SAND2017-8667C

An Introduction to Statistical Inverse Problems and Bayesian Inference

Habib N. Najm

Sandia National Laboratories
Livermore, CA, USA

Parametric Uncertainty Summer School & Workshop
Budapest, Hungary
Jul 3-7, 2017

Acknowledgement

B.J. Debusschere, R.D. Berry, K. Sargsyan, C. Safta,
J. Prager, K. Chowdhary, M. Khalil, T. Casey
— Sandia National Laboratories, CA

R.G. Ghanem — U. South. California, Los Angeles, CA
Y.M. Marzouk — Mass. Inst. of Tech., Cambridge, MA

This work was supported by:

- US Dept of Energy Office of Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC)
- US Dept of Energy ASCR Applied Mathematics program.
- US Dept of Energy Office of Basic Energy Sciences, Div. of Chem. Sci., Geosci., & Biosci.

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525.

Outline

1. Inverse Problems
   - Least Squares Parameter Estimation
   - Regularization
   - Sparsity
2. Statistical Inverse Problems
   - Bayesian Inference
   - Examples
   - Regularization & Sparsity
3. Markov chain Monte Carlo
   - Metropolis-Hastings MCMC Algorithm
   - Examples
4. Approximate Bayesian Computation (ABC) Methods
5. Model Selection, Validation, Averaging
6. Closure

Inverse Problem Definition

Inverse problem :

$$f(x;\lambda) = y$$

Given x, y, solve for λ

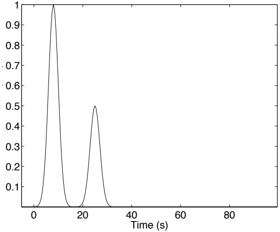
- $x \in \mathbb{R}^d$: independent coordinates, space, time, operating conditions
- $\lambda \in \mathbb{R}^n$: model parameters – objects of inference
  - Generally $\lambda(x) : \Omega \to \mathbb{R}^n$, infinite dimensional
- $f()$: forward model
  - e.g. polynomial fit model, PDE system, etc
- $y \in \mathbb{R}^m$: prediction observable, data
  - Data: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$

Challenges with Inverse Problems

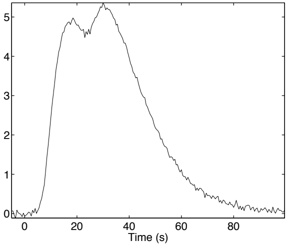
- Inverse problem solution is difficult
  - f^{-1} often non-local, non-causal
- Inverse problems are typically ill-posed:
  - No solution may match the data (existence)
  - Many solutions may match the data (uniqueness)
    - Dependence on initial guess on λ
  - Ill-conditioning or lack of stability
    - Small changes in y can lead to large changes in λ
    - Sensitivity to noise
  - Regularization

Challenges with – noise and ill-conditioning

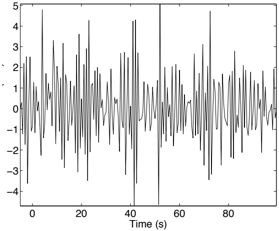
True Input



Forward Model + 5% noise



Inverse Problem Solution



Parameter Estimation and Inverse Problems
Aster, Borchers, and Thurber
Academic Press, 2004, 2012

Least-Squares Parameter Estimation

- Fit model $f()$; unknown parameters λ; measurement y
- Forward Problem:

$$f(\lambda) = y$$

- Estimate λ for best fit between $f(\lambda)$ and y :

$$\lambda_{\text{fit}} \quad = \quad f^{-1}(y)$$

- Inverse problem – solve using least-squares regression

$$\lambda_{\text{rms}} \quad = \quad \underset{\lambda}{\operatorname{argmin}}(||y - f(\lambda)||)$$

i.e. minimize the χ^2:

$$\chi^2 = \sum_{k=1}^{\mathcal{D}} \frac{((f(\lambda) - y)^2}{\sigma_k^2}$$

- Uncertainty estimation, e.g. with Support Planes method
  - χ^2 value decays with parameter variation away from optimum
  - Vary one parameter at a time away from λ_{rms}, refit, estimate stdv based on χ^2 decay below specified threshold

Issues with Least Squares (LS) Parameter Estimation

- Choice of optimal number of fit parameters (p)
  - χ^2 decreases with increased p
  - Danger of overfitting
- No general means for handling *nuisance* parameters
  - Other uncertain parameters in the problem
  - Not objects of inference
- LS best fit is the Maximum Likelihood Estimate (MLE) assuming Gaussian noise in the data
  - What about non-Gaussian noise?
- LS Estimation of Uncertainty in inferred parameter values relies on assumed linearity of the model in the parameters
- Uncertainty estimate does not provide general probabilistic characterization of parameters

Regularization for Deterministic Inverse Problem Solution

- Regularization allows enforcement of select constraints on the inverse problem solution
  - Smoothness
  - Positivity, ...
- Example: Tikhonov-type regularization:

$$\lambda = \underset{\lambda'}{\operatorname{argmin}} \left(\|f(\lambda') - y\|_2^2 + \alpha \|L\lambda'\|_2^2 \right)$$

- How to choose regularization form, L, α ?
  - Somewhat arbitrary
- Regularization introduces bias, destroys consistency
- What about uncertainty/confidence intervals in λ ?

The choice of norm

- The use of the L2-norm

$$
\begin{aligned}
||y - g(x,\theta)||_2^2 &= \frac{1}{N}\sum_{i=1}^{N}(y_i - g(x_i,\theta))^2 \\
||J(\theta)||_2^2 &= \frac{1}{M}\sum_{k=1}^{M}(J(\theta_k))^2
\end{aligned}
$$

  is not the only option for regression fitting or regularization
- Fitting:
  - Model-data misfit, Likelihood function
  - Reflect known data noise structure; Gaussian, Poisson, etc
  - The modeler's choice of metric for measuring misfit "distance" between data and model predictions
- Regularization
  - Optimization regularization term
  - Subjective choices; Prior information
  - Previous measurement

ℓ_1 norm fitting

- The ℓ_1-norm is of particular interest

$$
\begin{aligned}
||y - g(x,\theta)||_1 &= \frac{1}{N}\sum_{i=1}^{N} |y_i - g(x_i,\theta)| \\
||J(\theta)||_1 &= \frac{1}{M}\sum_{k=1}^{M} |J(\theta_k)|
\end{aligned}
$$

- The ℓ_1-norm is useful because it *automatically* identifies **sparsity** in the model, when
  - there is underlying sparsity
  - the model is linear in the parameters

Sparsity

- A sparse model is one that provides reliable predictions with only small number of its parameters being non-zero
  - Physical models: usually **sparse** in prediction of **smooth** observables
- Consider *e.g.* a chemical model for a hydrocarbon fuel
  - thousands of reactions $\Rightarrow$ thousands of parameters
- Not **all** these parameters are important for smooth quantities of interest
  - *e.g.* laminar flame burning speed S_L
- Full dimensionality for a chemical model with N reactions

$$S_L = f\left((A,n,E)_1, \cdots, (A,n,E)_N\right), \quad N \sim 10^4 \text{ (Hydrocarbon fuel)}$$

  Intrinsic dimensionality

$$S_L = g\left((A,n,E)_1, \cdots, (A,n,E)_K\right), \quad K \sim 10 \text{ (important reactions)}$$

- For linear models, ℓ_1-norm constrained ℓ_2 fitting allows identification of the underlying sparse structure of the model

Sparse regression

Model:

$$y = f(x) \simeq \sum_{k=0}^{K-1} c_k \Psi_k(x)$$

with $x \in \mathbb{R}^n$, Ψ_k max order p, and $K = (p+n)!/p!/n!$

- N samples $(x_1, y_1), \dots, (x_N, y_N)$
- Estimate K terms $c_0, \dots, c_{K-1}$, s.t.

$$\min ||\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}||_2^2$$

$$\text{where } \boldsymbol{y} \in \mathbb{R}^N,\ \boldsymbol{c} \in \mathbb{R}^K,\ \boldsymbol{A}_{ik} = \Psi_k(x_i),\ \boldsymbol{A} \in \mathbb{R}^{N \times K}$$

With $N << K \Rightarrow$ under-determined

- Need some form of regularization

Regularization – Compressive Sensing (CS)

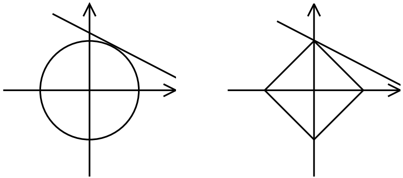
- ℓ_2-norm — Tikhonov regularization; Ridge regression:

$$\min \left\{ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 + \|\boldsymbol{c}\|_2^2 \right\}$$

- ℓ_1-norm — Compressive Sensing; LASSO; basis pursuit

$$\begin{aligned} \min \left\{ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 + \|\boldsymbol{c}\|_1 \right\} & \\ \min \left\{ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 \right\} & \quad \text{subject to } \|\boldsymbol{c}\|_1 \leq \epsilon \\ \min \left\{ \|\boldsymbol{c}\|_1 \right\} & \quad \text{subject to } \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 \leq \epsilon \end{aligned}$$

$\Rightarrow$ discovery of sparse signals



Statistical Inverse Problem

Motivation

- Empirical data D generally provides noisy measurements of y
- Best fit λ is uncertain
- Seeking a single best-fit answer contributes to ill-conditioning

Recasting as a statistical inverse problem improves conditioning

- Solve for a set of solutions, rather than a best fit answer
- Statistical formulation
  - Use statistical methods to estimate confidence intervals on λ
- Formulation as a **Bayesian** inverse problem – Bayesian inference
  - Use probability to describe degree of belief about λ
  - Discrepancy between model and data represented using statistical models
  - Build a data model mapping λ to D
  - Solve for $p(\lambda|D)$

Bayes formula for Parameter Inference

- Data Model (fit model with noise)
- Introduce random variable (field) $\epsilon(\omega)$ to model data misfit

$$y = f(\lambda, \epsilon)$$

- Bayes Formula:

$$p(\lambda, y) = p(\lambda|y)p(y) = p(y|\lambda)p(\lambda)$$

$$\underbrace{p(\lambda|y)}_{\text{Posterior}} = \frac{\overbrace{p(y|\lambda)}^{\text{Likelihood}}\;\overbrace{p(\lambda)}^{\text{Prior}}}{\underbrace{p(y)}_{\text{Evidence}}}$$

- Prior: knowledge of λ prior to data
- Likelihood: forward model and measurement noise
- Posterior: combines information from prior and data
- Evidence: normalizing constant for present context

Advantages of Bayesian Methods

- Formal means of logical inference and machine learning
- Means of incorporation of prior knowledge/measurements and heterogeneous data
- Full probabilistic description of uncertain parameters
- General means of handling nuisance parameters through marginalization
- Means of identification of *optimal* model complexity
  - Ockham's razor
  - Only as much complexity as is required by the physics, and no more
  - Avoid fitting to noise

The Prior

- Prior $p(\lambda)$ comes from
  - Physical constraints, prior data, Prior knowledge
- The prior can be **uninformative**
- It can be chosen to impose **regularization**
- Unknown aspects of the prior can be added to the rest of the parameters as hyperparameters

Examples:

- $\lambda \sim U(1,5)$ – Uniform distribution between 1 and 5
- $\lambda \sim N(\mu, \sigma^2)$
  - Normal distribution with mean μ and standard deviation σ
  - (μ, σ) hyper/nuisance parameters to be inferred from data

Note:

- The prior can be crucial when there is little information in the data
- When there is sufficient information in the data, the data can overrule the prior

Construction of the Likelihood $p(y|\lambda)$

- Where does probability enter the mapping $\lambda \to y$ in $p(y|\lambda)$?
- Through a presumed error model:
- Example:
  - Model:

$$y_m = f(\lambda)$$

  - Data: y
  - Error between data and model prediction: ϵ

$$y = f(\lambda) + \epsilon$$

- Model this error as a random variable
- Example
  - Error is due to instrument measurement noise
  - Instrument has Gaussian errors, with no bias

$$\epsilon \sim N(0, \sigma^2)$$

Construction of the Likelihood $p(y|\lambda)$ – cont'd

For any given λ, this implies

$$y|\lambda, \sigma \sim N(f(\lambda), \sigma^2)$$

or

$$p(y|\lambda, \sigma) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y - f(\lambda))^2}{2\sigma^2}\right)$$

Given N measurements $(y_1, \dots, y_N)$, and presuming independent identically distributed (*iid*) noise

$$\begin{aligned} y_i &= f(\lambda) + \epsilon_i \\ \epsilon_i &\sim N(0, \sigma^2) \\ L(\lambda) = p(y_1, \dots, y_N|\lambda, \sigma) &= \prod_{i=1}^{N} p(y_i|\lambda, \sigma) \end{aligned}$$

Construction of the Likelihood $p(y|\lambda)$ – cont'd

It is useful to use the log-Likelihood

$$\ln L(\lambda) = -\frac{1}{2} N \ln \sigma^2 - \frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \left[\frac{y_i - f(\lambda)}{\sigma} \right]^2$$

Frequently, signal noise amplitude is not constant
e.g. σ varies with signal amplitude
then

$$\ln L(\lambda) = -\frac{1}{2} \sum_{i=1}^{N} \ln \sigma_i^2 - \frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \left[\frac{y_i - f(\lambda)}{\sigma_i} \right]^2$$

Construction of the Likelihood $p(y|\lambda)$ – cont'd

Recall that the weighted least-squares data mis-fit is given by

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - f(\lambda)}{\sigma_i} \right]^2$$

and the best-fit estimate of λ is

$$\lambda_{\mathrm{rms}} = \underset{\lambda}{\operatorname{argmin}}(\chi^2(\lambda))$$

Minimizing χ^2 is equivalent to maximizing the likelihood
Maximum Likelihood Estimate (MLE):

$$\lambda_{\mathrm{MLE}} \equiv \lambda_{\mathrm{rms}}$$

Exploration of the likelihood provides for a more general examination of quality of fit than χ^2

Likelihood Modeling

- This is frequently the *core* modeling challenge
  - Error model: a statistical model for the discrepancy between the forward model and the data
  - composition of the error model with the forward model
- Error model composed of discrepancy between
  - data and the truth – (data error)
  - model prediction and the truth – (model error)
- Mean bias and correlated/uncorrelated noise structure
- Hierarchical Bayes modeling, and dependence trees

$$p(\phi, \theta | D) = p(\phi | \theta, D) p(\theta | D)$$

- Choice of observable – constraint on Quantity of Interest?

Experimental Data

- Empirical data error model structure can be informed based on knowledge of the experimental apparatus
- Both bias and noise models are typically available from instrument calibration
- Noise PDF structure
  - A counting instrument would exhibit Poisson noise
  - A measurement combining many noise sources would exhibit Gaussian noise
- Noise correlation structure
  - Point measurement
  - Field measurement

Posterior

$$p(\lambda|y) \propto p(y|\lambda)p(\lambda)$$

Continuing the above *iid* Gaussian likelihood example, consider also an *iid* Gaussian prior on λ with

$$\lambda \sim N(m, s^2)$$

$$p(\lambda) = \frac{1}{\sqrt{2\pi}\, s} \exp\left(-\frac{(\lambda - m)^2}{2s^2}\right)$$

Posterior cont'd

Then the posterior is

$$p(\lambda|y) \propto_\lambda \; e^{-||y-f(\lambda)||} \; e^{-||\lambda-m||}$$

and the log posterior is

$$\ln p(\lambda|y) \;\; = \;\; -||y - f(\lambda)|| - ||\lambda - m|| + C_\lambda$$

Thus, the maximum a-posteriori (MAP) estimate of λ is equivalent to the solution of the regularized least-squares problem

$$\underset{\lambda}{\operatorname{argmin}}(||y - f(\lambda)|| + ||\lambda - m||)$$

The prior plays the role of a regularizer

Line fitting example

Consider the fitting of a straight line

$$y_m = ax + b$$

to data $D = \{(x_i, y_i),\ i = 1, \dots, N\}$.
Consider an (improper) uninformative prior

$$\pi(a, b) = \mathrm{Const}$$

providing no prior information on (a, b).
Assume *iid* additive unbiased Gaussian noise in y with a given constant noise variance σ^2, thus the data model is:

$$y = ax + b + \epsilon, \qquad \epsilon \sim N(0, \sigma^2)$$

with no noise in the independent variable x.

Line fitting example

Presuming σ known, we have the likelihood,

$$L(a,b) = p(D|a,b) = \prod_{i=1}^{N} p(y_i|a,b)$$

where

$$p(y_i|a,b) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\right)$$

and, per Bayes formula, the posterior density $p(a,b|D)$ is

$$p(a,b|D) \quad = \quad \frac{p(D|a,b)\pi(a,b)}{p(D)} \propto p(D|a,b)\pi(a,b)$$

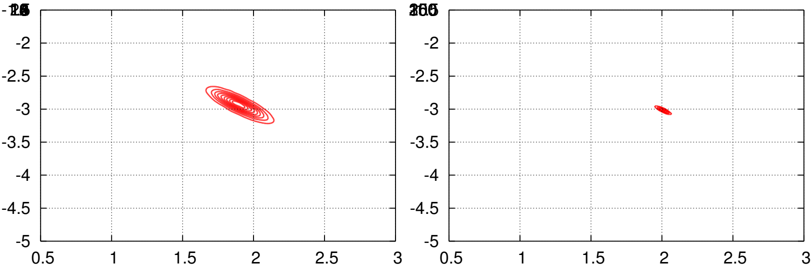
Line fitting example – cont'd

The posterior on (a, b) is the two-dimensional Multivariate Normal (MVN) distribution

$$
\begin{aligned}
p(a, b|D) &\propto (2\pi\sigma^2)^{-N/2} \prod_{i=1}^{N} \exp\left(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\right) \\
&\propto (2\pi\sigma^2)^{-N/2} \exp\left(-\sum_{i=1}^{N} \frac{(y_i - ax_i - b)^2}{2\sigma^2}\right)
\end{aligned}
$$

Linear model, Gaussian noise, σ-given, and a Gaussian or constant-uninformative prior.

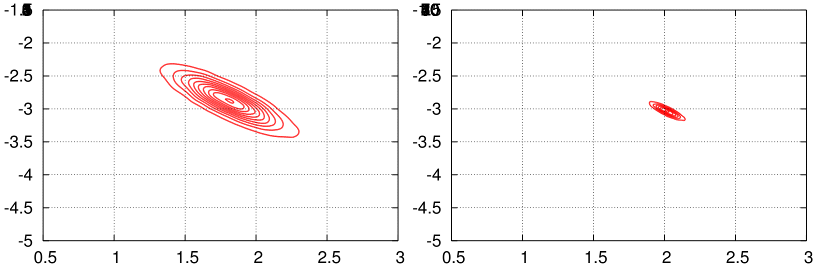
Line fitting example – Effect of data size on $p(a, b|D)$ Low data noise: $\sigma = 0.25$



$N = 20$ $N = 200$

- More data $\Rightarrow$ more accurate parameter estimates

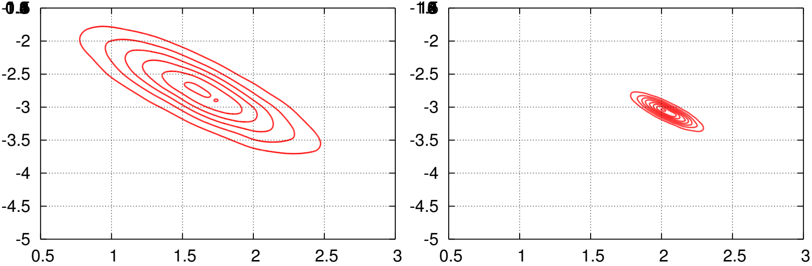
Line fitting example – Effect of data size on $p(a,b|D)$ Medium data noise: $\sigma = 0.5$



$N = 20$ $N = 200$

- More data $\Rightarrow$ more accurate parameter estimates
- Higher noise amplitude $\Rightarrow$ higher uncertainty

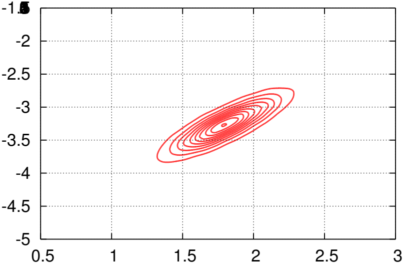
Line fitting example – Effect of data size on $p(a,b|D)$ High data noise: $\sigma = 1.0$



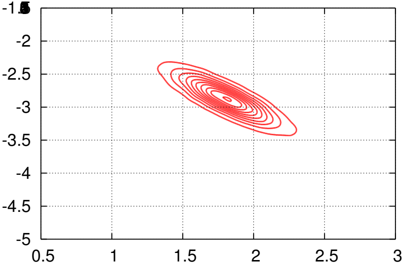
$N = 20$ $N = 200$

- More data $\Rightarrow$ more accurate parameter estimates
- Higher noise amplitude $\Rightarrow$ higher uncertainty

Line fitting example – Effect of data range on $p(a, b|D)$ Medium data noise: $\sigma = 0.5$



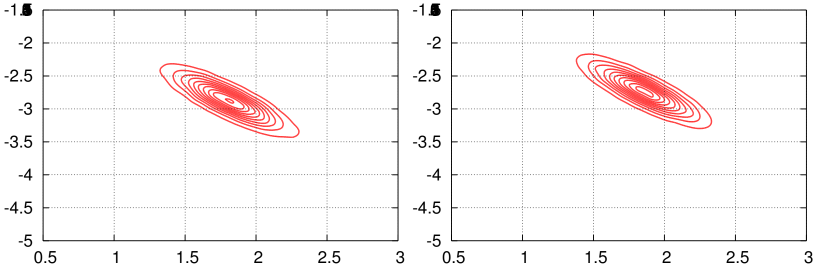
$x \in [-2, 0]$



$x \in [0, 2]$

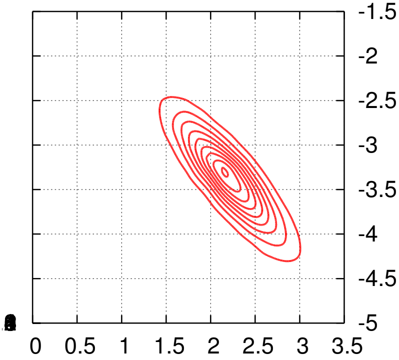
- Posterior correlation structure depends on subjective details of the experiment

Line fitting – Effect of data realization on $p(a, b|D)$
Medium data noise: $\sigma = 0.5$

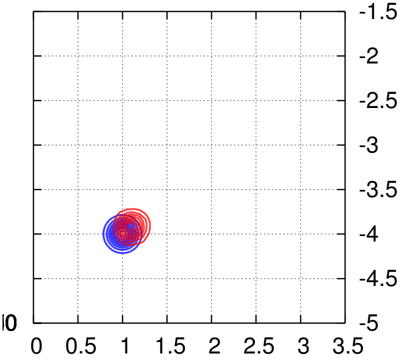


- Posterior depends on specific measured data set
- Two data sets, each with $N = 20$

Line fitting example – prior vs. data-size
20 data points

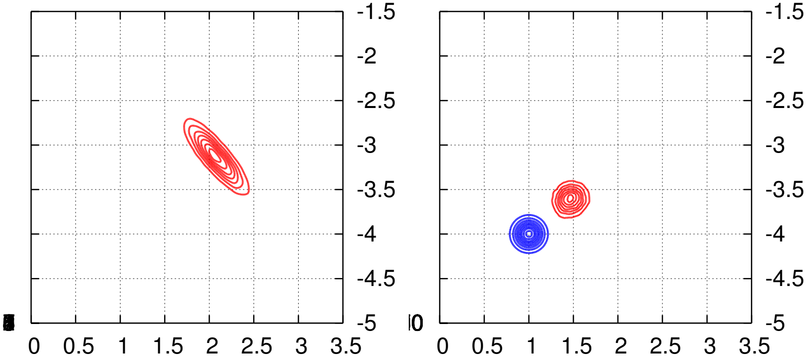


Constant uninformative prior



Gaussian prior

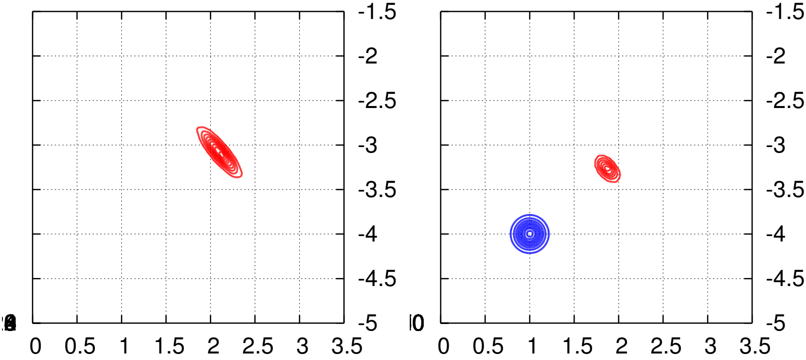
Line fitting example – prior vs. data-size 80 data points



Constant uninformative prior

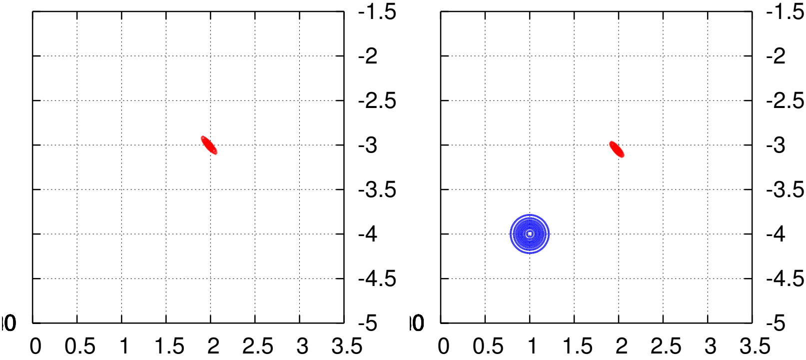
Gaussian prior

Line fitting example – prior vs. data-size
200 data points



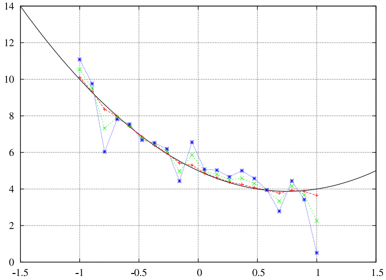
Constant uninformative prior   Gaussian prior

Line fitting example – prior vs. data-size
2000 data points



Constant uninformative prior | Gaussian prior

Bayesian inference illustration: noise↑ ⇒ uncertainty↑



- data: $y = 2x^2 - 3x + 5 + \epsilon$
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$, $\sigma = \{0.1, 0.5, 1.0\}$
- Fit model $y = ax^2 + bx + c$

Marginal posterior density $p(a, c)$:

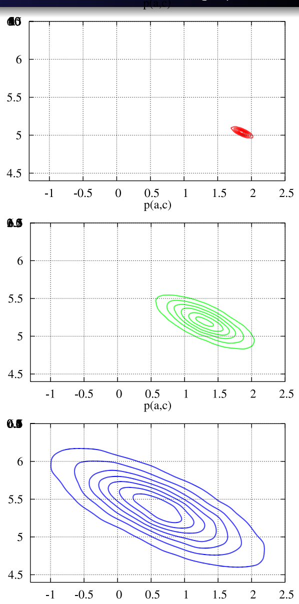
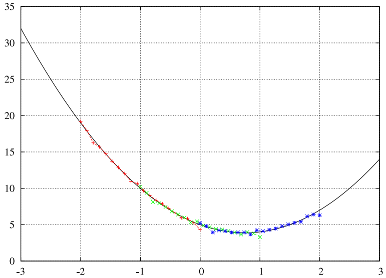
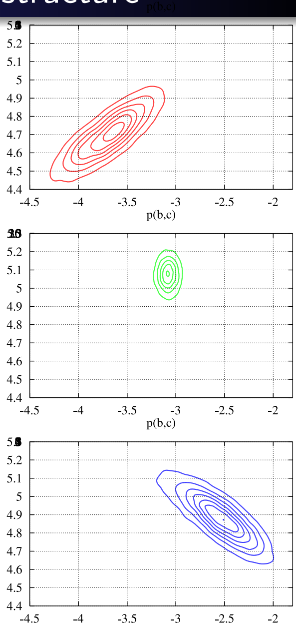


Illustration: Data range ⇒ correlation structure

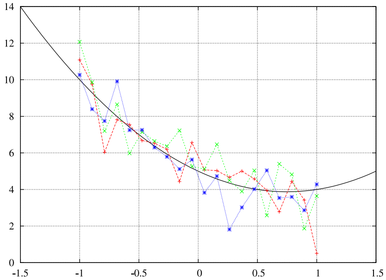


- data: $y = 2x^2 - 3x + 5 + \epsilon$
- $\epsilon \sim \mathcal{N}(0, 0.04)$
- ranges: $x \in \{[-2,0],[-1,1],[0,2]\}$
- Fit model $y = ax^2 + bx + c$

Marginal posterior density $p(b,c)$:

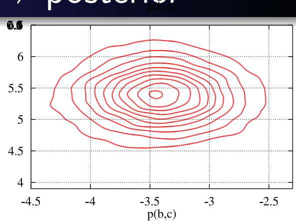


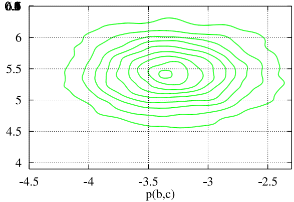
Bayesian illustration: Data realization ⇒ posterior

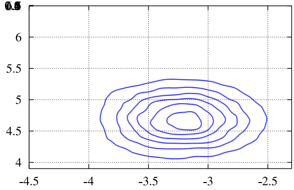


- data: $y = 2x^2 - 3x + 5 + \epsilon$
- $\epsilon \sim \mathcal{N}(0,1)$
  - 3 different random seeds
- Fit model $y = ax^2 + bx + c$

Marginal posterior density $p(b,c)$:







Bayesian Regression

- Bayes formula

$$p(\boldsymbol{c}|D) \propto p(D|\boldsymbol{c})\pi(\boldsymbol{c})$$

- Bayesian regression: prior as a regularizer, *e.g.*
  - Log Likelihood $\Leftrightarrow$ $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2$
  - Log Prior $\Leftrightarrow$ $\|\boldsymbol{c}\|_p^p$
- Laplace sparsity priors $\pi(c_k|\alpha) = \frac{1}{2\alpha} e^{-|c_k|/\alpha}$
- LASSO (Tibshirani 1996) ... formally:

$$\min \left\{ \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{c}\|_2^2 + \lambda\|\boldsymbol{c}\|_1 \right\}$$

Solution $\sim$ the posterior mode of $\boldsymbol{c}$ in the Bayesian model

$$y \sim \mathcal{N}(\boldsymbol{A}\boldsymbol{c}, I_N), \qquad c_k \sim \frac{1}{2\alpha} e^{-|c_k|/\alpha}$$

- Bayesian LASSO (Park & Casella 2008)

Bayesian Compressive Sensing (BCS)

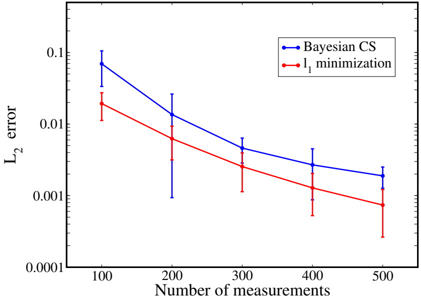
- BCS (Ji 2008; Babacan 2010)— hierarchical priors:
  - Gaussian priors $\mathcal{N}(0, \sigma_k^2)$ on the c_k
  - Gamma priors on the σ_k^2

  $\Rightarrow$ Laplace sparsity priors on the c_k
- Evidence maximization establishes ML estimates of the σ_k
  - many of which are found $\approx 0 \quad \Rightarrow \quad c_k \approx 0$
  - iteratively include terms that lead to the largest increase in the evidence
- iterative BCS (iBCS) (Sargsyan 2012):
  - adaptive iterative order growth
  - BCS on order-p Legendre-Uniform PC
  - repeat with order-$p+1$ terms added to surviving p-th order terms

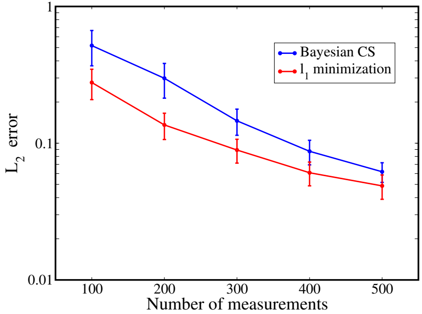
CS and BCS

Corner-peak Genz function

- $f(x) = (1 + \sum_{i=1}^{n} a_i x_i)^{-(n+1)}; \quad a_i \propto 1/i^2$
- Legendre-Uniform PC, 10^{th}-order/5d; 5^{th}-order/10d



5d

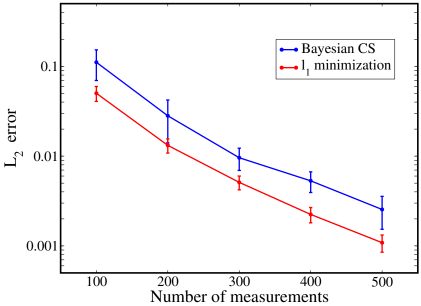


10d

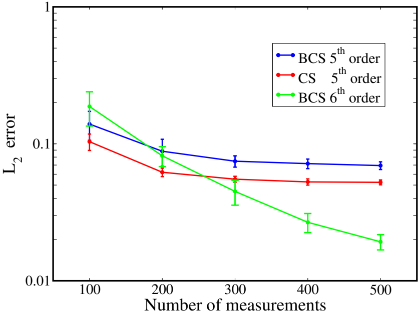
CS and BCS

Oscillatory Genz function

- $f(x) = \cos(2\pi r + \sum_{i=1}^{n} a_i x_i); \quad a_i \propto 1/i^2; \quad r = 0$
- Legendre-Uniform PC, 10^{th}-order/5d; $(5,6)^{th}$-order/10d

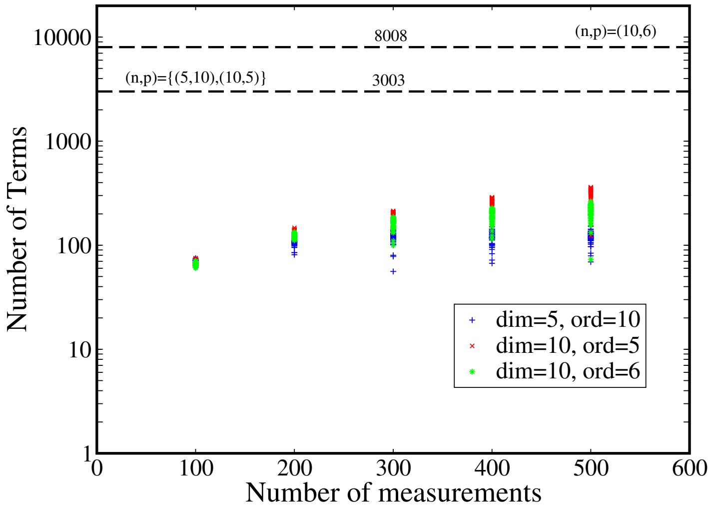


5d

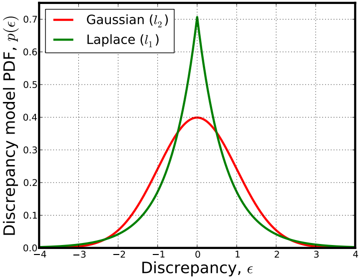


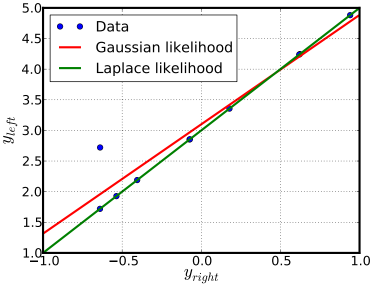
10d

Oscillatory function – BCS number of terms



ℓ_1 norm fitting – Robustness to outliers





- Using ℓ_1-norm fitting, or Laplace likelihood, provides significant robustness to outliers
- The ℓ_1-norm effectively minimizes the number of significant error terms
  - Neglects occasional outlier with large error

Exploring the Posterior – MCMC

- Given any sample λ, the un-normalized posterior probability can be easily computed

$$p(\lambda|y) \propto p(y|\lambda)p(\lambda)$$

- Explore posterior w/ Markov Chain Monte Carlo (MCMC)
  - Metropolis-Hastings algorithm:
    - Random walk with proposal PDF & rejection rules
  - Computationally intensive, $\mathcal{O}(10^5)$ samples
  - Each sample: evaluation of the forward model
    - Surrogate models
- Evaluate moments/marginals from the MCMC statistics

Metropolis-Hastings MCMC sampling of density $\pi(x)$

Algorithm:

- Given a starting point x_0 and proposal density $p(y|x_n)$
- Draw a proposed sample y from proposal density
- Calculate acceptance ratio

$$\alpha(x_n, y) = \min\left\{1, \frac{\pi(y)q(x_n|y)}{\pi(x_n)q(y|x_n)}\right\}$$

- Put

$$x_{n+1} = \begin{cases} y, & \text{with probability } \alpha(x_n, y) \\ x_n, & \text{with probability } 1 - \alpha(x_n, y) \end{cases}$$

Note:

- If $q(y|x_n) \propto \pi(y)$ then $\alpha = 1$
- q does not have to be symmetric.
- π need be evaluated only up to a multiplicative constant

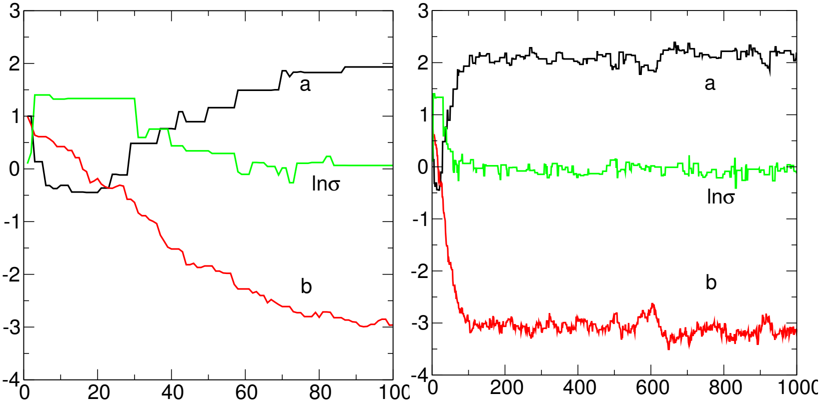
Adaptive Metropolis

- Idea: learn a better proposal $q(y|x)$ from past samples.
  - Learn an appropriate proposal **scale**.
  - Learn an appropriate proposal **orientation** and anisotropy; this is *essential* in problems with strong correlation in π
- Adaptive Metropolis scheme of [Haario *et al.* 2001]:
  - Covariance matrix at step n

    $$C_n^* = s_d \mathbf{Cov}\left(x_0, \ldots, x_n\right) + s_d \epsilon I_d$$

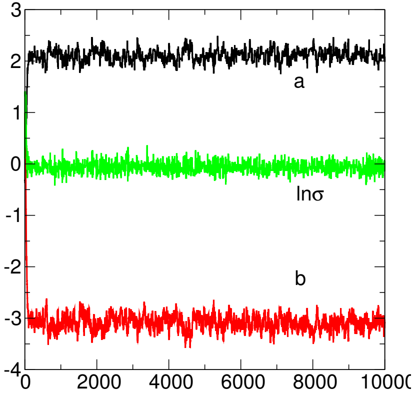
    where $\epsilon > 0$, d is the dimension of the state, and $s_d = 2.4^2/d$ (scaling rule-of-thumb).
  - Proposals are Gaussians centered at x_n.
  - Use fixed covariance C_0 for the first n_0 steps, then use C_n^*.
  - Chain is not Markov.
  - Nonetheless, one can prove that the chain converges to π
- Other adaptive MCMC ideas have been developed

Line fitting example – MCMC – $(a, b, \ln\sigma)$ samples



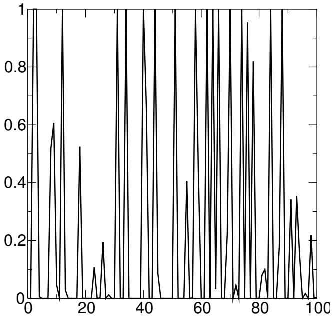
- Initial transient "Burn-in" period, ≈ 100 steps
- Problem and initial condition dependent

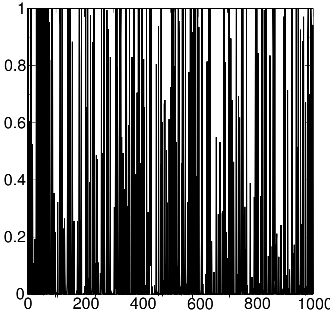
Line fitting example – MCMC – $(a, b, \ln\sigma)$ samples



- Visual inspection reveals “good mixing”
- No significant long-term correlation or periodicity

Line fitting example – MCMC – acceptance probability

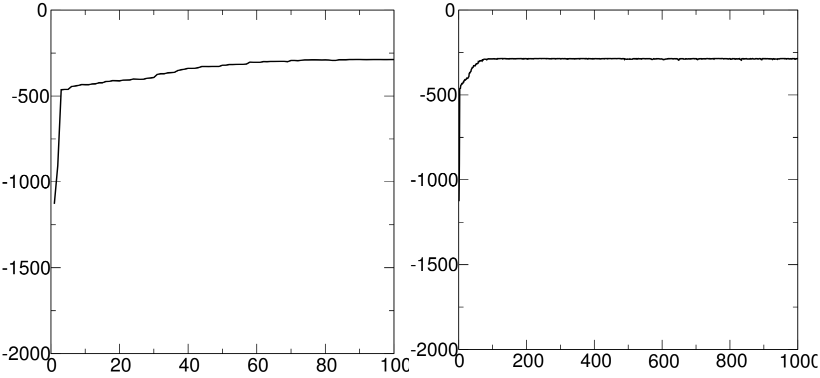




- An average acceptance probability of ~ 0.2 is "good"
- A typical compromise between accepting most samples
    - not moving much, strong correlation

  and rejecting most samples
    - moving too far off, wasted CPU time in rejections

Line fitting example – MCMC – posterior density



- Chain finds high posterior density (HPD) region
- stays there generating many random samples

MCMC practicalities

Effective use of MCMC still requires some (problem-specific) experience. Some useful rules of thumb:

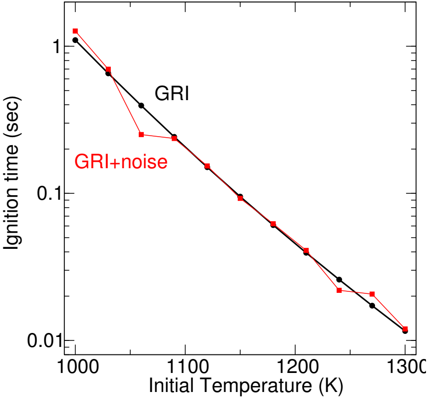
- Adaptive schemes are not a panacea.
- Whenever possible, parameterize the problem in order to minimize posterior correlations.
- What to do, if anything, about "burn-in?"
- Visual inspection of chain components is often the first and best convergence diagnostic.
- Also look at:
  - autocorrelation plots
  - multivariate potential scale reduction factor (MPSRF, Gelman & Brooks)
  - and other diagnostics.
- Optimal acceptance rates? Maybe ... $\sim$0.2
  - But in practice it's best to explore chain diagnostics

Chemical Rate Parameter Estimation example

Synthetic ignition data generated using a detailed model+noise

- Ignition using GRImech3.0 methane-air chemistry
- Ignition time versus Initial Temperature
- Multiplicative noise error model
- 11 data points:

$$\tau_i^d = \tau^{\mathrm{GRI}}(T_i^o)\,(1 + \sigma\epsilon_i)$$
$$\epsilon \sim N(0,1)$$

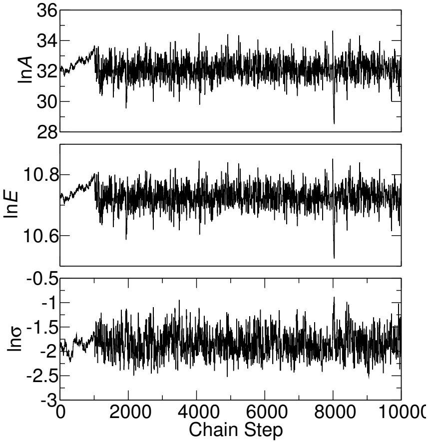


Fitting with a simple chemical model

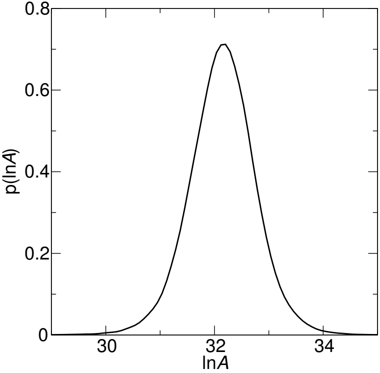
- Fit a global single-step irreversible chemical model

$$CH_4 + 2O_2 \rightarrow CO_2 + 2H_2O$$
$$\mathfrak{R} = [CH_4][O_2]k_f$$
$$k_f = A\exp(-E/R^oT)$$

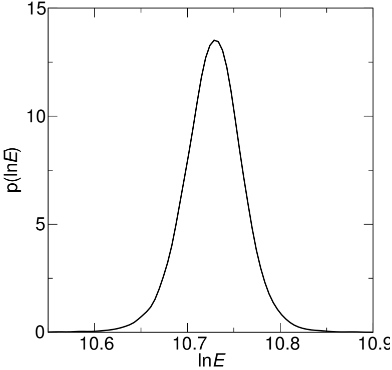
- Infer 3-D parameter vector $(\ln A, \ln E, \ln\sigma)$
- Good mixing with adaptive MCMC when start at MLE



Marginal Posteriors on $\ln A$ and $\ln E$

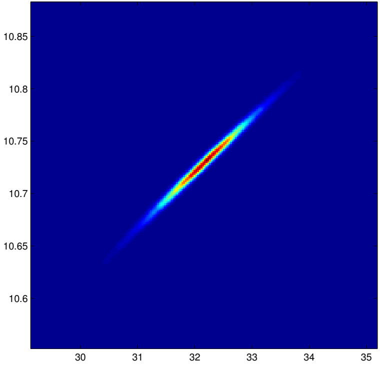


$$\ln A = 32.15 \pm 3 \times 0.61$$

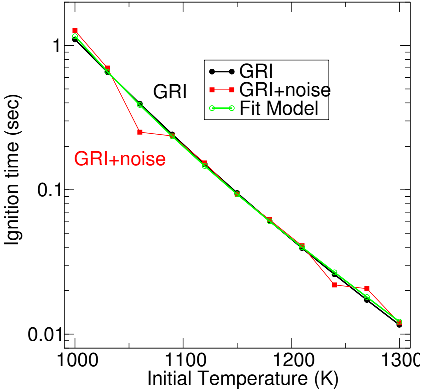


$$\ln E = 10.73 \pm 3 \times 0.032$$

Bayesian Inference Posterior and Nominal Prediction



Marginal joint posterior on $(\ln A, \ln E)$ exhibits strong correlation



Nominal fit model is consistent with the true model

Approximate Bayesian Computation (ABC)

- Data model: $y = f(x, \lambda) + \epsilon_{\mathrm{d}}, \qquad \epsilon_{\mathrm{d}} \sim N(0, \sigma^2)$ and $\alpha \equiv (\lambda, \sigma)$
- Full Likelihood: $L(\alpha) = p(D|\alpha) = p(y_{\mathrm{d}}|\alpha)$
- Often, the likelihood cannot be formulated or is too costly to compute, e.g.

$$
\begin{aligned}
L(\alpha) &:= L^*(\alpha)Z(\alpha) \quad \text{where } Z(\alpha) \text{ is unknown} \\
L(\alpha) &:= \int L^*(\alpha, u)\mathrm{d}u \quad \text{where } u \text{ is high dimensional}
\end{aligned}
$$

Resolution:

- Bypass computation of Likelihood
- Generate replicate data samples z from the data model
- Employ a pseudo-likelihood based on a kernel density that enforces select constraints on the predictions z
    - Constraint employs some distance measure between y_d and z

ABC Likelihood

With $\rho(\mathcal{S})$ being a metric of the statistic $\mathcal{S}$, use the kernel function as an ABC likelihood:

$$L_{\mathrm{ABC}}(\alpha) = \frac{1}{\epsilon} K\left(\frac{\rho(\mathcal{S})}{\epsilon}\right)$$

where ϵ controls the severity of the consistency control

Example, enforce the mean data prediction

$$\mathcal{S}(y) = \mathrm{E}(y) = \mu_y$$

with $z = z(\alpha)$, and

$$\rho(\mathcal{S}) := \mu_z(\alpha) - \mu_{y_d}$$

Propose the Gaussian kernel density:

$$L_\epsilon(\alpha) = \frac{1}{\epsilon\sqrt{2\pi}} \exp\left(-\frac{(\mu_z(\alpha) - \mu_{y_d})^2}{2\epsilon^2}\right)$$

Model UQ

- No model of a physical system is strictly true
- The probability of a model being strictly true is zero
- Given limited information, some models may be relied upon for describing the system

Let $\mathcal{M} = \{M_1, M_2, ...\}$ be the set of all models

- $p(M_k|I)$ is the probability that M_k is the model behind the available information
    - Model Plausibility
- Parameter estimation from data is conditioned on the model

$$p(\theta|D, M_k) = \frac{p(D|\theta, M_k)\pi(\theta|M_k)}{p(D|M_k)}$$

Bayesian Model Comparison

Evidence (marginal likelihood) for M_k:

$$p(D|M_k) = \int p(D|\theta, M_k)\pi(\theta|M_k)\mathrm{d}\theta$$

Bayes Factor B_{ij}:

$$B_{ij} = \frac{p(D|M_i)}{p(D|M_j)}$$

Plausibility of M_k:

$$p(M_k|D, \mathcal{M}) = \frac{p(D|M_k)\ \pi(M_k|\mathcal{M})}{\sum_s p(D|M_s)\pi(M_s|\mathcal{M})} \qquad k = 1, \ldots$$

Posterior odds:

$$\frac{p(M_i|D, \mathcal{M})}{p(M_j|D, \mathcal{M})} = B_{ij}\ \frac{\pi(M_i|\mathcal{M})}{\pi(M_j|\mathcal{M})}$$

Marginal Likelihood example

- Consider Fitting with data from a truth model

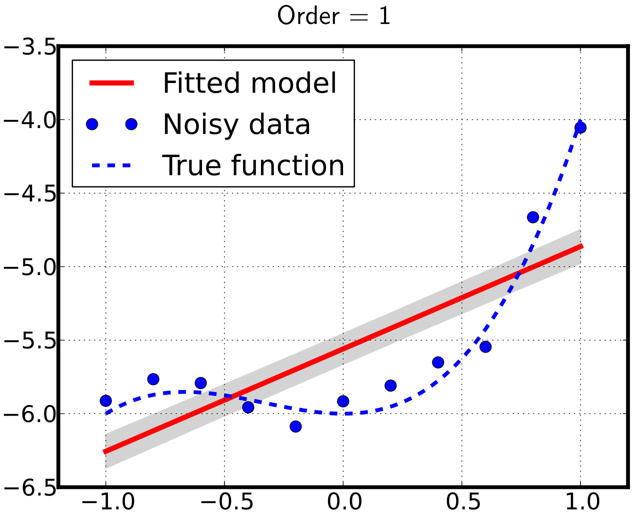
$$y_t = x^3 + x^2 - 6$$

- Gaussian *iid* additive noise model with fixed variance s
- Bayesian regression with a Gaussian Likelihood, *iid* and given s
- Consider a set of Legendre Polynomial expansion models, order 1-10

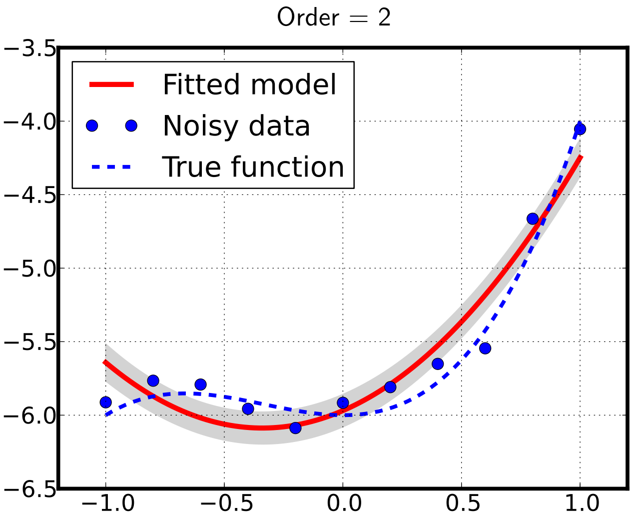
$$y_m = \sum_{k=0}^{P} c_k \psi_k(x)$$

- Uniform priors $[-D, D]$ on all coefficients

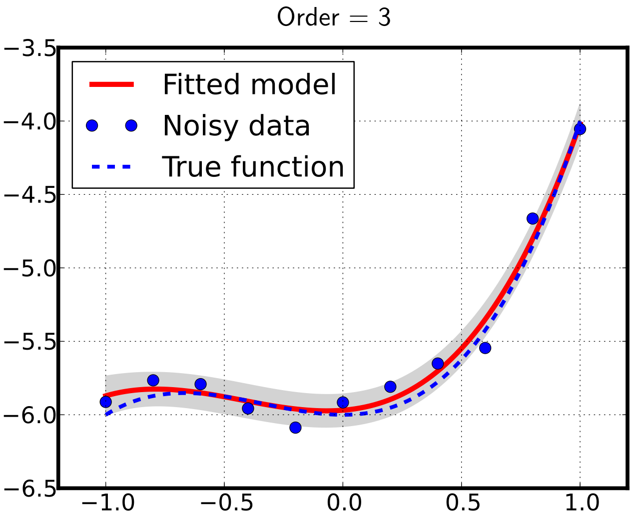
Too much model complexity leads to overfitting



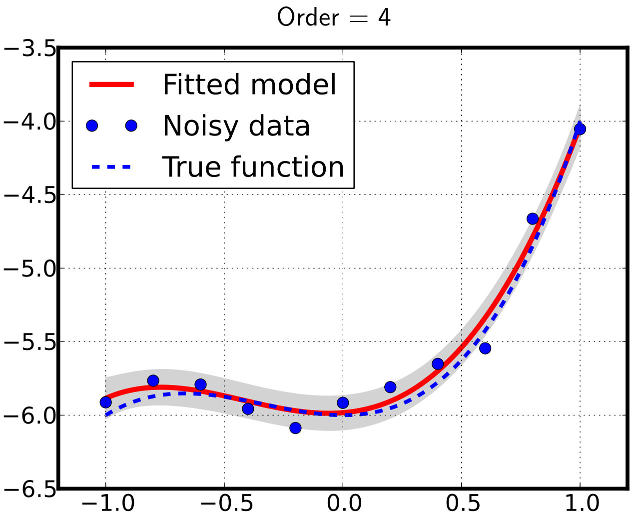
Too much model complexity leads to overfitting



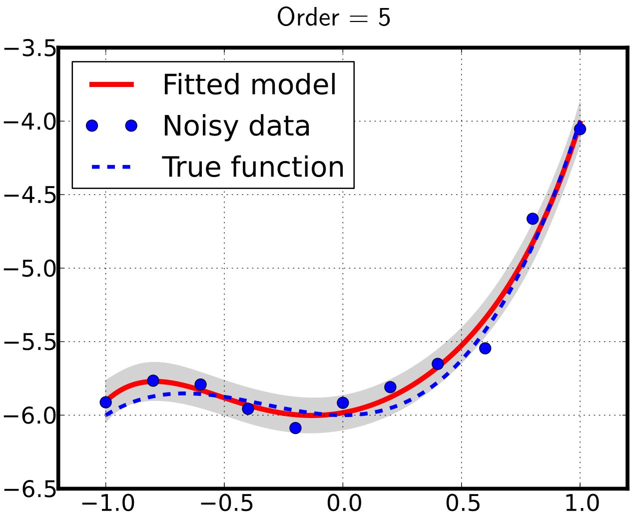
Too much model complexity leads to overfitting



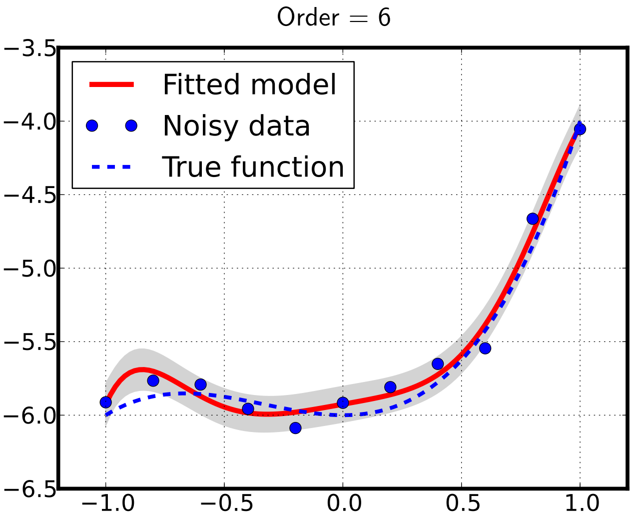
Too much model complexity leads to overfitting



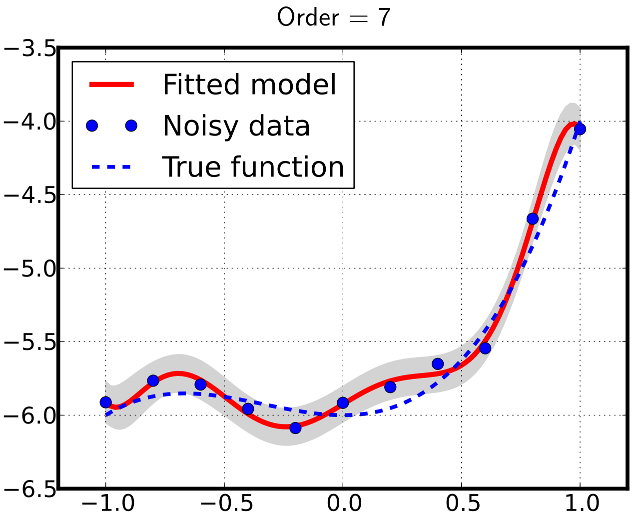
Too much model complexity leads to overfitting



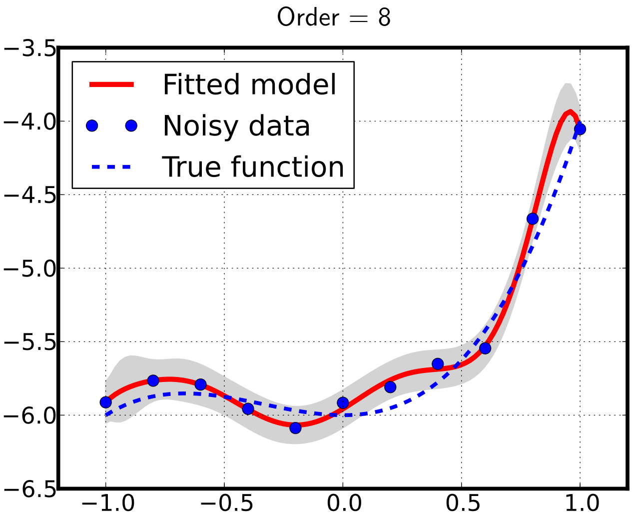
Too much model complexity leads to overfitting



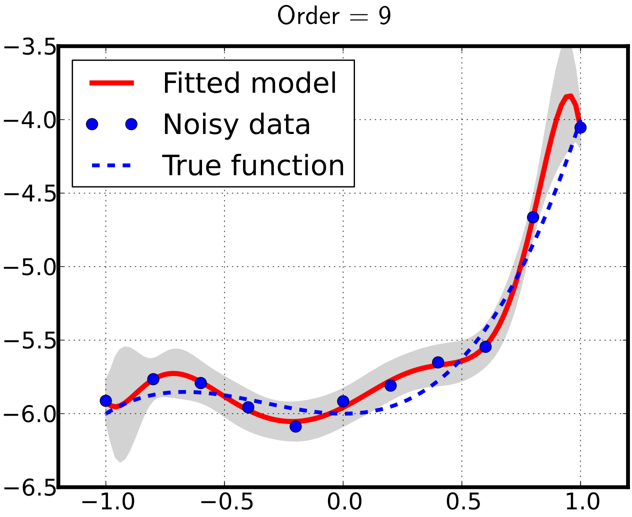
Too much model complexity leads to overfitting



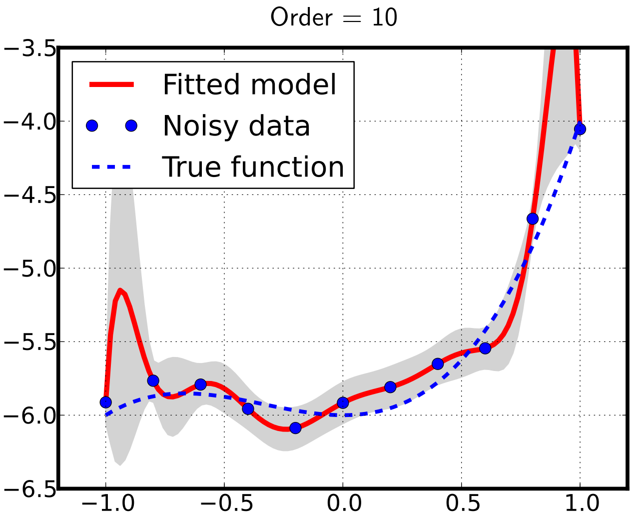
Too much model complexity leads to overfitting



Too much model complexity leads to overfitting



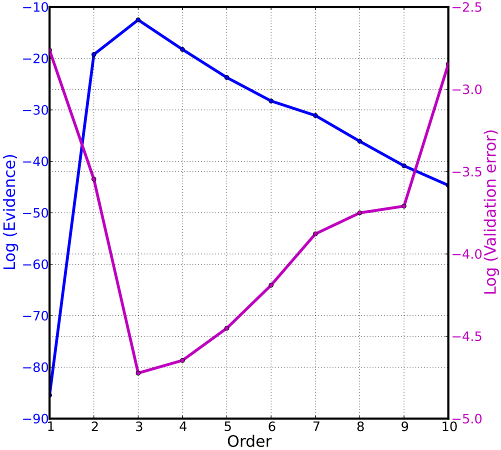
Too much model complexity leads to overfitting



Evidence and Validation Error

Log Evidence:

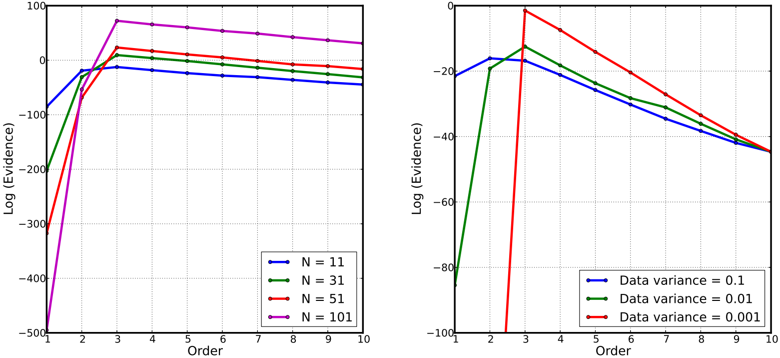
$$\ln p(D|M_k)$$



- Validation error – ℓ_2 error for a random set of 1000 points
  - Minimal at 3rd-order
- Log evidence: sum of two scores, balances complexity & fit
  - Peaks at 3rd order

Muto & Beck 2008

Evidence – Discrimination among Models



- Discrimination among models is more clear-cut with higher amount of data D and/or less data noise

Prediction

Consider that a model

$$y_m = f(x, \lambda)$$

was fitted according to

$$y = f(x, \lambda) + \epsilon, \quad \epsilon \sim N(0, \sigma^2),$$

providing:

- The posterior $p(\lambda, \sigma|D)$
- The marginal posterior $p(\lambda|D)$

Define:

- Pushed forward posterior (PFP) distribution : $p(y_m|x, D)$
- Posterior predictive (PP) distribution : $p(y|x, D)$

Pushed forward posterior (PFP)

- PFP distribution $p(y_m|x, D)$
- Push-forward of the marginal posterior measure on λ through $f(x, \lambda)$
- PFP random process

$$
\begin{aligned}
Y_m(x, \omega) &= f(x, \lambda(\omega)) \\
&\sim p(y_m|x, D)
\end{aligned}
$$

- The PFP provides the uncertain prediction by the calibrated model
  - Forward UQ
  - Mean prediction $E[Y_m]$
  - Predictive variance $V[Y_m]$

Posterior Predictive (PP)

Posterior Predictive distribution $p(y|x, D)$

- With $\alpha \equiv (\lambda, \sigma)$,

$$p(y|x, D) = \int p(y|x, \alpha, D)p(\alpha|D)\mathrm{d}\alpha$$

PP random process

$$\begin{aligned} Y^{PP}(x, \omega) &= \mathrm{E}_\alpha[Y(x, \omega)] \\ &\sim p(y|x, D) \end{aligned}$$

provides the marginal prediction of the data. Where

$$Y(x, \omega) = f(x, \lambda) + \epsilon(\omega, \sigma)$$

is the PP data predictor

- Posterior predictive check – evaluate distance between the PP and the actual/empirical distribution of the data

Validation

- Validity is a statement of model utility for predicting a given observable under given conditions
- Inspection of model utility requires accounting for uncertainty
- Statistical tool-chest for model validation
  - Cross-validation
  - Bayes Factor
  - Model Plausibility
  - Posterior Odds
  - Posterior predictive:

$$p(\tilde{D}|D, M_k) = \int p(\tilde{D}|\theta, M_k) p(\theta|D, M_k) d\theta$$

Model Averaging

- When multiple models are acceptable, and no model is a clear winner, model averaging can be used to provide a prediction of interest
- If prediction errors among models are uncorrelated, then averaging is expected to reduce prediction errors
  - Not likely if models are dependent, or if they have comparable large bias errors in a given observable of interest
- Bayesian Model Averaging

$$p(\phi|D,\mathcal{M}) = \sum_{k=1}^{N} p(\phi|D,M_k)p(M_k|D,\mathcal{M})$$

where

$$\mathcal{M} = \{M_1,\ldots,M_N\}$$

Closure

- Inverse problems are ubiquitous in science and engineering
- Where possible, employing the Bayesian framework provides for more robust, reliable and informed solutions
- Bayesian inversion facilitates subsequent prediction with uncertainty
- Bayesian model selection strategies are relevant to the identification of parsimonious models that explain empirical data