Week 11: Additive Models

Francisco Richter Ernst Wit

Introduction to Data Science (MSc)

1 Generalized Linear Models (GLMs): Foundations

Generalized Linear Models extend linear regression to non-Gaussian responses in the exponential family via a link function. They unify logistic regression, Poisson regression, Gamma regression, and others under a single framework with a common fitting algorithm (IRLS).

1.1 Components

- **Random component**: Y follows an exponential family distribution with $\mathbb{E}[Y] = \mu$ and $\mathrm{Var}(Y) = a(\phi)\, V(\mu)$.
- **Systematic component**: linear predictor $\eta = \mathbf{x}^T \boldsymbol{\beta}$.
- **Link**: $g(\mu) = \eta$ (canonical link when $g(\mu) = \theta$).

1.2 Variance Functions

Different exponential-family choices imply different variance functions $V(\mu)$:

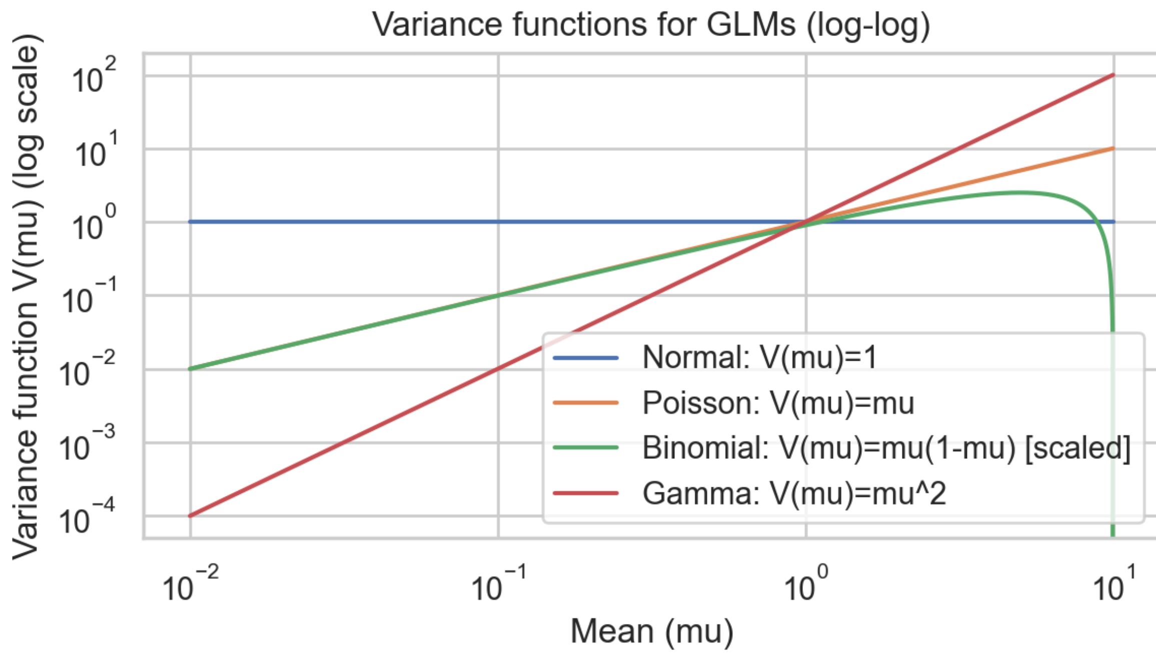

Figure 1: Variance functions for Normal, Poisson, Binomial, and Gamma families (log–log axes).

1.3 IRLS in a Nutshell

For GLMs, maximum likelihood estimation proceeds by Iteratively Reweighted Least Squares (IRLS), solving at iteration t a weighted least squares system with weights $w_i^{(t)}$ and working response $z_i^{(t)}$ until convergence.

1.4 GLMs in Practice

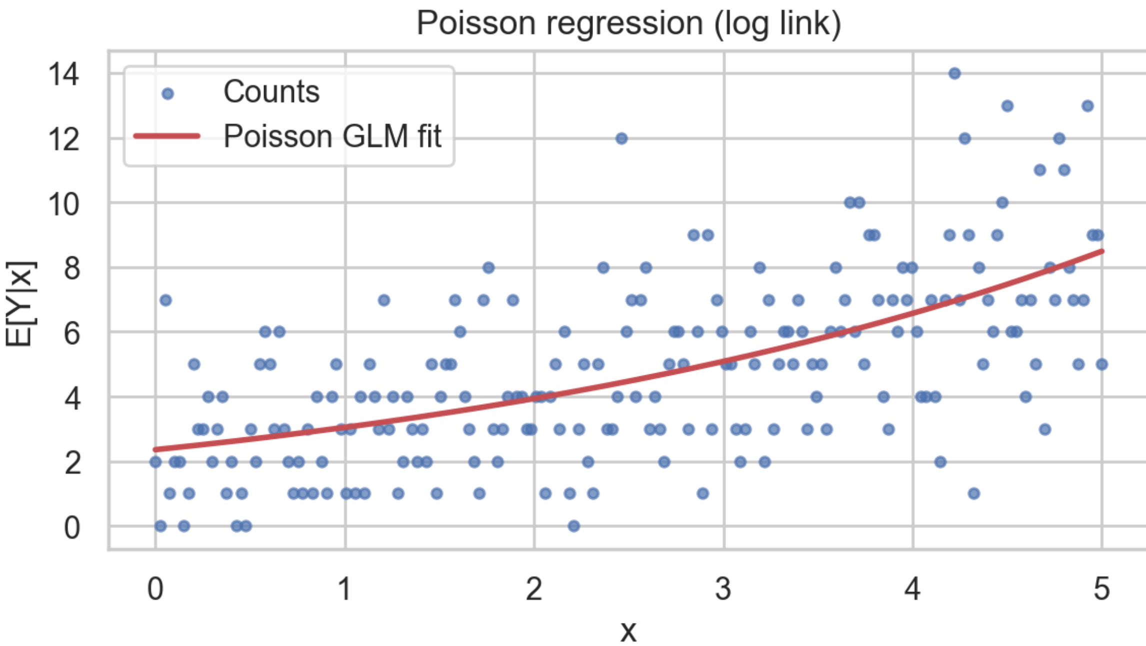

Figure 2: Poisson GLM fit (log link) on synthetic counts: mean curve captures exponential mean–variance structure.

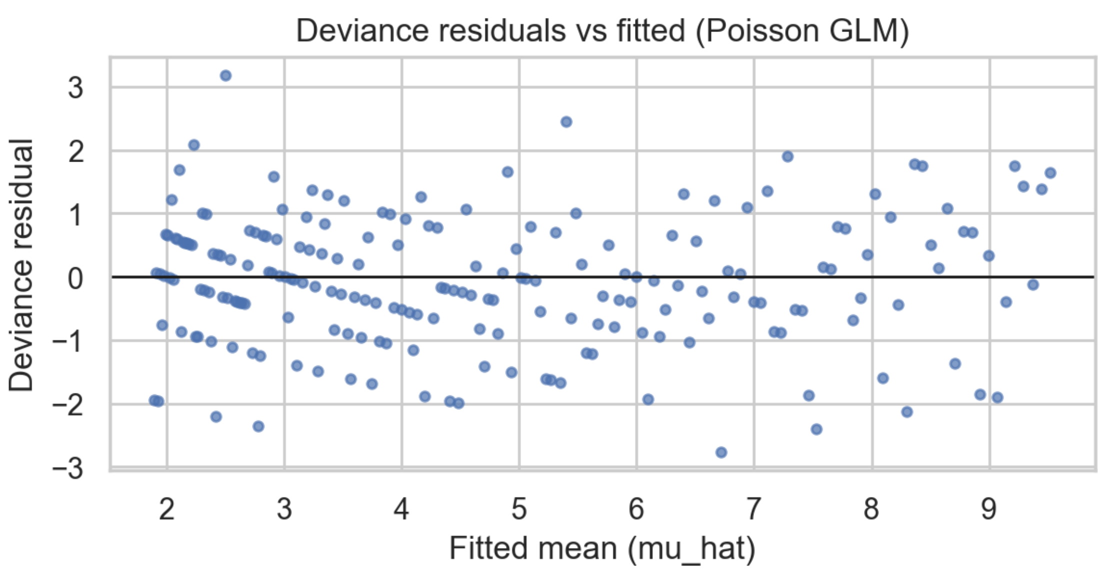

Figure 3: Deviance residuals vs fitted values: pattern-free clouds support model adequacy; structure suggests misspecification.

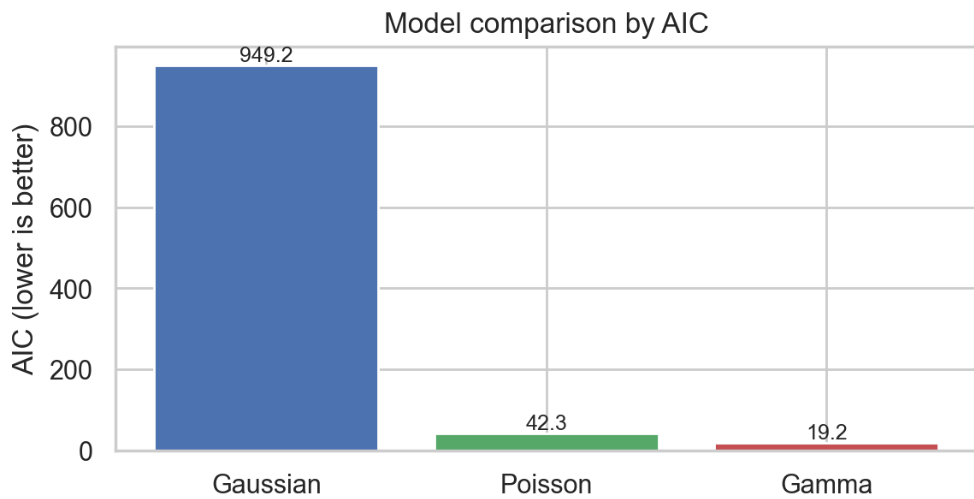

Figure 4: AIC comparison across Gaussian, Poisson, and Gamma GLMs on the same data.

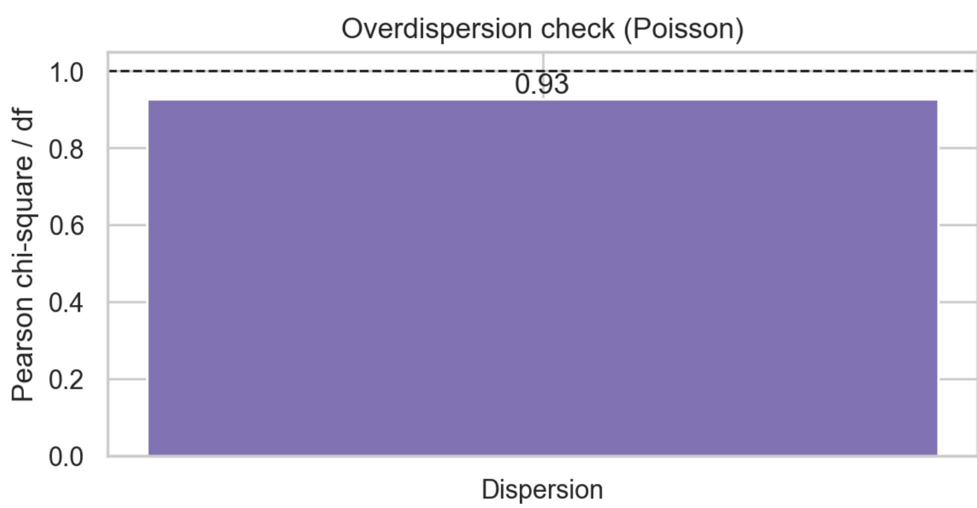

Figure 5: Overdispersion check (Poisson): Pearson chi-square divided by df; values > 1 indicate extra-Poisson variability.

2 Introduction to Additive Models

Additive models provide a flexible framework for modeling nonlinear relationships while maintaining interpretability. They extend linear models by allowing smooth functions of predictors rather than just linear terms, making them particularly valuable in data science for exploratory analysis and prediction.

The fundamental insight underlying additive models is the recognition that many real-world relationships are inherently nonlinear, yet we still desire interpretable models that can provide meaningful insights into how individual predictors affect the response. Instead of assuming that the effect of each predictor is strictly linear, additive models allow each predictor to have a smooth, potentially nonlinear effect on the response variable. The "additive" nature of these models means that these individual effects simply add together, preserving the interpretability that makes linear models so appealing while gaining the flexibility needed to capture complex patterns in modern datasets.

3 Generalized Additive Models (GAMs)

3.1 Model Definition

Definition 3.1 (Generalized Additive Model)**.** *A GAM has the form:*

$$g(E[Y]) = \alpha + \sum_{j=1}^{p} f_j(X_j)$$

where $g(\cdot)$ is a link function, α is the intercept, $f_j(\cdot)$ are smooth functions of the predictors X_j, and Y follows an exponential family distribution.

GCV approximates leave-one-out cross-validation without refitting by adjusting the training residual sum of squares by $(1 - \mathrm{tr}(\mathbf{H}_\lambda)/n)^{-2}$. It is computationally attractive and often effective for Gaussian GAMs, though it can sometimes undersmooth in the presence of correlated errors or outliers. K-fold cross-validation is a robust alternative at higher computational cost.

Workflow. Fix a rich basis for each smooth, select λ by REML or CV, refit, and report edf per term along with partial dependence plots and uncertainty bands.

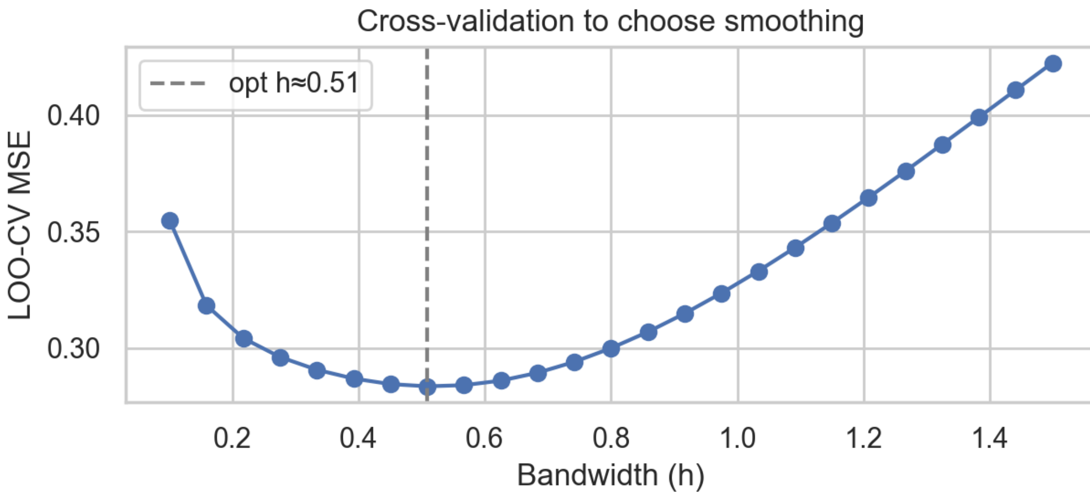

Figure 6: Cross-validation curve versus bandwidth (smoothing parameter proxy); the vertical line marks the optimal choice minimizing CV error.

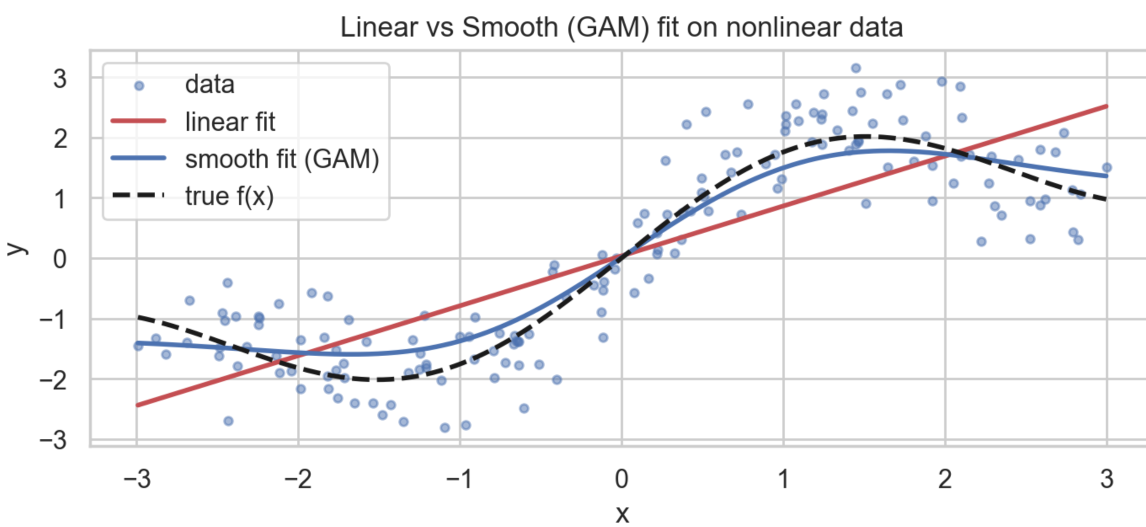

Figure 7: Nonlinear relationship captured by a smooth additive term compared to a linear fit. The smooth (GAM) fit tracks the true function while the linear fit underfits.

The structure of a GAM represents a natural generalization of both linear models and generalized linear models. The link function $g(\cdot)$ serves the same role as in GLMs, allowing us to model different types of response variables while maintaining the linear predictor structure. However, instead of restricting ourselves to linear combinations of the predictors, we allow each predictor X_j to enter the model through a smooth function $f_j(\cdot)$. This flexibility enables the model to capture complex nonlinear patterns while maintaining the additive structure that makes interpretation straightforward.

The smooth functions $f_j(\cdot)$ are typically estimated from the data rather than specified a priori, making GAMs a semi-parametric approach that combines the interpretability of parametric models with the flexibility of nonparametric methods. Each function f_j can be thought of as representing the partial effect of predictor X_j on the transformed response $g(E[Y])$, holding all other predictors constant.

3.2 Special Cases and Extensions

Several important models emerge as special cases of the GAM framework. When all smooth functions $f_j(X_j) = \beta_j X_j$ are linear, we recover the standard generalized linear model. Polynomial models arise when $f_j(X_j) = \sum_{k=1}^{d} \beta_{jk} X_j^k$, allowing for polynomial relationships of specified degree.

The Gaussian GAM uses the identity link function with normal errors, making it particularly suitable for continuous response variables where we expect smooth, nonlinear relationships. The logistic GAM employs the logit link for binary outcomes, extending logistic regression to allow for nonlinear effects of continuous predictors while maintaining the probabilistic interpretation of the model.

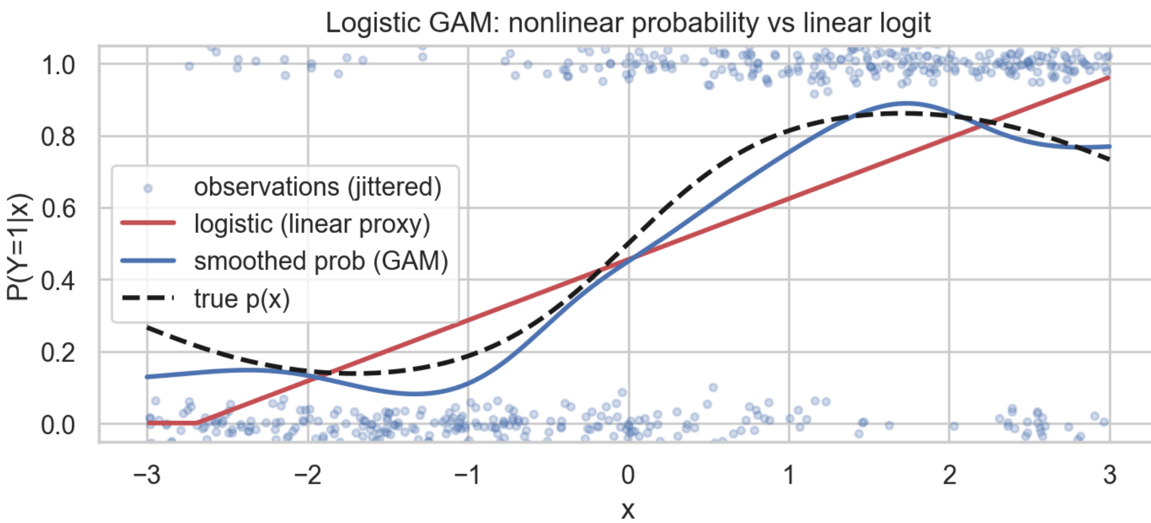

Figure 8: Logistic GAM: a smoothed probability curve (blue) recovers nonlinear log-odds patterns that a linear logit cannot capture, closely matching the true probability.

4 Smooth Functions and Basis Expansions

4.1 Theoretical Foundation of Basis Functions

The representation of smooth functions through basis expansions forms the mathematical foundation of additive models. Rather than attempting to estimate arbitrary smooth functions directly, we express each smooth function f_j as a linear combination of known basis functions:

$$f_j(x) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x)$$

where $b_{jk}(x)$ are basis functions and β_{jk} are coefficients to be estimated.

This approach transforms the problem of nonparametric function estimation into a parametric problem of estimating the coefficients β_{jk}. The choice of basis functions determines the types of smooth functions that can be represented, while the number of basis functions K_j controls the flexibility of the approximation. The art of additive modeling lies in selecting appropriate basis functions and determining the optimal number of basis functions to balance flexibility with overfitting.

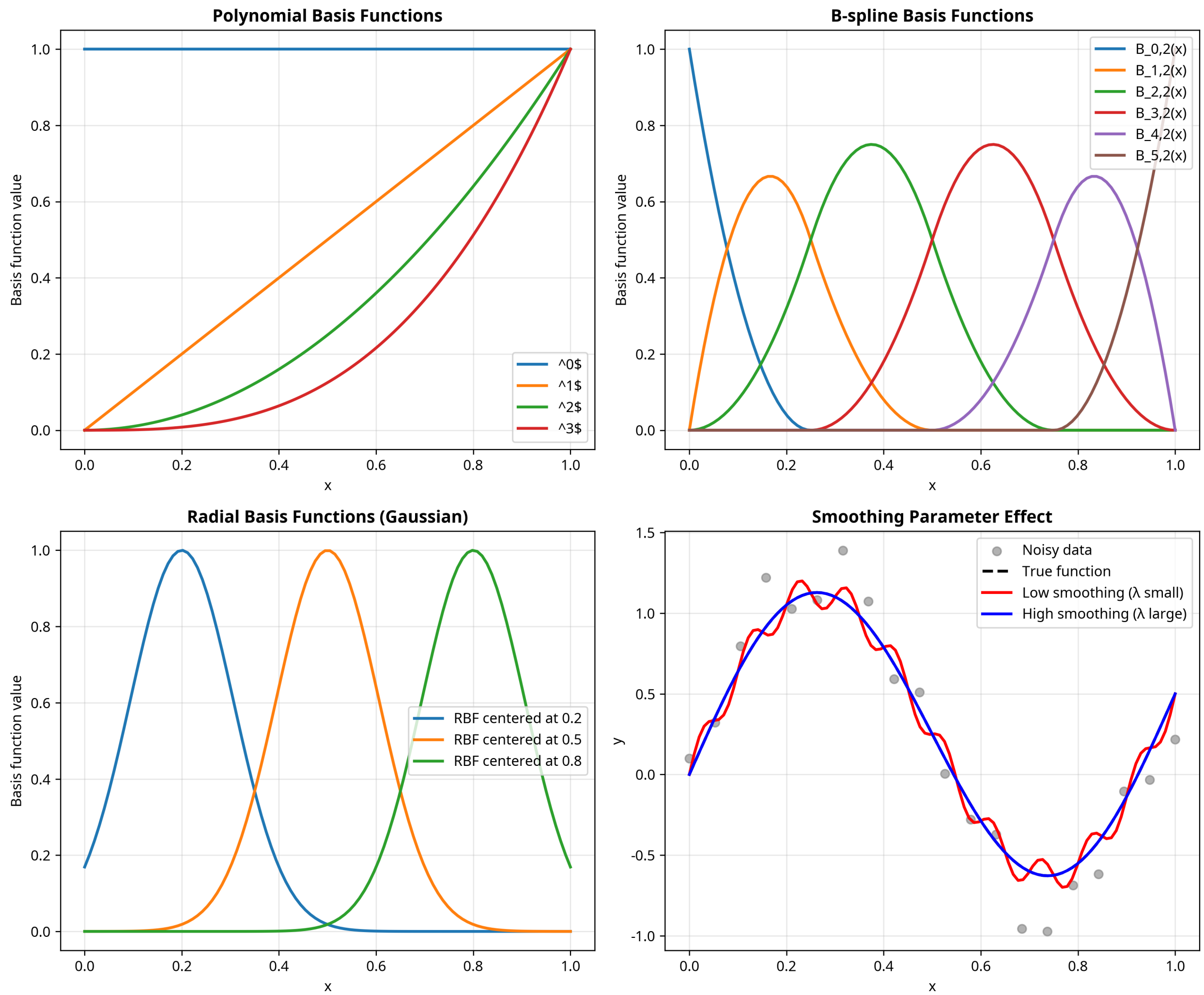

Figure 9: Basis Functions and Smoothing Concepts: (Top left) Polynomial basis functions showing increasing powers of x. (Top right) B-spline basis functions with local support properties. (Bottom left) Radial basis functions centered at different locations. (Bottom right) Effect of smoothing parameter λ on function estimation, demonstrating the bias-variance tradeoff in nonparametric estimation.

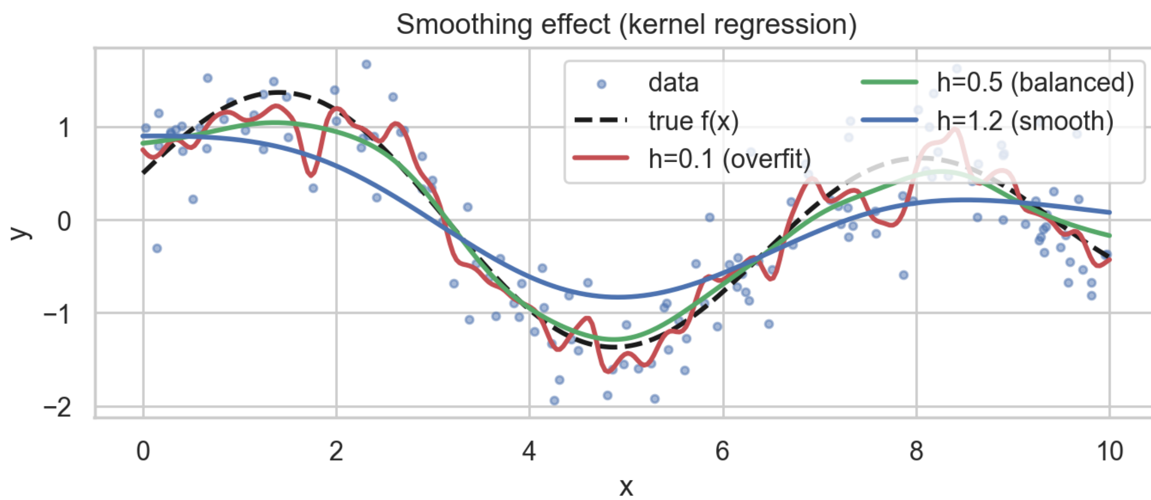

Figure 10: Effect of smoothing/bandwidth on the fitted function: small bandwidth overfits noise; large bandwidth oversmooths; a balanced choice captures the signal.

4.2 Polynomial Basis Functions

The polynomial basis represents the simplest approach to basis expansion, where $b_k(x) = x^{k-1}$ for $k = 1, 2, \ldots, K$. While conceptually straightforward, polynomial bases suffer from several limitations that make them less suitable for practical additive modeling. The global nature of polynomial functions means that changes in one region of the input space affect the entire function, leading to poor local adaptability. Additionally, polynomial bases can exhibit severe oscillatory behavior, particularly near the boundaries of the data range, a phenomenon known as Runge's phenomenon.

Despite these limitations, polynomial bases provide important theoretical insights into the approximation properties of basis expansions. The Weierstrass approximation theorem guarantees that any continuous function on a compact interval can be uniformly approximated by polynomials, establishing the theoretical foundation for basis function approaches.

4.3 B-spline Basis Functions

B-splines (basis splines) represent a significant advancement over polynomial bases, offering local support and numerical stability. A B-spline of degree d is defined recursively:

$$B_{i,d}(x) = \frac{x - t_i}{t_{i+d} - t_i} B_{i,d-1}(x) + \frac{t_{i+d+1} - x}{t_{i+d+1} - t_{i+1}} B_{i+1,d-1}(x)$$

where $\{t_i\}$ represents the knot sequence that determines the locations where the spline pieces connect.

The local support property of B-splines means that each basis function is non-zero only over a limited interval, typically spanning $d + 1$ knot intervals. This locality ensures that changes in the function in one region have minimal impact on other regions, providing much better numerical stability and interpretability compared to polynomial bases. The smoothness of the resulting spline function is determined by the degree d, with higher degrees producing smoother functions.

4.4 Radial Basis Functions

Radial basis functions (RBFs) offer an alternative approach where basis functions depend only on the distance from fixed center points: $b_k(x) = \phi(\|x - c_k\|)$. The most common choice is the Gaussian RBF, $\phi(r) = \exp(-r^2/\sigma^2)$, which provides smooth, bell-shaped basis functions centered at the points c_k.

RBFs are particularly effective for scattered data interpolation and can adapt well to irregular data distributions. The choice of centers c_k and the shape parameter σ significantly influences the approximation quality. Centers are often chosen to coincide with data points or placed on a regular grid, while the shape parameter controls the width of the basis functions and thus the smoothness of the resulting approximation.

5 Smoothing and Penalization

5.1 Penalized Likelihood

To control smoothness, we add penalty terms:

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\sum_{j=1}^{p} \lambda_j \boldsymbol{\beta}_j^T \mathbf{S}_j \boldsymbol{\beta}_j$$

where:

- $\ell(\boldsymbol{\beta})$ is the log-likelihood
- λ_j are smoothing parameters
- $\mathbf{S}_j$ are penalty matrices

For Gaussian responses with identity link, maximizing ℓ_p is equivalent to solving a penalized least squares problem. Writing the overall design matrix for all basis coefficients as $\mathbf{X}_\beta$ and the block-diagonal penalty as $\mathbf{S}_\lambda = \mathrm{diag}(\lambda_1\mathbf{S}_1, \ldots, \lambda_p\mathbf{S}_p)$, the normal equations become

$$(\mathbf{X}_\beta^T\mathbf{X}_\beta + \mathbf{S}_\lambda)\,\hat{\boldsymbol{\beta}} = \mathbf{X}_\beta^T\mathbf{y}.$$

The associated smoothing (or hat) matrix is

$$\mathbf{H}_\lambda = \mathbf{X}_\beta(\mathbf{X}_\beta^T\mathbf{X}_\beta + \mathbf{S}_\lambda)^{-1}\mathbf{X}_\beta^T,$$

which maps observed responses to fitted values $\hat{\mathbf{y}} = \mathbf{H}_\lambda\mathbf{y}$. The *effective degrees of freedom* (edf) of the fit is

$$\mathrm{edf}(\lambda) = \mathrm{tr}(\mathbf{H}_\lambda),$$

quantifying model complexity in the presence of smoothing. Larger penalties shrink coefficients and reduce edf, controlling variance at the cost of bias. From a Bayesian perspective, the penalty corresponds to a Gaussian prior $\boldsymbol{\beta}_j \sim N(\mathbf{0},\, (\lambda_j\mathbf{S}_j)^{-1})$.

5.2 Smoothing Splines

Definition 5.1 (Smoothing Spline)**.** *A smoothing spline minimizes:*

$$\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int [f''(x)]^2 dx$$

The solution is a natural cubic spline with knots at the data points.

This is a special case of the representer theorem: among all twice-differentiable functions, the minimizer of the penalized criterion lies in a finite-dimensional space spanned by spline basis functions determined by the inputs. The penalty $\int [f''(x)]^2\, dx$ enforces global smoothness by controlling curvature. As $\lambda \to 0$ the fit interpolates the data; as $\lambda \to \infty$ it approaches the least squares line.

5.3 P-splines

P-splines combine B-spline basis functions with difference penalties:

$$\text{Penalty} = \lambda \sum_{k=d+1}^{K} (\Delta^d \beta_k)^2$$

where Δ^d is the d-th order difference operator.

P-splines decouple flexibility (many, evenly spaced knots) from smoothness (controlled by the difference penalty). Using a rich B-spline basis (e.g., 20–40 knots) with a second- or third-order difference penalty yields estimates that are relatively insensitive to the precise knot placement while remaining computationally efficient due to the banded structure of B-spline design and penalty matrices.

6 Estimation Methods

6.1 Backfitting Algorithm

For Gaussian GAMs, the backfitting algorithm iteratively estimates each smooth function:

1. Initialize: $\hat{\alpha} = \bar{y}$, $\hat{f}_j = 0$ for all j
2. For $j = 1, 2, \dots, p$:

$$\text{partial residuals: } e_j = y - \hat{\alpha} - \sum_{k \neq j} \hat{f}_k(x_k)$$

$$\text{update: } \hat{f}_j = S_j(e_j)$$

   where S_j is a smoother
3. Repeat until convergence

Backfitting can be seen as Gauss–Seidel iteration on the normal equations obtained by treating each f_j as the image of a linear smoothing operator applied to its partial residuals. When each S_j is a linear smoother and the spectral radius of the combined operator is less than one, the algorithm converges to the unique solution that minimizes the penalized least squares criterion (see, e.g., Buja, Hastie and Tibshirani, 1989). Centering each $\hat{f}_j$ to have zero mean (or to be orthogonal to the intercept) ensures identifiability.

6.2 Penalized Iteratively Reweighted Least Squares (P-IRLS)

For non-Gaussian GAMs, extend IRLS with penalties:

1. Compute working response: $z = \eta + (y - \mu)/\mu'$
2. Compute weights: $w = (\mu')^2/V(\mu)$
3. Solve penalized weighted least squares:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{W}^{1/2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\|^2 + \sum_j \lambda_j \boldsymbol{\beta}_j^T \mathbf{S}_j \boldsymbol{\beta}_j$$

4. Update η, μ, and repeat

Collecting all coefficients in $\boldsymbol{\beta}$, the penalized normal equations at each iteration are

$$(\mathbf{X}^T \mathbf{W}\, \mathbf{X} + \mathbf{S}_\lambda)\, \hat{\boldsymbol{\beta}} = \mathbf{X}^T \mathbf{W}\, \mathbf{z}.$$

This yields the same algebraic form as in the Gaussian case, with $\mathbf{W}$ and $\mathbf{z}$ updated from the current mean μ and linear predictor η. The fitted values are $\hat{\eta} = \mathbf{X}\hat{\boldsymbol{\beta}}$ and the IRLS loop proceeds until changes in $\boldsymbol{\beta}$ (or deviance) are below tolerance.

Practical tips. Standardize continuous covariates before fitting; start with moderate smoothing (e.g., REML default) and inspect residual diagnostics; increase basis dimension (K) only if edf approaches the basis limit.

7 Smoothing Parameter Selection

7.1 Cross-Validation

Definition 7.1 (Generalized Cross-Validation (GCV))**.**

$$GCV(\lambda) = \frac{n\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{(n - tr(\mathbf{H}))^2}$$

where $\mathbf{H}$ *is the hat matrix and* $tr(\mathbf{H})$ *is the effective degrees of freedom.*

7.2 Restricted Maximum Likelihood (REML)

REML treats smoothing parameters as variance components:

$$\text{REML}(\lambda) = -\frac{1}{2}\left[\log|\mathbf{V}| + \log|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}| + \mathbf{y}^T\mathbf{P}\mathbf{y}\right]$$

In mixed model form, each smooth can be represented as a random effect with precision proportional to its smoothing parameter. Maximizing REML over variance components (and hence over λ) yields smoothing parameter estimates with good small-sample properties and natural uncertainty quantification.

7.3 Akaike Information Criterion

$$\mathrm{AIC} = -2\ell(\hat{\boldsymbol{\beta}}) + 2 \cdot \mathrm{edf}$$

where edf is the effective degrees of freedom.

Using edf in place of the raw parameter count accounts for the fact that smoothing penalizes flexibility. AIC targets out-of-sample predictive accuracy and is especially useful when comparing non-nested GAMs fit with different smooth structures.

7.4 Partial Effects and Interpretation

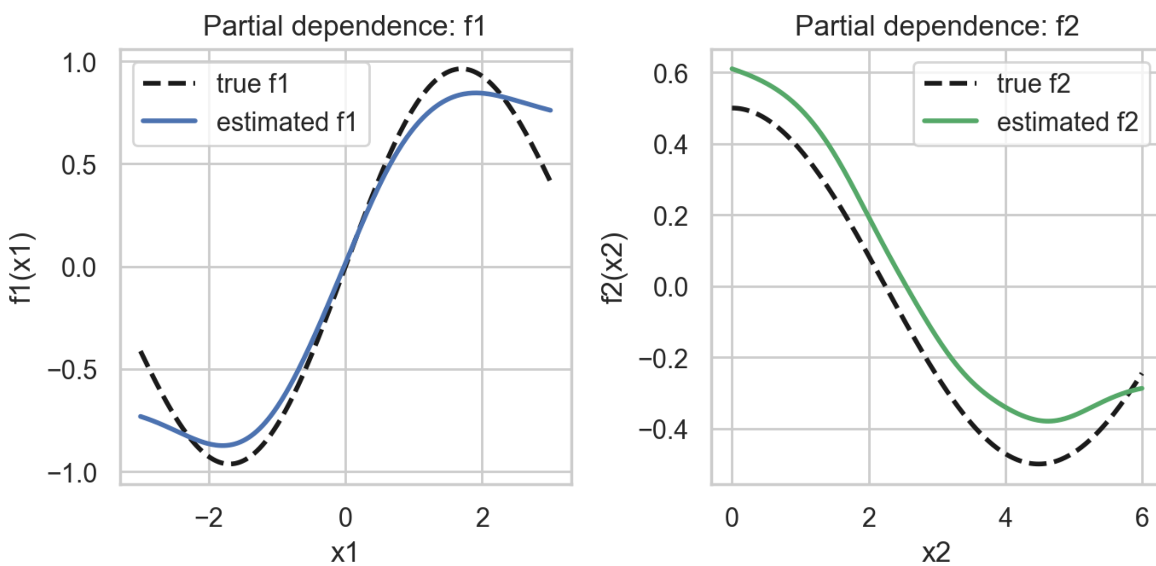

Figure 11: Partial dependence functions in a simple additive model: estimated smooths closely track the true underlying functions, enabling interpretation of each predictor's nonlinear effect.

8 Inference and Uncertainty

8.1 Confidence Intervals

For Gaussian GAMs, pointwise confidence intervals are:

$$\hat{f}(x) \pm z_{\alpha/2}\sqrt{\mathrm{Var}[\hat{f}(x)]}$$

The variance can be computed from the covariance matrix of the coefficients.

Under the penalized Gaussian framework, $\hat{\boldsymbol{\beta}}$ is approximately normal with covariance

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) \approx \sigma^2 \left(\mathbf{X}_\beta^T \mathbf{X}_\beta + \mathbf{S}_\lambda\right)^{-1} \mathbf{X}_\beta^T \mathbf{X}_\beta \left(\mathbf{X}_\beta^T \mathbf{X}_\beta + \mathbf{S}_\lambda\right)^{-1},$$

yielding pointwise standard errors for smooths via the delta method. For joint inference on an entire smooth, *simultaneous confidence bands* can be constructed using Scheffé-type adjustments or posterior simulation under the Bayesian interpretation of penalties.

Reporting. For each smooth, report edf, approximate F- or chi-square statistics, and 95% pointwise or simultaneous bands. Interpret effects on the scale of the linear predictor and, for GLMs, transform for presentation (e.g., odds scale).

8.2 Hypothesis Testing

8.2.1 Testing for Nonlinearity

Test $H_0 : f(x) = \beta x$ vs $H_1 : f(x)$ is nonlinear using:

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(\mathrm{df}_0 - \mathrm{df}_1)}{\mathrm{RSS}_1/\mathrm{df}_1}$$

8.2.2 Testing for Zero Effect

Test $H_0 : f(x) = 0$ using the effective degrees of freedom and residual variance.

More generally, approximate Wald or likelihood ratio tests can be formed for linear functionals of smooths, using edf to calibrate null distributions. In the mixed model formulation, standard inference for variance components (e.g., restricted likelihood ratio tests) can be used to assess whether a smooth term contributes beyond noise.

9 Model Selection and Diagnostics

9.1 Variable Selection

- **Stepwise selection**: Add/remove terms based on significance
- **Penalized selection**: Use penalties that can shrink functions to zero
- **Information criteria**: Compare models using AIC/BIC

9.2 Diagnostics

1. **Residual plots**: Inspect residuals vs fitted and vs each covariate for patterns indicating misspecification or under/oversmoothing.
2. **Q-Q plots**: Assess normality (for Gaussian GAMs) or use deviance/working residuals for non-Gaussian families.
3. **Influence and leverage**: The diagonal of $\mathbf{H}_\lambda$ generalizes leverage; large values flag influential observations.
4. **Partial residual plots**: Visualize the contribution of each smooth net of others to detect remaining structure.
5. **Concurvity**: Quantify nonlinear collinearity among smooths (e.g., via regressing one set of basis functions on another and reporting R^2); severe concurvity inflates uncertainty and destabilizes estimates.

9.3 Concurvity

Definition 9.1 (Concurvity)**.** *Concurvity is the nonlinear analog of collinearity - when one smooth function can be approximated by one or more other smooth functions.*

High concurvity can lead to unstable estimates and inflated standard errors.

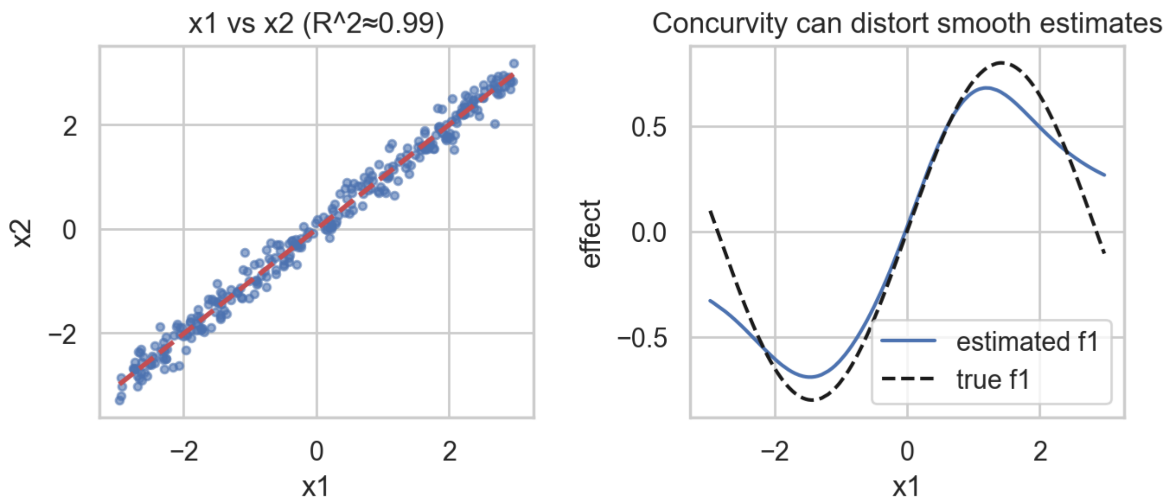

Figure 12: Concurvity illustration: highly related predictors (left) make it difficult to separate smooth effects (right), potentially distorting estimates.

10 Extensions and Variations

10.1 Varying Coefficient Models

Definition 10.1 (Varying Coefficient Model)**.**

$$Y = \sum_{j=1}^{p} f_j(Z)X_j + \epsilon$$

where the coefficients $f_j(Z)$ vary smoothly with a modifier variable Z. This extends linear interaction terms by allowing the effect of X_j to vary smoothly over Z, estimated via tensor product bases in (Z) and (optionally) X_j.

10.2 Functional Data Analysis

When predictors are functions:

$$Y = \alpha + \int f(t)X(t)dt + \epsilon$$

10.3 Spatial Smoothing

For spatial data, use tensor products or thin plate splines:

$$f(x, y) = \sum_{i,j} \beta_{ij} b_i(x) b_j(y)$$

Tensor product smooths combine marginal bases and penalties to adapt to different scales and smoothness along each dimension, while thin plate splines provide rotation-invariant smoothing with automatically placed basis functions.

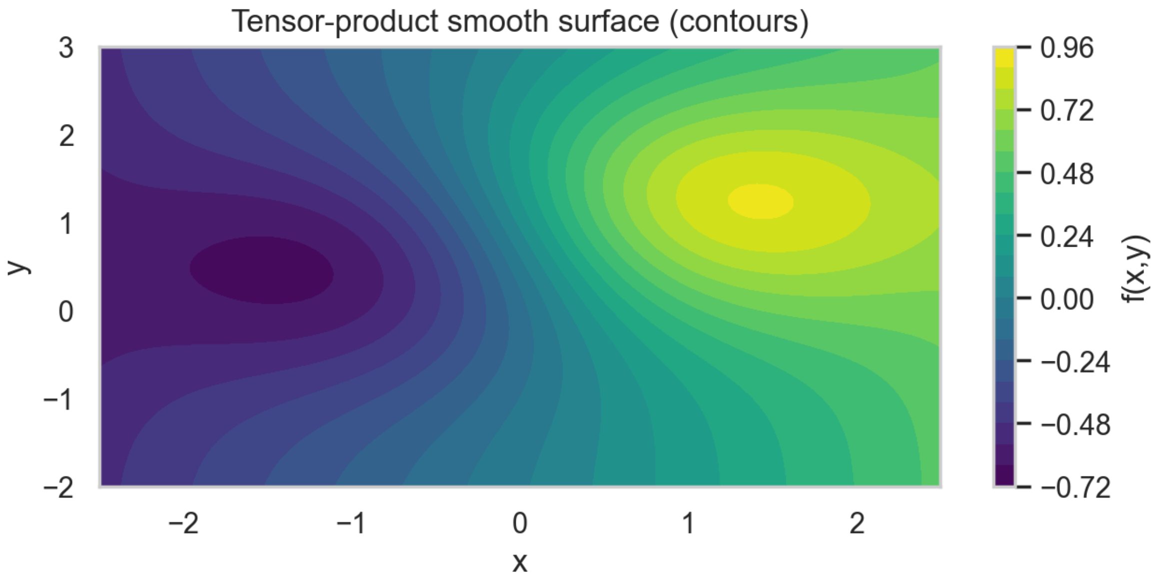

Figure 13: Tensor-product smooth surface: contours illustrate a smooth bivariate function constructed from marginal bases.

11 Computational Considerations

11.1 Efficient Algorithms

- **Wood's algorithm**: Efficient computation using QR decomposition
- **Sparse matrix methods**: Exploit sparsity in penalty matrices
- **Parallel computation**: Distribute smoothing across processors

11.2 Large Data Challenges

For big data:

- Use reduced rank smoothers
- Implement online/streaming algorithms
- Apply random sampling strategies

12 Applications in Data Science

12.1 Exploratory Data Analysis

GAMs excel at revealing nonlinear patterns:

- Identify threshold effects
- Discover seasonal patterns
- Uncover complex relationships

12.2 Predictive Modeling

- **Time series forecasting**: Capture nonlinear trends
- **Risk modeling**: Model nonlinear risk factors
- **Recommendation systems**: Capture user preference curves

12.3 Causal Inference

- Control for confounders nonlinearly
- Model dose-response relationships
- Estimate treatment effects across subgroups

13 Software and Implementation

Popular implementations include:

- **R**: mgcv, gam, VGAM packages
- **Python**: scikit-learn, statsmodels, pyGAM
- **Specialized**: GAMLSS, brms (Bayesian)

To illustrate the practical application of these concepts, consider a data scientist modeling website conversion rates as a function of user age and time spent on site using a logistic GAM. This application demonstrates how additive models can capture complex nonlinear relationships in business contexts, where the relationship between user characteristics and conversion probability may not follow simple linear patterns. The effective degrees of freedom for each smooth function provide insight into the complexity of the relationships discovered by the model, while the deviance explained measures the overall model performance compared to simpler alternatives.

14 Summary

This lecture provided a comprehensive introduction to Generalized Additive Models (GAMs), which extend linear models by allowing nonlinear relationships through smooth functions. We covered the mathematical foundations, estimation methods using penalized likelihood, and practical applications in data science contexts.

Key concepts include understanding how smooth functions capture nonlinear patterns, the role of smoothing parameters in controlling model complexity, and the interpretation of effective degrees of freedom. The connection between GAMs and other statistical methods provides important theoretical insights for understanding flexible modeling approaches.

Practical considerations such as model selection, diagnostics, and computational aspects are essential for proper application in data science projects. GAMs serve as a bridge between parametric and nonparametric methods, offering interpretable nonlinear modeling capabilities.

15 Exercises

1. **GAM Fundamentals:** Explain the difference between parametric and nonparametric components in GAMs. How do smoothing splines balance fit and smoothness? Describe the role of the smoothing parameter λ.

2. **Website Conversion Analysis:** A logistic GAM for conversion rates shows: age smooth (edf = 3.2), time smooth (edf = 5.8), deviance explained = 34%, AIC = 1247.3. Interpret each component and suggest model improvements.

3. **Effective Degrees of Freedom:** Calculate the effective degrees of freedom for a cubic smoothing spline with smoothing parameter $\lambda = 0.1$ applied to 100 data points. What does this value tell us about model complexity?

4. **Model Selection:** Compare three models for predicting house prices: (1) linear regression, (2) GAM with smooth terms for area and age, (3) GAM with interaction surface. Design a model selection strategy using AIC, cross-validation, and residual analysis.

5. **Smoothing Parameter Selection:** Implement generalized cross-validation (GCV) for selecting the smoothing parameter in a univariate smoothing spline. Compare results with REML estimation and discuss the trade-offs.

6. **GAM Diagnostics:** Develop a comprehensive diagnostic framework for GAMs including residual analysis, concurvity detection, and smooth function assessment. What specific issues should you check for in GAM residuals?

7. **Tensor Product Smooths:** Extend the basic GAM framework to include interaction effects using tensor products. Model sales data as a function of price and advertising with a smooth interaction surface.

8. **Computational Aspects:** Compare the computational complexity of fitting GAMs using backfitting versus penalized iteratively reweighted least squares. Discuss scalability considerations for large datasets.