Lecture 3: Exponential Family Distributions

STATS305B: Applied Statistics II

Scott Linderman

January 13, 2025

Recap

Last time...

- Logistic Regression
- Maximum Likelihood Estimation via Gradient Descent, Newton's Method, and IRLS
- Regularization and Converge Rates

Outline

Today...

- Definition and Examples
- The Log Normalizer
- Maximum Likelihood Estimation
- Mean Parameterization
- KL Divergence and Deviance Residuals

Exponential Families

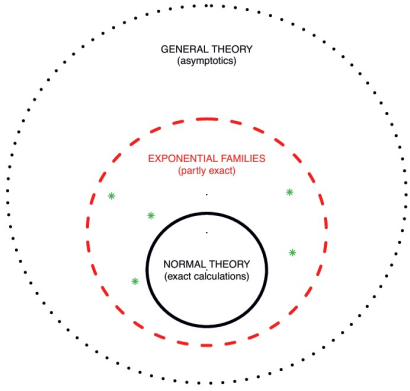


Figure 1 Three levels of statistical modeling.

From Efron (2022).

Definition

Exponential family distributions have densities of the form,

$$p(y;\eta) = h(y)\exp\left\{\langle t(y),\eta\rangle - A(\eta)\right\},$$

where

- $h(y) : \mathscr{Y} \to \mathbb{R}_+$ is the **base measure**,
- $t(y) \in \mathbb{R}^T$ are the **sufficient statistics**,
- $\eta \in \mathbb{R}^T$ are the **natural parameters**, and
- $A(\eta) : \mathbb{R}^T \to \mathbb{R}$ is the **log normalizing** function (aka the **partition function**).

Definition

The log normalizer ensures that the density is properly normalized,

$$A(\eta) = \log \int h(y) \exp\left\{\langle t(y), \eta \rangle\right\} \mathrm{d}y$$

The domain of the exponential family is the set of valid natural parameters, $\Omega = \{\eta : A(\eta) < \infty\}$. An exponential family is a family of distributions defined by base measure h and sufficient statistics t, and it is indexed by natural paremeters $\eta \in \Omega$.

Example: Poisson distribution

Take the Poisson pmf,

$$\begin{aligned}\mathrm{Po}(y;\lambda) &= \frac{1}{y!}\lambda^y e^{-\lambda} \\ &= \frac{1}{y!}\exp\{y\log\lambda - \lambda\} \\ &= h(y)\exp\left\{y\eta - A(\eta)\right\}\end{aligned}$$

where

- the base measure is $h(y) = \frac{1}{y!}$
- the sufficient statistics are $t(y) = y$
- the natural parameter is $\eta = \log\lambda$
- the log normalizer is $A(\eta) = \lambda = e^{\eta}$

Bernoulli distribution

Exercise: Write the Bernoulli distribution in exponential family form. Recall that its pmf is

$$\text{Bern}(y; p) = p^y (1-p)^{1-y}$$

What are the base measure, the sufficient statistics, the natural parameter, the log normalizer, and the domain?

Submit your answers here,

https://tinyurl.com/stats305blec03

Categorical distribution

Finally, take the categorical pmf for $Y \in \{1, \ldots, K\}$,

$$\begin{aligned}
\text{Cat}(y; \boldsymbol{\pi}) &= \prod_{k=1}^{K} \pi_k^{\mathbb{I}[y=k]} \\
&= \exp\left\{ \sum_{k=1}^{K} \mathbb{I}[y=k], \log \pi_k \right\} \\
&= \exp\left\{ \langle \boldsymbol{e}_y, \log \boldsymbol{\pi} \rangle \right\} \\
&= h(y) \exp\left\{ \langle t(y), \boldsymbol{\eta} \rangle - A(\boldsymbol{\eta}) \right\}
\end{aligned}$$

where

- the base measure is $h(y) = \mathbb{I}[y \in \{1, \ldots, K\}]$
- the sufficient statistics are $t(y) = \boldsymbol{e}_y$, the one-hot vector representation of y

Categorical distribution

- the natural parameter is $\boldsymbol{\eta} = \log \boldsymbol{\pi} = (\log \pi_1, \ldots, \log \pi_K)^\top \in \mathbb{R}^K$
- the log normalizer $A(\boldsymbol{\eta}) = 0$
- the domain is $\Omega = \mathbb{R}^K$

The Log Normalizer

The **cumulant generating function** — i.e., the log of the moment generative function — is a difference of log normalizers,

$$\begin{aligned}\log \mathbb{E}_\eta[e^{\langle t(Y),\theta\rangle}] &= \log \int h(y) \exp\left\{\langle t(y), \eta+\theta\rangle - A(\eta)\right\} \mathrm{d}y \\ &= \log e^{A(\eta+\theta)-A(\eta)} \\ &= A(\eta+\theta) - A(\eta) \\ &\triangleq K_\eta(\theta)\end{aligned}$$

Its derivatives (with respect to θ and evaluated at zero) yield the cumulants. In particular,

- $\nabla_\theta K_\eta(0) = \nabla A(\eta)$ yields the first cumulant of $t(Y)$, its mean
- $\nabla^2_\theta K_\eta(0) = \nabla^2 A(\eta)$ yields the second cumulant, its covariance

Higher order cumulants can be used to compute skewness, kurtosis, etc.

Gradient of the log normalizer

We can also obtain this result more directly.

$$\begin{aligned}
\nabla A(\eta) &= \nabla \log \int h(y) \exp\left\{\langle t(y), \eta \rangle\right\} \mathrm{d}y \\
&= \frac{\int h(y) \exp\left\{\langle t(y), \eta \rangle\right\} t(y) \, \mathrm{d}y}{\int h(y) \exp\left\{\langle t(y), \eta \rangle\right\} \mathrm{d}y} \\
&= \int p(y \mid \eta) \, t(y) \, \mathrm{d}y \\
&= \mathbb{E}_\eta[t(Y)]
\end{aligned}$$

Again, the gradient of the log normalizer yields the **expected sufficient statistics**,

Hessian of the log normalizer

The Hessian of the log normalizer yields the **covariance of the sufficient statistics**,

$$\begin{aligned}\nabla^2 A(\eta) &= \nabla \int p(y \mid \eta)\, t(y)\, \mathrm{d}y \\ &= \int p(y \mid \eta)\, t(y)\, (t(y) - \nabla A(\eta))^\top \, \mathrm{d}y \\ &= \mathbb{E}[t(Y) t(Y)^\top] - \mathbb{E}[t(Y)]\, \mathbb{E}[t(Y)]^\top \\ &= \mathrm{Cov}[t(Y)]\end{aligned}$$

Maximum Likelihood Estimation

Suppose we have $y_i \overset{\text{iid}}{\sim} p(y; \eta)$ for a minimal exponential family distribution with natural parameter η. The log likelihood is,

$$\begin{aligned}\mathscr{L}(\eta) &= \sum_{i=1}^{n} \log p(y_i; \eta) \\ &= \left\langle \sum_{i=1}^{n} t(y_i), \eta \right\rangle - nA(\eta) + c\end{aligned}$$

The gradient is

$$\nabla \mathscr{L}(\eta) = \sum_{i=1}^{n} t(y_i) - n\nabla A(\eta),$$

and the Hessian is $\nabla^2 \mathscr{L}(\eta) = -n\nabla^2 A(\eta)$.

Maximum Likelihood Estimation

Since the log normalizer is convex, all local optima are global. If the log normalizer is *strictly* convex, the MLE will be unique.

Setting the gradient to zero and solving yields the stationary conditions for the MLE,

$$\nabla A[\hat{\eta}_{\text{MLE}}] = \mathbb{E}[t(Y); \hat{\eta}_{\text{MLE}}] = \frac{1}{n}\sum_{i=1}^{n} t(y_i).$$

When ∇A is invertible, the MLE is unique,

$$\hat{\eta}_{\text{MLE}} = [\nabla A]^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} t(y_i)\right).$$

Even if ∇A is not invertible, maximum likelihood estimation amounts to matching empirical means of the sufficient statistics to corresponding natural parameters.

Asymptotic normality

Recall that the MLE is asymptotically normal with variance given by the inverse Fisher information.

For an exponential family distribution, the Fisher information is,

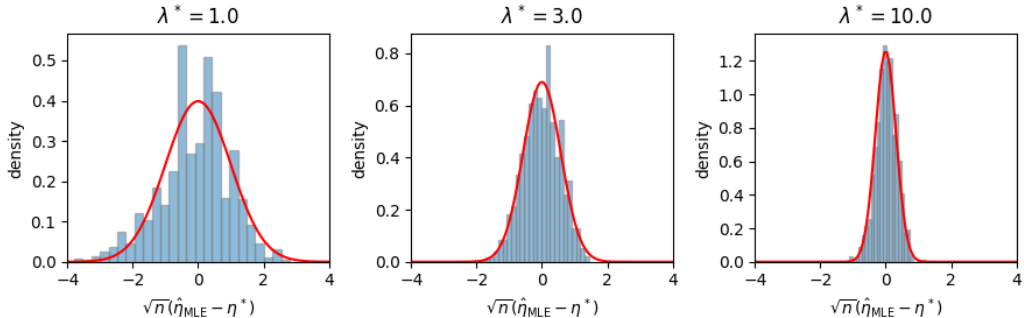
$$\mathcal{I}(\eta) = -\mathbb{E}[\nabla^2 \log p(y_i; \eta)] = \nabla^2 A(\eta) = \mathrm{Cov}_\eta[t(Y)].$$

Thus,

$$\sqrt{n}(\hat{\eta}_{\mathrm{MLE}} - \eta) \to \mathrm{N}(0, \mathcal{I}(\eta)^{-1}) = \mathrm{N}(0, \mathrm{Cov}_\eta[t(Y)]^{-1})$$

Example: MLE for the Poisson distribution

Suppose $Y_i \overset{\text{iid}}{\sim} \text{Po}(\lambda)$ for $i = 1, \ldots, n$. The natural parameter of the Poisson distribution is $\eta = \log \lambda$, and the maximum likelihood estimate is $\hat{\eta}_{\text{MLE}} = \log\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)$. The Fisher information matrix is the variance, $\mathscr{I}(\eta) = \text{Var}_{\eta}[Y] = e^{\eta}$.



Minimal Exponential Families

The Hessian of the log normalizer gives the covariance of the sufficient statistic. Since covariance matrices are positive semi-definite, the *log normalizer is a convex function* on Ω.

If the covariance is strictly positidive definite — i.e., if the minimum eigenvalue of $\nabla^2 A(\eta)$ is strictly greater than zero for all $\eta \in \Omega$ — then the log normalizer is *strictly* convex. In that case, we say that the exponential family is **minimal**

Question: Is the exponential family representation of the categorical distribution above a minimal representation? If not, how could you encode it in minimal form?

Mean Parameterization

When constructing models with exponential family distributions, like the generalized linear models below, it is often more convenient to work with the **mean parameters** instead. for a d-dimensional sufficient statistic, let,

$$\mathcal{M} \triangleq \left\{\mu \in \mathbb{R}^d : \exists\, p \text{ s.t. } \mathbb{E}_p[t(Y)] = \mu\right\}$$

denote the set of mean parameters realizable *by any distribution p*.

Two facts:

1. The gradient mapping $\nabla A : \Omega \mapsto \mathcal{M}$ is *injective* (one-to-one) if and only if the exponential family is minimal.
2. The gradient is a *surjective* mapping from mean parameters to the *interior* of $\mathcal{M}$. All mean parameters in the interior of $\mathcal{M}$ (excluding the boundary) can be realized by an exponential family distribution. (Mean parameters on the boundary of $\mathcal{M}$ can be realized by a limiting sequence of exponential family distributions.)

Mean Parameterization

Together, these facts imply that the gradient of the log normalizer defines a *bijective* map from Ω to the interior of $\mathscr{M}$ for minimal exponential families.

For minimal families, we can work with the mean parameterization instead,

$$p(y; \mu) = h(y) \exp\left\{\langle t(y), [\nabla A]^{-1}(\mu)\rangle - A([\nabla A]^{-1}(\mu))\right\}.$$

for mean parameters μ in the interior of $\mathscr{M}$.

MLE for the Mean Parameters

Alternatively, consider the maximum likelihood estimate of the mean parameter $\mu \in \mathcal{M}$. Before doing any math, we might expect the MLE to be the empirical mean. Indeed, that is the case. To simplify notation, let $\eta(\mu) = [\nabla A]^{-1}(\mu)$. The log likelihood,

$$\mathcal{L}(\mu) = \left\langle \sum_{i=1}^{n} t(y_i), \eta(\mu) \right\rangle - nA(\eta(\mu)) + c$$

has gradient,

$$\begin{aligned}
\nabla \mathcal{L}(\mu) &= \left(\frac{\partial \eta}{\partial \mu}(\mu) \right) \left[\sum_{i=1}^{n} t(y_i) - n \nabla A(\eta(\mu)) \right] \\
&= \left(\frac{\partial \eta}{\partial \mu}(\mu) \right) \left[\sum_{i=1}^{n} t(y_i) - n\mu \right],
\end{aligned}$$

MLE for the Mean Parameters

where $\frac{\partial \eta}{\partial \mu}(\mu)$ is the Jacobian of inverse gradient mapping at μ. Assuming the Jacobian is positive definite, we immediately see that,

$$\hat{\mu}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} t(y_i).$$

Now back to the Jacobian… applying the inverse function theorem, shows that it equals the inverse covariance matrix,

$$\frac{\partial \eta}{\partial \mu}(\mu) = \frac{\partial [\nabla A]^{-1}}{\partial \mu}(\mu) = [\nabla^2 A([\nabla A]^{-1}(\mu))]^{-1} = \text{Cov}_{\eta(\mu)}[t(Y)]^{-1},$$

which is indeed positive definite for minimal exponential families.

Asymptotic normality

We obtain the Fisher information of the mean parameter μ by left and right multiplying by the Jacobian,

$$\begin{aligned}\mathscr{I}(\mu) &= \left(\frac{\partial \eta}{\partial \mu}(\mu)\right)^\top \mathscr{I}(\eta(\mu)) \left(\frac{\partial \eta}{\partial \mu}(\mu)\right) \\ &= \mathrm{Cov}_{\eta(\mu)}[t(Y)]^{-1} \mathrm{Cov}_{\eta(\mu)}[t(Y)] \mathrm{Cov}_{\eta(\mu)}[t(Y)]^{-1} \\ &= \mathrm{Cov}_{\eta(\mu)}[t(Y)]^{-1}.\end{aligned}$$

Thus, the MLE of the mean parameter is asymptotically normal with covariance determined by the inverse Fisher information, $\mathscr{I}(\mu)^{-1} = \mathrm{Cov}_{\eta(\mu)}[t(Y)]$. More formally,

$$\sqrt{n}\,(\hat{\mu}_{\mathrm{MLE}} - \mu^\star) \to \mathrm{N}(0, \mathrm{Cov}_{\eta(\mu)}[t(Y)])$$

As usual, to derive confidence intervals we plug in the MLE to evaluate the asymptotic covariance.

Asymptotic normality

Compare this result to the asymptotic covariances we computed in Lecture 1 for the Bernoulli distribution. Recall that for $X_i \overset{\text{iid}}{\sim} \text{Bern}(\theta)$, where $\theta \in [0, 1]$ is the mean parameter, we found,

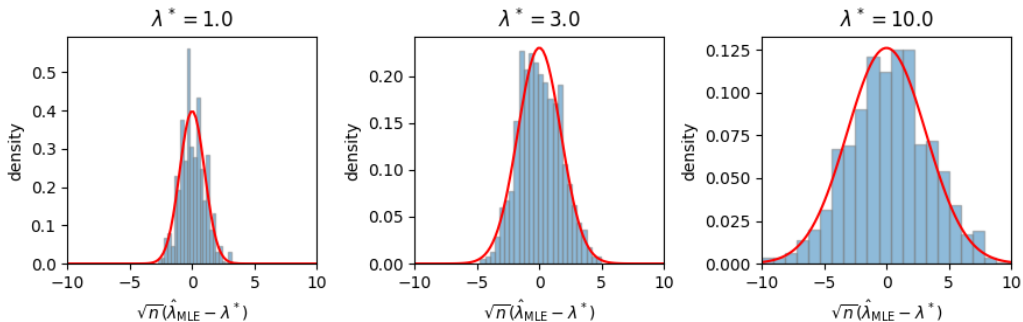
$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta^\star) \to \text{N}\left(0, \text{Var}_\theta[X]\right).$$

Now we see that this is a general property of exponential family distributions.

Revisiting the Poisson example

Revisiting the example above, here we have $\hat{\lambda}_{\text{MLE}} = \frac{1}{n}\sum_i y_i$ and $\mathscr{I}(\lambda) = \text{Var}_\lambda[Y]^{-1} = \frac{1}{\lambda}$. We expect,

$$\sqrt{n}(\hat{\lambda}_{\text{MLE}} - \lambda) \to \text{N}(0, \mathscr{I}(\lambda)^{-1}) = \text{N}(0, \lambda).$$



Conjugate duality

The log normalizer is a convex function. Its conjugate dual is,

$$A^*(\mu) = \sup_{\eta \in \Omega} \left\{ \langle \mu, \eta \rangle - A(\eta) \right\}$$

We recognize this as the maximum likelihood problem mapping expected sufficient statistics μ to natural parameters η. For minimal exponential families, the supremum is uniquely obtained at $\eta(\mu) = [\nabla A]^{-1}(\mu)$. The conjugate dual evaluates to the log likelihood obtained at $\eta(\mu)$.

It turns out the conjugate dual is also related to the entropy; in particular, for any μ in the interior of $\mathscr{M}$,

$$A^*(\mu) = -\mathbb{H}[p_{\eta(\mu)}],$$

Conjugate duality

where $\eta(\mu) = [\nabla A]^{-1}(\mu)$ for minimal exponential families. To see this, note that

$$\begin{aligned} -\mathbb{H}[p_{\eta(\mu)}] &= \mathbb{E}_{p(\eta(\mu))}[\log p(X; \eta(\mu))] \\ &= \mathbb{E}_{p(\eta(\mu))}[\langle t(X), \eta(\mu)\rangle - A(\eta(\mu))] \\ &= \langle \mu, \eta(\mu)\rangle - A(\eta(\mu)) \\ &= A^*(\mu). \end{aligned}$$

Moreover, for minimal exponential families, the gradient of A^* provides the inverse map from mean parameters to natural parameters,

$$\nabla A^*(\mu) = \arg\max_{\eta \in \Omega} \left\{ \langle \mu, \eta\rangle - A(\eta) \right\} = [\nabla A]^{-1}(\mu).$$

Conjugate duality

Finally, the log normalizer has a variataional representation in terms of its conjugate dual,

$$A(\eta) = \sup_{\mu \in \mathcal{M}} \left\{ \langle \mu, \eta \rangle - A^*(\mu) \right\}.$$

For more on conjugate duality, see Wainwright and Jordan (2008), ch. 3.6.

KL Divergence

The **Kullback-Leibler (KL) divergence**, or **relative entropy**, between two distributions is,

$$D_{\mathrm{KL}}(p \,\|\, q) = \mathbb{E}_p\left[\log \frac{p(Y)}{q(Y)}\right].$$

It is non-negative and equal to zero if and only if $p = q$. The KL divergence is *not* a distance because it is not a symmetric function of p and q. (generally, $D_{\mathrm{KL}}(p \,\|\, q) \neq D_{\mathrm{KL}}(q \,\|\, p)$.)

When p and q belong to the same exponential family with natural parameters η_p and η_q, respectively, the KL simplifies to,

$$\begin{aligned} D_{\mathrm{KL}}(p \,\|\, q) &= \mathbb{E}_p\left[\langle t(Y), \eta_p\rangle - A(\eta_p) - \langle t(Y), \eta_q\rangle + A(\eta_q)\right] \\ &= \langle \mathbb{E}_p[t(Y)], \eta_p - \eta_q\rangle - A(\eta_p) + A(\eta_q) \\ &= \langle \nabla A(\eta_p), \eta_p - \eta_q\rangle - A(\eta_p) + A(\eta_q). \end{aligned}$$

This form highlights that the KL divergence between exponential family distributions is a special case of a **Bregman divergence** based on the convex function A.

Example: Poisson Distribution

Consider the Poisson distribution with known mean λ. In the example above, we cast it as an exponential family distribution with - sufficient statistics $t(y) = y$ - natural parameter $\eta = \log \lambda$ - log normalizer $A(\eta) = e^{\eta}$

The KL divergence is,

$$\begin{aligned} D_{\mathrm{KL}}(p \parallel q) &= \langle e^{\eta_p}, \eta_p - \eta_q \rangle - e^{\eta_p} + e^{\eta_q} \\ &= \lambda_p \log \frac{\lambda_p}{\lambda_q} - \lambda_p + \lambda_q \end{aligned}$$

Deviance

Rearranging terms, we can view the KL divergence as a remainder in a Taylor approximation of the log normalizer,

$$A(\eta_q) = A(\eta_p) + (\eta_q - \eta_p)^\top \nabla A(\eta_p) + D_{\mathrm{KL}}(p \,\|\, q)\,.$$

From this perspective, we see that the KL divergence is related to the Fisher information,

$$\begin{aligned} D_{\mathrm{KL}}(p \,\|\, q) &\approx \frac{1}{2}(\eta_q - \eta_p)^\top \nabla^2 A(\eta_p)(\eta_q - \eta_p) \\ &= \frac{1}{2}(\eta_q - \eta_p)^\top \mathscr{I}(\eta_p)(\eta_q - \eta_p), \end{aligned}$$

up to terms of order $\mathcal{O}(\|\eta_p - \eta_q\|^3)$.

Deviance

Thus, while the KL divergence is not a distance metric due to its asymmetry, it is approximately a squared distance under the Fisher information metric,

$$2D_{\mathrm{KL}}\left(p \parallel q\right) \approx \|\eta_q - \eta_p\|^2_{\mathscr{I}(\eta_p)}.$$

We call this quantity the **deviance**. It is simply twice the KL divergence.

Deviance Residuals

In a normal model, the standarized residual is $\frac{\hat{\mu}-\mu}{\sigma}$. We can view this as a function of the deviance between two normals,

$$\frac{\hat{\mu}-\mu}{\sigma} = \operatorname{sign}(\hat{\mu}-\mu)\sqrt{2D_{\mathrm{KL}}\left(\hat{\mu} \parallel \mu\right)}$$

where we have used the shorthand notation

$$D_{\mathrm{KL}}\left(\mu \parallel \hat{\mu}\right) \triangleq D_{\mathrm{KL}}\left(\mathrm{N}(\mu, \sigma^2) \parallel \mathrm{N}(\hat{\mu}, \sigma^2)\right).$$

The same form generalizes to other exponential families as well, with the **deviance residual** between the true and estimated mean parameters defined as,

$$r_{\mathrm{D}}(\hat{\mu}, \mu) = \operatorname{sign}(\hat{\mu}-\mu)\sqrt{2D_{\mathrm{KL}}\left(\hat{\mu} \parallel \mu\right)}.$$

Deviance Residuals

One can show that deviance residuals tend to be closer to normal than the more obvious Pearson residuals,

$$r_{\mathrm{P}}(\hat{\mu}, \mu) = \frac{\hat{\mu} - \mu}{\sqrt{\mathrm{Var}[t(Y); \hat{\mu}]}}.$$

For more on deviance residuals, see Efron (2022), ch. 1.

Revisiting the Poisson example

Finally, let's revisit the Poisson example one again. We already computed the KL divergence between two Poisson distributions above,

$$D_{\mathrm{KL}}\left(\mathrm{Po}(\hat{\lambda}) \,\|\, \mathrm{Po}(\lambda)\right) = \hat{\lambda} \log \frac{\hat{\lambda}}{\lambda} - \hat{\lambda} + \lambda,$$

so the deviance residual is,

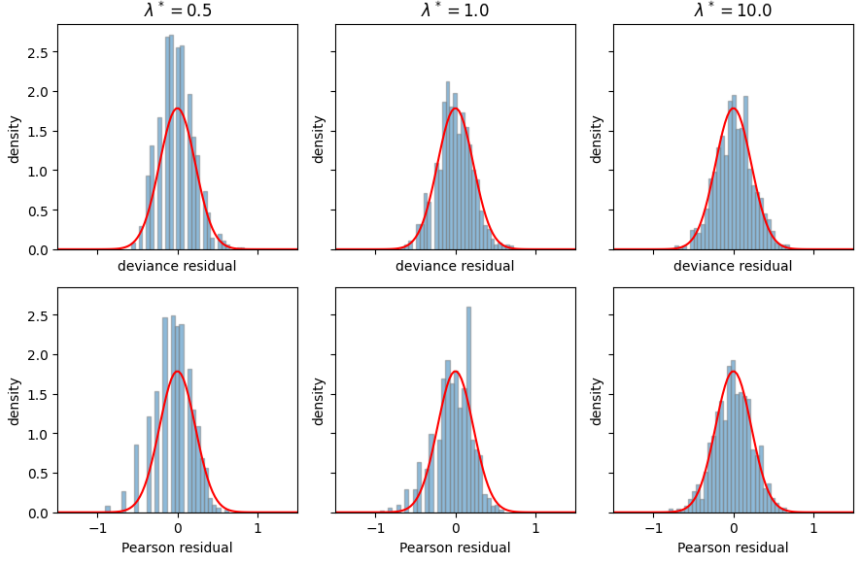
$$r_{\mathrm{D}}(\hat{\lambda}, \lambda) = \mathrm{sign}(\hat{\lambda} - \lambda)\sqrt{2\left(\hat{\lambda} \log \frac{\hat{\lambda}}{\lambda} - \hat{\lambda} + \lambda\right)}.$$

Compare this to the Pearon residual,

$$r_{\mathrm{P}}(\hat{\lambda}, \lambda) = \frac{\hat{\lambda} - \lambda}{\sqrt{\hat{\lambda}}}$$

Let's compare these residuals in simulation.

Revisiting the Poisson example



Conclusion

Exponential family distributions have many beautiful properties, and we've only scratched the surface.

- We'll see other nice properties when we talk about building probabilistic models for joint distributions of random variables using exponential family distributions, and conjugate relationships between exponential families will simplify many aspects of Bayesian inference.
- We'll also see that inference in exponential families is closely connected to convex optimization — we saw that today for the MLE! — but for more complex models, the optimization problems can still be computationally intractable, even though its convex. That will motivate our discussion of variational inference later in the course.

Armed with exponential family distributions, we can start to build more expressive models for categorical data. First up, generalized linear models!