A Comparison of Regularization Methods for Gaussian Processes

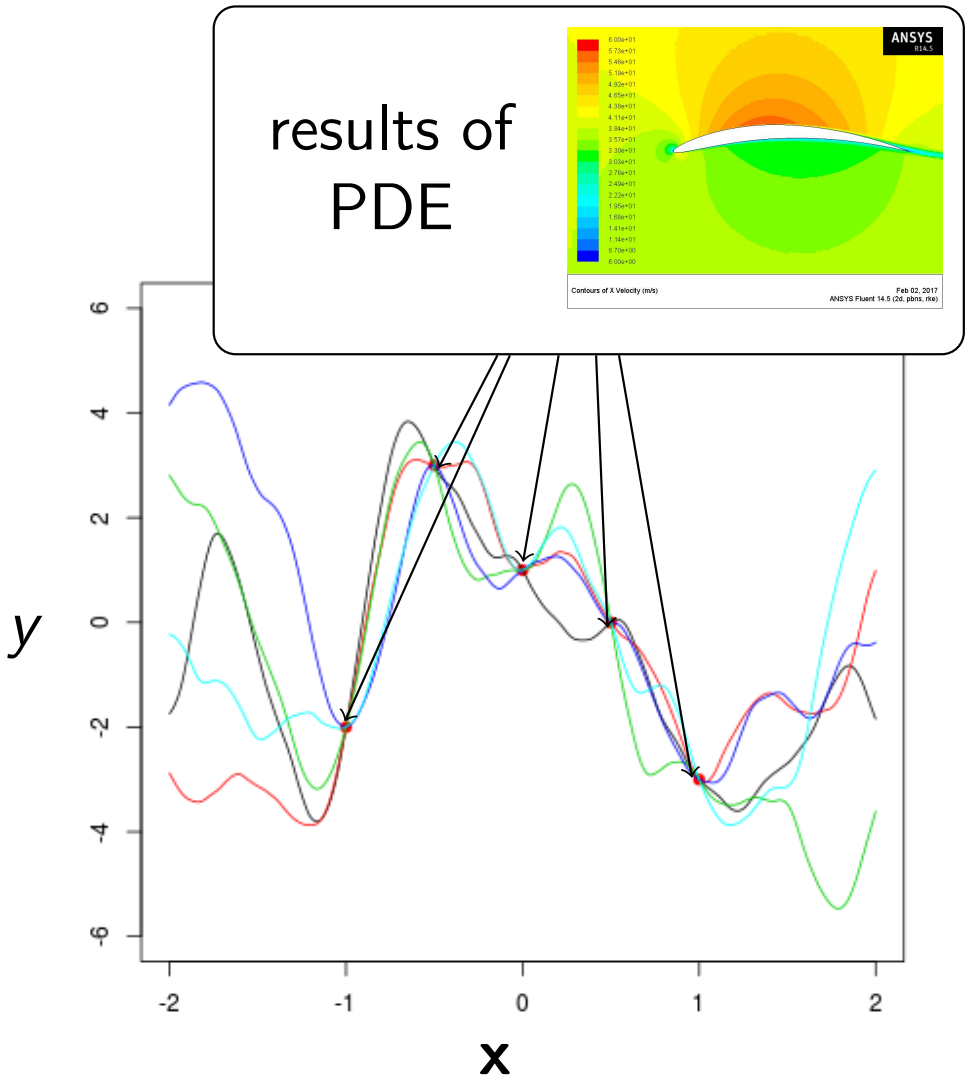
Rodolphe Le Riche[1,2] , Hossein Mohammadi[3] ,
Nicolas Durrande[2] , Eric Touboul[2] , Xavier Bay[2]

[1] CNRS LIMOS, France
[2] Ecole des Mines de Saint Etienne, France
[3] Univ. of Exeter, UK

May 2017

SIAM OP17 conference, Vancouver BC

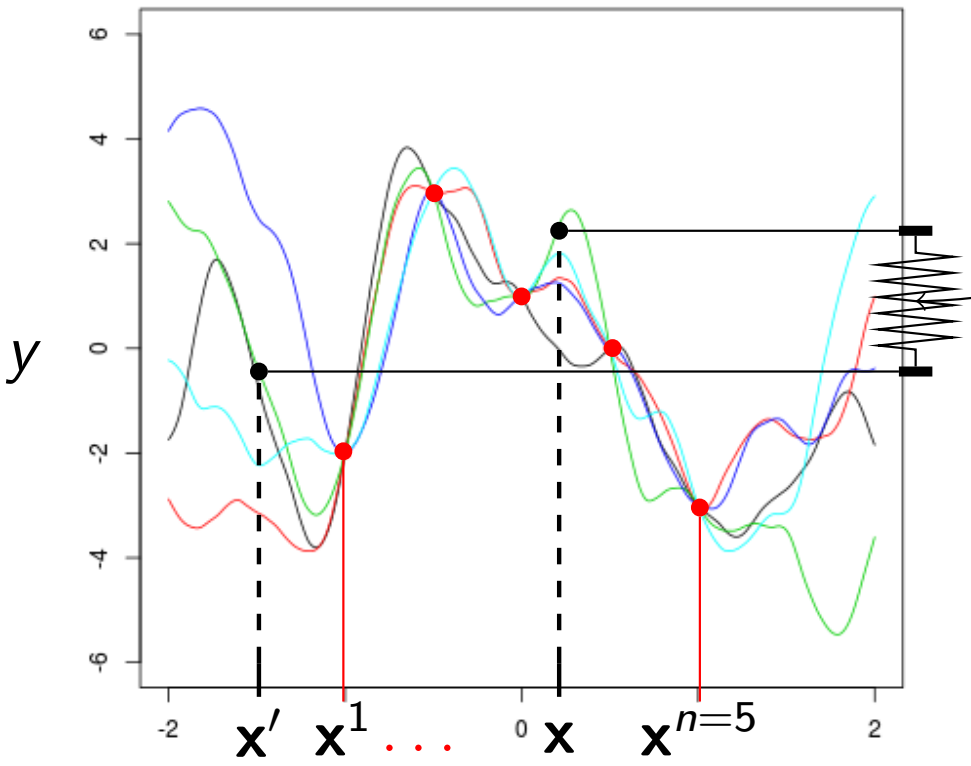
From PDEs to Gaussian Processes



- PDEs parameterized by $\mathbf{x}$ (a shape, boundary conditions ...) and post-processed to yield $y(\mathbf{x})$ (drag, lift, ...).
- To describe the many probable $y(x)$ between the observed/simulated $(\mathbf{x}^i, y(\mathbf{x}^i))$: *conditional Gaussian Processes*, GPs.

GPs and the covariance matrix

Every Gaussian Process involves a covariance matrix $\mathbf{C}$, made of $\mathbf{C}_{ij} = \mathrm{Cov}\big(Y(\mathbf{x}^i), Y(\mathbf{x}^j)\big)$, that must be inverted.



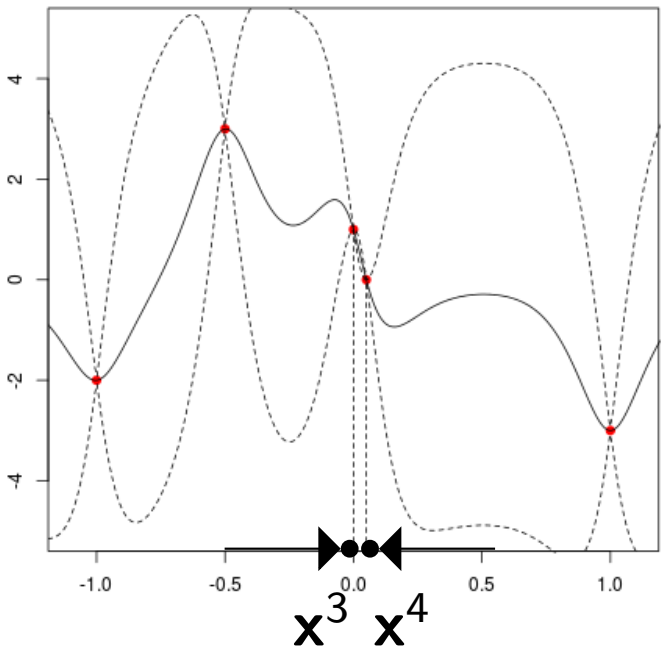
covariance function

(of non conditional GP),

$k(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}(Y(\mathbf{x}), Y(\mathbf{x}'))$

Motivations (1)

But two issues happen. Firstly, points get too close to each other:



the 2 columns become similar as $\mathbf{x}^3 \to \mathbf{x}^4$,

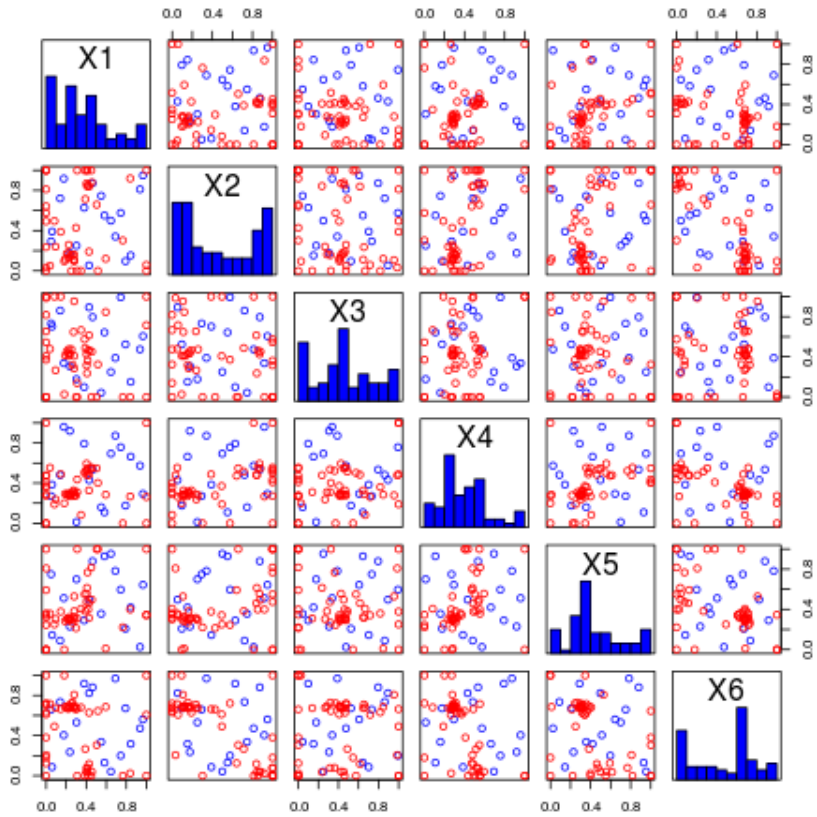
$$\begin{bmatrix} \dots & k(\mathbf{x}^1, \mathbf{x}^3) & k(\mathbf{x}^1, \mathbf{x}^4) & \dots \\ \dots & k(\mathbf{x}^2, \mathbf{x}^3) & k(\mathbf{x}^2, \mathbf{x}^4) & \dots \\ & & \dots & \\ \dots & k(\mathbf{x}^n, \mathbf{x}^3) & k(\mathbf{x}^n, \mathbf{x}^4) & \dots \end{bmatrix}$$

which makes $\mathbf{C}$ ill-conditioned.

Motivations (2)

Examples of points too close:

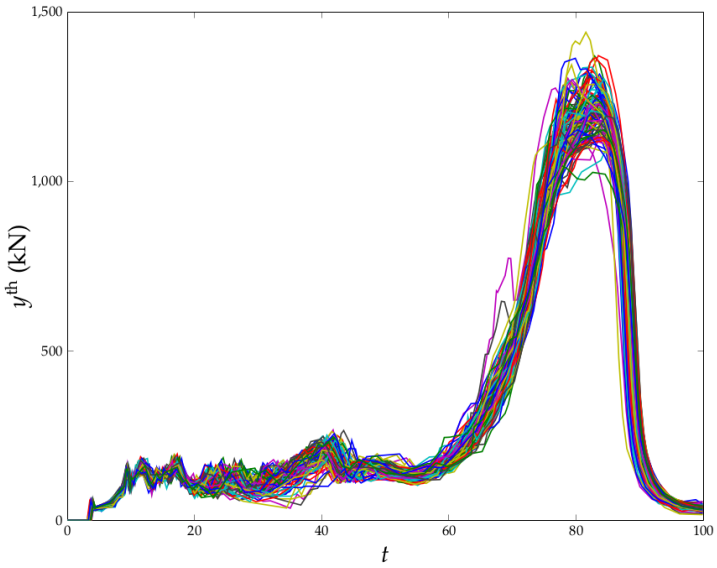
- (right) optimization using GPs with the EGO algorithm [D.R. Jones et al. 1998] on the 6D Hartmann function. Note the clusters of points.
- data spatially distributed as human beings (e.g., phones) is strongly clustered (towns vs. low density areas).

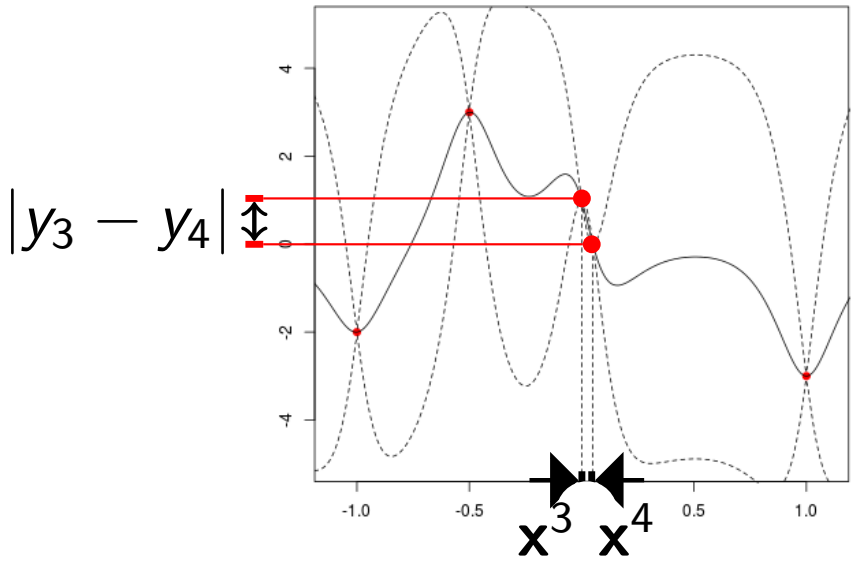


Ill-conditioning even happens without close points: additive kernels and rectangular design patterns, periodic kernels. Cf. HAL report no. 01264192 accompanying the talk (same title).

Motivations (3)

If points are too close, just delete the extra ones? Second issue: they may carry different information.





Examples:

- Repeated performance measures on human patients.
- PDE of chaotic phenomena, expl. of crash, force vs. time for infinitesimal perturbations of the mesh [from M. Maliki, *Adaptive surrogate models for the reliable lightweight design of automotive body structures*, PhD, 2016].

Overview of this work

- Covariance matrix ill-conditioning is endemic: GP softwares always include a regularization strategy, nugget (a.k.a. Tikhonov regularization, ridge regression) or pseudo-inverse.
- What is the difference between nugget and pseudo-inverse? Do they always have the required interpolation properties?

$\Rightarrow$ Clear differences between pseudo-inverse and nugget arise when pushing ill-conditioning to true singularity of $\mathbf{C}$: close points are merged, almost redundant points are made redundant.

$\Rightarrow$ Propose another regularization approach, the distribution-wise GP.

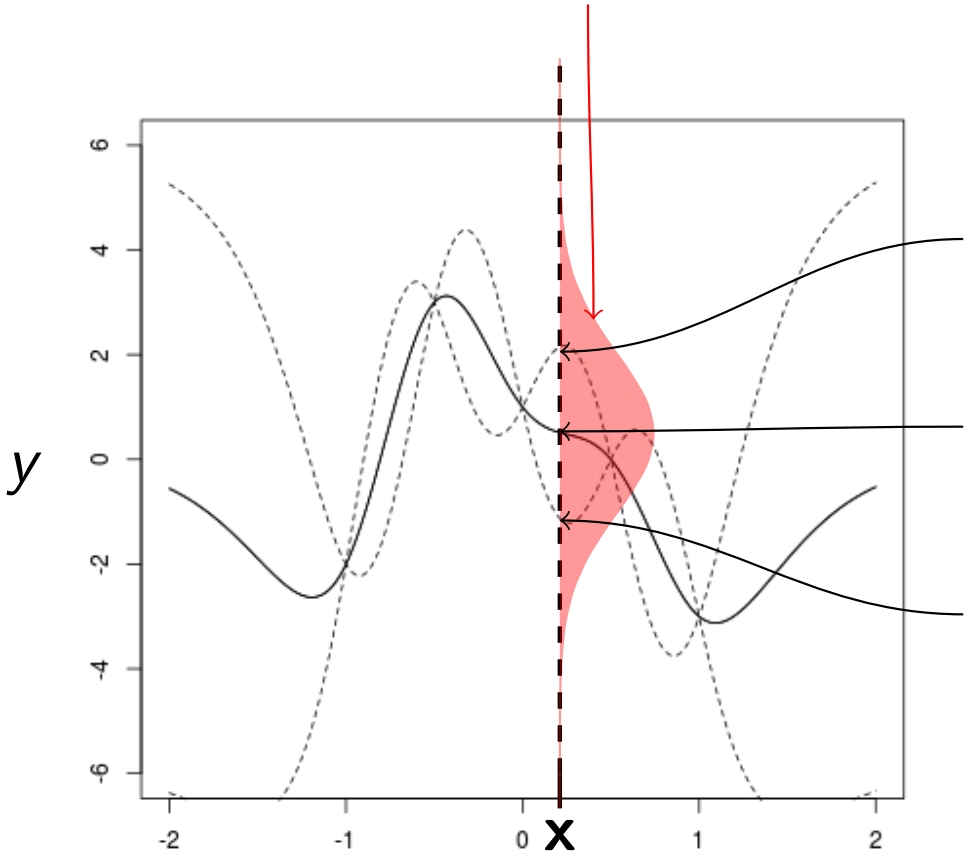
Related studies

In order to deal with **C** ill-conditioning in Gaussian Processes, the following approaches have been proposed.

- **Matrix regularization** (more general than GPs): nugget and pseudo-inverse, see later.
- **Choice of the design points, X**: [Salagame and Barton, *Factorial Des. for Spat. Correl. Reg.*, J. of Appl. Stats., 97], [Rennen, *Subset select. from large datasets for krg. modeling*, SMO, 09], ...
- **Choice of the covariance function**, $k(\mathbf{x}, \mathbf{x}')$: [Davis and Morris, *6 factors which affect the condit. nb of mat. associated with krg*, Math. Geol., 97], [Belsley, *Regression Diagnostics*, 2005], ...
- **GPs without C inversion**: [Gibbs, *Bayesian GPs for Reg. and Classif.*, PhD, 97].

Conditional Gaussian Processes: density

$$Y(\mathbf{x}) \mid \mathbf{x}^1, y(\mathbf{x}^1), \ldots, \mathbf{x}^n, y(\mathbf{x}^n) \sim \mathcal{N}(m(\mathbf{x}), c(\mathbf{x}, \mathbf{x})) \,, \begin{cases} \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d \\ y \in \mathbb{R} \end{cases}$$



$$m(\mathbf{x}) + \underbrace{c(\mathbf{x}, \mathbf{x})}_{\equiv v(\mathbf{x})}$$

$$m(\mathbf{x})$$

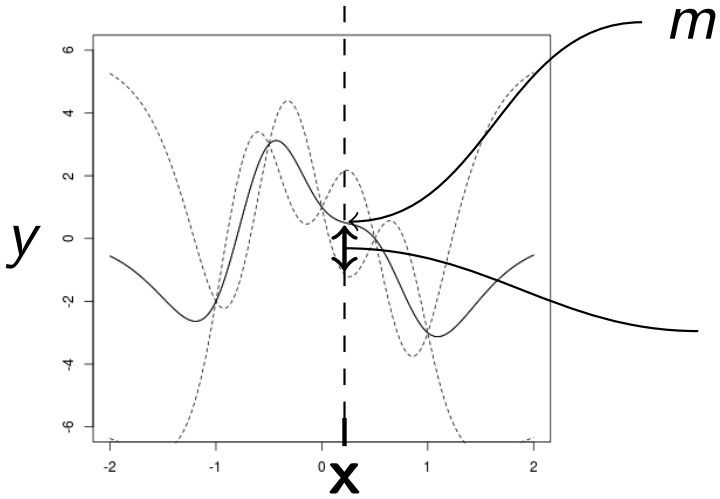
$$m(\mathbf{x}) - c(\mathbf{x}, \mathbf{x})$$

Cf. [Rasmussen and Williams, *Gaussian Processes for Machine Learning*, 2006]

Gaussian Processes: conditional mean and covariance

$$Y(\mathbf{x}) \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(m(\mathbf{x}), c(\mathbf{x}, \mathbf{x}))$$

(reminders in gray)



$$m(\mathbf{x}) = \mathbf{c}(\mathbf{x})^\top \mathbf{C}^{-1} \begin{pmatrix} y(\mathbf{x}^1) \\ \dots \\ y(\mathbf{x}^n) \end{pmatrix} \quad \mathbf{y}$$

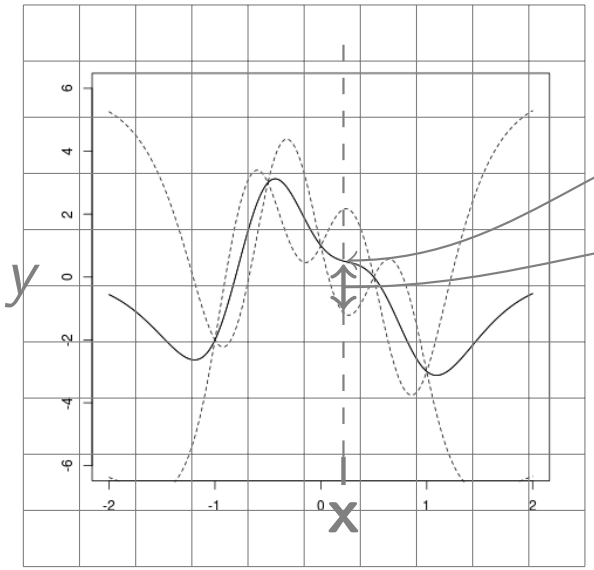
$$c(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{c}(\mathbf{x})^\top \mathbf{C}^{-1} \mathbf{c}(\mathbf{x}')$$

$$(\text{at } \mathbf{x}' = \mathbf{x}\ ,\ c(\mathbf{x}, \mathbf{x}) \equiv v(\mathbf{x}))$$

(next: the covariances $\mathbf{c}(\mathbf{x})$ and $\mathbf{C}$)

Gaussian Processes: covariance vector and matrix

$$Y(\mathbf{x}) \mid \mathbf{X}, \mathbf{y} \sim \mathcal{N}(m(\mathbf{x}), c(\mathbf{x}, \mathbf{x}))$$



$$m(\mathbf{x}) = \mathbf{c}(\mathbf{x})^\top \mathbf{C}^{-1} \mathbf{y}$$

$$c(\mathbf{x}, \mathbf{x}) = k(\mathbf{x}, \mathbf{x}) - \mathbf{c}(\mathbf{x})^\top \mathbf{C}^{-1} \mathbf{c}(\mathbf{x})$$

Covariance *vector* :

$$\mathbf{c}(\mathbf{x})^\top = [k(\mathbf{x}^1, \mathbf{x}), \ldots, k(\mathbf{x}^n, \mathbf{x})]$$

Covariance *matrix* : $\mathbf{C} =$

$$\begin{bmatrix} k(\mathbf{x}^1, \mathbf{x}^1) & k(\mathbf{x}^1, \mathbf{x}^2) & \ldots & k(\mathbf{x}^1, \mathbf{x}^n) \\ k(\mathbf{x}^2, \mathbf{x}^1) & k(\mathbf{x}^2, \mathbf{x}^2) & \ldots & k(\mathbf{x}^2, \mathbf{x}^n) \\ & \ldots & & \\ k(\mathbf{x}^n, \mathbf{x}^1) & k(\mathbf{x}^n, \mathbf{x}^2) & \ldots & k(\mathbf{x}^n, \mathbf{x}^n) \end{bmatrix}$$

(covariance *function*)

$$\mathbf{X} = \{\mathbf{x}^1, \ldots, \mathbf{x}^n\}, \quad m(\mathbf{X}) \equiv \begin{pmatrix} m(\mathbf{x}^1) \\ \ldots \\ m(\mathbf{x}^n) \end{pmatrix} = \mathbf{C}\mathbf{C}^{-1}\mathbf{y} \in Im(\mathbf{C})$$

$\mathbf{C}$ Eigenanalysis

Image space

$$Im(\mathbf{C}) = \text{span}(\overbrace{\mathbf{V}^1 \dots \mathbf{V}^r}^{\mathbf{V}})$$

Null space (non empty, $\mathbf{C}$ singular)

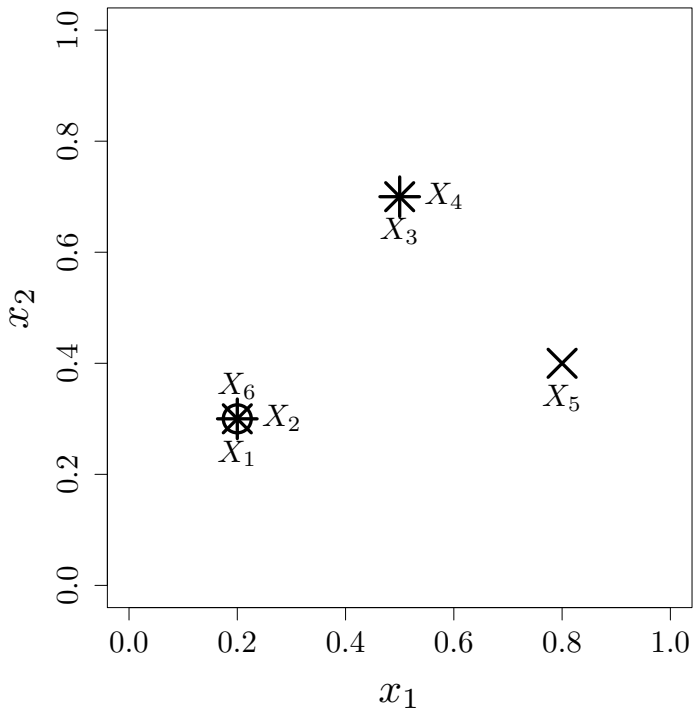
$$Null(\mathbf{C}) = \text{span}(\overbrace{\mathbf{W}^1 \dots \mathbf{W}^{n-r}}^{\mathbf{W}})$$

$$\mathbf{C} = [\mathbf{V}\ \mathbf{W}] \begin{bmatrix} diag\,(\boldsymbol{\lambda})_{r\times r} & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(n-r)\times r} & \mathbf{0}_{(n-r)\times(n-r)} \end{bmatrix} [\mathbf{V}\ \mathbf{W}]^\top$$

Definition of redundant points

Redundant points definition: the points responsible for the linear dependency in the columns of $\mathbf{C}$. Their indices are the non-zero off-diagonal terms of $\mathbf{V}\mathbf{V}^\top$.

Expl. where points $\{1, 2, 6\}$ and $\{3, 4\}$ are redundant (actually repeated):



$$\mathbf{V}\mathbf{V}^\top = \begin{bmatrix} 0.33 & 0.33 & 0.00 & 0.00 & 0.00 & 0.33 \\ 0.33 & 0.33 & 0.00 & 0.00 & 0.00 & 0.33 \\ 0.00 & 0.00 & 0.50 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.50 & 0.50 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 \\ 0.33 & 0.33 & 0.00 & 0.00 & 0.00 & 0.33 \end{bmatrix}$$

$$k(\mathbf{x}, \mathbf{x'}) = \exp\left(-\frac{(x_1 - x_1')^2}{2 \times .25^2}\right) \times \exp\left(-\frac{(x_2 - x_2')^2}{2 \times .25^2}\right)$$

Definitions and properties framed with a green background are, to the authors' knowledge, "new".

Kriging with Pseudo-Inverse (1)

E.g., in [Siefert et al., *MAPS software*].

Moore-Penrose Pseudo-Inverse

$$\mathbf{C}^{\dagger} = [\mathbf{V}\ \mathbf{W}] \begin{bmatrix} diag\left(\frac{1}{\lambda}\right)_{r\times r} & \mathbf{0}_{r\times(n-r)} \\ \mathbf{0}_{(n-r)\times r} & \mathbf{0}_{(n-r)\times(n-r)} \end{bmatrix} [\mathbf{V}\ \mathbf{W}]^{\top}$$

$$m(\mathbf{x}) = \mathbf{c}(\mathbf{x})^{\top} \underbrace{\cancel{\mathbf{C}^{-1}}}_{\text{(singularity)}} \mathbf{y} \quad \Longrightarrow \quad \boxed{m^{PI}(\mathbf{x}) = \mathbf{c}(\mathbf{x})^{\top}\mathbf{C}^{\dagger}\mathbf{y}}$$

(idem with $c(\mathbf{x},\mathbf{x})$)

PI kriging property 1: the PI kriging prediction at data points $\mathbf{X}$ is the orthogonal projection of the observations onto $Im(\mathbf{C})$,

$$m^{PI}(\mathbf{X}) = \mathbf{V}\mathbf{V}^{\top}\mathbf{y}$$

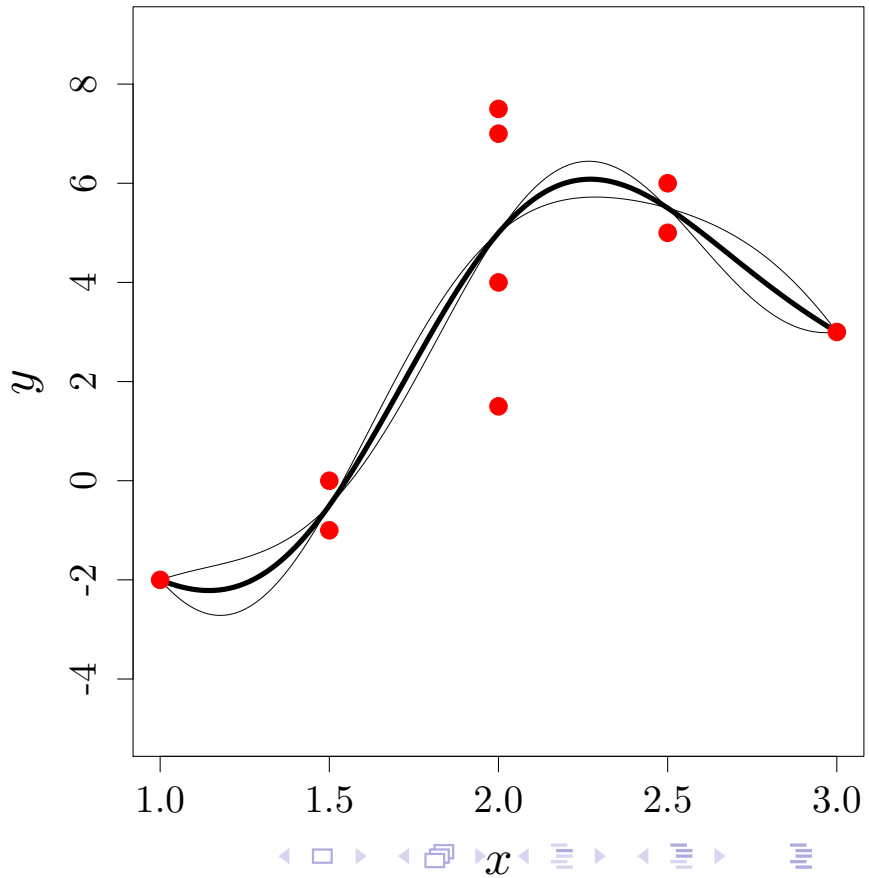
(proof straightforward, cf. report)

$m^{PI}()$ is all the kriging model can express.

Kriging with Pseudo-Inverse (2)

PI kriging properties 2&3: At data points, repeated or not, the PI kriging averages the output values and the PI variance is zero.

Proof of Property 2: exhibit the projection matrix which is made of averaging formula at repeated points. Proof of Property 3: direct application of PI property $\mathbf{C}\mathbf{C}^{\dagger}\mathbf{C} = \mathbf{C}$.



Kriging with nugget (1)

The most often used regularization approach, a.k.a. Tikhonov regularization [A. Tikhonov, 1943], ridge regression [A. Hoerl, 1962].
In GPs, see for expl. [Andrianakis and Challenor, *The effect of the nugget on Gaussian process emulators of computer models*, Comput. Stats. & Data Anal., 12], [Ranjan et al., *A computationally stable approach to GP interpolation of deterministic computer simulation*, Technometrics, 11], [Gramacy and Lee, *Cases for the nugget in modeling computer experiments*, Stats. & Computing, 12], ...

Principle: **add τ^2 to the diagonal**

$$m(\mathbf{x}) = \mathbf{c}(\mathbf{x})^\top \underbrace{\cancel{\mathbf{C}^{-1}}}_{\text{(singularity)}} \mathbf{y} \quad \Longrightarrow \quad \boxed{m^{Nug}(\mathbf{x}) = \mathbf{c}(\mathbf{x})^\top (\mathbf{C} + \tau^2 \mathbf{I})^{-1} \mathbf{y}}$$

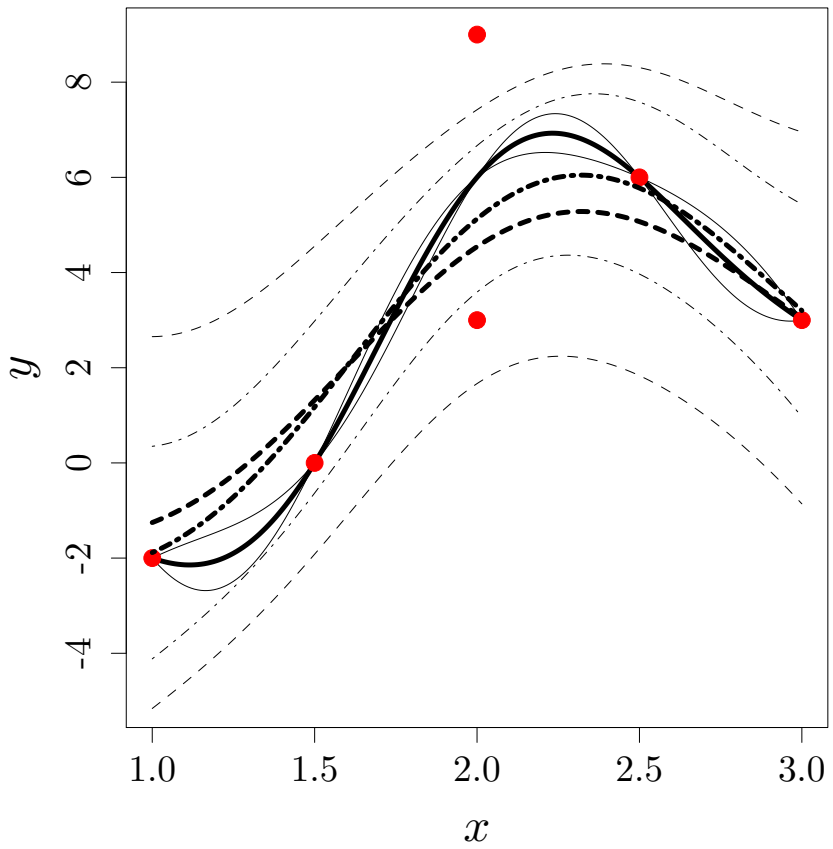
(idem with $c(\mathbf{x}, \mathbf{x})$)

Kriging with nugget (2)

$$\mathbf{C} \quad \Longrightarrow \quad (\mathbf{C} + \tau^2 \mathbf{I})$$

With nugget, kriging no longer interpolates and has a non zero variance at data points.

Right plot: PI (solid lines), nugget estimated by *maximum likelihood* (dashed lines) and nugget estimated by *cross-validation* (dash-dotted lines)



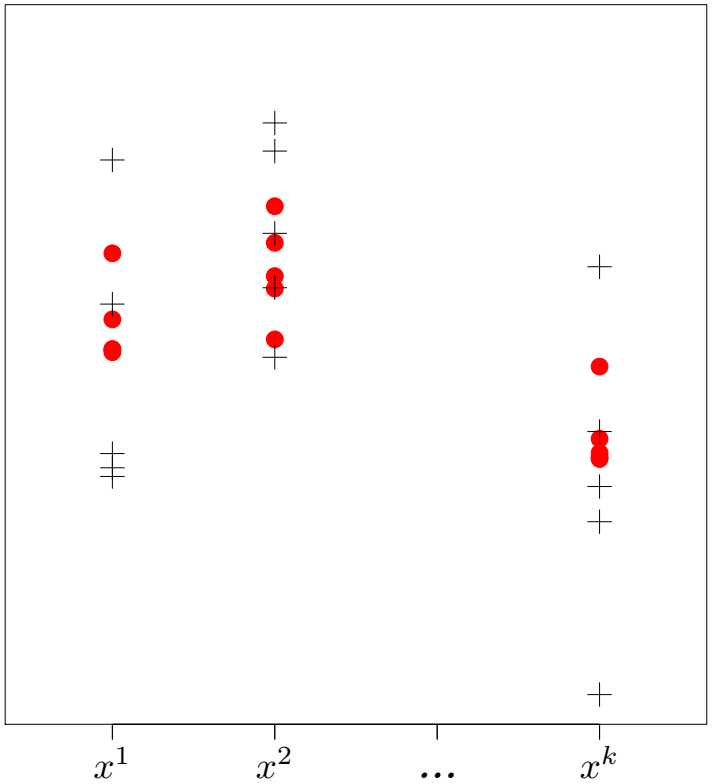
Nugget estimation

$\mathbf{C} \quad \Longrightarrow \quad (\mathbf{C} + \tau^2 \mathbf{I}) \quad ,$ an intuitive result:

Nugget ML property: the nugget estimated by maximum likelihood, $\widehat{\tau}^2$, increases with the spread at repeated points.

Proof: eigendecomposition and monotonicities in the log-likelihood w.r.t. spread and nugget.

On the right, $\widehat{\tau^+}^2$ associated to $+$, $\widehat{\tau}^2$ to $\bullet$, and $\widehat{\tau^+}^2 > \widehat{\tau}^2$.



GP model-data discrepancy

Definition: Model-data discrepancy,

$$discr = \frac{\|\mathbf{y} - m^{PI}(\mathbf{X})\|^2}{\|\mathbf{y}\|^2} = \frac{\|\mathbf{W}\mathbf{W}^\top \mathbf{y}\|^2}{\|\mathbf{y}\|^2} \text{ if } r < n$$
$$= 0 \text{ if } r = n$$

- $0 \leq discr \leq 1$
- How to change $\mathbf{y}$ to reduce *discr* (if applicable): step along

$$-\nabla_{\mathbf{y}} \|\mathbf{y} - m^{PI}(\mathbf{X})\|^2 = -\mathbf{W}\mathbf{W}^\top \mathbf{y}$$

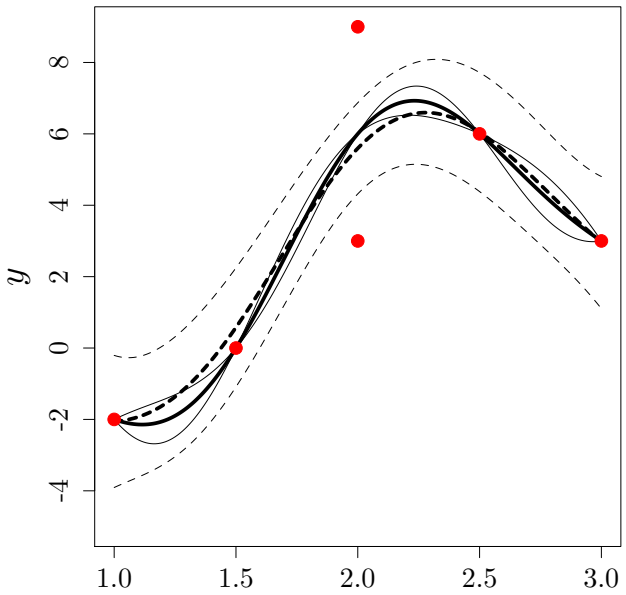
PI vs. Nugget GP regularization (1)

PI - nugget equivalence property: as the nugget decreases to 0, the mean and covariance of GPs regularized by PI and nugget tend towards each other.

The proof involves eigendecomposition and the following property:

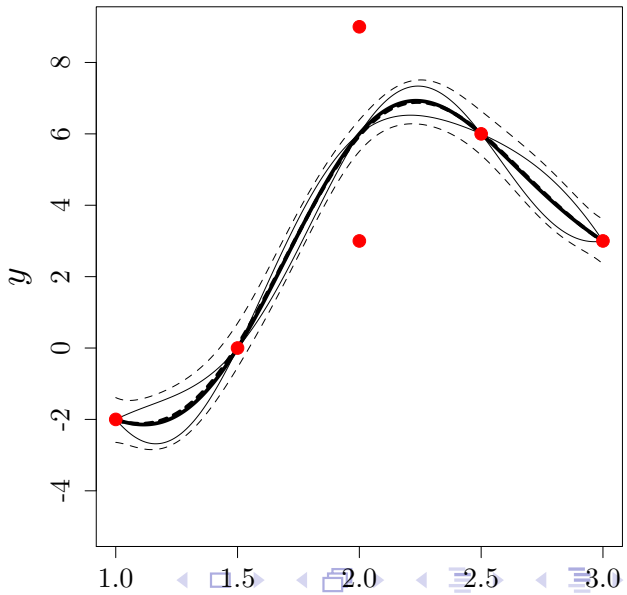
Covariance vector property: $\mathbf{c}(\mathbf{x})$ is perpendicular to the null space of $\mathbf{C}$.

Proof based on the positive-definitness of the extended cov. matrix at $(\mathbf{x}, \mathbf{X})^\top$



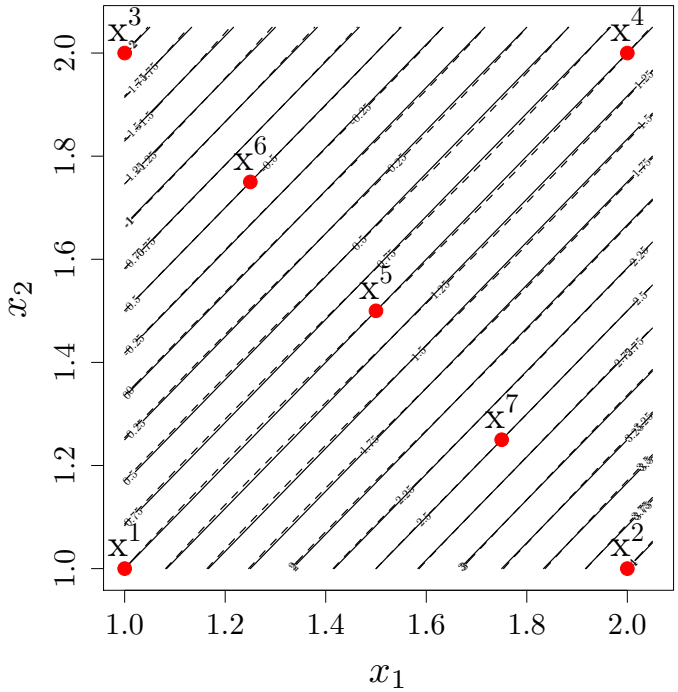
$\tau^2 = 1$ ←

$\tau^2 = 0.1$ →



PI vs. Nugget GP regularization (2)

- As the model-data discrepancy decreases ($\mathbf{C}$ remains singular), the nugget maximizing the (regularized) maximum likelihood tends to 0 $\Rightarrow$ PI and nugget regularizations become equivalent.
- A non-zero discrepancy affects $m^{Nug}()$ throughout the $\mathcal{X}$ space while it only affects $m^{PI}()$ at the redundant points.

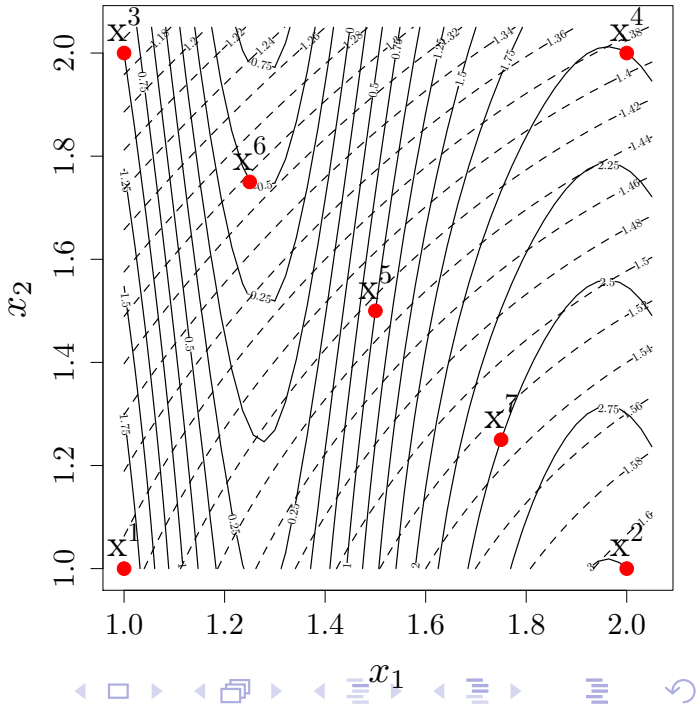


additive kernel and function, $discr = 0$, $\widehat{\tau}^2 = 10^{-2}$

$\longleftarrow$

$\longrightarrow$

y_3 changed, no longer additive, $\widehat{\tau}^2 \approx 1.91$

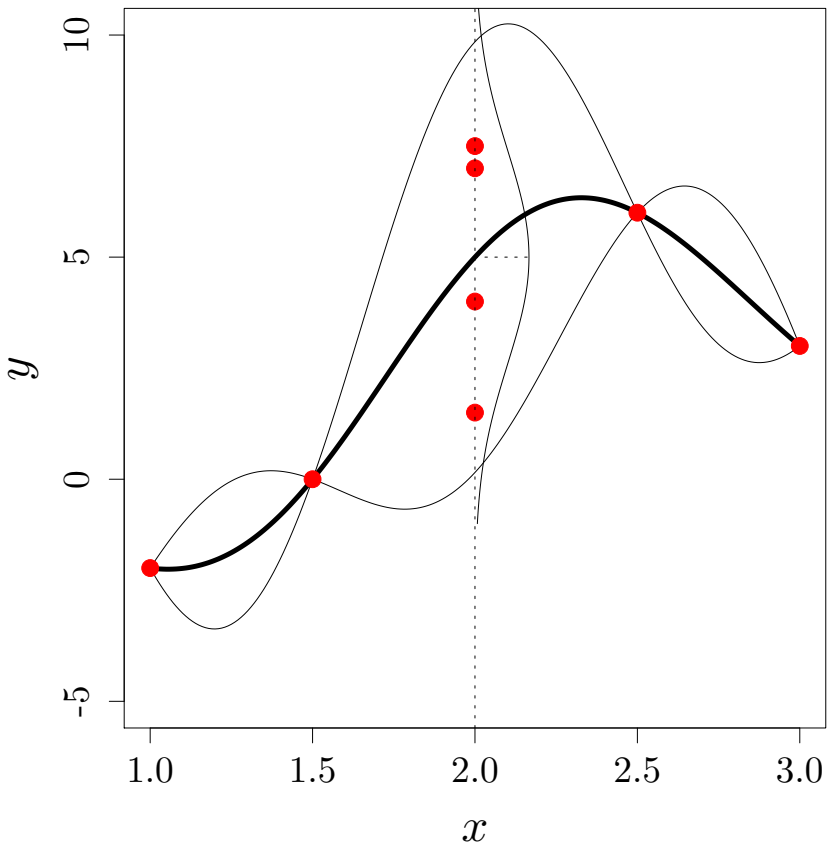


Interpolating repeated points?

The point-wise definition of interpolation does not apply. We wish a GP model

- whose trajectories pass through uniquely defined data points ($c(\mathbf{x}^i, \mathbf{x}^i) = 0$) and
- whose process mean and variance equals the empirical mean and variance at repeated points.

$\Rightarrow$ interpolate distributions.



Distribution-wise GP (1)

Derived and used differently in [Titsias, *Variational learning of inducing var. in sparse GPs*, JMLR, 2009]

Derivation: assume that the observations are actually random, $y(\mathbf{x}^i) \to Z(\mathbf{x}^i)$. k number of unique points (repeated points counted once, e.g., $k = 5$, $n = 8$ on previous plot). Use the law of total expectation and variance,

$$m^{Dist}(\mathbf{x}) = \mathbb{E}_Z\left(\mathbb{E}_\Omega(Y(\mathbf{x})|y(\mathbf{x}^i) = Z(\mathbf{x}^i)\ ,\ 1 \leq i \leq k\right) = \mathbb{E}_Z\left(\mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1}\mathbf{Z}\right) = \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1}\mathbb{E}(\mathbf{Z})$$

$$c^{Dist}(\mathbf{x},\mathbf{x}) = \mathbb{E}_Z \mathrm{Var}_\Omega\left(Y(\mathbf{x})|y(\mathbf{x}^i) = Z(\mathbf{x}^i)\ ,\ 1 \leq i \leq k\right) + \mathrm{Var}_Z\left(\mathbb{E}_\Omega(Y(\mathbf{x})|y(\mathbf{x}^i) = Z(\mathbf{x}^i)\ ,\ 1 \leq i \leq k\right) = k(\mathbf{x},\mathbf{x}) - \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1}\mathbf{c}_Z(\mathbf{x}) + \mathbf{c}_Z(\mathbf{x})^\top \mathbf{C}_Z^{-1}\mathrm{Var}(\mathbf{Z})\,\mathbf{C}_Z^{-1}\mathbf{c}_Z(\mathbf{x})$$

(compare to $m(\mathbf{x}) = \mathbf{c}(\mathbf{x})^\top \mathbf{C}^{-1}\mathbf{y}$ and $c(\mathbf{x},\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - \mathbf{c}(\mathbf{x})^\top \mathbf{C}^{-1}\mathbf{c}(\mathbf{x})$)

Distribution-wise GP (2)

Interpolation property of distribution-wise GPs: they interpolate the means and variances at data points,

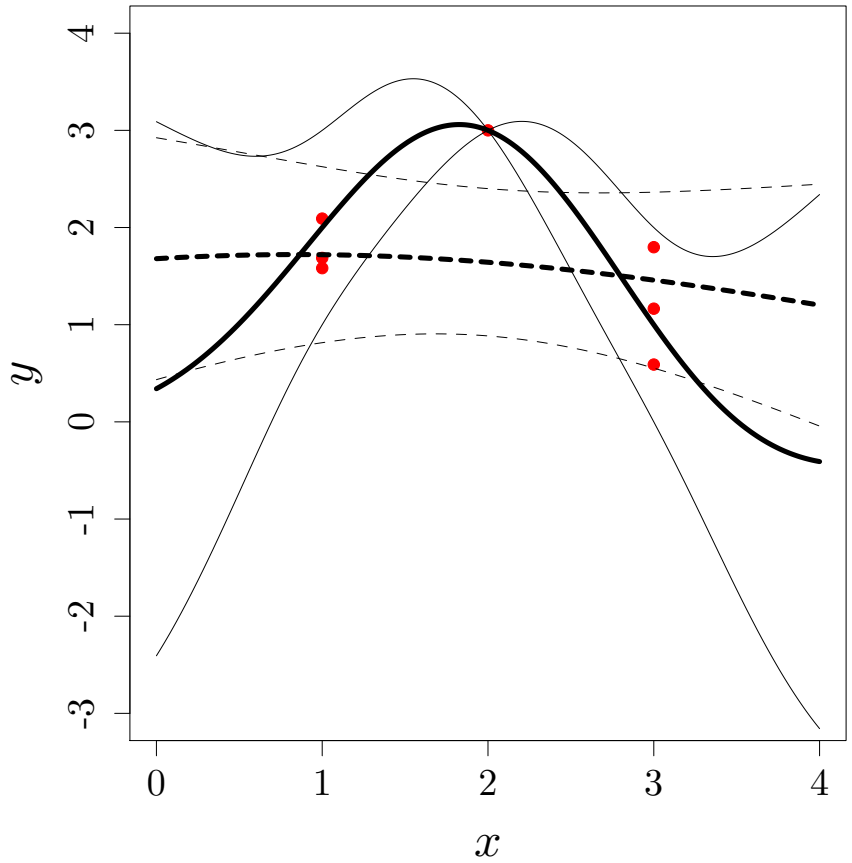
$$m^{Dist}(\mathbf{x}^i) = \mathbb{E}(Z_i) \quad , \quad c^{Dist}(\mathbf{x}^i, \mathbf{x}^i) = \mathrm{Var}(Z_i)$$

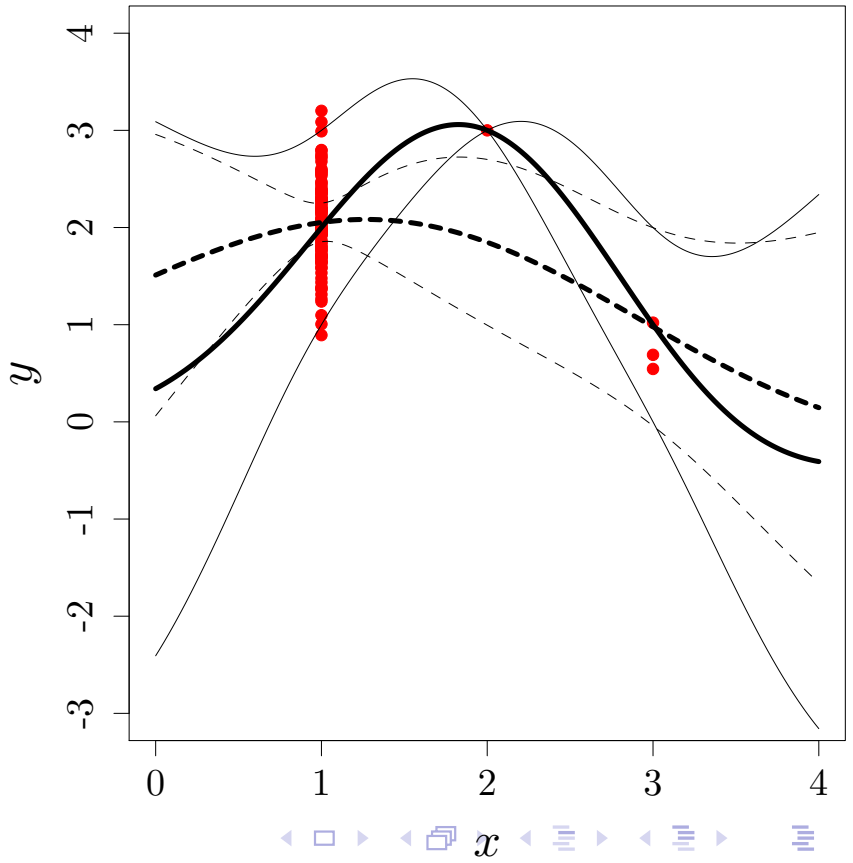
Distribution-wise GPs scale as $\mathcal{O}(k^3)$, which may be cheaper than the usual $\mathcal{O}(n^3)$ cost of traditional GPs.

Implementation: use the empirical means for $\mathbb{E}(\mathbf{Z})$ and the diagonal of empirical (uncorrelated) variances for $\mathrm{Var}(\mathbf{Z})$.

Distribution-wise GP versus nugget regularization

$$\mathbf{z} \sim \mathcal{N}\left(\begin{bmatrix}2\\3\\1\end{bmatrix}, \begin{bmatrix}0.25 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0.25\end{bmatrix}\right)$$, observe how distribution-wise GP (solid) is independent of the number of samples n while the variance of GP regularized by nugget (dashes) tends to 0 as $n \nearrow$:





Concluding remarks

- We have given new algebraic results for comparing nugget and pseudo-inverse as regularization strategies in Gaussian Processes.
- Proofs and further discussion in our report, *An analytic comparison of regularization methods for Gaussian Processes*, HAL Technical report no. hal-01264192, Jan. 2016 version 1 $\rightarrow$ May 2017 version 4.
- Possible continuations: investigate data points clustering for distribution-wise GP as a way to deal with large number of data points; investigate the use of heteroskedastic nugget as a regularization strategy (using e.g., developments of [Binois et al., *Practical heteroskedastic GP model. for large simul. exp.*, ArXiv TR, 2016]).