\boldsymbol{y}_I, \sigma^2\boldsymbol{K}_I\,(\boldsymbol{K}_I + \sigma^2\,\mathbf{I})^{-1}\right)$$

- The model underfits and is under-confident:

$$p(f_*|x_*, \boldsymbol{y}_I) \sim \mathcal{N}(\mu_*, \sigma_*^2)$$
$$\mu_* = \boldsymbol{K}_{*,I}\,(\boldsymbol{K}_I + \sigma^2\,\mathbf{I})^{-1}\boldsymbol{y}\;\;, \qquad \sigma_*^2 = \boldsymbol{K}_{*,*} - \boldsymbol{K}_{*,I}\,(\boldsymbol{K}_I + \sigma^2\,\mathbf{I})^{-1}\boldsymbol{K}_{*,I}^\top$$

- Training scales with m^3, predicting with m and m^2 (mean and var)
- Baseline approximation: we want higher accuracy and confidence

Subset Of Regressors

- Finite linear model with peculiar prior on the weigths:

$$f_* = \boldsymbol{K}_{*,I}\,\boldsymbol{\alpha}_I \ , \quad \boldsymbol{\alpha}_I \sim \mathcal{N}(0, K_I^{-1}) \qquad \Rightarrow \qquad f_* = \boldsymbol{K}_{*,I}\,\boldsymbol{K}_I^{-1}\boldsymbol{f}_I \ , \quad \boldsymbol{f}_I \sim \mathcal{N}(0, K_I)$$

- Posterior now benefits from all of $\boldsymbol{y}$:

$$\begin{aligned} q(\boldsymbol{f}_I|\boldsymbol{y}) &\propto \mathcal{N}(\boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1}\boldsymbol{f}_I|\boldsymbol{y}, \sigma^2\,\mathbf{I}) \cdot \mathcal{N}(\boldsymbol{f}_I|0, \boldsymbol{K}_I), \\ &\sim \mathcal{N}\left(\boldsymbol{K}_I\,[\boldsymbol{K}_{I,\cdot}\,\boldsymbol{K}_{I,\cdot}^\top + \sigma^2\,\boldsymbol{K}_I]^{-1}\boldsymbol{K}_{I,\cdot}\,\boldsymbol{y}, \sigma^2\,\boldsymbol{K}_I\,[\boldsymbol{K}_{I,\cdot}\,\boldsymbol{K}_{I,\cdot}^\top + \sigma^2\,\boldsymbol{K}_I]^{-1}\boldsymbol{K}_I\right) \end{aligned}$$

- The conditional distribution of f_* is degenerate!

$$p(f_*|\boldsymbol{f}_I) \sim \mathcal{N}\left(\boldsymbol{K}_{*,I}\,\boldsymbol{K}_I^{-1}\boldsymbol{f}_I, \mathbf{0}\right)^\top$$

- The predictive distribution produces nonsense errorbars

$$\begin{aligned} \mu_* &= \boldsymbol{K}_{*,I}\ \left[\boldsymbol{K}_{I,\cdot}\boldsymbol{K}_{I,\cdot}^\top + \sigma^2\,\boldsymbol{K}_I\right]^{-1}\boldsymbol{K}_{I,\cdot}\,\boldsymbol{y} \ , \\ \sigma_*^2 &= \sigma^2\,\boldsymbol{K}_{*,I}\left[\boldsymbol{K}_{I,\cdot}\boldsymbol{K}_{I,\cdot}^\top + \sigma^2\,\boldsymbol{K}_I\right]^{-1}\boldsymbol{K}_{*,I}^\top \end{aligned}$$

- Under the prior, only functions with m degrees of freedom

Projected Process (Seeger et al)

- Basic principle: likelihood approximation

$$p(\boldsymbol{y}|\boldsymbol{f}_I) \sim (\boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1} \boldsymbol{f}_I, \sigma^2 \mathbf{I})$$

- Leads to exactly the same posterior as for Subset of Regressors
- But the conditional distribution is now non-degenerate (process approximation)

$$p(f_*|\boldsymbol{f}_I) \sim \mathcal{N}\left(\boldsymbol{K}_{*,I} \boldsymbol{K}_I^{-1} \boldsymbol{f}_I, \boldsymbol{K}_{*,*} - \boldsymbol{K}_{*,I} \boldsymbol{K}_I^{-1} \boldsymbol{K}_{*,I}\right)^\top$$

- Predictive distribution with same mean as Subset of Regressors, but with way under-confident predictive variance!

$$\mu_* = \boldsymbol{K}_{*,I} \left[\boldsymbol{K}_{I,\cdot} \boldsymbol{K}_{I,\cdot}^\top + \sigma^2 \boldsymbol{K}_I\right]^{-1} \boldsymbol{K}_{I,\cdot} \boldsymbol{y}$$

$$\sigma_*^2 = \boldsymbol{K}_{*,*} - \boldsymbol{K}_{*,I} \boldsymbol{K}_I^{-1} \boldsymbol{K}_{*,I}^\top + \sigma^2 \boldsymbol{K}_{*,I} \left[\boldsymbol{K}_{I,\cdot} \boldsymbol{K}_{I,\cdot}^\top + \sigma^2 \boldsymbol{K}_I\right]^{-1} \boldsymbol{K}_{*,I}^\top$$

Augmented Subset Of Regressors

- For each x_*, augment f_I with f_*; new active set $I*$
- Augmented posterior: $q\left(\left[\begin{array}{c}\boldsymbol{f}_I\\ f_*\end{array}\right]\middle|\,\boldsymbol{y}\right)$
- ... at a cost of $\mathcal{O}(nm)$ per test case: need to compute $K_{*,\cdot}\, K_{I,\cdot}^\top$
- aSoR:

$$\mu_* = \boldsymbol{K}_{*,\cdot}\, \left[\boldsymbol{Q} + \frac{\boldsymbol{v}_*\,\boldsymbol{v}_*^\top}{c_*}\right]^{-1} \boldsymbol{y}$$

$$\sigma_*^2 = \boldsymbol{K}_{*,*} - \boldsymbol{K}_{*,\cdot}\, \left[\boldsymbol{Q} + \frac{\boldsymbol{v}_*\,\boldsymbol{v}_*^\top}{c_*}\right]^{-1} \boldsymbol{K}_{*,\cdot}^\top$$

with the ususal approximate covariance:

$$\boldsymbol{Q} = \boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,\cdot} + \sigma^2\, \mathbf{I}$$

with the difference between actual and projected covariance of f_* and f:

$$\boldsymbol{v}_* = \boldsymbol{K}_{*,\cdot}^\top - \boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,*}$$

with the difference between the prior variance of f_* and the projected:

$$c_* = \boldsymbol{K}_{*,*} - \boldsymbol{K}_{I,*}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,*}$$

Dude, where's my prior?

The Priors

The equivalent prior on $[\boldsymbol{f}, f_*]^\top$ is $\mathcal{N}(0, \boldsymbol{P})$ with:

$$\boldsymbol{Q} = \boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,\cdot}$$

Subset of Regressors:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,*} \\ \boldsymbol{K}_{I,*}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,\cdot} & \boldsymbol{K}_{I,*}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,*} \end{bmatrix}$$

Projected Process

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,*} \\ \boldsymbol{K}_{I,*}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,\cdot} & \boldsymbol{K}_{*,*} \end{bmatrix}$$

Nyström: (positive definiteness!)

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{K}_{*,\cdot}^\top \\ \boldsymbol{K}_{*,\cdot} & \boldsymbol{K}_{*,*} \end{bmatrix}$$

Ed and Zoubin's funky thing

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{Q} + \boldsymbol{\Lambda} & \boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,*} \\ \boldsymbol{K}_{I,*}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,\cdot} & \boldsymbol{K}_{*,*} \end{bmatrix}$$

$$\boldsymbol{\Lambda} = \mathrm{diag}\left(\boldsymbol{K}_\cdot\right) - \mathrm{diag}\left(\boldsymbol{Q}\right)$$

Augmented Subset of Regressors:

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{Q} + \frac{\boldsymbol{v}_* \boldsymbol{v}_*^\top}{c_*} & \boldsymbol{K}_{*,\cdot}^\top \\ \boldsymbol{K}_{*,\cdot} & \boldsymbol{K}_{*,*} \end{bmatrix}$$

with:

$$\boldsymbol{v}_* = \boldsymbol{K}_{*,\cdot}^\top - \boldsymbol{K}_{I,\cdot}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,*} \;, \qquad c_* = \boldsymbol{K}_{*,*} - \boldsymbol{K}_{I,*}^\top \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,*}$$

More on Ed and Zoubin's Method

- Here's a way of looking at it: the prior is a posterior process

$$f_* | \boldsymbol{f}_I = \mathcal{N}(\boldsymbol{K}_{*,I}\, \boldsymbol{K}_I^{-1} \boldsymbol{f}_I, \boldsymbol{K}_{*,*} - \boldsymbol{K}_{*,I}\, \boldsymbol{K}_I^{-1} \boldsymbol{K}_{*,I}^\top) \ ,$$

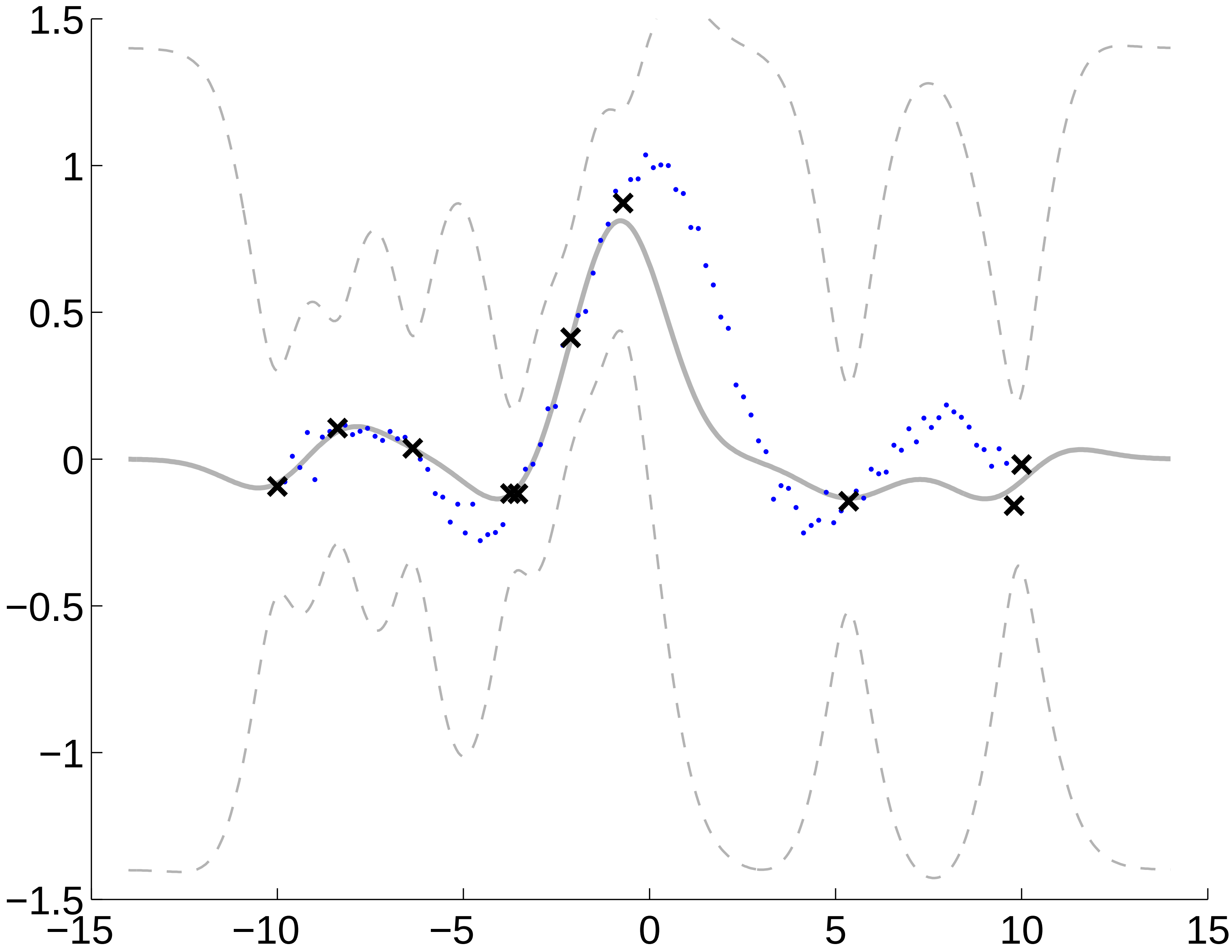
  ... well, almost: $E[f_+, f_* | \boldsymbol{f}_I] = 0$

- And then of course $\boldsymbol{f}_I \sim \mathcal{N}(0, \boldsymbol{K}_I)$
- The corresponding prior is

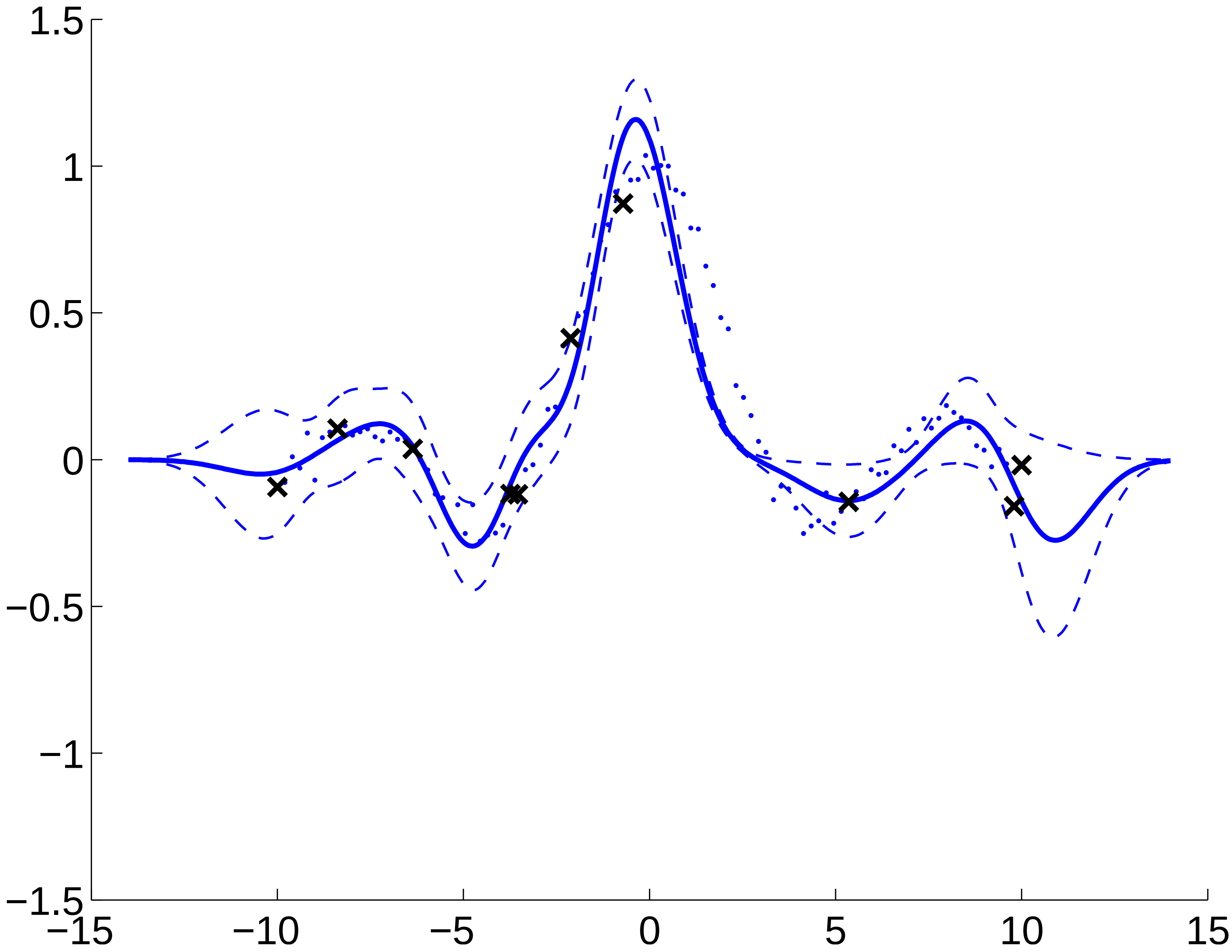
$$p(\boldsymbol{f}) = \mathcal{N}(0, \boldsymbol{K}_{*,*}\mathbf{I} + \boldsymbol{Q} - \mathrm{diag}(\boldsymbol{Q})) \ , \qquad \boldsymbol{Q} = \boldsymbol{K}_{I,\cdot}\, \boldsymbol{K}_I^{-1} \boldsymbol{K}_{I,\cdot}^\top$$

- With a bit of algebra you recover the marginal likelihood and the predictive distribution
- I finished this 30 minutes ago, which is why I won't show figures on it! (well, I now may)
- but ...

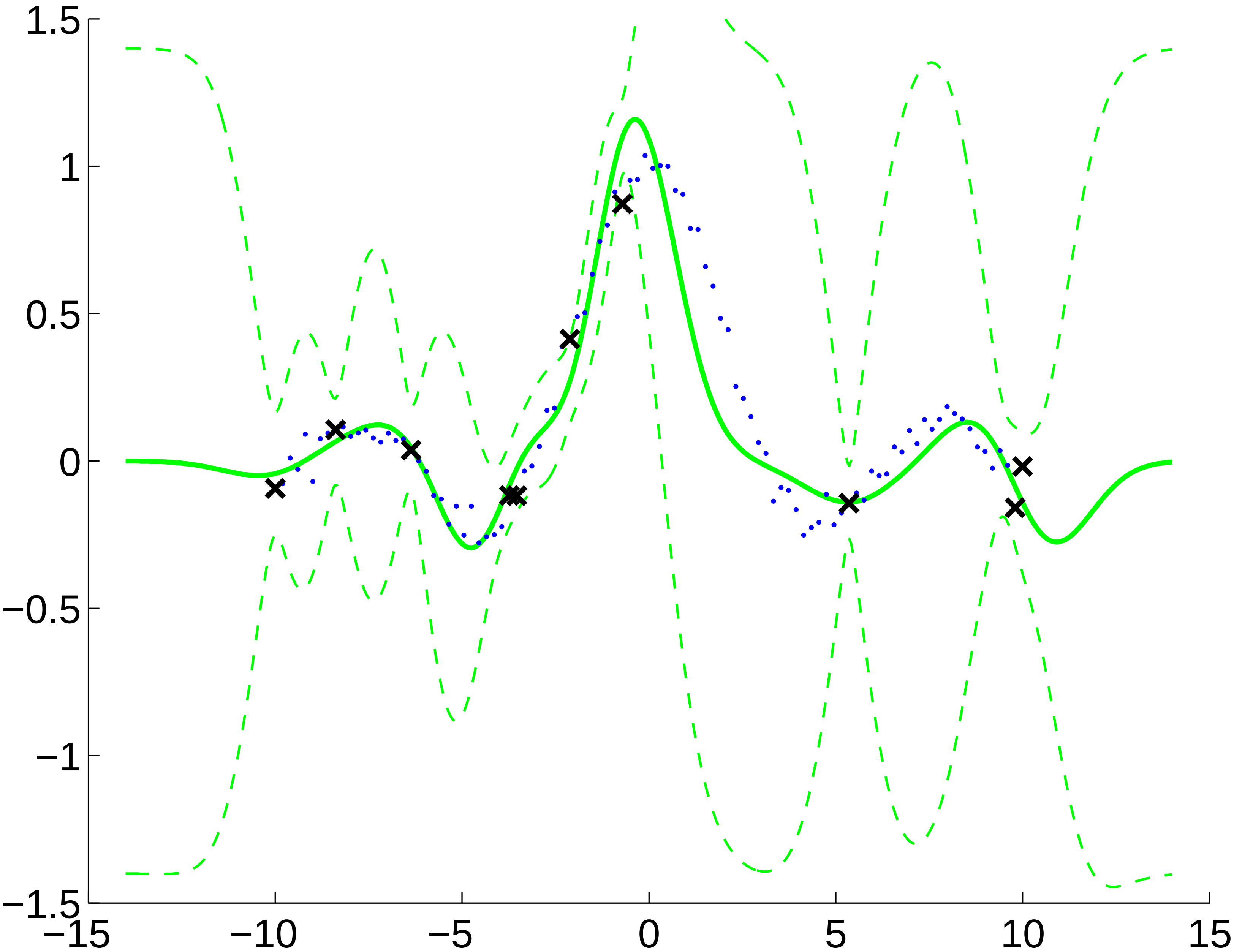
Naïve Process Approximation



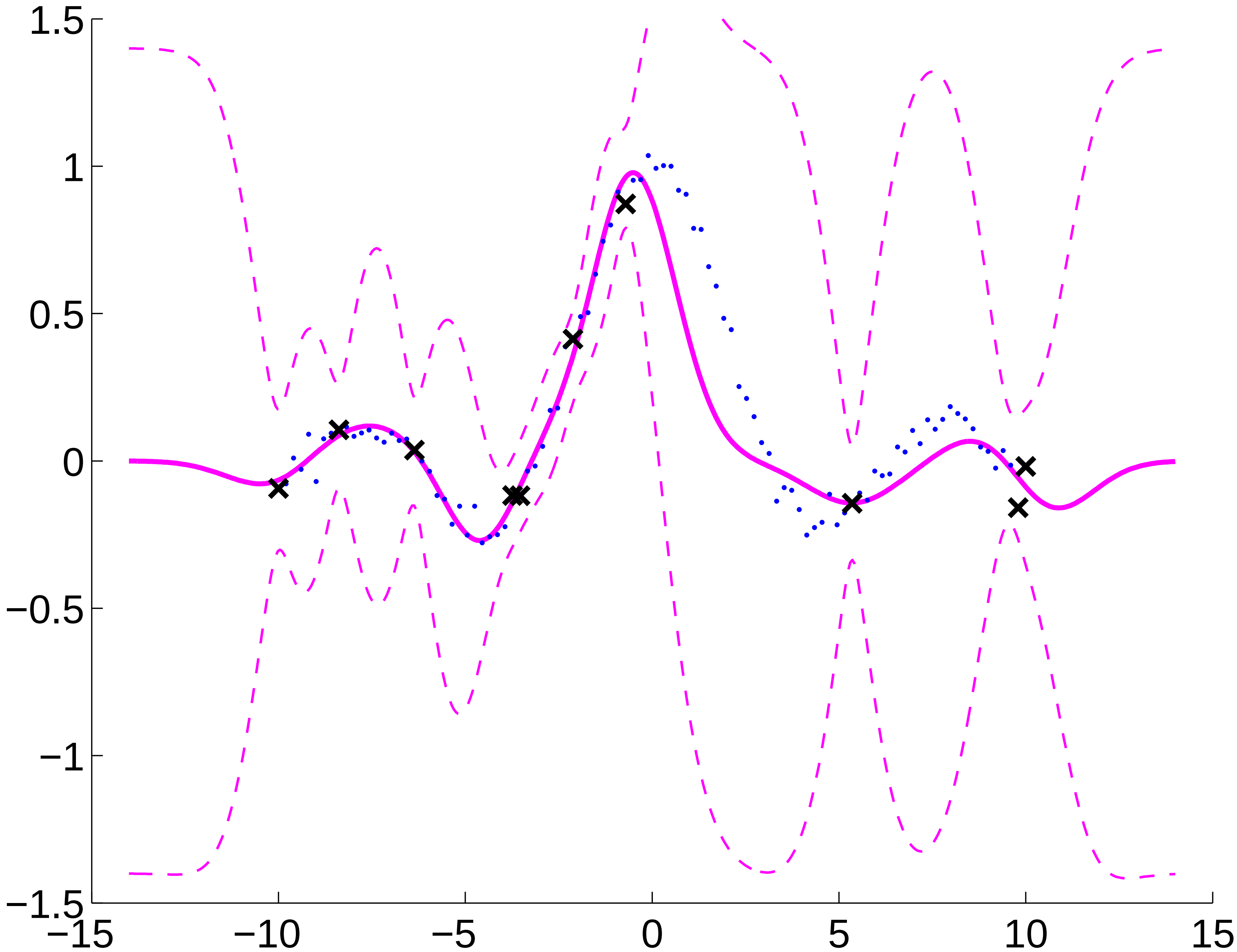
Subset of Regressors (degenerate)



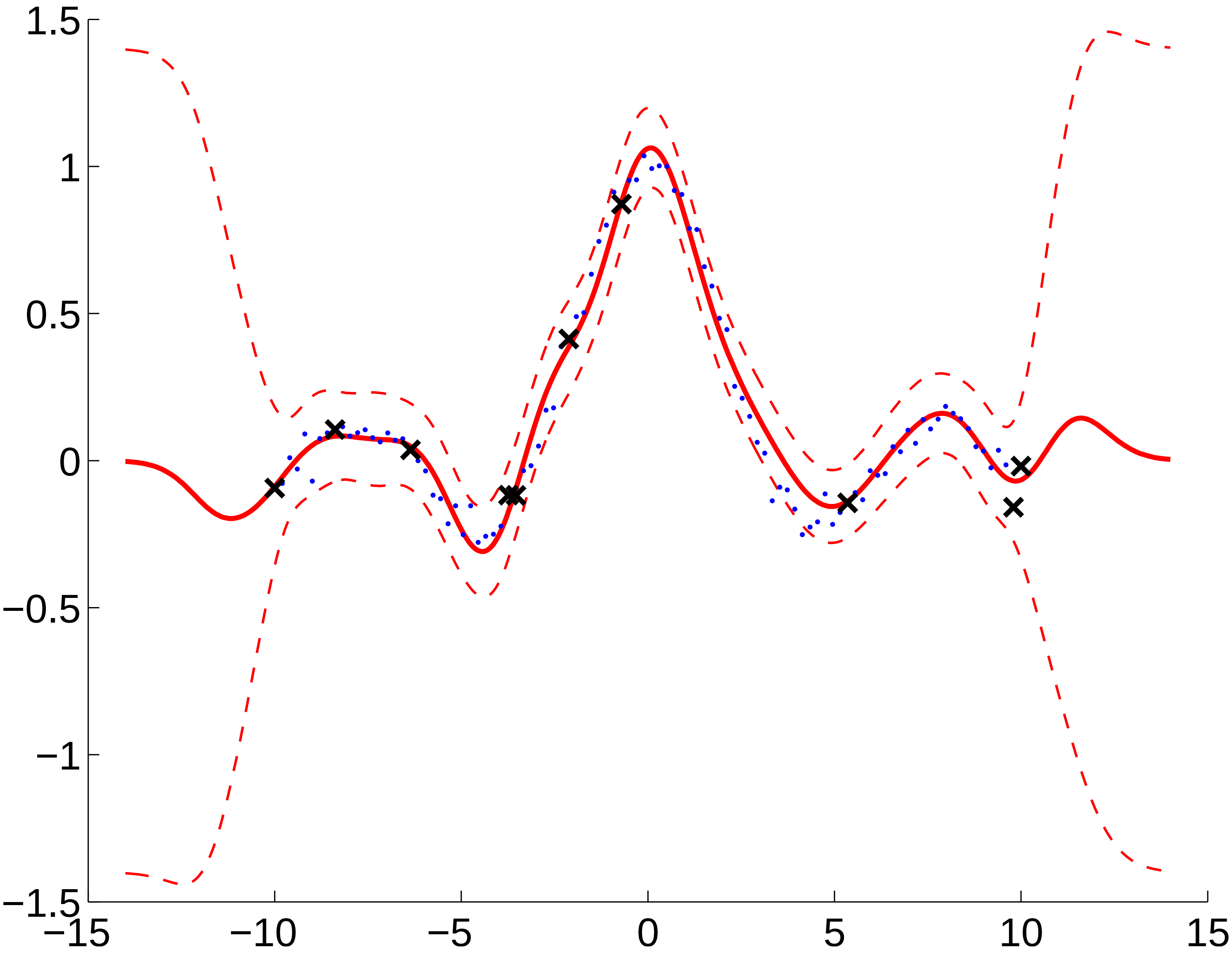
Projected Process Approximation



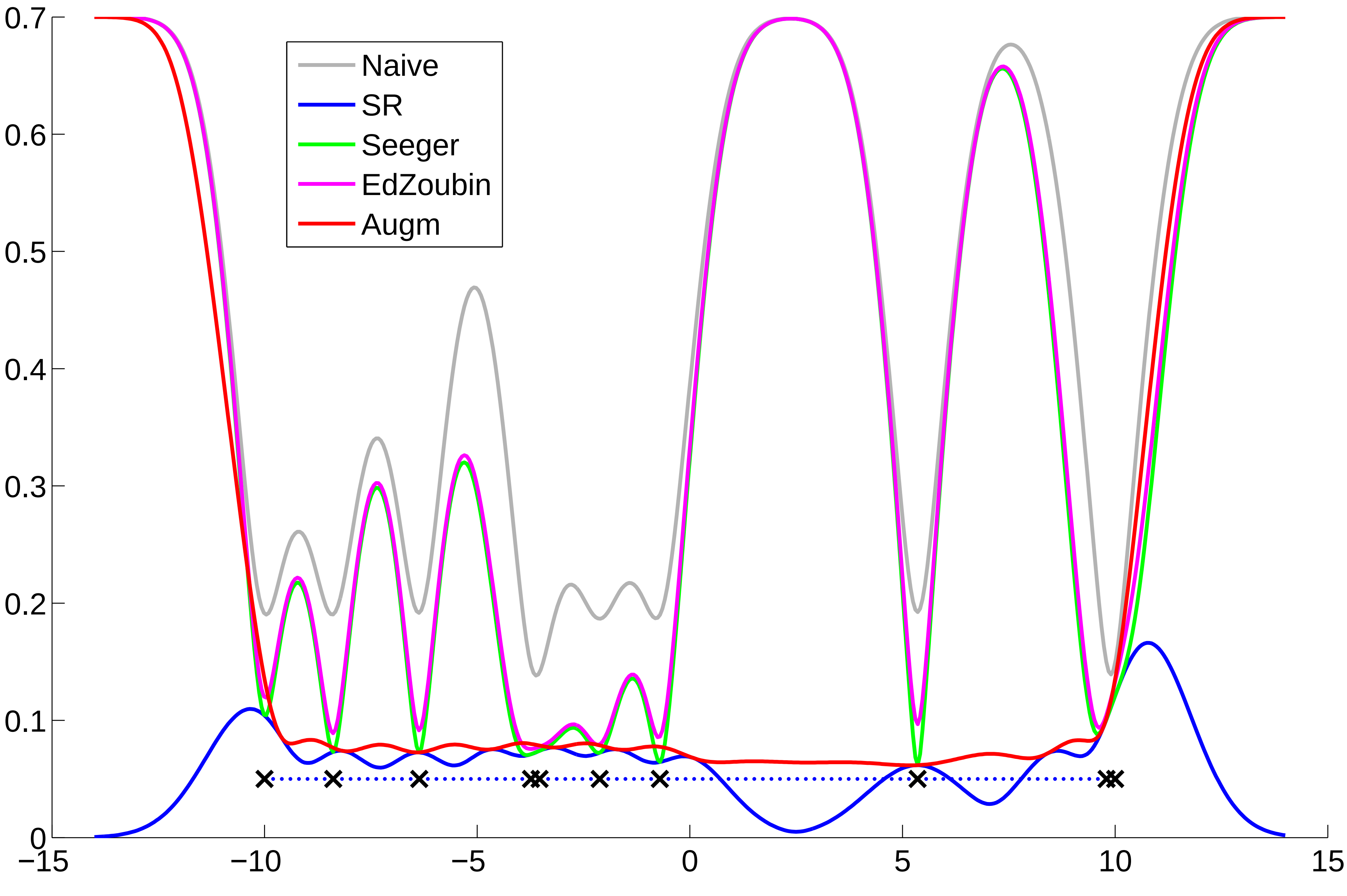
Ed and Zoubin's Projected Process Method



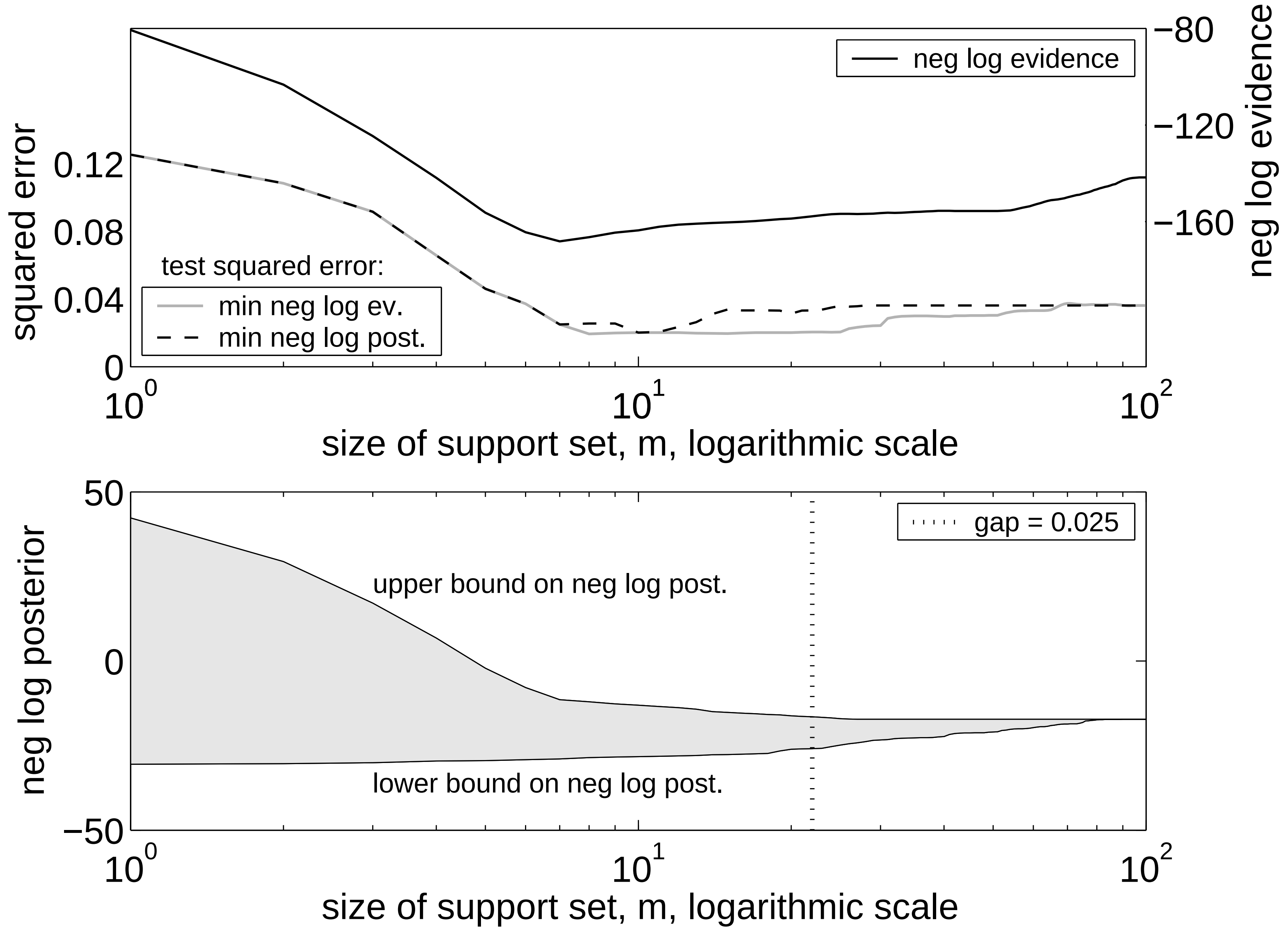
Augmented SoR (pred scales with *nm*)



Comparing the Predictive Uncertainties



Smola and Bartett's Greedy Selection

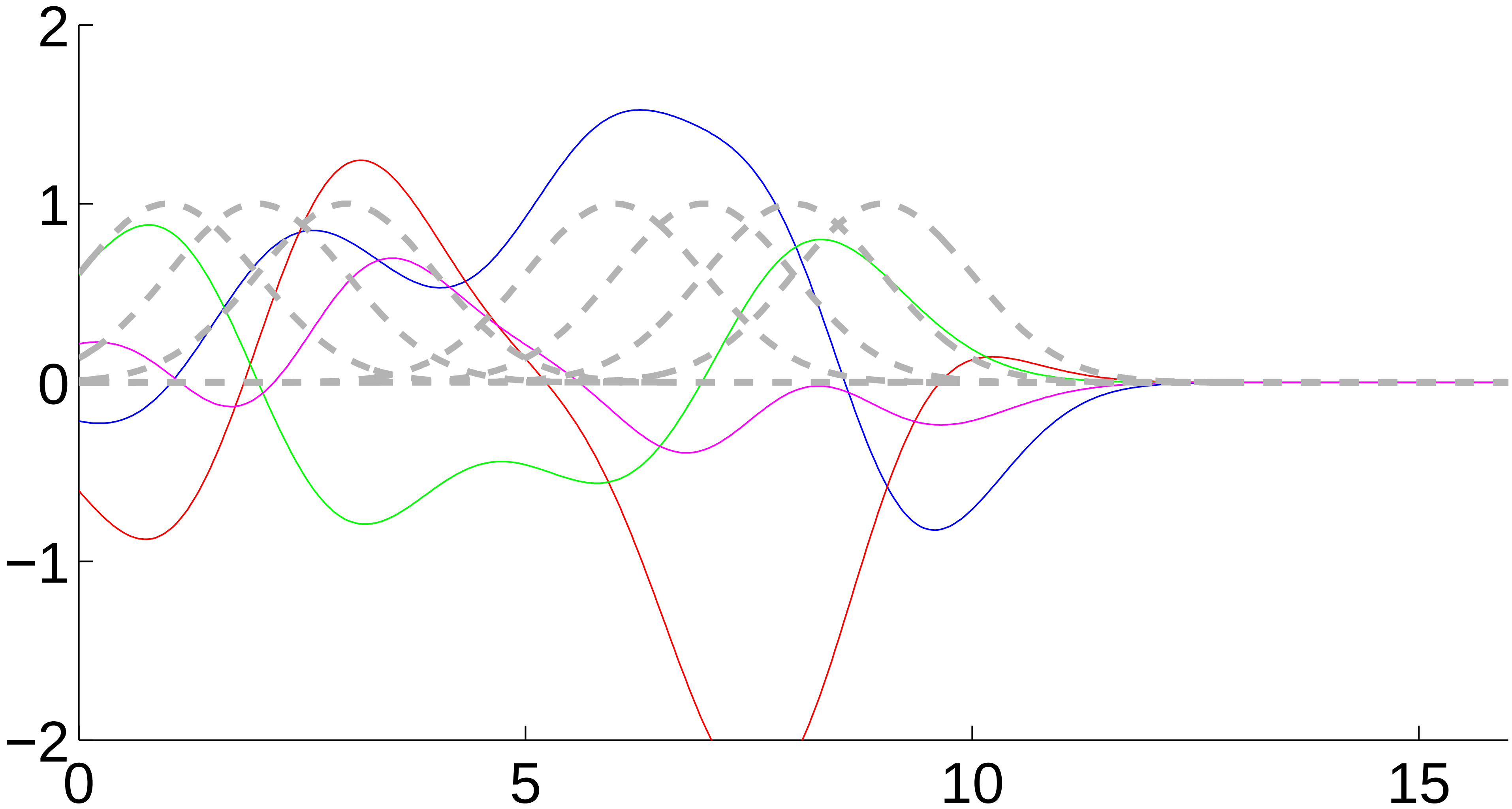


Wrap Up

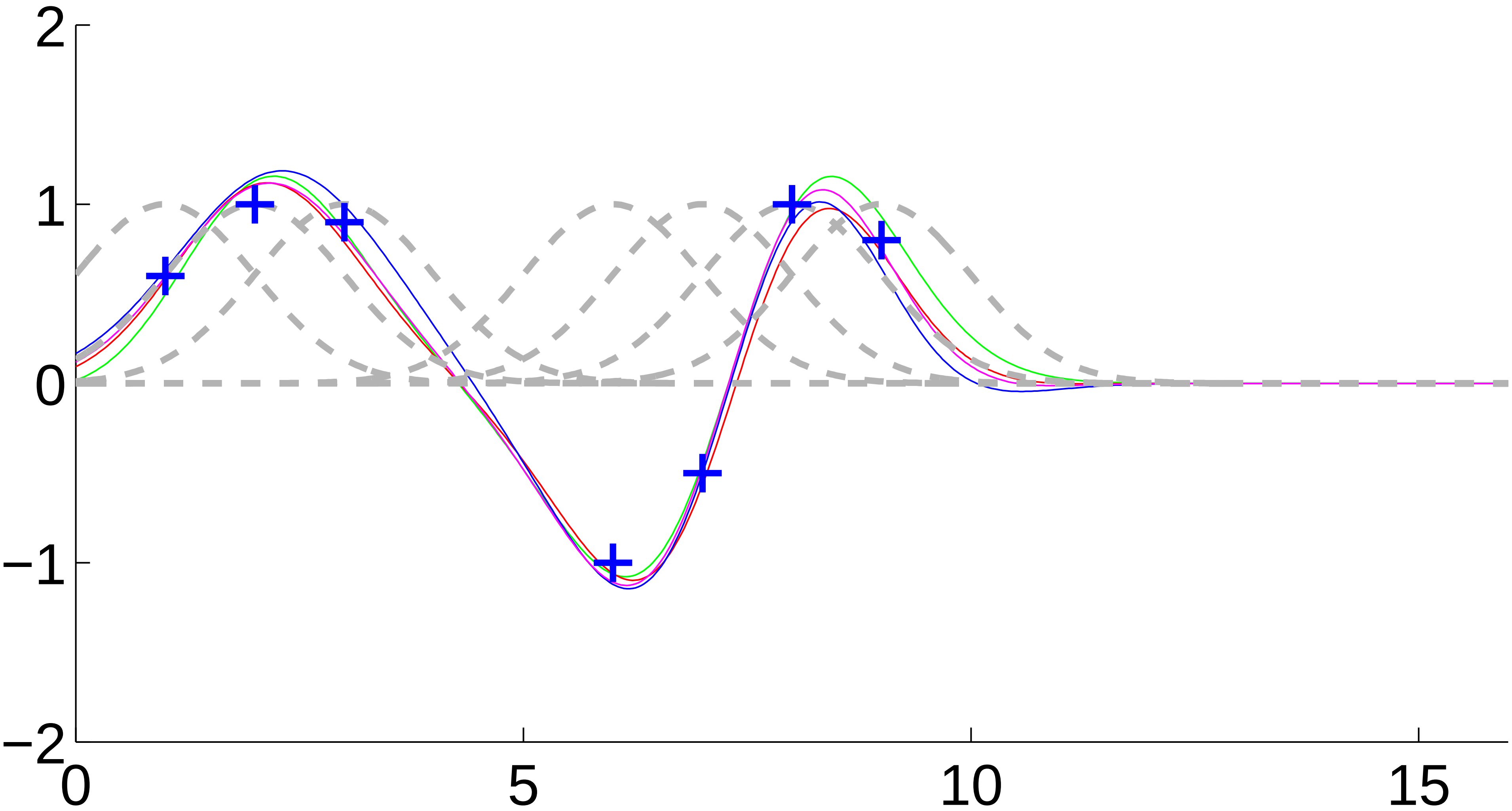
- Training: from $\mathcal{O}(n^3)$ to $\mathcal{O}(nm^2)$
- Predicting: from $\mathcal{O}(n^2)$ to $\mathcal{O}(m^2)$ (or $\mathcal{O}(nm)$)
- Be sparse if you must, but only then
- Beware of over-fitting prone greedy selection methods
- Do worry about the prior implied by the approximation!

Appendix: Healing the RVM by Augmentation
(joint work with Carl Rasmussen)

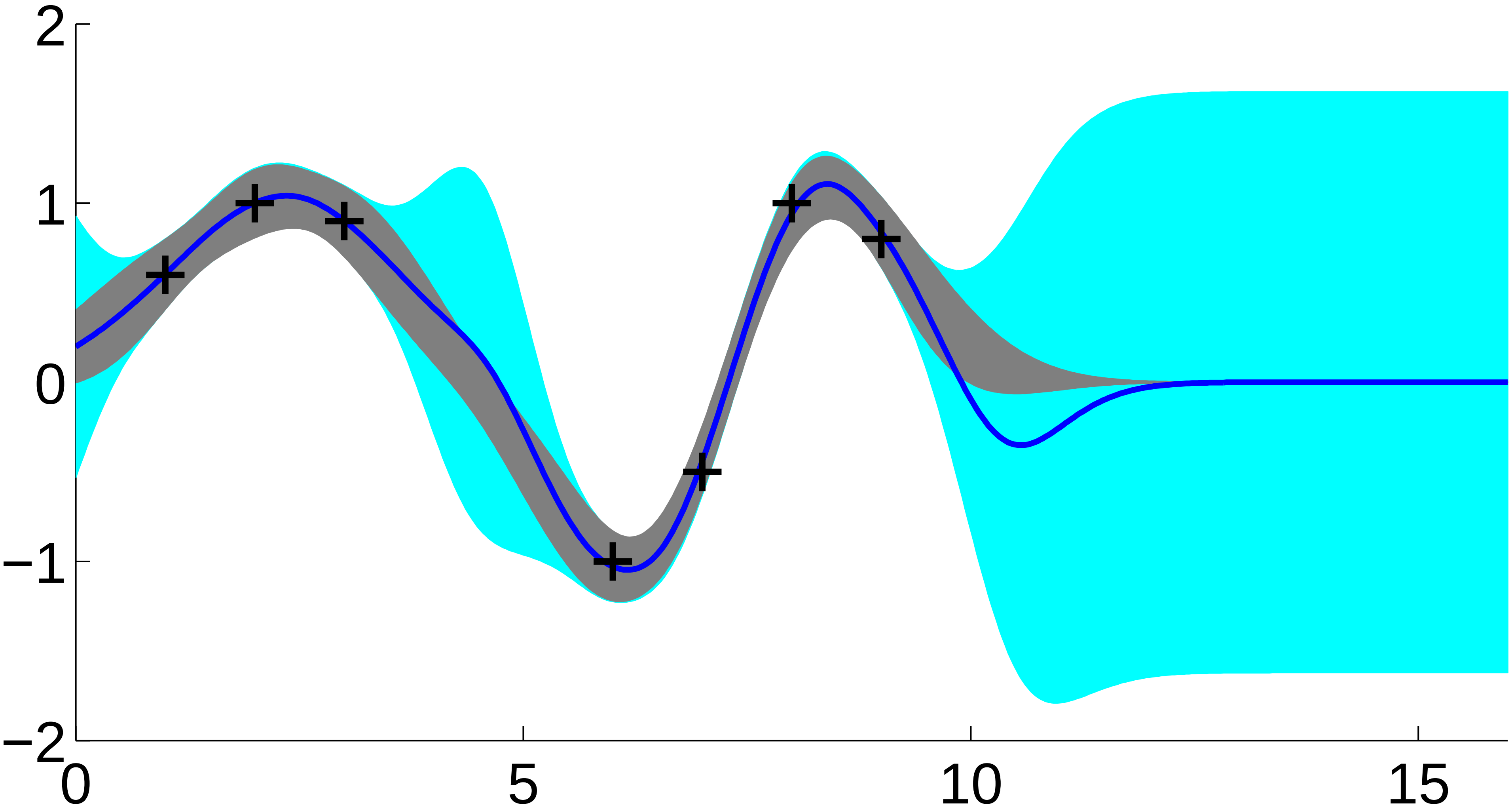
Finite Linear Model



A Bad Probabilistic Model



The Healing: Augmentation



Augmentation?

- Train once your m-dimensional model
- At each new test point add a new basis function
- Update the $m+1$-dimensional model (update posterior)
- Testing is now more expensive

Wait a minute ...

I don't care about probabilistic predictions!

Another Symptom: Underfitting

Abalone

	Squared error loss			Absolute error loss			- log test density loss		
	RVM	RVM*	GP	RVM	RVM*	GP	RVM	RVM*	GP
Loss:	0.138	0.135	0.092	0.259	0.253	0.209	0.469	0.408	0.219
RVM	$\cdot$	not sig.	< 0.01	$\cdot$	0.07	< 0.01	$\cdot$	< 0.01	< 0.01
RVM*		$\cdot$	0.02		$\cdot$	< 0.01		$\cdot$	< 0.01
GP			$\cdot$			$\cdot$			$\cdot$

Robot Arm

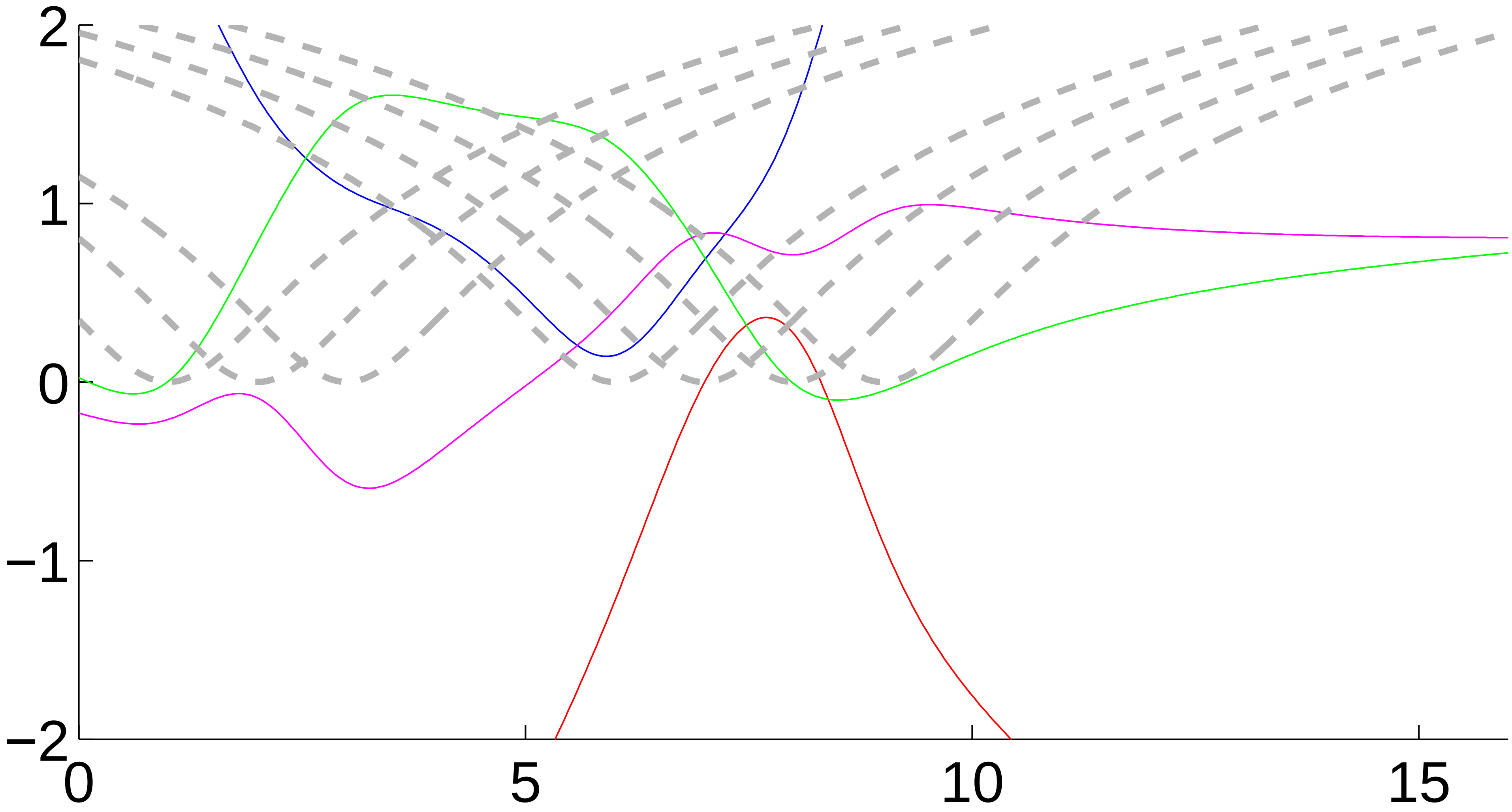
	Squared error loss			Absolute error loss			- log test density loss		
	RVM	RVM*	GP	RVM	RVM*	GP	RVM	RVM*	GP
Loss:	0.0043	0.0040	0.0024	0.0482	0.0467	0.0334	-1.2162	-1.3295	-1.7446
RVM	$\cdot$	< 0.01	< 0.01	$\cdot$	< 0.01	< 0.01	$\cdot$	< 0.01	< 0.01
RVM*		$\cdot$	< 0.01		$\cdot$	< 0.01		$\cdot$	< 0.01
GP			$\cdot$			$\cdot$			$\cdot$

- GP (Gaussian Process): infinitely augmented linear model
- Beats finite linear models in all datasets I've looked at

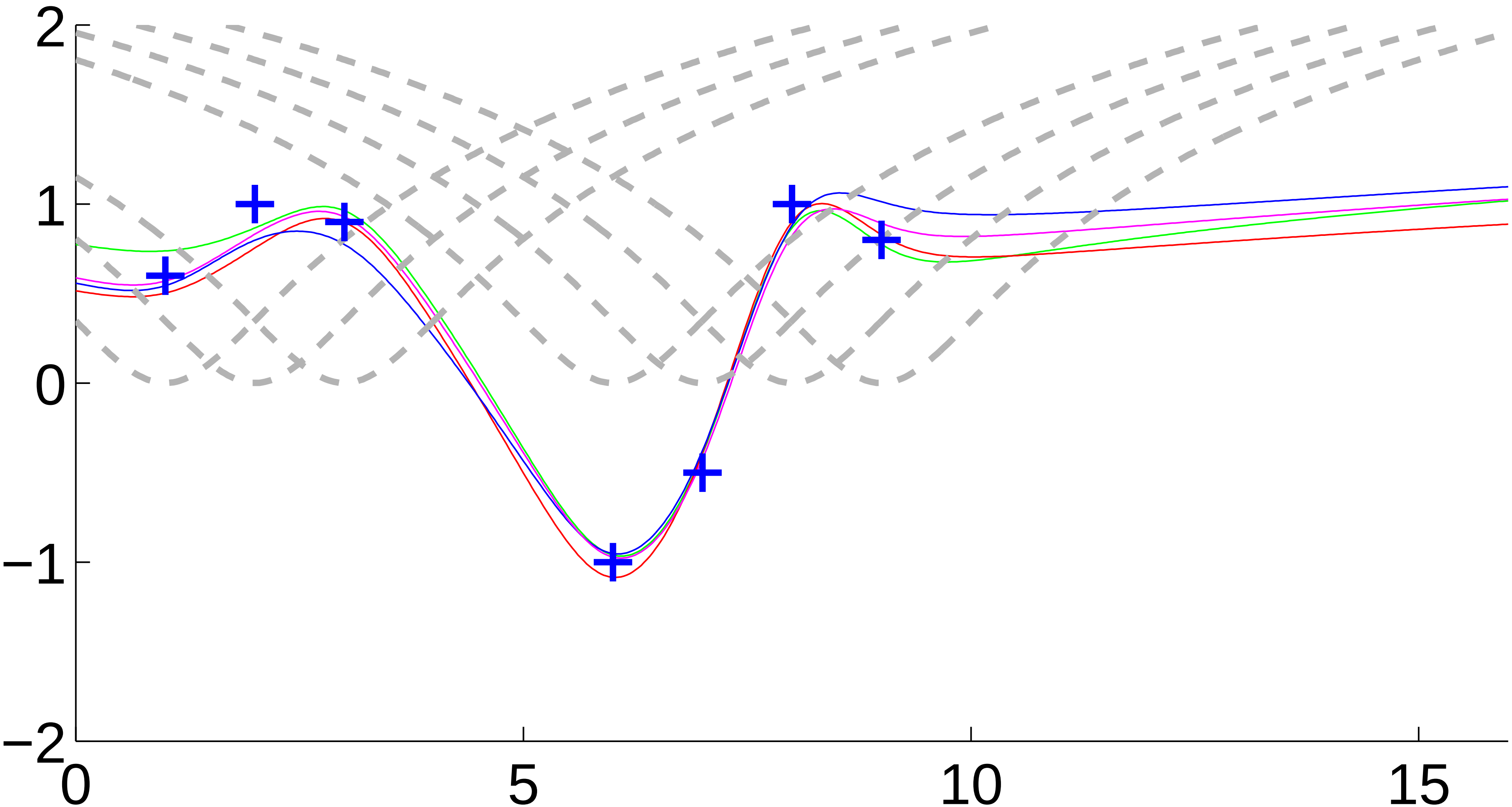
Interlude

None of this happens with non-localized basis functions

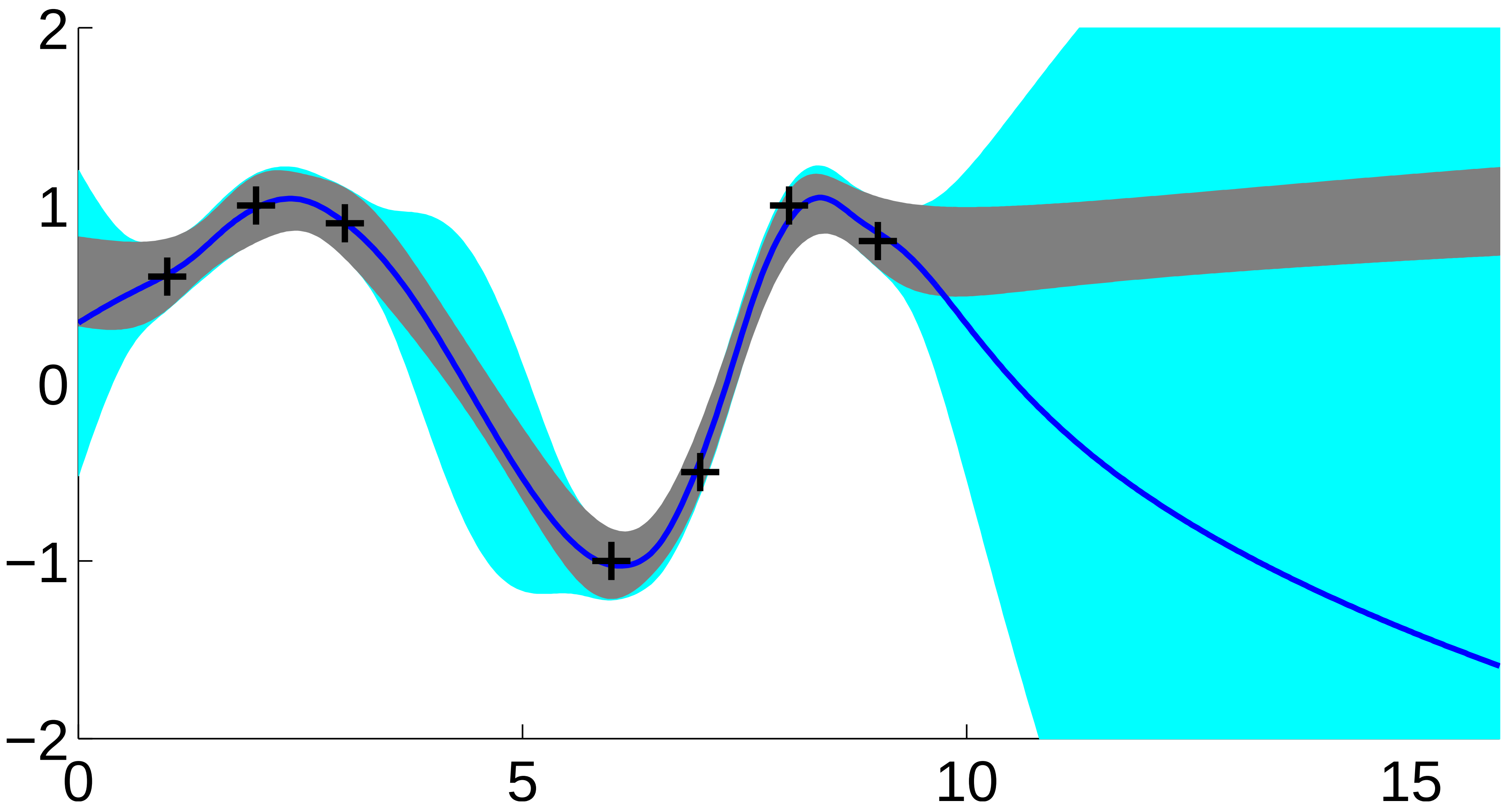
Finite Linear Model



A Bad Probabilistic Model



The Healing: Augmentation



Appendix: Augmentation in Sparse GPs

- $\mathcal{O}(nm^2)$ sparse approx. to Gaussian Processes (Smola and Bartlett, 2001)
- Augmentation: same training, more expensive testing
- Better mean based and probabilistic performance

		non-augmented			*augmented*		
method	tr. neg ev.	MAE	MSE	NTL	MAE	MSE	NTL
SGGP	–	0.0481	0.0048	-0.3525	0.0460	0.0045	-0.4613
SGEV	-1.1555	0.0484	0.0049	-0.3446	0.0463	0.0045	-0.4562
HPEV-rand	-1.0978	0.0503	0.0047	-0.3694	0.0486	0.0045	-0.4269
HPEV-SGEV	-1.3234	0.0425	0.0036	-0.4218	0.0404	0.0033	-0.5918
HPEV-SGGP	-1.3274	0.0425	0.0036	-0.4217	0.0405	0.0033	-0.5920

2000 *training* - 2000 *test*

SGEV	-1.4932	0.0371	0.0028	-0.6223	0.0346	0.0024	-0.6672
HPEV-rand	-1.5378	0.0363	0.0026	-0.6417	0.0340	0.0023	-0.7004

36000 *training* - 4000 *test*

Thanks a lot to Sheffield and to Neil!