Discrete Profiles: Statisticians's View

Anthony Davison

and

Timmy Tse

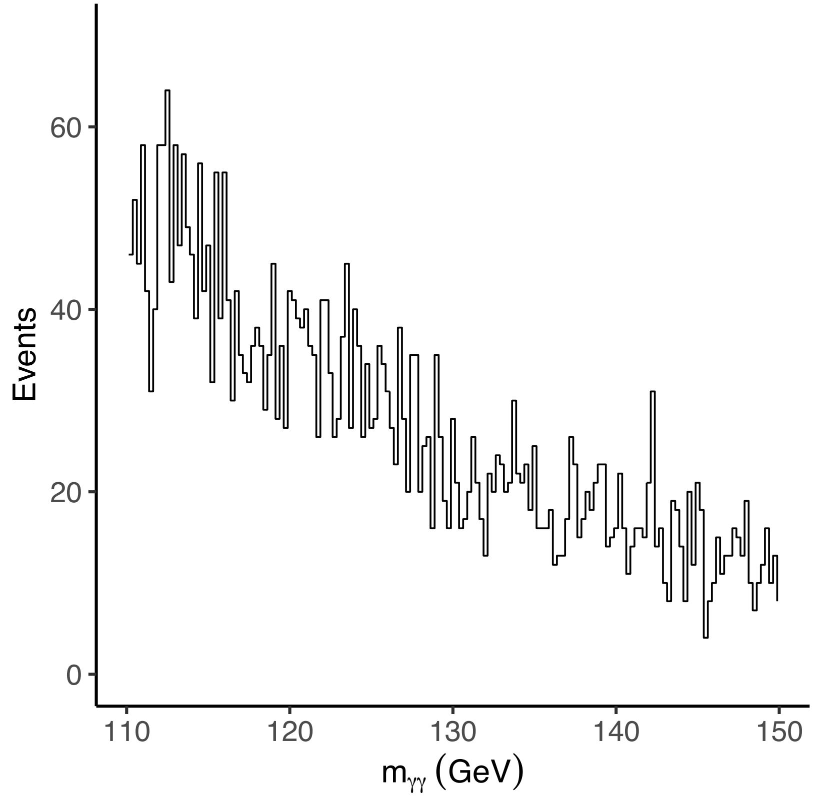
http://stat.epfl.ch

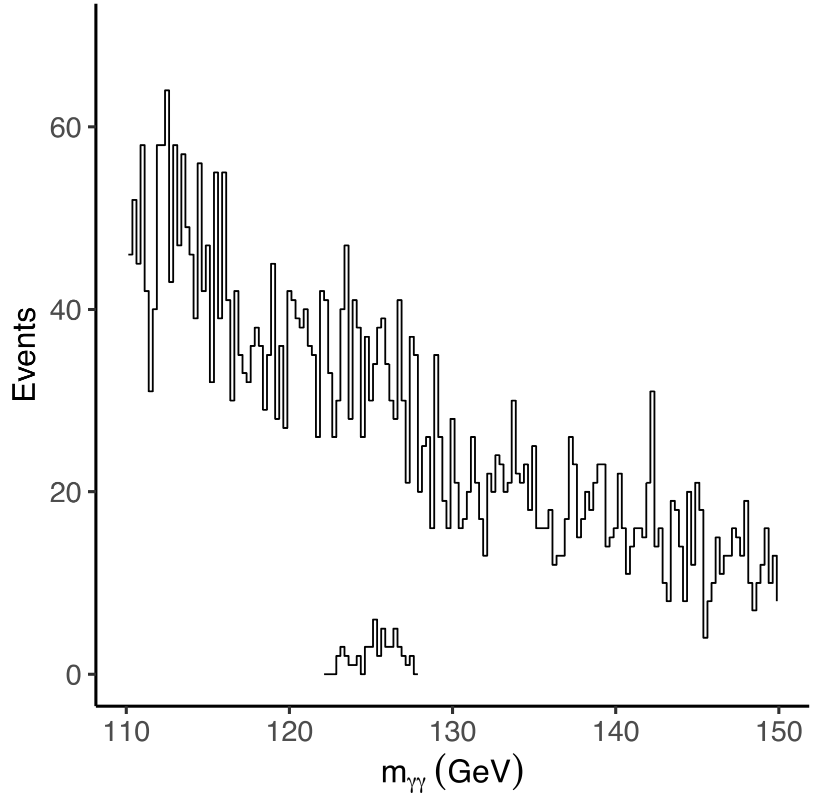
Basic setup

- Detecting a signal (interest) against an unknown background (nuisance).
- Simple model: vector $y = (y_1, \ldots, y_n)$ of independent Poisson variables with mean vector

$$\mu + \psi z,$$

  where μ represents the background, z is a known vector and interest parameter is ψ.
- Is $\psi > 0$?





Classical likelihood theory

- Parametric probability model $f(y;\psi,\lambda)$ depends on interest parameter ψ (usually scalar) and nuisance parameter λ
- Both parameters vary continuously in subsets of $\mathbb{R}^d$
- Test $\psi=\psi_0$ using **likelihood ratio statistic**

$$W_p(\psi_0) = 2\{\log f(y;\widehat{\psi},\widehat{\lambda}) - \log f(y;\psi_0,\widehat{\lambda}_{psi_0})\}$$

  where $(\widehat{\psi},\widehat{\lambda})$ is overall MLE and $(\psi_0,\widehat{\lambda}_{\psi_0})$ is MLE when $\psi=\psi_0$.
- Under classical regularity conditions (especially that the model is correct!) and scalar ψ,

$$W(\psi_0) \dot{\sim} \chi_1^2,$$

  when ψ_0 is the true value of ψ, leading to profile likelihood $(1-\alpha)$ confidence set

$$\left\{\psi : -2\log f(y;\psi,\widehat{\lambda}_\psi) \le -2\log f(y;\widehat{\psi},\widehat{\lambda}) + c_1(1-\alpha)\right\}.$$

- If the model is somewhat mis-specified, this interval remains approximately valid, provided the mis-specification is 'orthogonal to ψ' (Battey and Reid, 2024, PNAS) — difficult to guarantee.

Discrete profiling

- Density functions $f_m(y;\psi,\lambda_m)$ for M models with respective nuisance parameters λ_m.
- The discrete profile likelihood ratio statistic is

$$W_d(\psi_0) = 2\left\{\max_{\psi,\lambda_m,m} \log f_m(y;\psi,\lambda_m) - \max_{\lambda_m,m} f_m(y;\psi_0,\lambda_m)\right\},$$

  and the Taylor expansions leading to the limiting χ^2 distribution do not hold.
- Clearly $W_d(\psi_0)$ is stochastically smaller than $W(\psi_0)$, i.e.,

$$\mathrm{P}\left\{W_d(\psi_0) \le w\right\} \ge \mathrm{P}\left\{W(\psi_0) \le w\right\}, \quad w > 0,$$

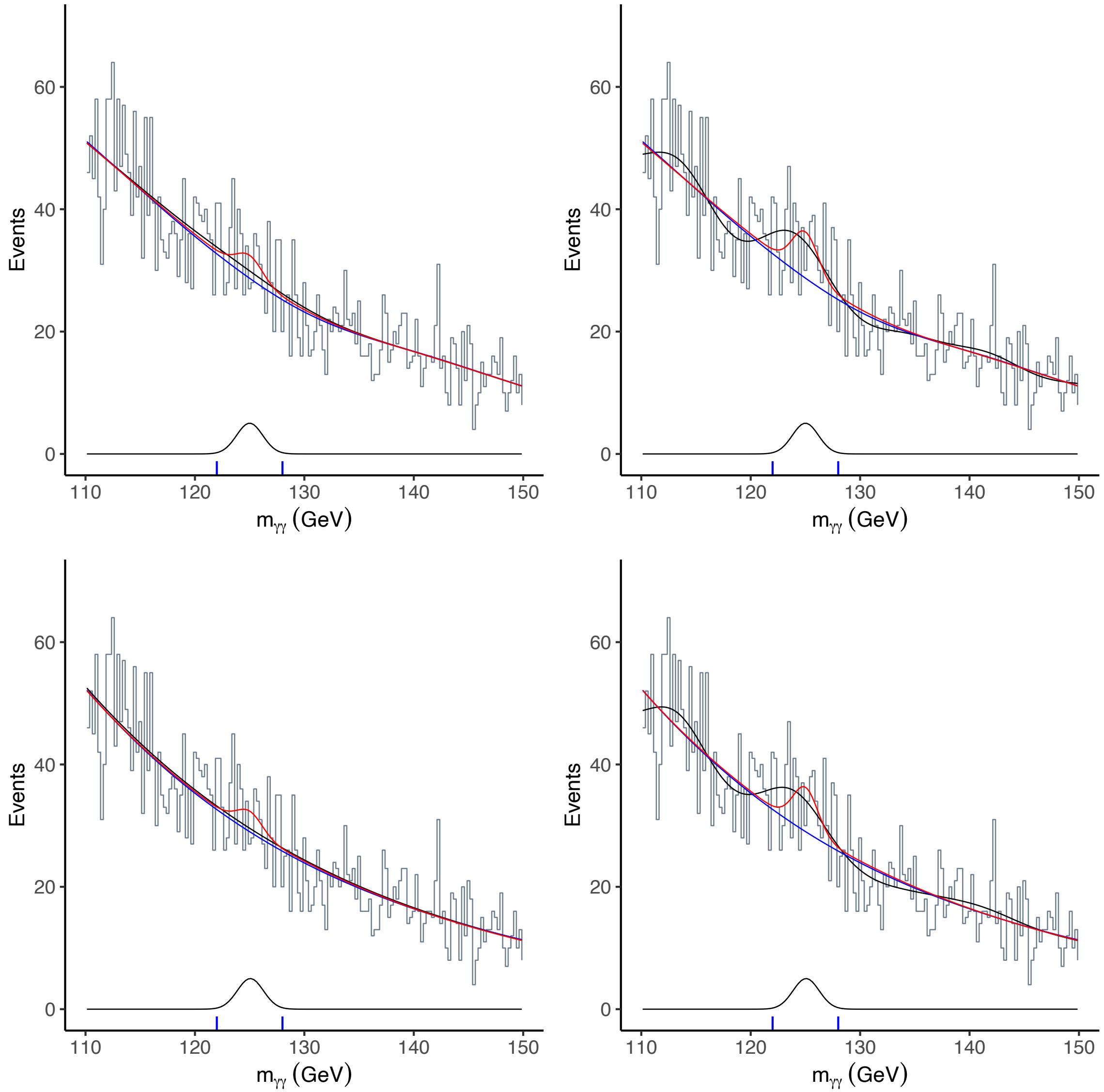
  but does this matter?
- Can we use χ^2 asymptotics anyway?
- Should we adjust for the different dimensions of λ_m?

Choice of model

- Parametrically: $\mu = X\lambda$ or $\mu = \exp(X\lambda)$ or ... for some $n \times p$ matrix X (polynomials, inverse polynomials, ...).
  - How to choose the link function?
  - How to choose X? Fit improves as p increases, perhaps uselessly (bias-variance tradeoff).
- Semi-parametrically: use a **generalized additive model (GAM)** (Wood, 2017).
  - X is a matrix of splines and λ is penalised;
  - degree of penalisation is determined from data by maximising a marginal likelihood;
  - highly efficient code available (`R` package `mgcv`);
  - but flexibility could be a curse?
- Next slide shows GAM fits of Poisson model using identity link function (top) and log link function (bottom), for
  - $X\lambda$ for entire dataset,
  - reduced background fits of $X\lambda$ to data outside blue ticks,
  - full model fits of $X\lambda + z\psi$ to all data.
- Log link function fit gets a preliminary estimate $\tilde{\mu}$ of μ, then writes

$$\mu + z\psi = \mu(1 + z\psi/\mu) \approx \mu \exp(z'\psi) = \exp(X\lambda + z'\psi), \quad z' = z/\tilde{\mu}.$$

Some GAM fits



GAM deviances

Link	Model	No signal added: Residual deviance	Residual edf	Spline edf	ψ	Signal added: Residual deviance	Residual edf	Spline edf	ψ
Identity	b	204.97	155.7	3.3		193.30	150.5	8.5	
	rb	172.82	131.9	3.1		173.31	131.9	3.1	
	b+s	202.70	155.0	3.0	$9.73_{6.88}$	204.24	155.1	2.9	$20.87_{7.13}$
Log	b	207.28	158.0	1.0		193.91	151.4	7.6	
	rb	175.64	133.7	1.3		176.16	133.7	1.3	
	b+s	205.01	157.0	1.0	$8.49_{5.58}$	206.83	157.0	1.0	$18.12_{5.34}$

Results of Poisson GAM fits to data with and without added signal.

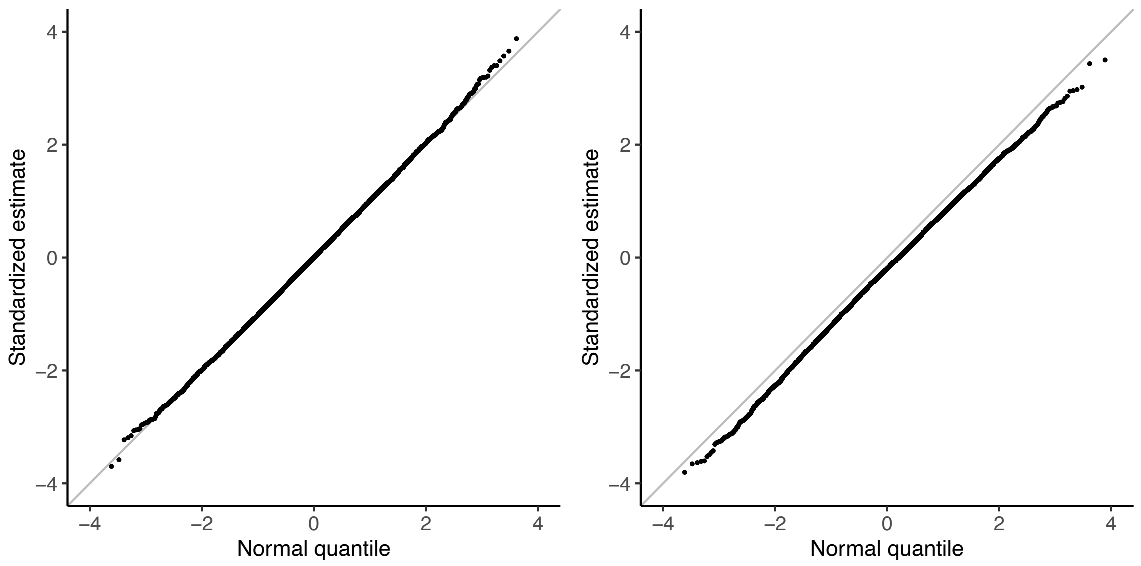
- The models use identity or log link functions, and terms for background (b), reduced background (rb) or background and signal (b+s).
- Standard errors for estimates of the signal size ψ are shown as subscripts.
- The **deviance** for a model is equivalent to

$$-2\max\ell,$$

and the **residual deviance** should be approximately $\chi^2_{\text{residual df}}$ for a well-fitting model.

GAM summary

- Interesting alternative for fitting background
- Isotonic splines could allow for monotone background (have not attempted this)
- Flexibility is
  - a blessing (no need to choose X or p);
  - a curse (presence of a signal pollutes background fit)
- Inference has to be based on estimate of ψ (not likelihood ratio), since deviance can increase when signal added — negative likelihood ratio statistic!
- Even with a GAM there is bias with the wrong link function:



Theoretical discussion

- Explicit calculations possible for simple Gaussian model

$$y \sim \mathcal{N}_n(\mu + z\psi_0, \sigma^2 I_n) \quad \text{but we fit linear model} \quad y \sim \mathcal{N}_n(X\lambda + z\psi, \sigma^2 I_n).$$

- Define standard projection matrices in linear regression,
  - $P = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ (projects $\mathbb{R}^n$ onto the space orthogonal to the columns of X),
  - $Q = I_n - z(z^{\mathrm{T}}Pz)^{-1}z^{\mathrm{T}}P$ (ditto for (X, z)).
- In this case the deviance for a given ψ is proportional to $S_{\mathrm{p}}(\psi) = (y - z\psi)^{\mathrm{T}}P(y - z\psi)$,

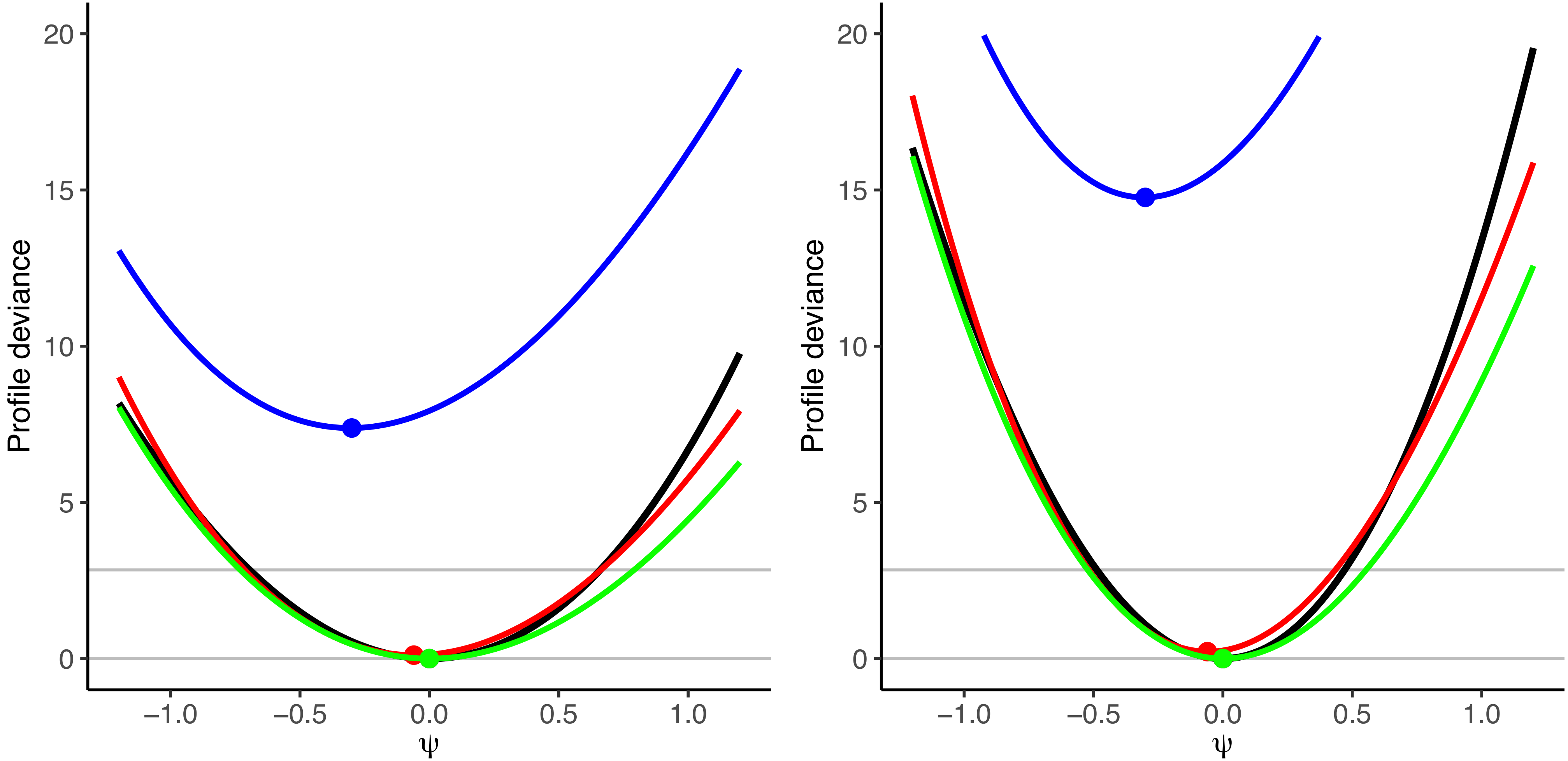
$$\widehat{\psi} = (z^{\mathrm{T}}Pz)^{-1}z^{\mathrm{T}}Py \sim \mathcal{N}(\psi_0 + z^{\mathrm{T}}P\mu/z^{\mathrm{T}}Pz, \sigma^2/z^{\mathrm{T}}Pz),$$

and the terms of

$$\mathrm{E}\left\{S_{\mathrm{p}}(\psi)\right\} = \sigma^2\left\{\mu^{\mathrm{T}}PQ\mu/\sigma^2 + (n-p) + \frac{(\psi - \psi_0 - z^{\mathrm{T}}P\mu/z^{\mathrm{T}}Pz)^2}{\sigma^2/z^{\mathrm{T}}Pz}\right\}$$

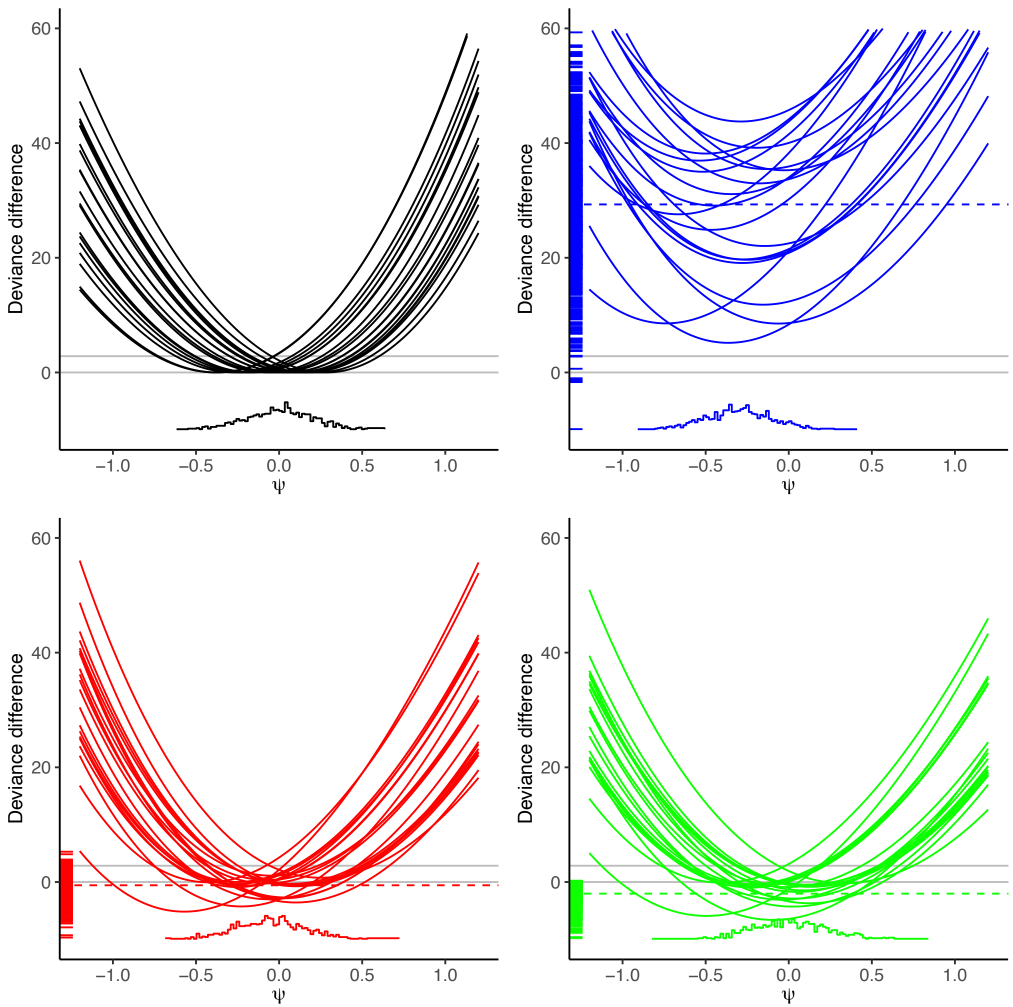
can be interpreted as
  - upward bias when μ cannot be explained by (X, z),
  - noise degrees of freedom, decreased by estimation of λ,
  - profile for ψ minimised at $\psi_0 + z^{\mathrm{T}}P\mu/z^{\mathrm{T}}Pz$, so biased if $P\mu \neq 0$.

Illustration for noiseless Poisson model



Profiles under true model with log link, and (wrong) fits with identity link and linear, quadratic and cubic polynomials.

Illustration for Poisson model



20 simulated profiles under true model with log link, and (wrong) fits with identity link and linear, quadratic and cubic polynomials.

Model selection

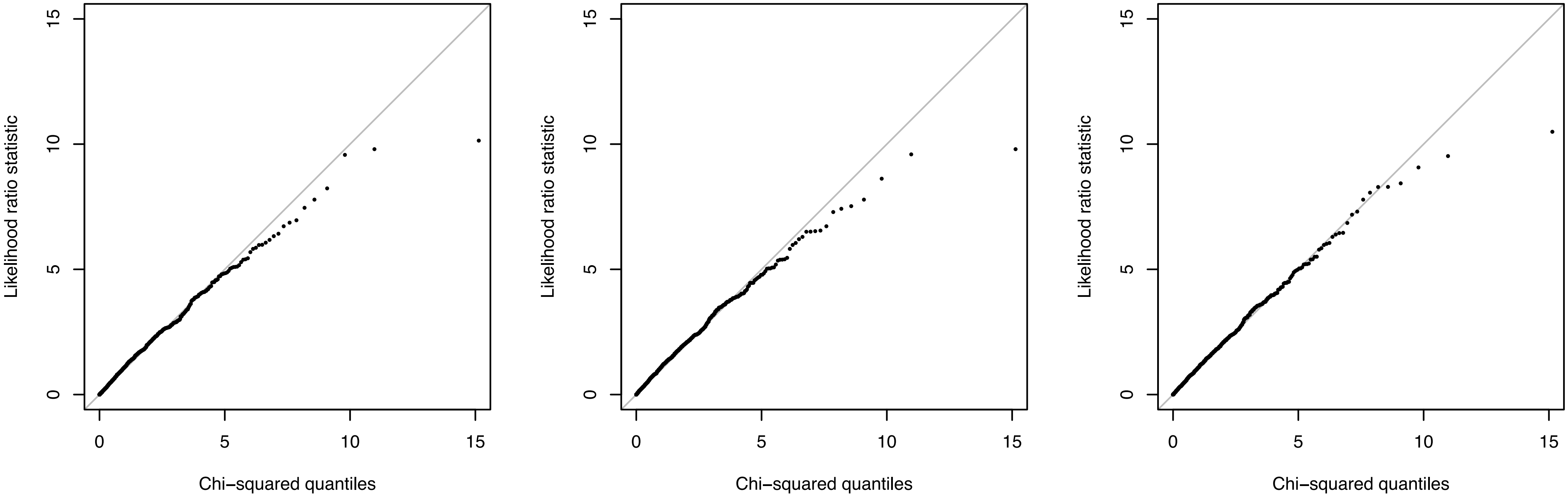
- Useful to distinguish
  - **True model**: unknown
  - **Correct models**: contain the true model, but have useless extra parameters
  - **Wrong models**: do not contain the true model
- To paraphrase George Orwell:

  All models are wrong, but some are wronger than others:

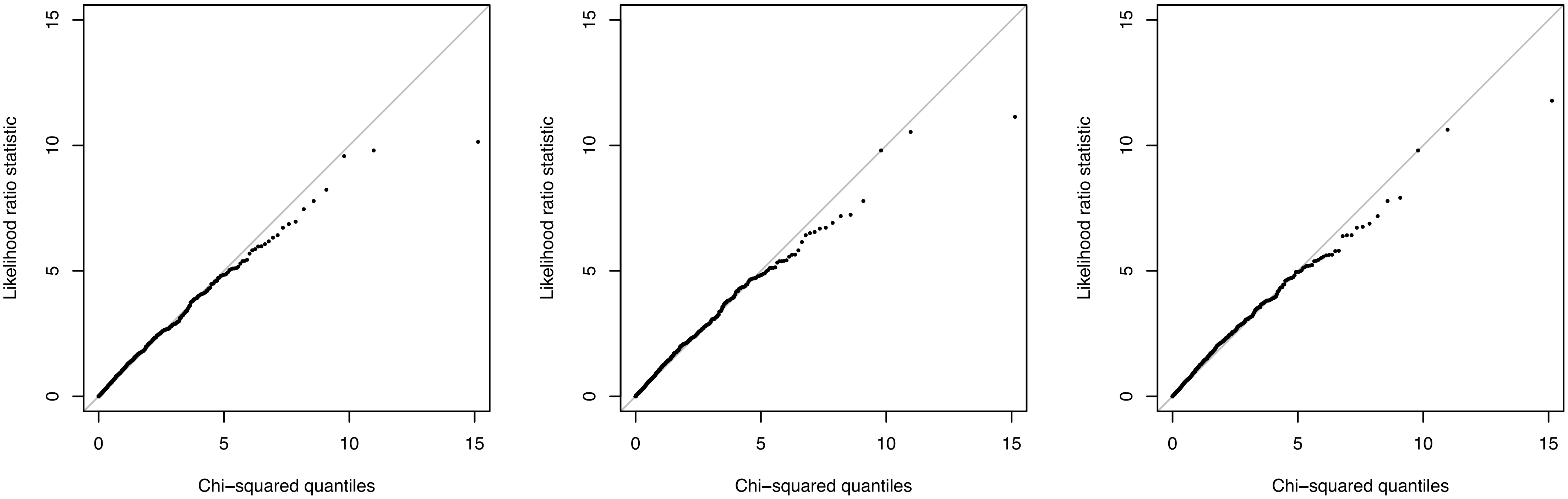
  - the linear fit is badly wrong
  - the quadratic fit shows visible bias but variance only slightly larger than the ideal fit
  - the cubic fit shows negligible bias but clearly increased variance
- Looks like
  - we should add p to the deviance to counteract the estimation effect, not $2p$ (as with AIC)
  - but this will not account for the added variance.
- Simulations with these models and profile LR statistics seem ambiguous ...

Illustration for Poisson model

- Comparison of 10000 simulated discrete profile LR statistics: original scale (left), with p added (centre), with $2p$ added (right), with χ^2_1 quantiles.
- Adding $2p$ gets closest to the 'nominal' χ^2_1.
- The true model is among those fitted.

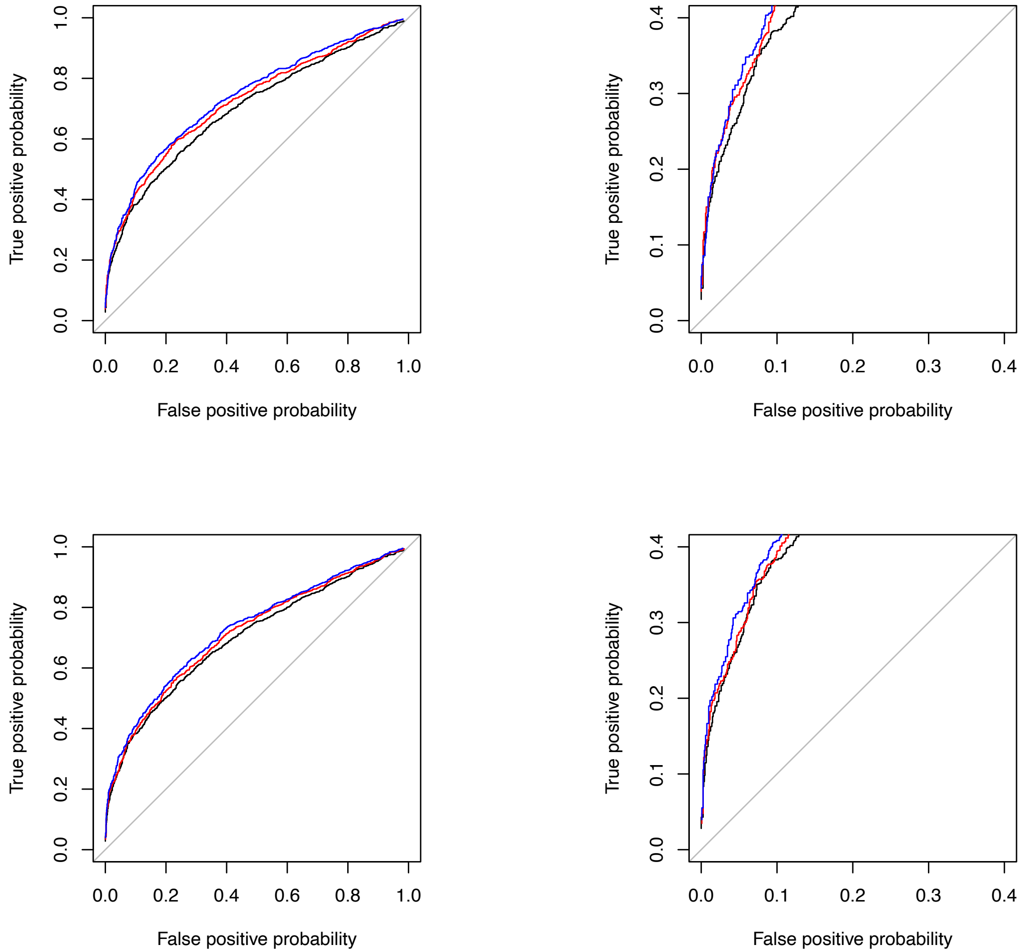


- Comparison of 10000 simulated discrete profile LR statistics: original scale (left), with p added (centre), with $2p$ added (right), with χ_1^2 quantiles.
- Adding nothing or p seems best — visible distortion on the right?
- The true model is not among those fitted.



Power

- ROC curves for discrete profile likelihood ratios: original, with p added, with $2p$ added.
- Top: true model included; bottom: true model not included.
- Right: zoom of left.



Conclusions?

- Discrete profiling
  - seems like a useful approach to inference on an interest parameter ψ,
  - but ψ needs to be the same for all models (possibility of bias)
  - very wrong models are (almost) irrelevant
  - power is lost when using classical χ^2 asymptotics?
  - looks as though adding $2p$ for estimation gives highest power (but it's a close thing)
- Fitting (isotonic) GAMs might be a way forward — but how good is classical asymptotic theory in this case?